Assignment 2

MACHINE LEARNING

Fergus Walsh*

Due 5th March, Epiphany Term, 2021

Executive Summary

This enquiry examines data collected in 2005–2006 from 649 students in two secondary schools in Alentejo, Portugal: the students' self-reported details about their behaviours, background and academic performance were collated with their final Portuguese language exam results. These exam results were converted into two categories, pass and fail with a pass mark of 10⁄20. The data is explored in more detail in section 1.

A variety of machine learning techniques were employed to explore the relationships between the predictor variables provided by the students and their exam results, and also to develop a model to accurately predict a new student's result based on such information. The first method used was a single classification tree (section 2.1.1), despite being easy to interpret, these models were unable to accurately model the relationships between the predictor varaibles and exam results and unable to produce useful predictons. These models should not be used in any further analysis of the data. The second method (section 2.1.2) used was a random forest, which effectively modelled the importance of each predictor variable and was able to provide effective predictons. This model was used in analysis of the data (section 3.2) and is recommended for any future analysis or similar scenarios. A neural network was used to predict future students' exam performance (section 2.2) and was found to be the most accurate method for predicting students' results. However, this method is unable to show the importance of the predictors and as such is unsuitable for detailed analysis of the data. The modelling techniques are compared in section 3.1.

Finally, the data seemed to show that a student's home and social life, and their desire to pursue higher education were the most important predictors of exam results, and that extra academic support or socio-economic background had little influence on predicting a student's result.

1 Introduction

This enquiry will examine the Student Performance data set from the UCI Machine Learning Repository; the data set records the end-of-year exam results of 395 students in their mathematics exams and 649 students in their Portuguese language exams, in addition to school, familial, socio-economic and behavioural statistics.[1]

The variables of the data set are summarised in Table 1. The data was collected during the 2005–2006 academic year from students at the Gabriel Pereira and Mousinho da Silveira state-maintained secondary schools in Alentejo, Portugal. The three exam marks (`G1`, `G2`, `G3`) and the absence figures (`absences`) were provided by the two schools for each student, while the other variables were collected from questionnaires filled in by the students themselves.[2]

This enquiry will examine the results of the 649 students whose Portuguese exam marks were collected, and attempt to discern which factors are significant predictors of whether a student passes or fails their exam. Although the results for both mathematics and Portuguese are available for 382 out of the total 1,031 students surveyed, only

[*] R code used can be found at `github.com/FergusJPWalsh/Master-of-Data-Science/blob/main/MLAssignment2FJPW.R`

[1] Cortez 2014; Dua and Graff 2017.

[2] Cortez and Silva 2008, p. 2.

the data for the 649 students whose Portuguese language results are available will be considered in this enquiry, for the sake of simplicity and brevity. Naturally, the data set with the greater number of observations (i.e. the Portuguese exam results) was chosen in the belief that it would lead to a better model.

It is here necessary to acknowledge that Cortez and Silva, who published the data set, include an article in which they analyse the data. In their analysis, the data for both mathematics and Portuguese results were examined using a variety of modelling methods, both with and without `G1` and `G2`, each student's marks from the first and second academic terms, which naturally serve as very good predictors for `G3`, their final mark. Again, for simplicity's sake, `G1` and `G2` will not be included in this enquiry, even though the resulting models will have a worse performance.[3]

Many of the variables are categorical, including those that might more naturally be recorded continuously (such as `famsize`, `traveltime` and `studytime`). As in France, the Portuguese exam marks are given out of 20, with 0 the lowest score and 20 the highest. In order to use the final exam results in a classification analysis a new variable `pass` was created using the `G3` variable, where `Pass` is a mark > 10 and `Fail` ≤ 10.[4] This new variable, along with the other categorical variables, was then transformed using the `as.factor()` function so that R would correctly interpret the values in a classification setting.

