Predicting the Final Course Grade Based on Course Data

1 Introduction

In this research paper, we will explore the data from a course in Data Science at Åbo Akademi University.

Based on course data, we will train two machine-learning models to predict the students' final grades.

2 Data processing

The data consists of the anonymized information of 107 students. The data includes three mini-projects, three peer reviews, three quizzes, and a total score at the end of the course.

In addition to the information on assignments students have performed during the course, course logs are provided. 36 course logs range over the 9 weeks of the course and have four statuses.

Status0 is course, lecture, and content related. It tracks how the amount of times the student has viewed different parts of the course page and lesson activity. The nature of the lesson activity is not specified.

Status1 is assignment related. It tracks activity regarding quizzes on the course page, such as started quiz attempts and submissions.

Status 2 is grade-related. It tracks the viewing of the grades and user profiles.

Status3 is forum-related and tracks the creations and viewings of posts on the course forum, among other things.

2.1 Averages on course logs

To examine the course logs and their correlation to the final grade the students received, average values for each status for the student were created. This would significantly reduce the dimensionality for analysis where status logs are included.

By calculating the average for each activity for each student, 36 logs are reduced to four averages.

2.2 Correlation Analysis

For each remaining variable, a Pearson Correlation Analysis was performed. The analysis compares each variable with the target, the final grade of the course.

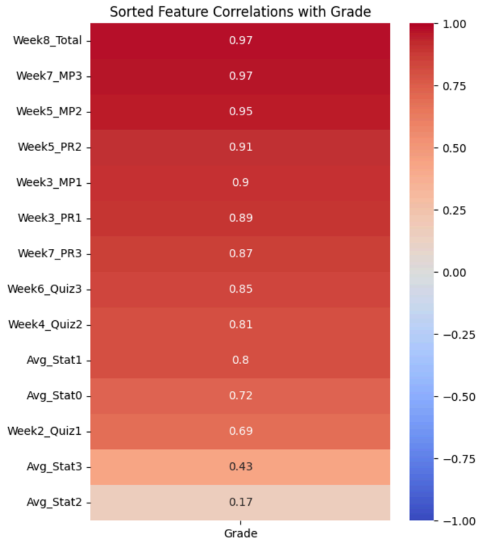


Figure 1: A Pearson Correlation Analysis

Week8_Total and Week7_MP3 have the highest correlation of all the variables, closely followed by Week5_MP2. In the course context, the total points at week 8 equals the final grade. Thus, it will be removed from the final data frame.

The mini-projects also impact the final grade significantly, as they award students more points than the peer reviews and the quizzes.

Only the averages for status 0 and 1 significantly impact the grade. The averages for statuses 2 and 3 have a low impact.

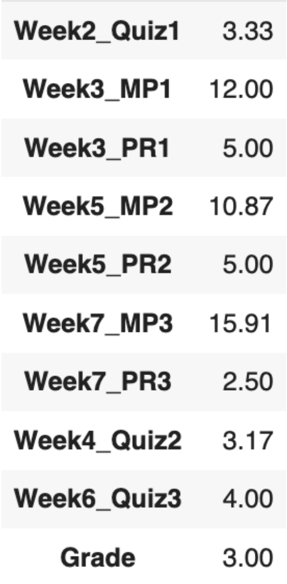
Given the ambiguity surrounding the accumulation of different statuses, which represent diverse activities that in the context of the course do not exhibit a direct correlation with course success, these variables were excluded from the final analysis

2.3 Visual Data Analysis

The remaining variables are:

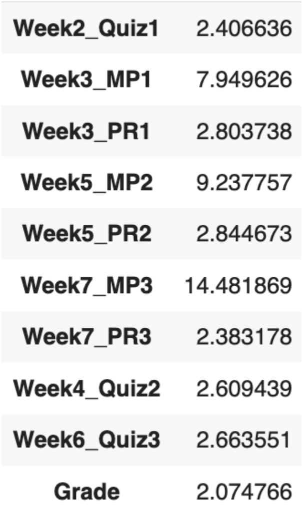
- 3 mini projects
- 3 peer reviews
- 3 quizzes
- final course grade

The medians and means for each variable are provided in the tables.



