PREDICTING FINAL GRADES OF PORTUGUESE HIGH SCHOOL STUDENTS IN MATHS AND PORTUGUESE

Aryana Collins Jackson

30/3/18

COMP9060 Applied Machine Learning

Assignment 1: Report

Table of Contents

[ABSTRACT 3](#_Toc510804199)

[INTRODUCTION 3](#_Toc510804200)

[RELATED RESEARCH 3](#_Toc510804201)

[ALGORITHM/MODEL DETAIL 4](#_Toc510804202)

[Datasets 4](#_Toc510804203)

[Algorithms 5](#_Toc510804204)

[EMPIRICAL EVALUATION 5](#_Toc510804205)

[Exercise 1: *Students' Habits and Grades Prediction* by Jose Ignacio Hervás Díaz 5](#_Toc510804206)

[Exercise 2: *Basic EDA and Final Grade Prediction* by Dmitriy Batogov 7](#_Toc510804207)

[Model 1: Portuguese dataset 12](#_Toc510804208)

[Regression 12](#_Toc510804209)

[Classification 14](#_Toc510804210)

[Model 2: Maths dataset 16](#_Toc510804211)

[Regression 17](#_Toc510804212)

[Classification 20](#_Toc510804213)

[Model 3: Portuguese and Maths 22](#_Toc510804214)

[Regression 22](#_Toc510804215)

[Regression with Hervás Díaz but in R 25](#_Toc510804216)

[Classification 26](#_Toc510804217)

[CONCLUSION AND FUTURE WORK 28](#_Toc510804218)

[Conclusions 28](#_Toc510804219)

[Future Work 29](#_Toc510804220)

[BIBLIOGRAPHY 30](#_Toc510804221)

ABSTRACT

In order to predict students final grades and better understand what factors impact academic performance, two datasets from Cortez and Silva[[1]](#endnote-1) were analysed. These datasets were merged using Pandas in one instance and R in another. Predictions were done in three different ways for all three datasets: prediction of passes and fails; high passes, low passes, and fails; and the individual grades themselves.

By using existing code[[2]](#endnote-2),[[3]](#endnote-3) as a basis for the prediction of individual grades, regression and classification were used. Linear regression was found to be the best option in terms of accuracy and complexity of the model and decision trees, random forests, and support vector classification models were found to be the best depending on the dataset and the level of prediction. Unfortunately, in all datasets, many of the variables proved to be irrelevant and did not explain the target variable, the final grade. Thus, with feature pruning, most of the variables were eliminated without sacrificing accuracy. This made for more accurate models but also means that a lot of external factors affecting students do not actually impact their academic performance.

The models are evaluated at the end of this report which compare and contrast the various models and the levels of prediction as well as five-fold cross-validation and nested cross-validation. Future work will include more in-depth analysis that draws direct correlations between some of the variables and the final grades of the students. This report however focuses on the comparison of models.

INTRODUCTION

In Portugal, secondary school students struggle mostly with the two core subjects, maths and Portuguese. Past academic performance is an indicator of final grades, but there may be other factors that influence academic standing, which is why these datasets were composed. There is a 20-point grading scale, where 0 is the lowest grade and 20 is the perfect score.[[4]](#endnote-4)

The predictions were done in three different ways to assess various kinds of accuracy: binary classification (pass/fail), three-level classification (high passes, low passes, and fails), and individual grades. These used classification and regression, which are both supervised learning techniques.

Regression models used included linear regression, decision tree regression, lasso, and ridge. Classification models used included decision trees, support vectors, logistic regression, and random forests. These were evaluated across all three datasets and all three levels of classification. Nested cross-validation was compared with non-nested with the support vector model to determine if there was a difference between the two.

Accuracy achieved was over 90% in some cases. However, one hundred percent accuracy is difficult to achieve. There are still other factors that will affect academic performance, such as illnesses, missing an assessment, etc.[[5]](#endnote-5)

RELATED RESEARCH

Cortez and Silva first used this data in 2008. Computations were originally done in R with twenty runs of a 10-fold cross-validation. Four models (decision trees, random forests, neural networks, and support vector machines) and various input selections (for example, with and without previous grades) were tested. The best solution was found with a Naive Bayes method with an accuracy of 74%.[[6]](#endnote-6)

Majeed and Junejo use a different dataset dealing with students. This one uses a different dataset but similar techniques, including pre-processing (removing classes with a small number of instances and missing data).There are more records in this dataset (2500), which makes it more complicated. Again, four algorithms are used, but they are ID3, K-Nearest Neighbor, Naïve Bayes, and Rule Induction method. They use RapidMiner for prediction of individual grades; there is no binary classification or five-level classification.[[7]](#endnote-7)

Lulu Cheng of the Dublin Institute of Technology also studied this material with her dissertation. In it, she explores the dataset with SPSS instead. Her research includes a lot of exploratory analysis, which was useful when understanding the breakdown of the instances included.[[8]](#endnote-8) However, this analysis does not dwell much on such analysis and instead focuses on the prediction models.

ALGORITHM/MODEL DETAIL

## Datasets

The files used were taken from the UCI Machine Learning Repository and were donated in 2014 by Silva and Cortez. They include a text file containing information on the variables contained within the sets, a CSV file containing all maths students (395 instances), a CSV file containing all Portuguese students (649 instances), and an R script for merging the two (the merged file contains 382 instances).[[9]](#endnote-9) The script ensured duplicates among both CSV files were eliminated and only instances which include both maths and Portuguese were left.

The variables in all three files are as follows:

1. school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
2. sex - student's sex (binary: "F" - female or "M" - male)
3. age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: "U" - urban or "R" - rural)
5. famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
6. Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)
7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9. Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at\_home" or "other")
10. Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at\_home" or "other")
11. reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
12. guardian - student's guardian (nominal: "mother", "father" or "other")
13. traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16. schoolsup - extra educational support (binary: yes or no)
17. famsup - family educational support (binary: yes or no)
18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. activities - extra-curricular activities (binary: yes or no)
20. nursery - attended nursery school (binary: yes or no)
21. higher - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
23. romantic - with a romantic relationship (binary: yes or no)
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. health - current health status (numeric: from 1 - very bad to 5 - very good)
30. absences - number of school absences (numeric: from 0 to 93)
31. G1 - first period grade (numeric: from 0 to 20)
32. G2 - second period grade (numeric: from 0 to 20)
33. G3 - final grade (numeric: from 0 to 20, output target)

The target variable is G3, the final grade.

## Algorithms

The algorithms were split up into regression models and classification models. For regression, decision tree regression, linear regression, ridge, and lasso models were evaluated. For classification, the decision tree classifier and support vector classification were evaluated.

EMPIRICAL EVALUATION

The data used here has been used previously in several model-building exercises both by the authors Silva and Cortez, but by others. On Kaggle, two individuals created their own models using regression to classify final grades of students in Python.[[10]](#endnote-10),[[11]](#endnote-11) A third individual documented his own experiments in modelling in R with the dataset.[[12]](#endnote-12) Another individual used Kaggle to explore a different, but similar dataset dealing with student grades.[[13]](#endnote-13) All four of these examples served as a basis and an inspiration for the final models created.

## Exercise 1: *Students' Habits and Grades Prediction* by Jose Ignacio Hervás Díaz[[14]](#endnote-14)

The code taken from Hervás Díaz begins by merging the two CSV files together in Python rather than in R. A distribution of the grades for period 1 (G1), period 2 (G2), and the final (G3) are below in Figure 1. The grades for semester 2 appear to be the most normally distributed, but none of them are too far off.