Pages 3–5 show exploratory plots of a select number of variables from the data set: Figure 1 shows a histogram of `G3`, which follows a normal distribution centred just above the pass mark with a mean score of 11.906 (3 d.p.). This equates to 549 students passing and 100 failing (Figure 1b). Figure 2 details the distributions of the `school` and `address` variables: 423 of the surveyed students attended Gabriel Pereira school (65%) and 226 Mousinho da Silveira (35%), likewise 197 students (30%) reported that they lived in a rural location and 452 in an urban location (70%). It should be noted that Cortez and Silva do not provide details as to the sizes of the two schools' student body (Is one school over-represented? Why were more students from Gabriel Pereira surveyed than from Mousinho da Silveira?) or whether these schools are located in an urban or rural location. Such details could be divined by correlation from other variables such as `studytime` but the lack of contextual information here is worth bearing in mind. Figures 3 and 4 detail the distributions of students' mothers' and fathers' education and career-sector. Only a few categories were used in the survey, and as such it is no suprise that `other` is the most popular in each case. Figure 4 shows the education levels of both parents, and the mostly even distribution across classes `1-4` for both sexes is remarkable. It may not be wrong to infer from these plots that the students at both schools come a cross-section of socio-economic classes, but again such broader contextual information is lacking. Figure 5 shows students' self-reported alcohol consumption habits on the weekends and weekdays. While both plots show an inverse power relationship, the difference is most stark in Figure 5b, where a larger majority of students consume very little alcohol during the work-week.

Finally, Figure 6a plots the distribution of `failures`, how many times a student has previously taken (and failed) the Portuguese exam. Unsuprisingly the vast majority of students (549 out of 649, 85%) are taking the exam for the first time. Figure 6b is a histogram of the students' absences from school: 244 students have never missed a school day, the mean is 3.659 times per student and two students were recorded absent for 32 and 30 days.

2 Analysis & Modelling

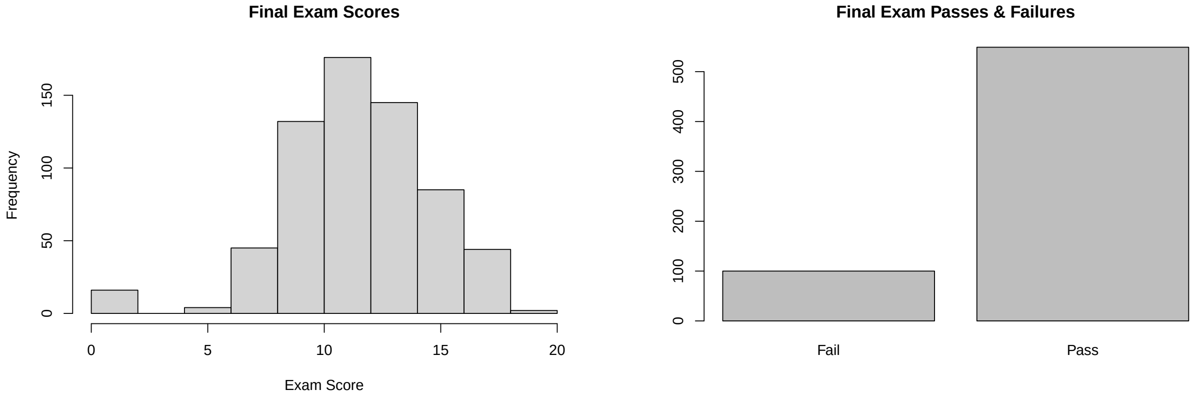
2.1 Classification Tree Models

In the introduction, it was noted that the continous variable `G3` was transformed into a binary categorical variable `pass` with values `Pass` and `Fail`. While a linear regression of the predictors onto `G3` would allow for the full range of possible marks to be modelled, due to all but three predictors being reported as categorical variables, the proliferation of dummy variables would make such a model cumbersome to interpret. For the purposes of this enquiry, a classification tree appears to be an ideal modelling method, even if it necessitates reducing the full range of marks to pass and fail.[5] Indeed, were many of the predictors recorded as quantitative variables, a linear regression might outperform a classi-

[3] Cortez and Silva 2008, pp. 5–6.

[4] Taking 10 as the pass-mark follows Cortez and Silva's method, but another value could also be chosen. Cortez and Silva 2008, p. 3.

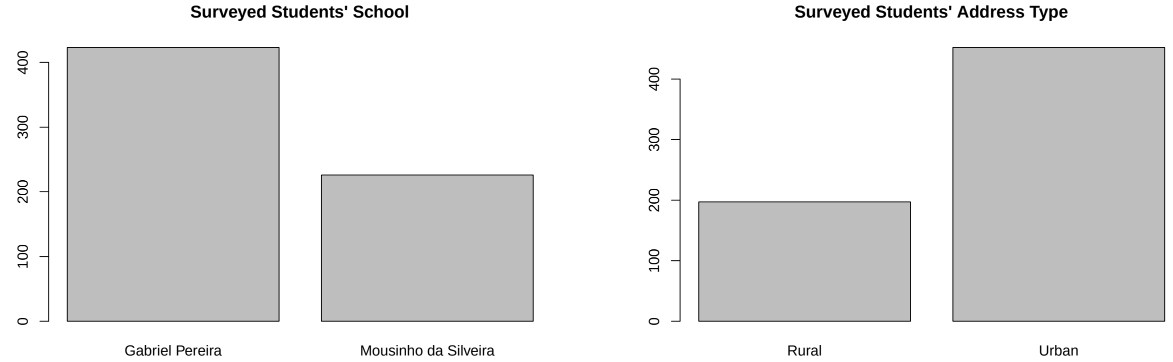
[5] Note, however, Cortez and Silva's analysis includes a five-level response classification and a regression tree for the full range of marks. Cortez and Silva 2008, pp. 1–2.