Week2_Quiz1	3.33
Week3_MP1	12.00
Week3_PR1	5.00
Week5_MP2	10.87
Week5_PR2	5.00
Week7_MP3	15.91
Week7_PR3	2.50
Week4_Quiz2	3.17
Week6_Quiz3	4.00
Grade	3.00

Figure 2: Median for each variable



Week2_Quiz1	2.406636
Week3_MP1	7.949626
Week3_PR1	2.803738
Week5_MP2	9.237757
Week5_PR2	2.844673
Week7_MP3	14.481869
Week7_PR3	2.383178
Week4_Quiz2	2.609439
Week6_Quiz3	2.663551
Grade	2.074766

Figure 3: Mean for each variable

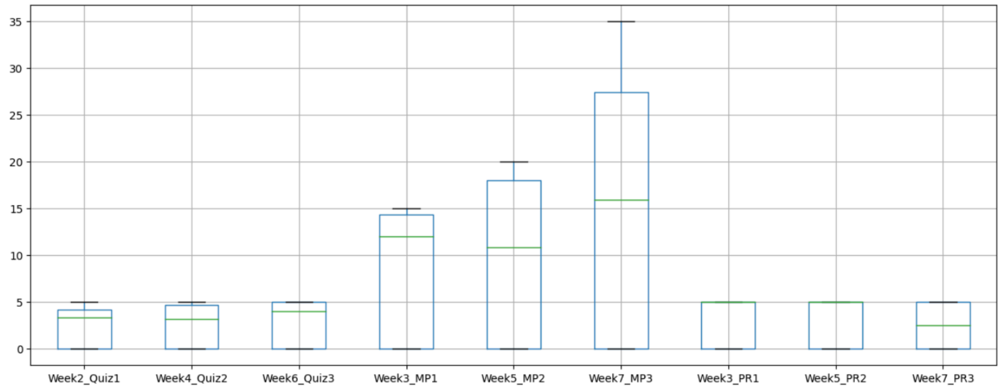


Figure 4: Box plots of the variables

Students generally exhibited high performance in quizzes and peer reviews. However, the higher point allocation for mini-projects resulted in a more substantial impact on the final grade.

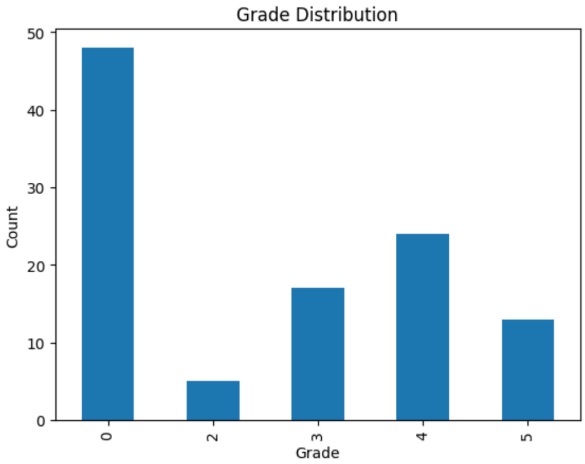
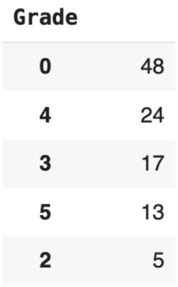


Figure 5: Distributions of grades in test data



Grade	
0	48
4	24
3	17
5	13
2	5

Figure 6: Distribution of grades in test data

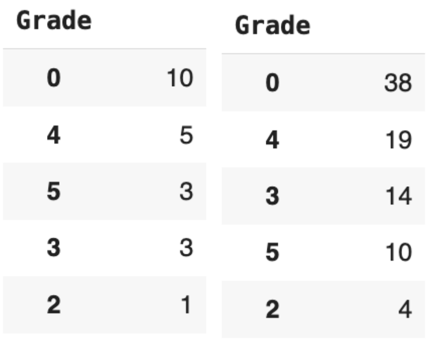
The test data does not include students who received grade 1. Thus we cannot train models that can predict grade 1.

Only five students received grade 2, which can lead to performance issues later on.