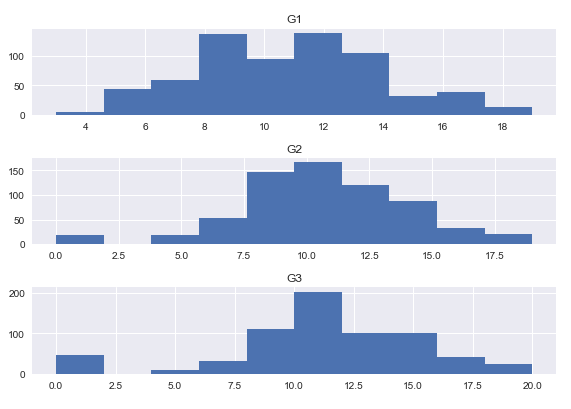


Figure 1

Unfortunately, when running the correlation analysis, only the academic performance variables (G1, G2, and G3) were found to have correlation statistics above a 0.75. Using just these values to predict G3, decision tree, linear regression, ridge, and lasso models have been evaluated for accuracy with five-fold cross validation:

Models performance in: G3

------------------------

DecisionTreeRegressor: -0.911312460027

LinearRegression: -0.0854373128615

Ridge: -0.08312185997

Lasso: -0.0713203347238

Models performance in: G2

------------------------

DecisionTreeRegressor: -1.04519561275

LinearRegression: 0.0205275512496

Ridge: 0.0220816367504

Lasso: -0.0457151878163

Models performance in: G1

------------------------

DecisionTreeRegressor: -0.918643936089

LinearRegression: 0.0688165762324

Ridge: 0.0702065387208

Lasso: -0.0552923936671

The values refer to the mean cross-validation score; the desired score is as close to zero as possible. The variable following “Models performance in:” means that that model has dropped that variable and the model has used the other two to predict G3. The strongest model is the linear regression model for the dataset that has dropped G2.

The same procedure is completed except the variable following “Models performance in:” is the variable that is dropped and then predicted.

Models performance in: G2

------------------------

DecisionTreeRegressor: 0.709300167637

LinearRegression: 0.834767268169

Ridge: 0.835090828569

Lasso: 0.843387346555

Models performance in: G3

------------------------

DecisionTreeRegressor: 0.588098964874

LinearRegression: 0.79766963931

Ridge: 0.798160300943

Lasso: 0.824593486283

These values are much higher, which confirms the theory that the grades of the previous period are a good indicator of how the grades of the next period will be.

All models are returning good results, but the linear regression model is chosen as it’s the simplest and best performing. The following results are testing whether G1 predicts G2:

Mean squared error: 3.63  
Variance score: 0.71

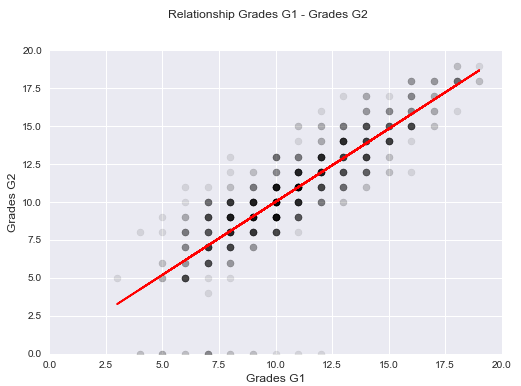


Figure 2

The mean squared error is quite small, as is variance. Figure 2 confirms that there is a very strong correlation.

The same procedure is completed for G2 and G3:

Mean squared error: 2.79

Variance score: 0.83



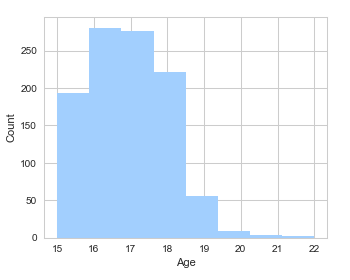
Figure 3

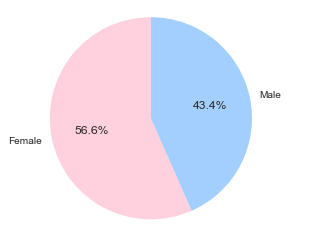
The model using G2 to predict G3 is stronger. The MSE is even smaller, which means G2 is a better predictor of G3 than G1 is. The simple linear regression model is accurately predicting the final grades with little error.

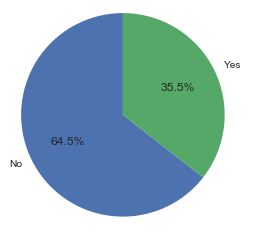
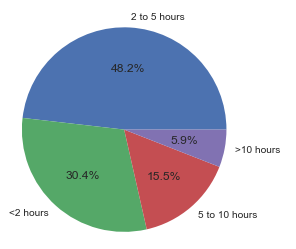
However, although the model is good, Hervás Díaz has established that only previous academic performance predicts future academic performance.

## Exercise 2: Basic EDA and Final Grade Prediction by Dmitriy Batogov[[15]](#endnote-15)

Hervás Díaz’ code is largely based on another project by Batogov. This code follows a similar structure, although only the dataset on Portuguese students is used.

First, some exploratory data analysis is completed. The breakup of the dataset is shown in Figures 4-12.



Figure 4 (Sex) Figure 5 (Age)

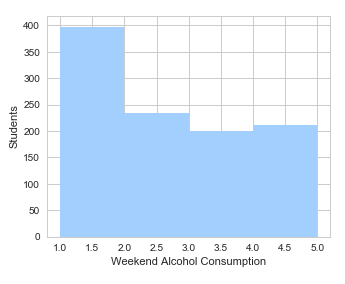
Figure 6 (Study time) Figure 7 (Romantic relationship: yes or no)



Figure 8 Figure 9

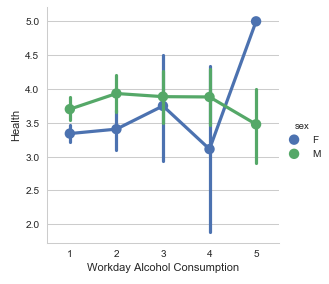
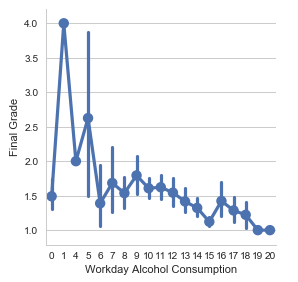


Figure 10 Figure 11

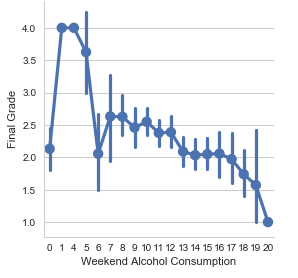


Figure 12

These visualisations show that there is good variation in the data to begin with based on a number of variables. Thus, the following analysis cannot be skewed too much due to variable and value skewness.