(a) Final exam scores (of out 20)

(b) Final exam results with 10 as pass-mark.

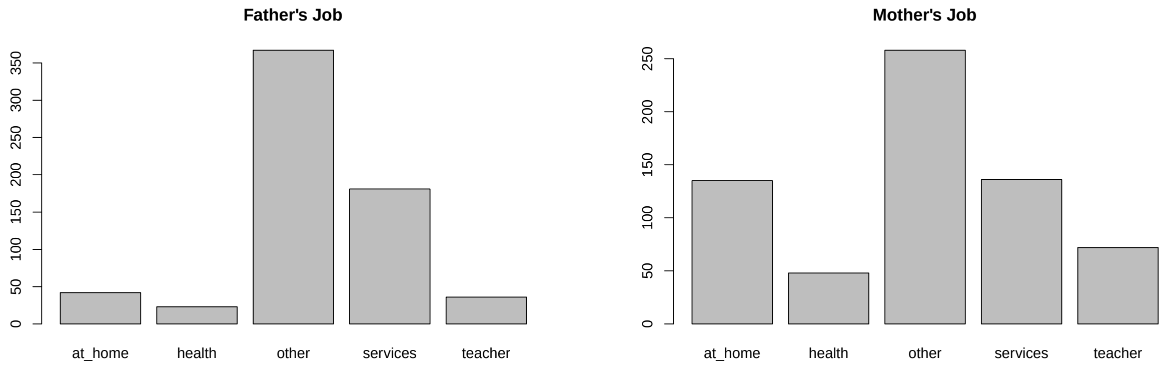
Figure 1: Result variables `G3` and `pass`.



(a) Surveyed students’ school.

(b) Surveyed students’ address type.

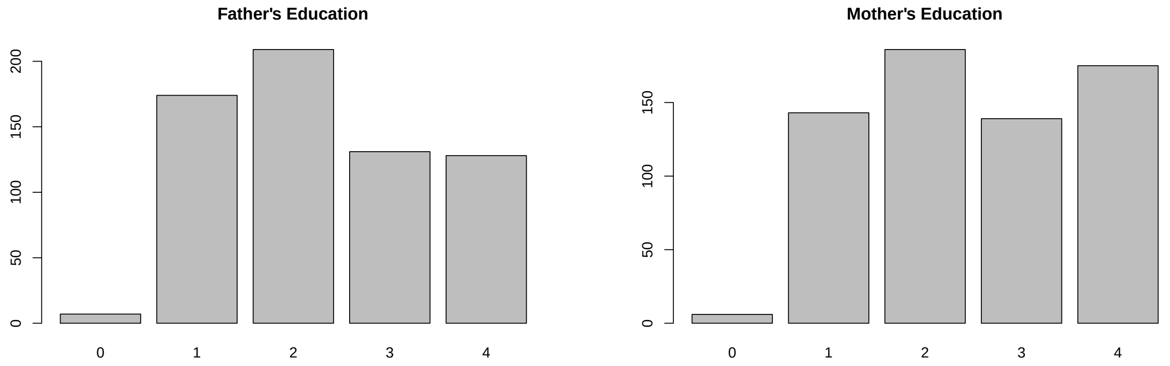
Figure 2: Variables `school` and `address`.



(a) Father's job. See Table 1 for definitions of categories.

(b) Mother's job. See Table 1 for definitions of categories.