3 Splitting data for testing and training

The data set is split into training and testing sets with a ratio of 20 % test and 80 % testing.

Stratify ensures that the different grades are proportionately represented in all data sets.



Grade	
0	10
4	5
5	3
3	3
2	1

Grade	
0	38
4	19
3	14
5	10
2	4

Figures 7 and 8: Distributions of grades in test and train data sets

As a result, we get two data sets, where the distribution of final grades is even.

At this point, the seed is manually set to three values to evaluate the performance of the models: 15, 42, and 70. This ensures some additional randomization of the train and test sets.

4 Model Training

As grade is not a continuous value, this excludes some models. In this analysis, numerical grades represent discrete categories. They are ordinal, as the order is meaningful.

The chosen models suit classification problems of this type.

4.1 K-nearest Neighbors

K-nearest Neighbors (KNN) is a non-parametric supervised learning method with one parameter K. It solves classification problems by comparing the neighboring data points.

The parameter K defines how many neighbors will be checked. StratifiedKFold was chosen as a cross-validation tactic with the split 5, as there are only 5 2's in the data set. For this data set, the maximum split 5 ensures a better generalization.

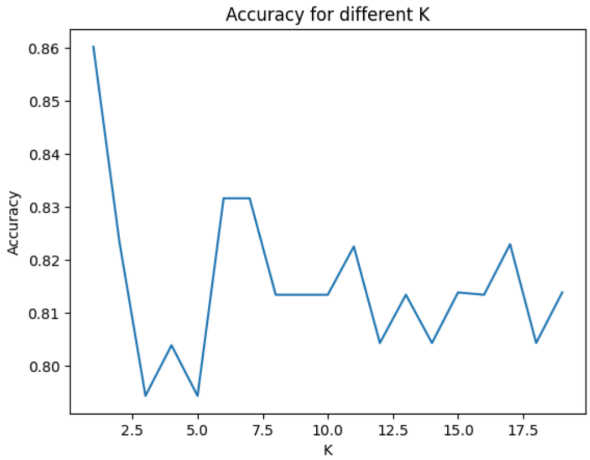


Figure 9: Best K's with a split of 5

An odd number of K's should be chosen to avoid ties when the model makes comparisons to neighbors while labeling.

Figure 9 displays the outcome of the StratifiedKFold with a split of 5. K=1 generally gives the best accuracy for our model, but it might be prone to have a high bias.

The next best value is 7. Both 1 and 7 are chosen to be examined in the analysis.