Like the code in the previous section, regression modelling was evaluated with four different models. The result is the mean of five-fold cross-validation for each model. Decision tree regression is strongest as shown below:

DecisionTreeRegressor: 0.538619139686

LinearRegression: 0.784308010943

Ridge: 0.784475562887

Lasso: 0.802533173483

When feature importance is run on the decision tree model to determine what features are used, the following list is returned:

1. Feature G2 (0.817393)

2. Feature absences (0.063030)

3. Feature goout (0.010598)

4. Feature age (0.009323)

5. Feature traveltime (0.007440)

6. Feature paid\_yes (0.006906)

7. Feature Dalc (0.006252)

8. Feature school\_MS (0.006234)

9. Feature Fedu (0.005784)

10. Feature G1 (0.005716)

11. Feature Walc (0.005409)

12. Feature sex\_M (0.005366)

13. Feature Medu (0.005241)

14. Feature famrel (0.004624)

15. Feature reason\_home (0.004341)

16. Feature health (0.004018)

17. Feature Mjob\_services (0.003844)

18. Feature schoolsup\_yes (0.003654)

19. Feature activities\_no (0.003650)

20. Feature failures (0.002387)

21. Feature Fjob\_services (0.002003)

22. Feature studytime (0.001663)

23. Feature nursery\_no (0.001505)

24. Feature freetime (0.001415)

25. Feature famsize\_LE3 (0.001312)

26. Feature reason\_reputation (0.001286)

27. Feature Fjob\_other (0.000739)

28. Feature address\_R (0.000703)

29. Feature famsup\_no (0.000660)

30. Feature reason\_other (0.000587)

31. Feature reason\_course (0.000550)

32. Feature Mjob\_health (0.000517)

33. Feature sex\_F (0.000507)

34. Feature famsize\_GT3 (0.000473)

35. Feature school\_GP (0.000450)

36. Feature Fjob\_at\_home (0.000440)

37. Feature Mjob\_other (0.000427)

38. Feature romantic\_no (0.000395)

39. Feature guardian\_mother (0.000366)

40. Feature Mjob\_teacher (0.000273)

41. Feature guardian\_father (0.000261)

42. Feature higher\_yes (0.000260)

43. Feature nursery\_yes (0.000257)

44. Feature paid\_no (0.000242)

45. Feature internet\_no (0.000224)

46. Feature Pstatus\_T (0.000164)

47. Feature Fjob\_teacher (0.000164)

48. Feature guardian\_other (0.000158)

49. Feature Fjob\_health (0.000149)

50. Feature famsup\_yes (0.000149)

51. Feature Pstatus\_A (0.000116)

52. Feature address\_U (0.000098)

53. Feature romantic\_yes (0.000096)

54. Feature internet\_yes (0.000093)

55. Feature higher\_no (0.000048)

56. Feature Mjob\_at\_home (0.000043)

57. Feature activities\_yes (0.000000)

58. Feature schoolsup\_no (0.000000)

G2 explains 81.74% of the G3 values. Hardly any of the other variables affect G3. The next most important feature is absences, which account for 6.30% of the G3 values. In order to determine what other variables influence the target variable, G1 and G2 are removed and the decision tree model is run again.

DecisionTreeRegressor: -0.778284231769

LinearRegression: 0.0124603819261

Ridge: 0.0131123743518

Lasso: -0.139664724063

Now, linear regression is showing as the strongest option. This confirms what Hervás Díaz found to be true – that linear regression best predicts when academic performance variables are removed. When feature importance is run again, the following list is returned.

1. Feature failures (0.164388)

2. Feature absences (0.148071)

3. Feature health (0.045753)

4. Feature studytime (0.045405)

5. Feature freetime (0.044860)

6. Feature Medu (0.038209)

7. Feature Walc (0.033597)

8. Feature paid\_no (0.029521)

9. Feature schoolsup\_no (0.027963)

10. Feature age (0.027752)

11. Feature Fedu (0.026270)

12. Feature famrel (0.025923)

13. Feature traveltime (0.025765)

14. Feature Dalc (0.024444)

15. Feature goout (0.023585)

16. Feature romantic\_yes (0.022021)

17. Feature higher\_yes (0.018595)

18. Feature Pstatus\_T (0.018188)

19. Feature sex\_M (0.014744)

20. Feature Mjob\_teacher (0.014113)

21. Feature Fjob\_services (0.013858)

22. Feature famsup\_yes (0.012551)

23. Feature reason\_course (0.010827)

24. Feature Fjob\_teacher (0.010603)

25. Feature Fjob\_at\_home (0.010593)

26. Feature Mjob\_at\_home (0.009666)

27. Feature reason\_reputation (0.009461)

28. Feature internet\_yes (0.008601)

29. Feature guardian\_mother (0.008265)

30. Feature romantic\_no (0.007518)

31. Feature guardian\_other (0.007043)

32. Feature school\_MS (0.006707)

33. Feature address\_U (0.006124)

34. Feature activities\_yes (0.005774)

35. Feature famsize\_LE3 (0.005652)

36. Feature nursery\_yes (0.005205)

37. Feature guardian\_father (0.004878)

38. Feature address\_R (0.003738)

39. Feature Mjob\_other (0.003732)

40. Feature activities\_no (0.003562)

41. Feature Mjob\_services (0.003385)

42. Feature reason\_other (0.003260)

43. Feature paid\_yes (0.003089)

44. Feature reason\_home (0.002565)

45. Feature internet\_no (0.002410)

46. Feature Fjob\_other (0.002143)

47. Feature famsize\_GT3 (0.002027)

48. Feature Fjob\_health (0.001963)

49. Feature sex\_F (0.001741)

50. Feature Mjob\_health (0.001266)

51. Feature nursery\_no (0.001021)

52. Feature famsup\_no (0.000994)

53. Feature school\_GP (0.000388)

54. Feature Pstatus\_A (0.000221)

Everything after 55 was at less than 0.000000.Now the most important feature, the number of past class failures, only accounts for 16.44% of final grade values. The resulting model would not be accurate.

## Model 1: Portuguese dataset

Using the above two analyses as well as Sklearn documentation[[16]](#endnote-16), Model 1 has been created and reflects the first model using both regression and classification. It uses only the dataset containing information on student performance in Portuguese.

Regression

First, G1 was used to predict G3 with the following results for five-fold cross-validation (with the same results for ten-fold):

Mean squared error: 3.3044

Variance score: 0.6829

Cross-val score: 0.6163

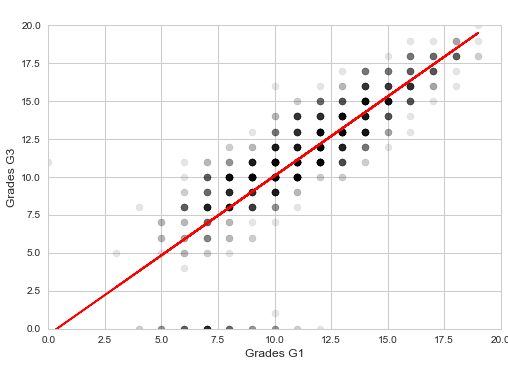


Figure 13

The same procedure was completed for G2 prediction G3. As shown, the prediction is stronger with a lower MSE, although higher variance and cross-validation score.

Mean squared error: 1.6285

Variance score: 0.8437

Cross-val score: 0.8124



Figure 14

The next strongest variables, “failures”, was used to predict G3. As expected from the results of the feature importance (in the code), the model is terrible. The MSE is quite high. Figure 15 shows why.

Mean squared error: 8.8090

Variance score: 0.1547

Cross-val score: -0.1250

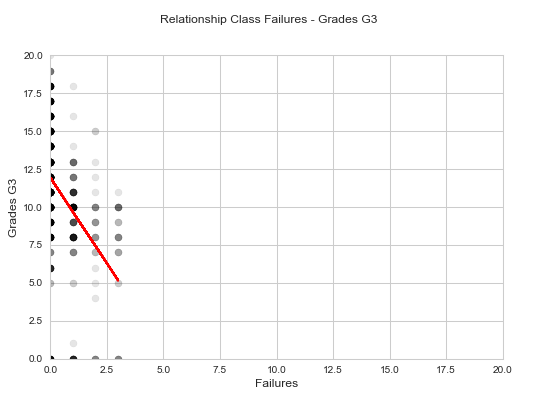


Figure 15

Classification

Classification was then completed in three levels: binary classification, three-level classification, and twenty-level classification (since the final grades are always whole numbers, individual grades can be classified easily).