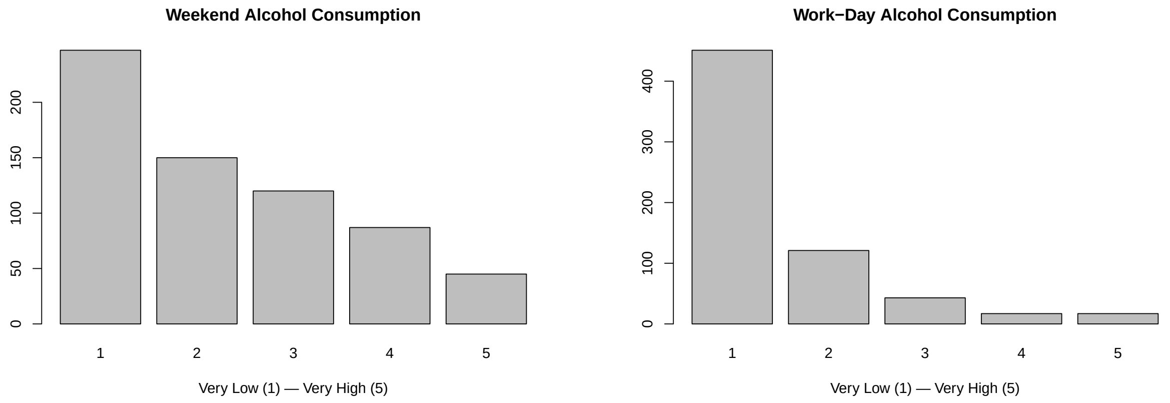
Figure 3: Variables `Fjob` and `Mjob`.



(a) Father's eduction. See Table 1 for definitions of categories.

(b) Mother's eduction. See Table 1 for definitions of categories.

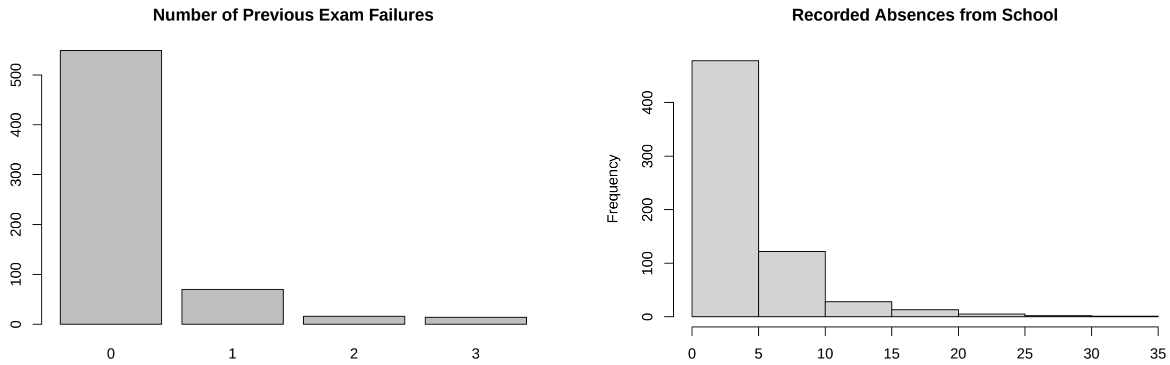
Figure 4: Variables `Fedu` and `Medu`.



(a) Weekend alcohol consumption.

(b) Week-day alcohol consumption.

Figure 5: Variables `Walc` and `Dalc`.



(a) Number of times a student has previously taken the exam.

(b) Recorded absences from school.

Figure 6: Variables `failures` and `absences`.

Variable	Description	Note
school	Student's school	GP: Gabriel Pereira, MS: Mousinho da Silveira
sex	Student's sex	F, M
age	Student's age	15–22
address	Home address type	U: urban, R: rural
famsize	Family size	GT3: larger than 3, LE3: 3 or fewer
Pstatus	Parent's cohabitation status	T: together, A: living apart
Medu	Mother's education	0–4 (See note below)
Fedu	Father's education	0–4
Mjob	Mother's job	teacher, health (health care or related), service (civil service or police), at_home, other
Fjob	Father's job	teacher, health (health care or related), service (civil service or police), at_home, other
reason	Reason for choosing this school	home (close to home), reputation, course (availablity of course or subject), other
guardian	Student's guardian	mother, father, other
traveltime	Home-school travel time	1: $<$ 15 min, 2: 15–30 min, 3: 30–60 min, 4: $>$ 60 min
studytime	Weekly study time	1: $<$ 2 hr, 2: 2–5 hr, 3: 3–5 hr, 4: $>$ 10 hr
failures	Number of times failed the exam	0–3
schoolsup	Extra education support given by school	yes, no
famsup	Extra education support given by family	yes, no
paid	Extra paid classes	yes, no
activities	Participant in extra-curricular activities	yes, no
nursery	Attended nursery school	yes, no
higher	Wants to pursue tertiary education	yes, no
internet	Internet access at home	yes, no
romantic	Engaged in a romantic relationship	yes, no
famrel	Quality of family relationships	1 (very bad) – 5 (very good)
freetime	Free time after school	1 (very low) – 5 (very high)
goout	Going out with friends	1 (very low) – 5 (very high)
Dalc	Work-day alcohol consumption	1 (very low) – 5 (very high)
Walc	Weekend alcohol consumption	1 (very low) – 5 (very high)
health	Health status	1 (very bad) – 5 (very good)
absences	Number of school absences	0–32
G1	First period mark	0–20
G2	Second period mark	0–20
G3	Final mark	0–20