4.1.2 One Neighbor

Seed 15 Accuracy: 0.909090
Seed 42 Accuracy: 0.818181
Seed 70 Accuracy: 0.863636

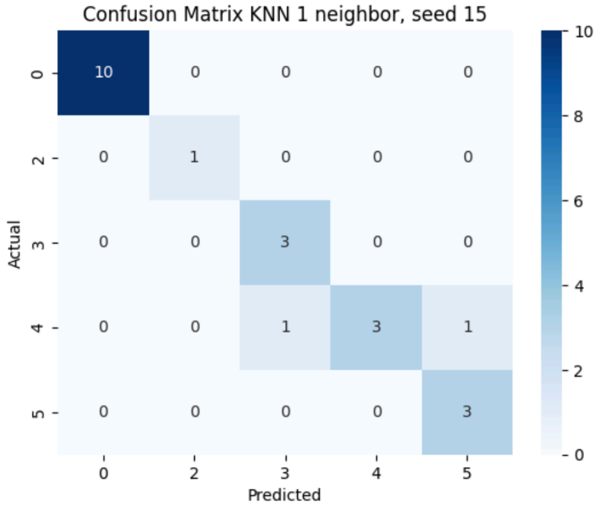


Figure 10: Confusion Matrix for KNN with 1 neighbor, seed 15

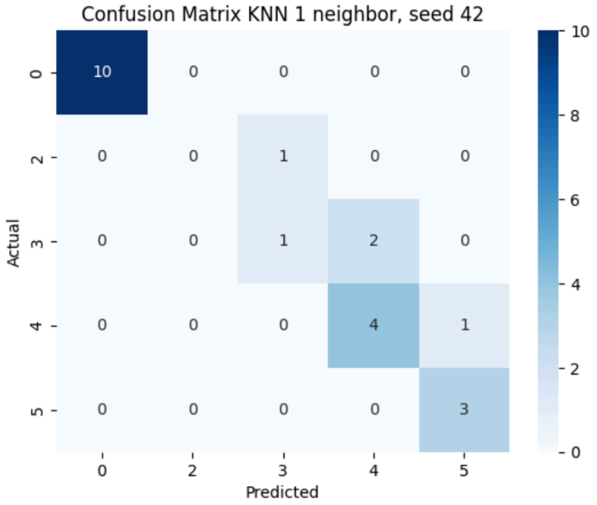


Figure 11: Confusion Matrix for KNN with 1 neighbor, seed 42

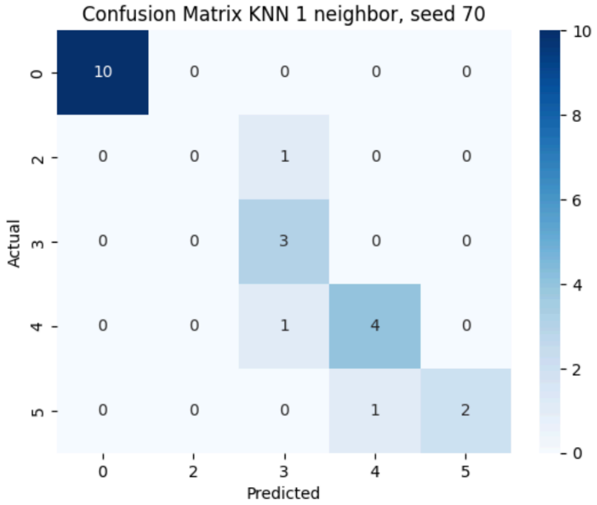


Figure 12: Confusion Matrix for KNN with 1 neighbor, seed 70

The model with one neighbor generally performs One of them was able to predict the grade 2 correctly.

The accuracy for all three seeds ranges from 82% to 91%. This is a high score, but a model with only one neighbor is prone to overfitting.

4.1.3 Seven Neighbors

To eliminate the risk of overfitting, the next best K was chosen from the StratifiedKFold analysis.