PASSES AND FAILS: BINARY CLASSIFICATION

The dataset was split up into grades above 10 (passes) and those 10 or below (fails). Like code in the previous section, the results were run to determine the best model for predicting a pass or fail. Confusion matrices were also produced (not included here) which show the Type 1 and Type 2 errors.

accuracy score model

0 0.876923 Logistic Regression

1 0.907692 Decision Tree

2 0.892308 Random Forest

3 0.907692 SVC

The decision tree and support vector classification models are the strongest at 90.77% accuracy. These four models have used ten-fold cross validation. However, there’s a better way to split up the data into training and testing samples. Nested cross-validation involves one split of the data into training and testing samples and another split of the data for each of those previous splits, hence the term “nested”. Nested cross-validation does not use the same data to train and test, which means the resulting score is less inflated and the model is not overfit. The accuracy scores produced are lower but more accurate.

The code was run again with five outer splits and four inner splits (those numbers were chosen due to the length of run time and RAM required). The support vector classification model was used.

Average difference of 0.005271 with std. dev. of 0.003955.

Cross val scores mean: 0.00527110323131

Nested scores mean: 0.922915246535

Non-nested scores mean: 0.927580893683

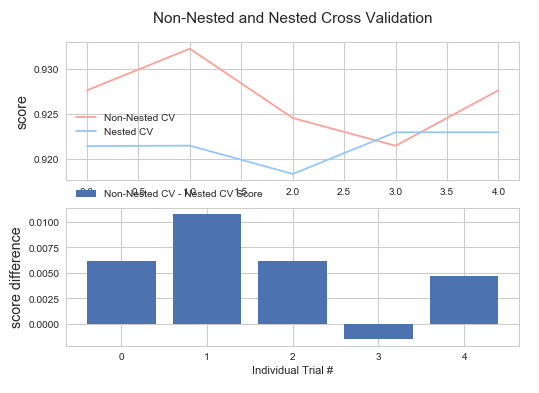


Figure 16

There was only a difference of 0.0053 between the nested and non-nested cross-validations. This may be due to the limited number of splits, however, the difference is still valuable. The accuracy of the Figure 16 visualises the information. The scores of nested and non-nested are close, but nested is slightly lower. This appears to be bad at first, but it means the score is more accurate.

The second graph shows no trend between the number of trials and the score difference. This may be because the number of trials is too low to see a general trend.

HIGH PASSES, LOW PASSES, AND FAILS: THREE-LEVEL CLASSIFICATION

This time, the data was split up into three sections: high passes (scores of 18 or 19), low passes (scores between 11 and 17), and fails (scores of 10 or below). Again, the confusion matrices were produced, but not shown here. Ten-fold cross-validation is used.

accuracy score model

0 0.815385 Logistic Regression

1 0.830769 Decision Tree

2 0.969231 Random Forest

3 0.907692 SVC

The random forest is producing an accuracy score that is much higher than the rest at 96.92%. Nested cross-validation was used again with five outer splits and four inner splits and the SVC model.

Average difference of 0.001236 with std. dev. of 0.003570.

Cross val scores mean: 0.00123553070245

Nested scores mean: 0.869025978944

Non-nested scores mean: 0.870570107858

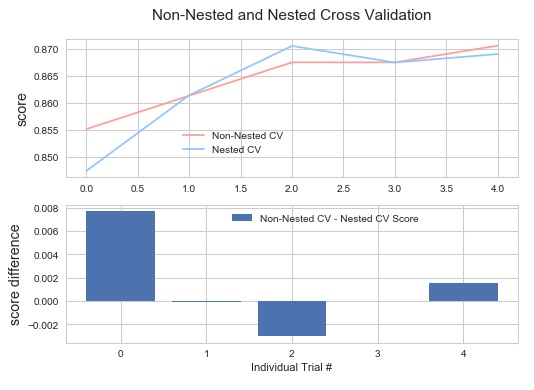


Figure 17

In this case, the scores between nested and non-nested cross-validation are closer. Figure 17 shows that they cross over a few times. The nested scores are more accurate however and do not include overfitting.

Again, the second visualisation in Figure 17 shows very little information, perhaps due to the low number of trials.

INDIVIDUAL GRADES: TWENTY-LEVEL CLASSIFICATION

The data was not split up for classification of individual grades. Each grade is a whole number, unlike percentage grades in other countries that can include decimals, so the classification is easier to do with this dataset.

accuracy score model

0 0.215385 Logistic Regression

1 0.323077 Decision Tree

2 0.553846 Random Forest

3 0.353846 SVC

Unfortunately, none of these models are much better than random choice. In fact, three of them are much worse. This is an indicator that the features in this set are very poor indicators of the target variable, G3. In the Future Work section of this report, this idea is expanded upon.

Nested-cross validation and cross-validation are compared again:

Average difference of 0.014148 with std. dev. of 0.007518.

Cross val scores mean: 0.0141483696508

Nested scores mean: 0.382147996667

Non-nested scores mean: 0.403697996918

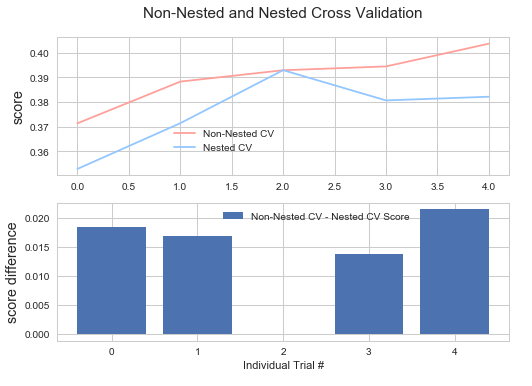


Figure 18

It is easier to see in Figure 18 that nested cross-validation scores are generally lower than the non-nested cross-validation scores. However, because all of them are so low, this information is not very valuable.

## Model 2: Maths dataset

The same regression and classification procedures were completed for the dataset involving students’ maths grades.

Regression

First, the four regression models were compared. The training data included the entire dataset minus the target variable G3.

DecisionTreeRegressor: 0.683411536806

LinearRegression: 0.784229447402

Ridge: 0.785574117351

Lasso: 0.809728494621

The linear regression and decision tree models are proving to be the best. When feature importances are run, the following list is returned.