Table 1: Variables of the Student Performance data set. (Cortez 2014, data description. Table adapted from Cortez and Silva 2008, Table 1)
Medu and Fedu are encoded as follows: 0: none, 1: primary school (4th grade), 2: primary school (5th–9th grade), 3: secondary school, 4: higher education.

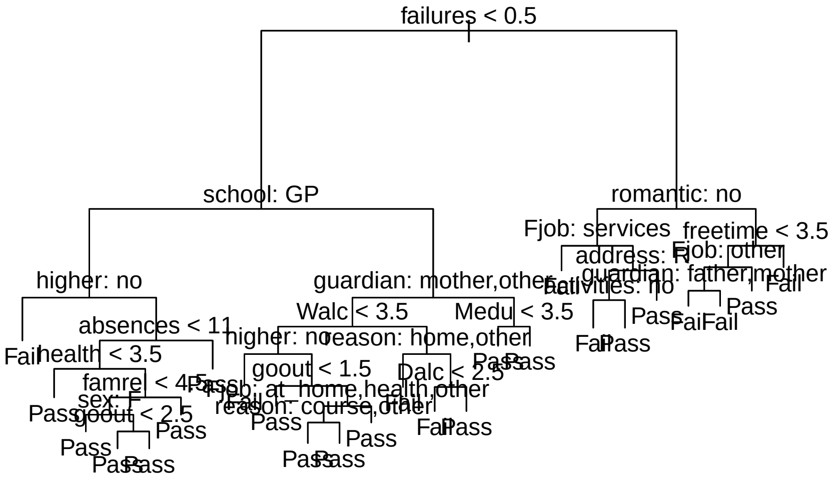


Figure 7: Diagram of unpruned classification tree.

	Predicted	
	Fail	Pass
Fail	20	20
Pass	14	195

Table 2: Error matrix of unpruned classification tree.

fication tree. However, as Cortez and Silva make clear, the survey methods they used to collect the data necessitated the use of categorical variables.[6]

2.1.1 Single Tree

First, the data was split into a training set (400 of 649 obeservations) and a testing set (the remaining 249 observations). The tree was then fitted to the training data, and is plotted in Figure 7. Table 2 shows an error matrix of predictions made with this unpruned tree, using the predictor variables of the test set, compared to the actual outcomes of the students in the test set. Of the 249 observations in the test set, 20 students failed and 195 passed; the model incorrectly predicted 14 students to fail when they in fact passed, and 20 to pass when they in fact failed. The test error rate is therefore $\frac{14+20}{249} = 0.137$, or 13%. It should be noted that classification trees exhibit high variability depending on the training data, and that if the data were to be refolded into new train-test splits, a tree different to that in Figure 7 would be generated with a different misclassification rate.

Naturally, such a complete tree may be suffer from overfitting; a pruned tree with fewer terminal nodes may have better predictive performance, have a better bias-variance trade-off and be easier to interpret.[7] Cost complexity pruning

[6] Cortez and Silva 2008, p. 2; “Trees can easily handle qualitative predictors without the need to create dummy variables.” James et al. 2013, p. 315.

[7] James et al. 2013, p. 307.

α Value	Nodes	Mean C.-V.	Mean Test Error	Test Error Range
$-\infty$	25	79.96	N/A	N/A
0	14	78.98	20%	13%
1	11	70.66	20%	12%
1.4	6	69.18	20%	11%
1.5	4	67.02	20%	13%
3	3	64.98	20%	12%
5.5	1	67.74	20%	13%

Table 3: Mean cross-validation statistics of the seven candidate trees as generated by the `cv.tree(FUN = prune.misclass)` function and mean classification error on the test set. Note that the unpruned tree T_0 is represented with an α value of $-\infty$ by `cv.tree()` which makes prediction impossible, though a single prediction using the full model is displayed in the error matrix of Table 2. Mean cross-validation statistics to 2 d.p., test errors to the nearest percent.

uses a tuning parameter α to generate a smaller candidate tree in the subset of the full tree, such that

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha|T|$$

is minimised, where $|T|$ is the number of terminal nodes in the tree.[8] As the tuning parameter increases, therefore, the tree looses terminal nodes. The `cv.tree(FUN = prune.misclass)` function in the 'tree' package generates seven candidate trees by varying the tuning parameter and computing the k-fold cross-validation error of each candidate tree.[9] Cross-validation error statistics vary each time the data is split into the k folds, and so to minimise variability in the results this process was repeated fifty times and the mean cross-validation error statistic was used to select the optimum tree. These mean cross-validation statistics, along with the seven α values and corresponding number of terminal nodes are shown in Table 3 and Figure 8.