Seed 15 Accuracy: 0.863636
Seed 42 Accuracy: 0.909090
Seed 70 Accuracy: 0.772727

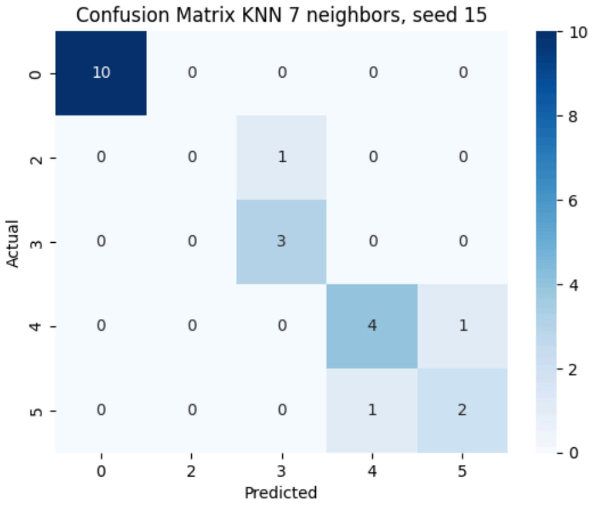


Figure 13: Confusion Matrix for KNN with 7 neighbors, seed 15

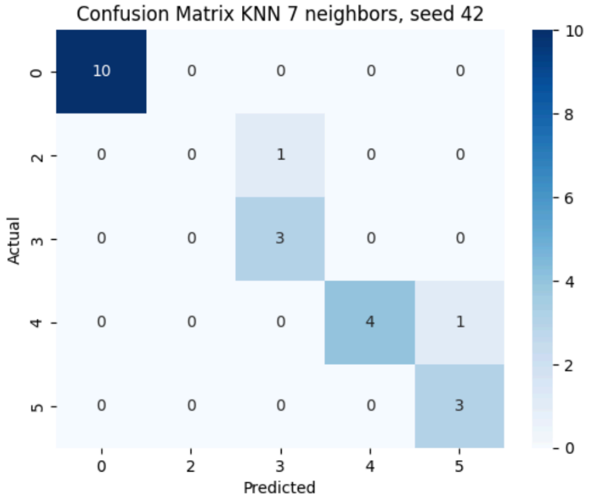


Figure 14: Confusion Matrix for KNN with 7 neighbors, seed 42

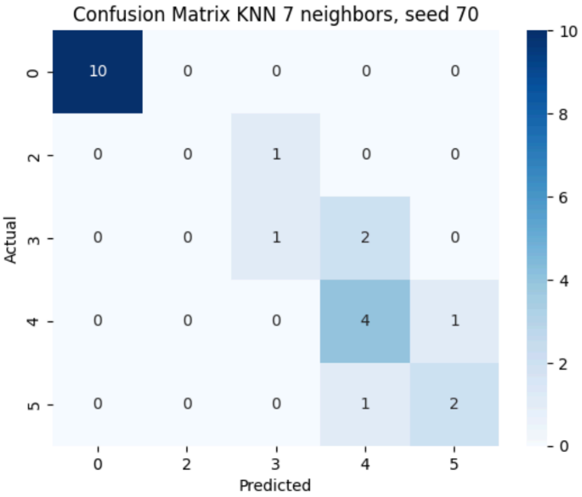


Figure 15: Confusion Matrix for KNN with 7 neighbors, seed 70

Seven neighbors did not increase the correct prediction of grade 2. For the case of seven neighbors, this can be explained by the fact that there are only four grade 2's to compare to in the training data. When seven neighbors are considered, grade 2 predictions are heavily impacted by grade 3 neighbors.

The model generally performs well, ranging from 77% to 91%. The performance is slightly reduced in comparison with the model with one neighbor.

4.2 Random Forest

The Random Forest model predicts by splitting data in decision trees. In this paper 125 decision trees are used.