1. Feature G2 (0.758438)

2. Feature absences (0.134064)

3. Feature age (0.024061)

4. Feature reason\_course (0.019730)

5. Feature famrel (0.012698)

6. Feature studytime (0.011445)

7. Feature failures (0.006365)

8. Feature goout (0.006051)

9. Feature health (0.005326)

10. Feature Mjob\_teacher (0.003869)

11. Feature G1 (0.003066)

12. Feature traveltime (0.001525)

13. Feature Walc (0.001305)

14. Feature Fjob\_other (0.001250)

15. Feature Mjob\_other (0.001054)

16. Feature nursery\_no (0.000938)

17. Feature Fjob\_health (0.000777)

18. Feature reason\_reputation (0.000756)

19. Feature Mjob\_at\_home (0.000578)

20. Feature reason\_other (0.000567)

21. Feature Dalc (0.000539)

22. Feature Fedu (0.000538)

23. Feature higher\_yes (0.000494)

24. Feature sex\_M (0.000488)

25. Feature address\_U (0.000474)

26. Feature address\_R (0.000451)

27. Feature Medu (0.000342)

28. Feature freetime (0.000321)

29. Feature famsize\_GT3 (0.000243)

30. Feature reason\_home (0.000206)

31. Feature activities\_no (0.000197)

32. Feature activities\_yes (0.000181)

33. Feature schoolsup\_no (0.000166)

34. Feature nursery\_yes (0.000161)

35. Feature famsup\_yes (0.000141)

36. Feature Mjob\_services (0.000126)

37. Feature guardian\_father (0.000117)

38. Feature Fjob\_teacher (0.000111)

39. Feature higher\_no (0.000097)

40. Feature schoolsup\_yes (0.000097)

41. Feature famsup\_no (0.000091)

42. Feature school\_GP (0.000091)

43. Feature romantic\_no (0.000091)

44. Feature sex\_F (0.000081)

45. Feature guardian\_mother (0.000081)

46. Feature famsize\_LE3 (0.000081)

47. Feature Fjob\_at\_home (0.000070)

48. Feature school\_MS (0.000060)

Everything after 48 was at less than 0.000000. Like the Portoguese dataset, the academic performance variables explain most of the target variable (G2 accounts for 75.84% of the G3 values). In order to see what other variables may be useful, the academic variables G1, G2, and G3 are removed and the models are compared again:

DecisionTreeRegressor: -0.900713475239

LinearRegression: -2.35542569822e+18

Ridge: 0.00829003725906

Lasso: -0.0365047676451

Now, the ridge and lasso models are the best. The results are completely different than above as expected. The next most important feature is the number of absences, which explains 23.32% of the target variable.

Using five-fold cross validation, G2 is used to predict G3 using linear regression, as it is one of the more simpler models while still accurate.

Mean squared error: 3.7940

Variance score: 0.8188

Cross-val score: 0.8100

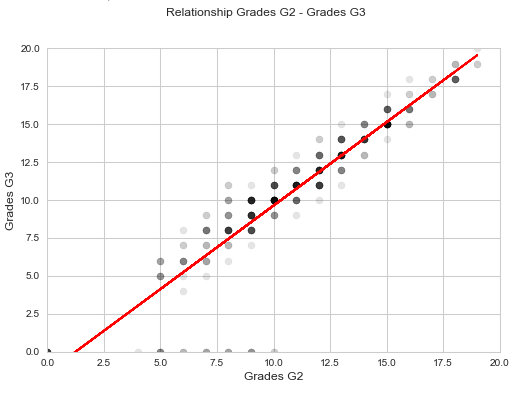


Figure 19

The MSE is quite low, and Figure 19 shows that there is an accurate prediction. G1 is then used to predict G3:

Mean squared error: 7.4879

Variance score: 0.6424

Cross-val score: 0.6188

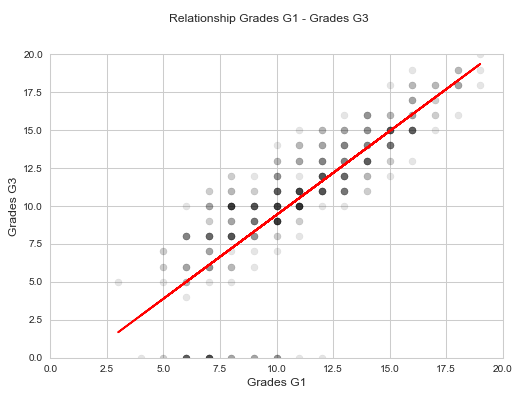


Figure 20

Unfortunately, the results are slightly worse. There is a much higher MSE. Figure 20 shows that while there is still a relationship, it is not as strong as G2 predicting G3.

The next most important feature, absences, is used to predict G3:

Mean squared error: 20.9119

Variance score: 0.0012

Cross-val score: -0.0732

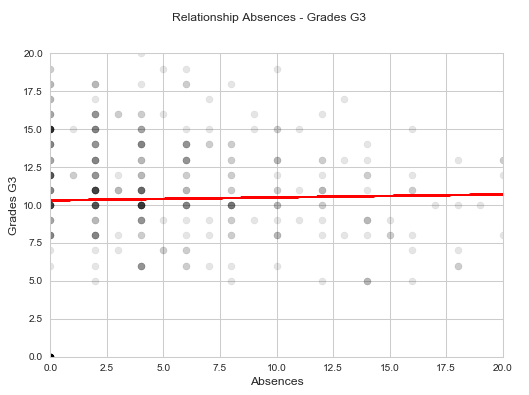


Figure 21

This model is terrible, much like the failures-G3 model in Model 1. The linear regression model and determined some relationship as shown in Figure 21, but the MSE is too high to be considered accurate.

Classification

Like Model 1, classification was completed in three levels: binary classification, three-level classification, and twenty-level classification.

Confusion matrices were produced in the code (not included here) which show the Type 1 and Type 2 errors.

PASSES AND FAILS: BINARY CLASSIFICATION

accuracy score model

0 0.925 Logistic Regression

1 0.850 Decision Tree

2 0.875 Random Forest

3 0.850 SVC

When using ten-fold cross-validation, logistic regression actually has the highest accuracy score at 92.5%. Again, nested cross-validation is compared to non-nested cross-validation. Five outer folds are used with four inner folds, and the SVC model is used for consistency with Model 1.

Average difference of 0.010084 with std. dev. of 0.006608.

Cross val scores mean: 0.0100835531758

Nested scores mean: 0.888579674294

Non-nested scores mean: 0.898734177215

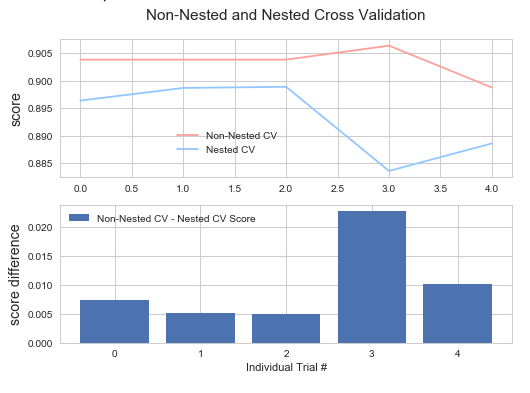


Figure 22

Nested cross-validation scores are lower with an average difference of 0.0101. Figure 22 shows that the accuracy of nested models are lower. As discussed earlier, this means that they are not as over fit.

HIGH PASSES, LOW PASSES, AND FAILS: THREE-LEVEL CLASSIFICATION

The data was split up into three sections: high passes (scores of 18 or 19), low passes (scores between 11 and 17), and fails (scores of 10 or below).

accuracy score model

0 0.850 Logistic Regression

1 0.900 Decision Tree

2 0.925 Random Forest

3 0.875 SVC

For ten-fold cross-validation, the random forest proves to be the most accurate at 92.5%. However, the nested cross-validation comparison will give a more accurate model:

Average difference of 0.006083 with std. dev. of 0.005243.