The function selected a tuning parameter of $\alpha = 3$ as the optimum pruned tree. To confirm that was in fact the optimum, the predictions of the seven candidate trees were tested against the test set. It was noted that the misclassification rate varied each time the training and testing sets were re-folded. Each candidate tree was therefore fitted and tested agains new test-train folds one-thousand times. As can also be seen in Table 3, the mean test error did in fact level-out at 20% for all the trees, but that for each train-test fold the misclassification rate could vary by ± 6%. The extreme instability of single trees has again been uncovered empirically: not only does the structure of the tree vary when the data is refolded as noted above, but when tested on new data the test error can vary between 14%–26%.

In addition, the optimum tree according to the `cv.prune` function has only three terminal nodes: the first variable used is `failures` with a split criterion of `< 0.5`, that is the most significant variable is whether a student has taken the exam before. However, the distribution of `failures` is very uneven (Figure 6a), which makes it hard to predict whether the majority of students who are taking the exam for the first time will pass or fail. The real-world utility of the single classification tree, namely its ease of interpretability by non-experts, is therefore lost, as this tree cannot offer predictions for 85% of students.

2.1.2 Random Forest

As observed above, a single classification tree has high variance which reduces its predictive performance on new data. Random forest methods aim to mitigate this while maintaining the low bias of a classification tree; multiple trees are fitted on the same training data, but each tree uses only a subset of the variables (termed m).[10] In this way, the method is a development of the bootstrap, and works similarly by training B trees on b bootstrapped samples from the training

[8] James et al. 2013, Eq. 8.4, p. 309.

[9] Documentation, Ripley 2019, p. 2.

[10] First introduced in, Breiman 2001.

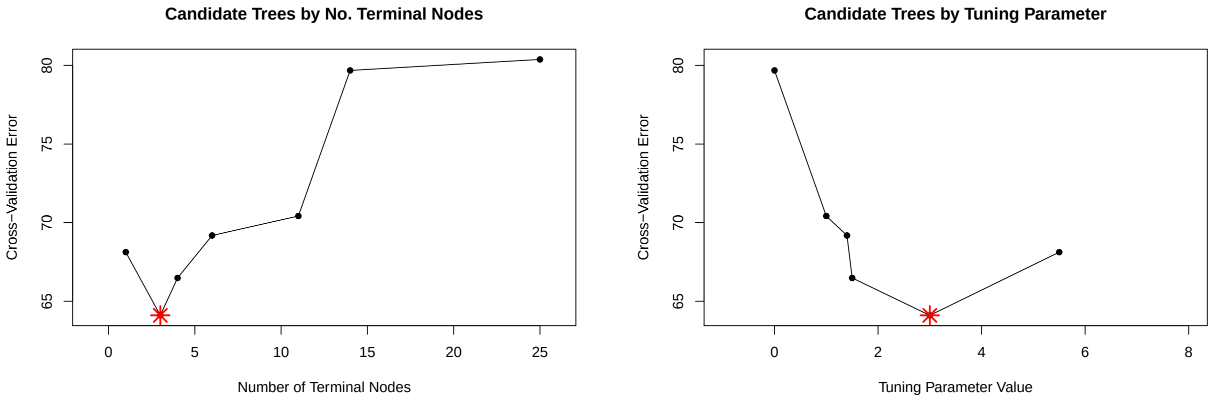


Figure 8: Mean cross-validation error of nine candidate trees plotted against number of terminal nodes (left) and tuning parameter value (right). In each case the tree with the minimum cross-validation error ($\alpha = 3, |T| = 3$) is marked with a red asterix.