Seed 15 Accuracy: 0.954545
Seed 42 Accuracy: 0.863636
Seed 70 Accuracy: 0.909090

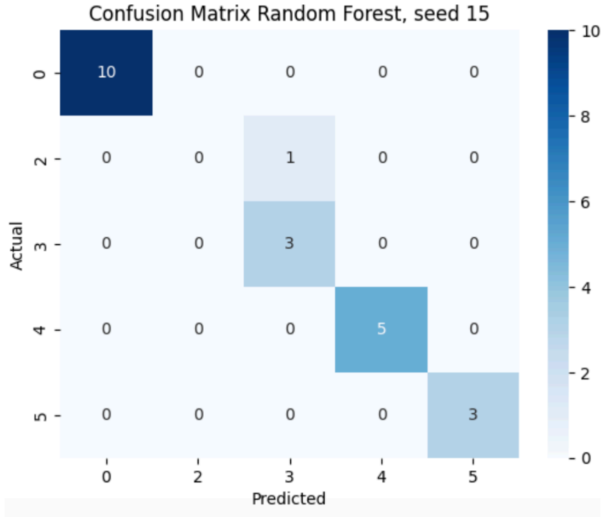


Figure 16: Confusion Matrix for Random Forest, seed 15

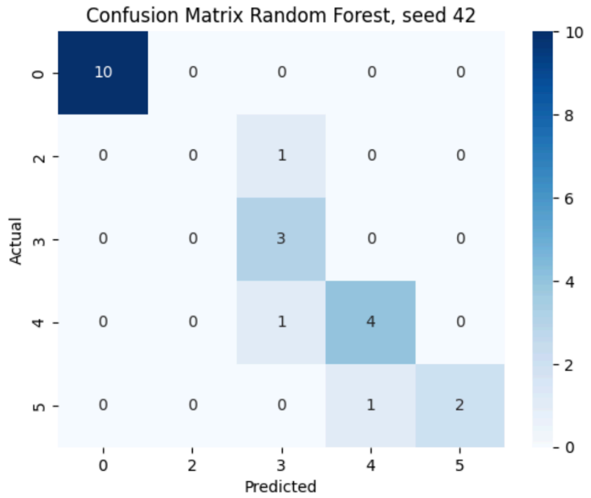


Figure 17: Confusion Matrix for Random Forest, seed 42

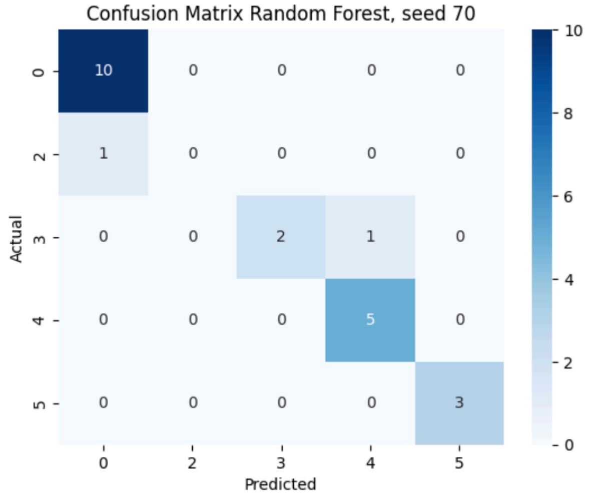


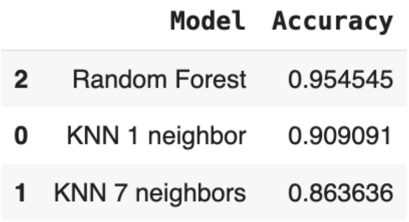
Figure 18: Confusion Matrix for Random Forest, seed 70

The Random Forest Approach yields the best accuracy, from 86% to 95%. Predicting grade 2 is also difficult for the Random Forest.

5 Comparisons

5.1 Seed 15

When randomizing the test data with the seed 15, we got the following results:



	Model	Accuracy
2	Random Forest	0.954545
0	KNN 1 neighbor	0.909091
1	KNN 7 neighbors	0.863636

Figure 19: Accuracy for models with seed 15

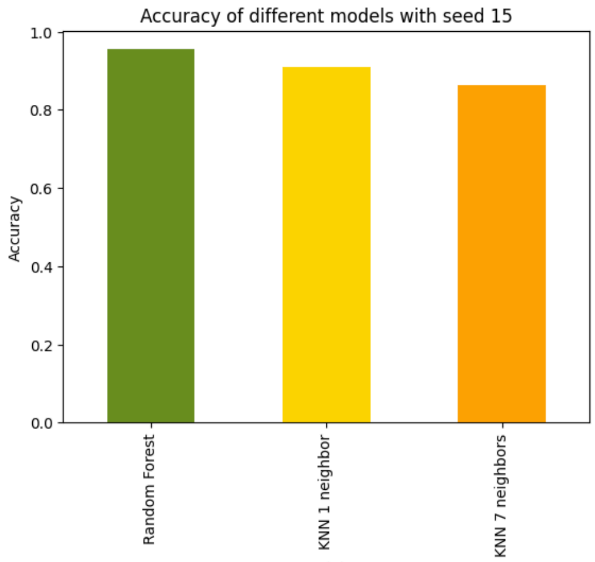
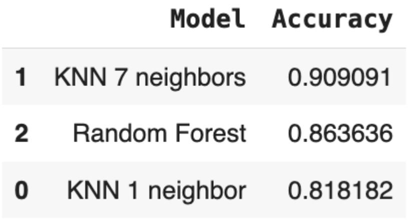


Figure 20: Accuracy for models with seed 15

With seed 15 to randomize training and test data, the Random Forest Model performed best, with an accuracy of 95.45%. The next best model was KNN with one neighbor.