Cross val scores mean: 0.00608317742133

Nested scores mean: 0.881003916718

Non-nested scores mean: 0.896202531646

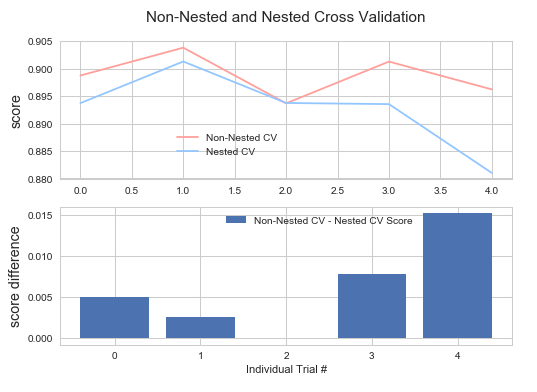


Figure 23

Unlike other cases, the second visualisation in Figure 23 shows that there may be some trend in the relationship between the number of outer folds and the accuracy score. More outer and inner folds would be needed to be sure though.

INDIVIDUAL GRADES: TWENTY-LEVEL CLASSIFICATION

For ten-fold cross-validation, the model comparison is below:

accuracy score model

0 0.200 Logistic Regression

1 0.450 Decision Tree

2 0.300 Random Forest

3 0.375 SVC

Like Model 1, these figures report terrible models. They are even worse. All of them are worse than random, which indicated that the features in the set do a very poor job of explaining G3. However, nested cross-validation is again compared for the SVC model:

Average difference of 0.004559 with std. dev. of 0.003685.

Cross val scores mean: 0.0045594018198

Nested scores mean: 0.409889713461

Non-nested scores mean: 0.410126582278

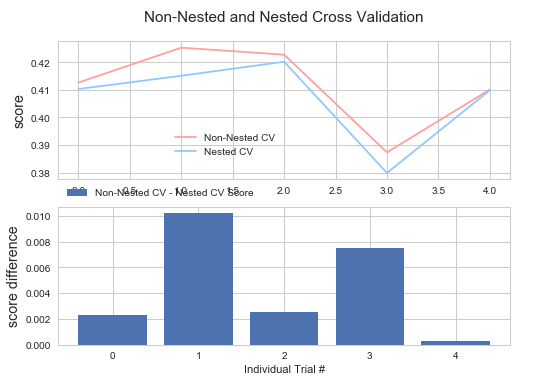


Figure 24

The scores are still terrible, there doesn’t appear to be a relationship between the trial number and the score difference, but as expected, the nested cross-validation scores are lower than the non-nested.

## Model 3: Portuguese and Maths

For the final model, the two datasets have been merged using the R script included in the original files rather than using Pandas as Hervás Díaz did. This decision was made because Cortez and Silva intended for the files to be merged this way. Also, this avoids duplication with Hervás Díaz’ code.

Regression

Using five-fold cross validation and the dataset that has excluded G3 from the training data, the following model comparison was produced:

DecisionTreeRegressor: 0.563185405569

LinearRegression: 0.71107278326

Ridge: 0.715680633524

Lasso: 0.767069639281

Unfortunately, these are significantly poorer than the comparisons from Models 1 and 2. Feature importances is then run to understand what’s going on:

1. Feature G2.y (0.803498)

2. Feature Unnamed: 0 (0.046628)

3. Feature famrel.y (0.028117)

4. Feature activities.y\_yes (0.012454)

5. Feature Walc.x (0.012292)

6. Feature guardian.x\_father (0.011187)

7. Feature failures.y (0.007493)

8. Feature sex\_F (0.006454)

9. Feature G1.y (0.006343)

10. Feature health.y (0.005620)

11. Feature address\_R (0.005532)

12. Feature traveltime.y (0.004519)

13. Feature famrel.x (0.004041)

14. Feature absences.y (0.003705)

15. Feature freetime.y (0.003679)

16. Feature G1.x (0.002894)

17. Feature Fedu (0.002677)

18. Feature Dalc.x (0.002423)

19. Feature Walc.y (0.002393)

20. Feature absences.x (0.002361)

21. Feature age (0.002277)

22. Feature traveltime.x (0.002230)

23. Feature freetime.x (0.002162)

24. Feature Medu (0.002092)

25. Feature famsize\_LE3 (0.001956)

26. Feature address\_U (0.001848)

27. Feature studytime.y (0.001672)

28. Feature studytime.x (0.001414)

29. Feature G3.x (0.001180)

30. Feature reason\_course (0.001139)

31. Feature paid.x\_yes (0.001076)

32. Feature Mjob\_services (0.000928)

33. Feature Mjob\_teacher (0.000803)

34. Feature Fjob\_health (0.000793)

35. Feature Fjob\_other (0.000550)

36. Feature activities.y\_no (0.000544)

37. Feature famsup.x\_no (0.000429)

38. Feature romantic.y\_yes (0.000407)

39. Feature reason\_reputation (0.000393)

40. Feature activities.x\_no (0.000353)

41. Feature failures.x (0.000330)

42. Feature reason\_home (0.000259)

43. Feature internet\_yes (0.000252)

44. Feature nursery\_yes (0.000227)

45. Feature school\_GP (0.000227)

46. Feature sex\_M (0.000151)

Everything after 46 was at less than 0.000000. As expected, the academic variable G2.y explains most of the data, 80.35%. However, what is making the models so poor is that the next most important feature, an unnamed one, explains just 4.66% of G3 values. This unnamed variable is explored more in the Future Work section.

Running the comparison again on the data without the academic performance variables G1, G2, and G3 yields the following results:

DecisionTreeRegressor: -0.899467653593

LinearRegression: -0.0399393498397

Ridge: 0.0340567099008

Lasso: -0.167274075028

Now it appears that the lasso model is the strongest. When feature importances are run, however, the number of past failures is shown as the highest. It only explains 18.56% of the final grades, which means the model would not be very accurate.

Like models 1 and 2, G1 is used to predict G3 with the following results:

Mean squared error: 7.4879

Variance score: 0.6424

Cross-val score: 0.6267

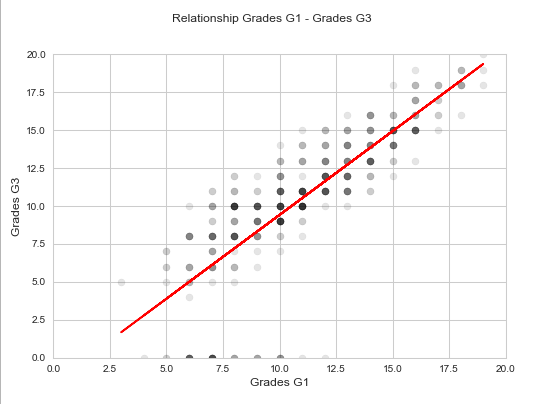


Figure 25

The MSE is higher than preferred, but Figure 25 shows that there is still a strong relationship. The same process is repeated with G2 and G3.

Mean squared error: 1.7612

Variance score: 0.7965

Cross-val score: 0.7772



Figure 26

The MSE is much lower, indicating a much better model. Figure 26 confirms this. Visually, it is apparent that there is less variation in the data. The procedure is completed for the next most important feature, failures, and G3.

Mean squared error: 7.3779

Variance score: 0.1474

Cross-val score: -0.1037

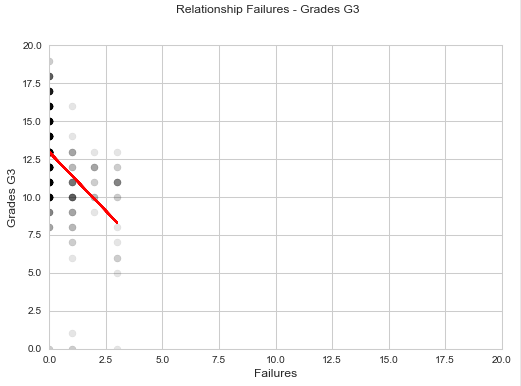


Figure 27

As with Models 1 and 2, this non-academic performance variable does a very poor job at prediction G3.