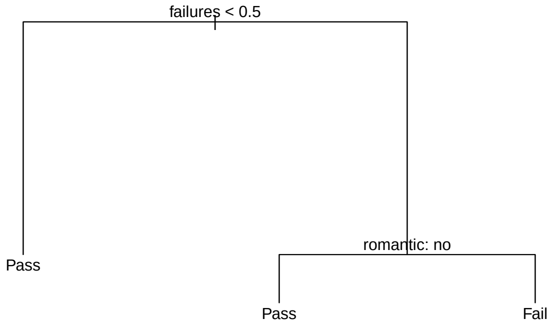


Figure 9: Classification tree pruned with α value = 3.

set, and then calculating the average of these trees ($\hat{f}_{bag}$):[11]

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$$

Applying this method to the Student Performance data set, a model was generated using two-thousand trees, which provide a sufficently large bootstrapped training set for the model.[12] An m value of $\sqrt{p}$ is recommended for classification models, and the `randomForest()` function automatically chose an $m = 5$ value ($\sqrt{36} = 5.48$).[13] Indeed, unlike many machine learning techniques, random forests do not require optimising of tuning parameters and in this scenario, with only thirty-six predictors, the the risk of overfitting to the training data is relatively low.[14]

[11] James et al. 2013, p. 317; Efron and Hastie 2016, p. 327; Friedman, Hastie, and Tibshirani 2009, p. 282.

[12] Efron and Hastie 2016, p. 328.

[13] James et al. 2013, p. 319; Documenation, Breiman et al. 2018, p. 18.

[14] "Using a very large value of B will not lead to overfitting." James et al. 2013, p. 317; "Becuase of the Law of Large Numbers they do not overfit."

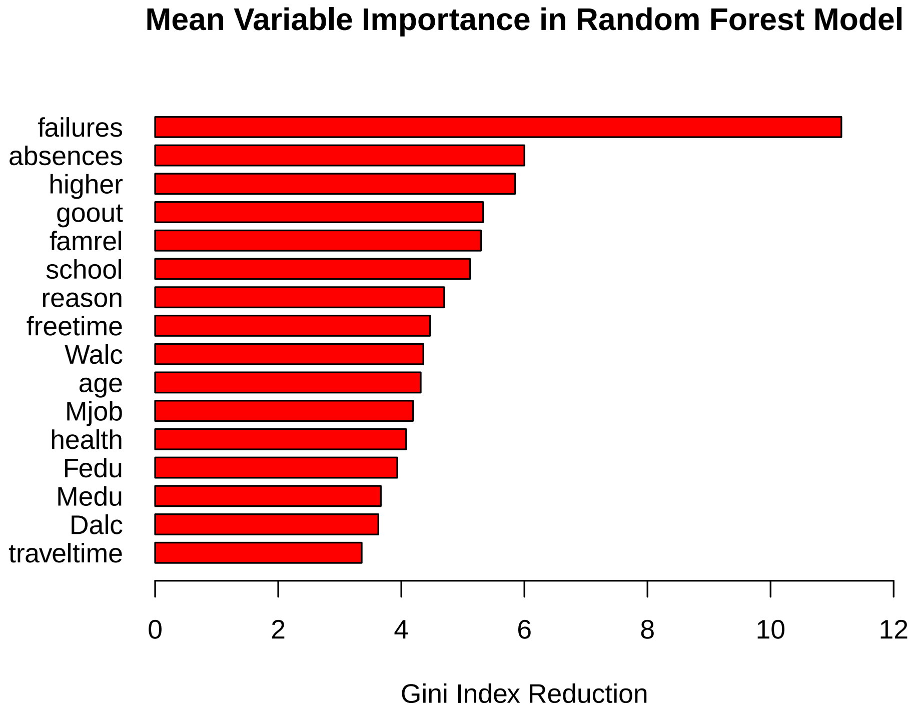


Figure 10: Mean fifteen most important variables in random forest model by reduction in Gini index.

The resulting model is not a tree, and hence cannot be displayed like the single tree in Figure 7 or 9, but the importance of each variable in the model's prediction can be visualised by using the decrease in Gini index per predictor averaged over all *B* trees.[15] Although random forests methods aim to reduce the variance of the model inherent in a single tree, due to the nature of the Student Performance data set (for example, the inverse log distributions of variables such as `failures` and `absences`) there is still variation in the variable importance values depending on the split of the train-test fold. Therefore, ten random forests were modelled and the mean Gini index reduction value for each variable was calculated, and the fifteen most influential predictors are plotted in Figure 10.[16]

The interpretation of these results will be postponed until section 3, but it should be noted here how `romantic` which featured as the second split in the pruned tree is only the ninth least important predictor variable in the random forest model, while `guardian`, `higher` and `Fjob` which were third-level predictors in the unpruned tree do not feature in the top fifteen predictors at all.

Finally, the test misclassification rate was calculated for each of the ten forests. The mean is 16%, while the values ranged between 13%–20%, a moderate increase in average predictive performance over the single tree.