5.2 Seed 42



	Model	Accuracy
1	KNN 7 neighbors	0.909091
2	Random Forest	0.863636
0	KNN 1 neighbor	0.818182

Figure 21: Accuracy for models with seed 42

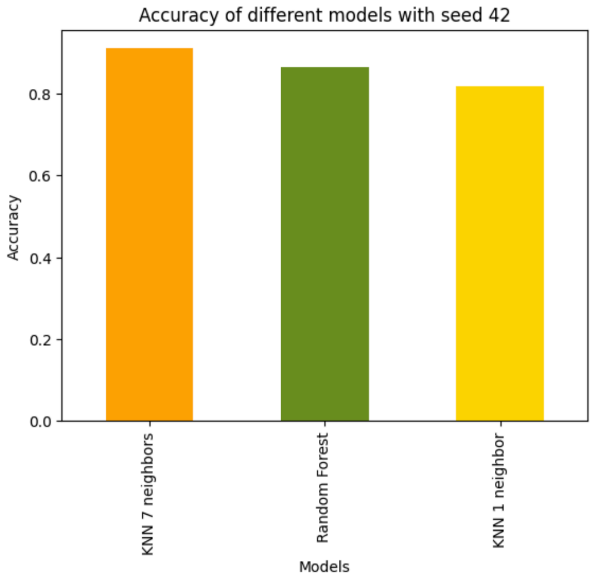
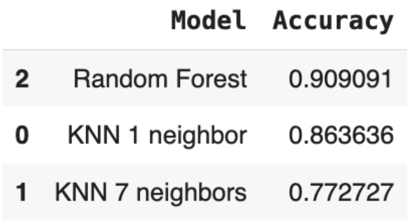


Figure 22: Accuracy for models with seed 42

When we used seed 42 for randomization, the KNN model with seven neighbors was the most accurate. The accuracy of the Random Forest model dropped significantly.

5.3 Seed 70



	Model	Accuracy
2	Random Forest	0.909091
0	KNN 1 neighbor	0.863636
1	KNN 7 neighbors	0.772727

Figure 23: Accuracy for models with seed 70

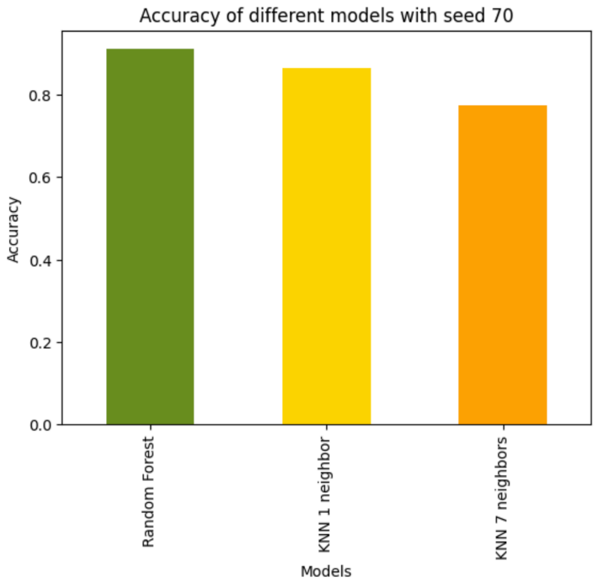


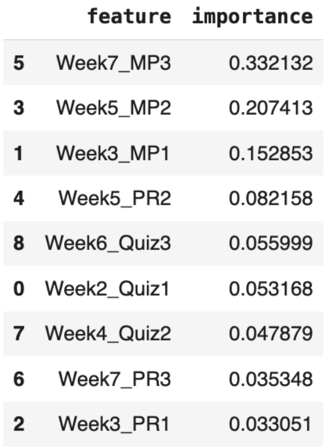
Figure 24: Accuracy for models with seed 70

In the last randomization with the seed 70, Random Forest performed best again, while KNN with 7 neighbors had the lowest accuracy.

The data shows variations within the models themselves when evaluating the performance. Generally, the Random Forest model had the highest accuracy ranging from 90.90% to 95.45%.