Regression with Hervás Díaz but in R

At this point, the code from Hervás Díaz comes back into play. His code is run again, except the file has been merged in R, not with Pandas. In his code, only three variables showed correlations above 0.75. In this iteration of the code, however, many more variables show correlations above 0.75 (not shown but included in the code).

The regression model is run again, only using the dataset minus the following variables with the low correlations to see the outcome.

goout.x goout.y --> 0.936608021198

famrel.y famrel.x --> 0.9703337168

freetime.x freetime.y --> 0.97442380099

studytime.x studytime.y --> 0.977999631884

The regression models are then compared. The variable following “Models performance in:” refers to the model leaving that variable out and predicting G3.

Models performance in: G3.y

------------------------

DecisionTreeRegressor: -0.427308646342

LinearRegression: 0.142081751089

Ridge: 0.217047512204

Lasso: 0.250043161244

Models performance in: G2.y

------------------------

DecisionTreeRegressor: -0.145887388637

LinearRegression: 0.252547595232

Ridge: 0.356072733699

Lasso: 0.316640673154

Models performance in: G1.y

------------------------

DecisionTreeRegressor: -0.412071088872

LinearRegression: -0.010294079965

Ridge: 0.229062319081

Lasso: 0.304716660409

The best model here is the linear regression when removing G1 (G2 predicting G3). This is in line with what Hervás Díaz found. Although he found only three variables with correlations and this code found many more, the results would be roughly the same in the rest of the model comparison.

Classification

Like Models 1 and 2, classification was completed in three levels: binary classification, three-level classification, and twenty-level classification. The confusion matrices produced in the code are not included here, but they show the Type 1 and Type 2 errors.

PASSES AND FAILS: BINARY CLASSIFICATION

accuracy score model

0 0.871795 Logistic Regression

1 0.846154 Decision Tree

2 0.923077 Random Forest

3 0.846154 SVC

The random forest has the highest accuracy with ten-fold cross-validation at 92.31%. As with the previous models, nested cross-validation would provide a more accurate accuracy score. The comparison of nested and non-nested cross validation is below.

Average difference of 0.003659 with std. dev. of 0.003570.

Cross val scores mean: 0.00365897974649

Nested scores mean: 0.916255482456

Non-nested scores mean: 0.921465968586

The average difference is only 0.0037, but as expected, the nested cross-validation scores are almost consistently lower (and more accurate) as shown in Figure 28.

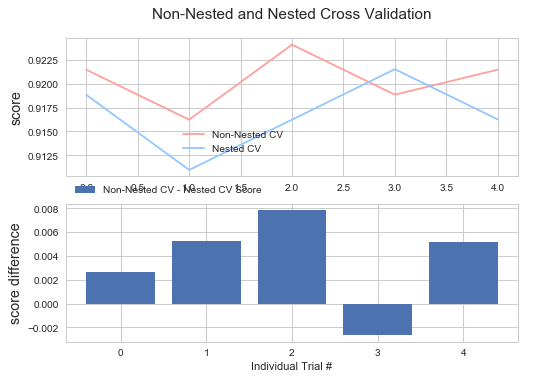


Figure 28

HIGH PASSES, LOW PASSES, AND FAILS: THREE-LEVEL CLASSIFICATION

Again, the data was split up into three sections: high passes, low passes, and fails. Confusion matrices were produced, but are not shown here. Ten-fold cross-validation is used to produce the following results:

accuracy score model

0 0.794872 Logistic Regression

1 0.820513 Decision Tree

2 0.846154 Random Forest

3 0.692308 SVC

These result are consistent with the binary classification results in that they are significantly lower than the results produced in Models 1 and 2. The random forest is again the strongest model at 84.62% accuracy.

Nested and non-nested cross-validation is compared again below.

Average difference of 0.004153 with std. dev. of 0.003588

Cross val scores mean: 0.00415323321392

Nested scores mean: 0.780180921053

Non-nested scores mean: 0.787958115183

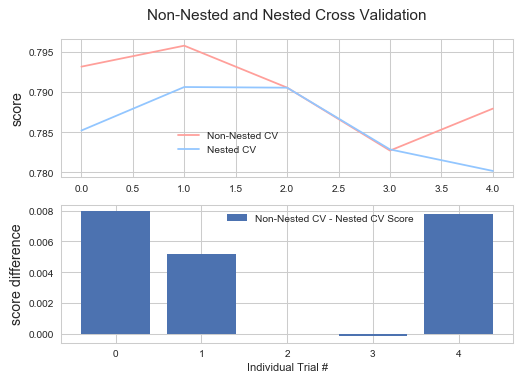


Figure 29

As with the rest of the results thus far, the average difference is low, but the nested cross-validation accuracy scores are lower than the non-nested scores. Interestingly here, the nested and non-nested scores are exactly the same or nearly the same (shown in Figure 29) during trials 2 and 3.

INDIVIDUAL GRADES: TWENTY-LEVEL CLASSIFICATION

The dataset is not split up for the final classification model; the prediction of individual grades.

accuracy score model

0 0.256410 Logistic Regression

1 0.230769 Decision Tree

2 0.333333 Random Forest

3 0.051282 SVC

The results are all less than 50%, indicating variables that are extremely irrelevant. As shown in the feature importances code, many of the attributes are irrelevant to the variable G3, so these results are expected.

When nested cross-validation is compared to non-nested, the following results are produced:

Average difference of 0.024044 with std. dev. of 0.007141

Cross val scores mean: 0.024044359098

Nested scores mean: 0.130975877193

Non-nested scores mean: 0.164921465969

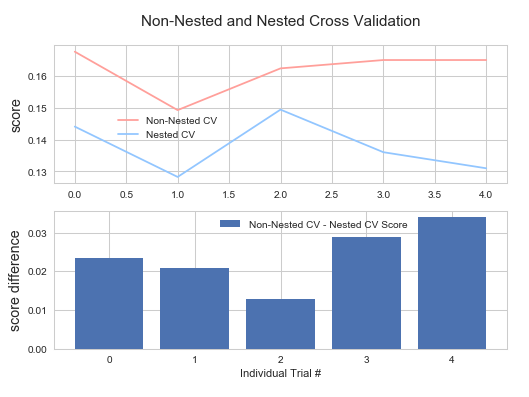


Figure 30

Nested cross-validation produces a lower accuracy score than non-nested, as expected. However, like the individual predictions from Models 1 and 2, these results are all useless because of the inaccuracy.

CONCLUSION AND FUTURE WORK

## Conclusions

Cortez and Silva’s original work explores what factors impacted student grades. The datasets included twenty-nine feature variables and one target variable, the final grade, for students in maths, Portuguese, and both. By exploring regression and classification models as well as three different levels of prediction, the following table was produced. It shows the recommended model given the dataset, the prediction type, and the model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Models 1, 2, and 3** | | |
| **Classification** | **Model Type** | *Portuguese* | *Maths* | *Both* |
| *Binary* | Decision Tree and SVC: 90.77% accuracy | Logistic Regression: 92.50% accuracy | Random Forest: 92.31% accuracy |
| *Three-level* | Random Forest: 96.92% accuracy | Random Forest: 92.50% accuracy | Random Forest: 84.66% accuracy |
| *Twenty-level* | Random Forest: 55.38% accuracy | Decision Tree: 45.00% accuracy | Random Forest: 33.33% accuracy |
| **Regression** | *G1* | Linear Regression: MSE 3.3044 | Linear Regression: MSE 7.4879 | Linear Regression: MSE 7.4879 |
| *G2* | Linear Regression: MSE 1.6285 | Linear Regression: MSE 3.7940 | Linear Regression: MSE 1.7612 |
| *Other* | Linear Regression: MSE 8.8090 | Linear Regression: MSE 20.9119 | Linear Regression: MSE 7.3779 |

*Figure 31*

This table summarises the findings in the rest of this report, aside from the differences between nested and non-nested cross-validation. Linear regression was chosen as the model of choice for all datasets for consistency purposes and also because it either had the best cross-validation score or the second best. It is one of the more simple models with less runtime as well. The mean square error was used as the metric to compare regression models. As seen from the table, using a factor other than G1 or G2 to predict G3 resulted in models that were not accurate.

This does not mean that there is no correlation between the other variables and the target variable G3. As Cheng explored in her dissertation, there is a relationship between the final grade and other variables, but they require other analysis such as ANOVAs to find.[[17]](#endnote-17)

As shown in the Hervás Díaz section in Model 3, it does not really matter whether the code is combined with Pandas or in R.

The classification models are produced acceptable results, except for the twenty-level classification (predicting individual final grades). This means that other features are great at determining passes or fails or three different levels, but cannot predict individual grades. Hervás Díaz and Batogov’s analyses ended with the conclusion that there were no correlations between any non-academic variables and the final grades. They were not entirely wrong, as they discovered that the features are poor at prediction individual grades, which is true, but they did not develop models that involved other kinds of prediction, as this report as done.

Ultimately, analysis such as this should be used by a school system or student coordinator to assist students in achieving academic success. As Cortez and Silva say, “There is a potential for an automatic on-line learning environment, by using a student prediction engine as part of a school management support system. This will allow the collection of additional features (e.g. grades from previous school years) and also to obtain a valuable feedback from the school professionals.[[18]](#endnote-18)

## Future Work

The dissertation by Cheng and the original paper by Cortez and Silva, as well numerous other papers on similar datasets, show that data of this kind can be explored and analysed almost to no end. The research is this paper are touching just the tip of the iceberg. Given more time, the following research could have been completed:

* A regression model improved from Batogov’s work according to feature importances and further feature pruning.
* More folds in nested cross-validation. Unfortunately the machine this code was run from was unable to process larger nests.
* Nested cross-validation could have been run on the regression models as well as other classification models aside from SVC.
* Nested cross-validation could have been explored in conjunction with feature pruning. It would have been interesting to see an iteration in process in which unimportant features were removed one at a time.
* The unnamed variables that appeared in Model 3 could have been explored further. There was no analysis done to find out what they were.
* As Cheng did, further analysis (such as ANOVAs) are useful to explore the correlations not only between the variables and the target variable but between the other variables themselves.

BIBLIOGRAPHY

A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification", *AAAI-98 workshop on learning for text categorization.*, vol. 752, no. 1, pp. 41-48, 1998.

D. Batogov, "Basic EDA and Final Grade Prediction", *Kaggle*, 2017. [Online]. Available: https://www.kaggle.com/dmitriy19/basic-eda-and-final-grade-prediction. [Accessed: 01-Apr 2018].

E. Majeed and K. Junejo, "Grade Prediction Using Supervised Machine Learning Techniques", *e Proceeding of the 4th Global Summit on Education 2016*, pp. 222-234, 2016. [Accessed: 31 Mar-2018].

J. Hervás Díaz, "Students' Habits and Grades Prediction", *Kaggle*, 2017. [Online]. Available: https://www.kaggle.com/zelzebu/students-habits-and-grades-prediction. [Accessed: 31- Mar 2018].

L. Cheng, "Exploring the Factors that Affect Secondary Student’s Mathematics and Portuguese Performance in Portugal", Masters, Dublin Institute of Technology, School of Computing, 2017. [Online]. Available: https://arrow.dit.ie/cgi/viewcontent.cgi?article=1103&context=scschcomdis. [Accessed: 31 Mar-2018].

M. Gregory, "Student performance in Portugal", *Machine Gurning*, 2016. [Online]. Available: http://www.machinegurning.com/rstats/student-performance/. [Accessed: 31- Mar- 2018].

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39 7. [Online]. Available: http://www3.dsi.uminho.pt/pcortez/student.pdf. [Accessed: 22-Mar 2018].

P. Cortez and A. Silva. Student Performance Dataset. UCI Machine Learning Repository (Donated 2014). [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Student+Performance>. [Accessed: 22-Mar-2018].

R. Ruizendaal, "Kaggle Kernels: Predicting Students’ Grades", *Towards Data Science*, 2017. [Online]. Available: <https://towardsdatascience.com/predicting-students-grades-on-kaggle> fd6ac9b1bfb9. [Accessed: 31- Mar- 2018].

"scikit-learn: machine learning in Python — scikit-learn 0.16.1 documentation", Scikit-learn.org, 2017. [Online]. Available: http://scikit-learn.org/. [Accessed: 06- Apr- 2018].

1. P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-397. [↑](#endnote-ref-1)
2. D. Batogov, "Basic EDA and Final Grade Prediction", *Kaggle*, 2017. [↑](#endnote-ref-2)
3. J. Hervás Díaz, "Students' Habits and Grades Prediction", Kaggle, 2017. [↑](#endnote-ref-3)
4. P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-6, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-397. [↑](#endnote-ref-4)
5. E. Majeed and K. Junejo, "Grade Prediction Using Supervised Machine Learning Techniques", e Proceeding of the 4th Global Summit on Education 2016, pp. 222-234, 2016. [↑](#endnote-ref-5)
6. P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. , Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-397. [↑](#endnote-ref-6)
7. E. Majeed and K. Junejo, "Grade Prediction Using Supervised Machine Learning Techniques", e Proceeding of the 4th Global Summit on Education 2016, pp. 222-234, 2016. [↑](#endnote-ref-7)
8. L. Cheng, "Exploring the Factors that Affect Secondary Student’s Mathematics and Portuguese Performance in Portugal", Masters, Dublin Institute of Technology, School of Computing, 2017. [↑](#endnote-ref-8)
9. P. Cortez and A. Silva. Student Performance Dataset. UCI Machine Learning Repository (Donated 2014). [↑](#endnote-ref-9)
10. D. Batogov, "Basic EDA and Final Grade Prediction", *Kaggle*, 2017. [↑](#endnote-ref-10)
11. J. Hervás Díaz, "Students' Habits and Grades Prediction", Kaggle, 2017. [↑](#endnote-ref-11)
12. M. Gregory, "Student performance in Portugal", *Machine Gurning*, 2016. [↑](#endnote-ref-12)
13. R. Ruizendaal, "Kaggle Kernels: Predicting Students’ Grades", *Towards Data Science*, 2017. [↑](#endnote-ref-13)
14. J. Hervás Díaz, "Students' Habits and Grades Prediction", *Kaggle*, 2017. [↑](#endnote-ref-14)
15. D. Batogov, "Basic EDA and Final Grade Prediction", *Kaggle*, 2017. [↑](#endnote-ref-15)
16. "scikit-learn: machine learning in Python — scikit-learn 0.16.1 documentation", *Scikit-learn.org*, 2017. [↑](#endnote-ref-16)
17. L. Cheng, "Exploring the Factors that Affect Secondary Student’s Mathematics and Portuguese Performance in Portugal", Masters, Dublin Institute of Technology, School of Computing, 2017, pp 35-36 [↑](#endnote-ref-17)
18. P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-397. [↑](#endnote-ref-18)