6 Important Features

The Random Forest was used to generate feature importance scores.



	feature	importance
5	Week7_MP3	0.332132
3	Week5_MP2	0.207413
1	Week3_MP1	0.152853
4	Week5_PR2	0.082158
8	Week6_Quiz3	0.055999
0	Week2_Quiz1	0.053168
7	Week4_Quiz2	0.047879
6	Week7_PR3	0.035348
2	Week3_PR1	0.033051

Figure 25: Important features

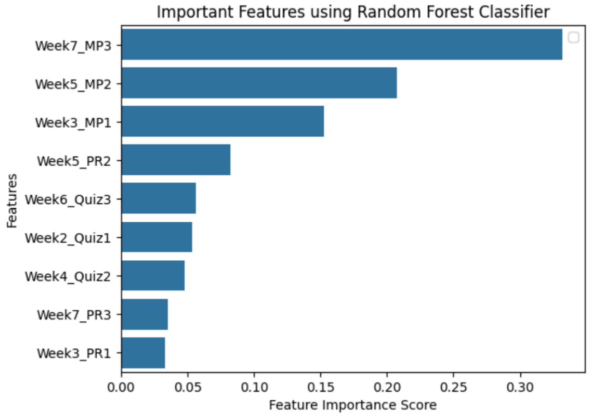


Figure 26: Important features

These scores were generated using the seed 15. All seeds had similar importance scores for the different features.

The results are aligned with the data in Figure 4. The third mini-project has the highest impact since it is worth 35 points. The second mini-project is worth 20 points, and the first 15 points. Peer reviews and quizzes are worth 5 points each, thus not impacting the final grade as significantly.

In comparison with the initial Pearson correlation analysis in Figure 1, the order is fairly the same, ranking Mini Project 3 and 2 at the top. The first peer review had a high correlation with the grade but was not ranked as high in the feature importance scores, as is the case for the third peer review.

7 Conclusions

In this paper, we have explored a data set of a course in machine learning to teach models to predict the final grade of students.

The different features were examined. As the different statuses did not provide specific information on the student's course activity, the columns were discarded even though status averages 0 and 1 correlated with the final grade.

If we had been able to get more precise information on the nature of the statuses, such as definite information that the student has finished a video lecture or opened links, etc., keeping the statuses in the analysis would have been more meaningful. We do not know if the student has finished several lectures, or just paused one lecture many times. In the same way, we cannot know if a student has opened a quiz or if it was submitted.

The correlation of status 0 and 1 with the grade is expected, as the students who completed the course have viewed the course page and done assignments, students with 0 points in all assignments are unlikely to have opened the course page to attempt to do assignments.

The dimensionality of the models could have been reduced further by combining the peer reviews and quizzes into averages. By keeping them separate, we saw differences in the correlations and importance of some of the assignments.

The data set for this assignment was small, consisting of the data from 107 students. This led to difficulties throughout the analysis, from splitting the test and training data to the model training.

The main issue of the data set was the lack of representation for grades 1 and 2, where grade 1 had no representation, and 2 only had five counts. Because of this, the models cannot predict grade 1 at all, and the accuracy for predicting grade 2 correctly is low. Stratify was used to ensure the proportionate representation of each available grade in both training and test sets.

As the data set was fairly small, three different seeds for randomization of the training and testing data sets were used to compare the performance results.

For the KNN model, StratifiedKFold was used to conclude the optimal number of neighbors K. As we got one neighbor as the best result and seven as the next best, both were used in the analysis. The model with one neighbor was the only model able to predict the right label for the 2 in the data set, even though K=1 did not predict the grade 2 correctly for all different randomized test sets.

The KNN model with K=7 used seven neighbors for comparison. This reduced the risk for overfitting but might lead to issues predicting the grade 2 in such a small data set, where there are only 4 possible neighbors to the grade 2 in each training data set.

The Random Forest used 125 n-estimators to predict the outcome of the final grade based on the features provided. It yielded a high accuracy but failed to predict grade 2 in the test set correctly for each randomized test data set.

Random Forest was also used to determine the most important features of the models, which were the three mini-projects that yield 15-35 course points each.

Lastly, the accuracies for the models were compared using different seeds for the training and test data. Random Forest generally had a high and stable accuracy, whereas KNN models had more variation for K=1 and K=7.

Based on the findings of this study Random Forest generally performed with better accuracy than the KNN models, even though KNN was more accurate for some of the randomization seeds.

The study would be more accurate with a larger data set with a broader representation of all the grade categories.

Additionally, a similar study with more informative course metadata could be interesting. With informative metadata, we could ask questions such as whether students perform well despite not finishing their video lectures or whether or not students who resubmitted assignments tended to succeed well.