Breiman 2001, p. 29; "In our experience random forests do remarkably well, with very little tuning required." Friedman, Hastie, and Tibshirani 2009, pp. 590, 596–597.

[15] James et al. 2013, pp. 319–320.

[16] The other variables' influence runs as follows (3 d.p.): `Pstatus` 0.817, `paid` 1.065, `schoolsup` 1.103, `internet` 1.352, `nursery` 1.495, `famsup` 1.509, `activities` 1.515, `famsize` 1.516, `romantic` 1.564, `sex` 1.664, `address` 1.749, `guardian` 2.181, `studytime` 3.123, `Fjob` 3.222

2.2 Neural Network

Neural networks, first developed in the 1980's, may best be described as multiple systems of linear models interacting with each other via a non-linear activation function.[17] Each layer of the network contains a layer of 'neurons' (h) of the type,

$$h_n = f(c_n + \mathbf{x}_n^T \mathbf{w}_n \mathbf{h}_{n-1})$$

where the predictor variables (as a matrix **x**) interact with an intercept c (here 'bias') and coefficents **w** (likewise a matrix and termed 'weights'), as in a linear model. However, each neuron contains a non-linear activation function (f) while the whole output of the previous layer (hence as a matrix $\mathbf{h}_{n-1}$) is also combined with the current layer. This enables the modelling of more intricate relationships than would be possible with a single linear model; the non-linear activation functions can enable neural networks to particularly excel in classification tasks, where calculating probabilites is central.[18]

Neural networks require that the data be formatted in tensor form (here a matrix as our data has only two dimensions: features and observations) and that all variables be hot-coded as binary categorical data.[19] This was achieved using the 'recipes' package in R which automatically prepares the data by reformatting it as required.[20] In addition to a training and testing fold, a neural network requires a third fold in the data for validating the training process, a process termed 'back-propagation.' This is becuase as the parameters of the network are fitted, the error in the network is calculated using a loss function using the validation set (cross-entropy in the case of classification), and this loss function is used by a regularisation algorithm to improve the fit of the network.[21] After selecting the appropriate loss function (to perform binary classification: `binary_crossentropy`) and choosing the activation functions for each layer (the recommended ReLU was used for the hidden layers and the sigmoid function for the output layer), the program Keras automatically performs these validations and back-propagations to optimise the model.[22] It is left to the user to design the hyper-parameters of the network and this will be the focus of the following discussion.

Neural networks are often overfit: unlike a linear least squares regression that must be fit to training data only once, the back-propagation method requires many cycles of re-fitting (termed 'epochs') to achieve high accuracy.[23] With each epoch, there is naturally a greater chance of the network becoming optimised to the validation data set at the expense of generalised performance on new data.[24] Moreover, in order to accomplish complex classification tasks such as image recognition, "the current collective wisdom suggests it is better to have an abundant number of hidden units, and control the model complexity instead by weight regularization," yet with a large number of hidden units, overfitting and over-parametrisation is very likely.[25]

Fortunately, the classification task posed by the Student Performance data set is not particularly challenging in a deep learning setting. *Pace* Efron and Hastie, it was judged that the advice of Chollet and Allaire was more suitable in this case: "simpler models are less likely to overfit than complex ones."[26] Although the authors use this aphorism to argue for the use of weight decay, ℓ_1 or ℓ_2 regularisation, due to the comparatively small size of the final network and its performance, it was not deemed necessary to use these methods to combat overfitting.

The architecture of the network was chosen in the following way, again following the advice of Chollet and Allaire: a large network was constructed that suffered from overfitting and over-parametrisation, then this network was pruned back to improve its general performance until the accuracy of the network began to suffer.[27] The main measure of fit is the change in validation loss and accuracy over successive epochs. Simply put, if the validation loss increases for each additional epoch, then the model has become overfit to the validation data.[28] This can be clearly seen in the first

[17] Goodfellow, Bengio, and Courville 2016, p. 164; "The knee-jerk response from statisticians was 'What's the big deal? A neural network is just a nonlinear model, not too different from many other generalizations of linear models."' Efron and Hastie 2016, pp. 351–352.

[18] Goodfellow, Bengio, and Courville 2016, p. 178.

[19] Chollet and Allaire 2018, pp. 77–79.

[20] Kuhn and Wickham 2020.

[21] Friedman, Hastie, and Tibshirani 2009, pp. 395–397; Chollet and Allaire 2018, p. 84.

[22] Falbel 2020; Goodfellow, Bengio, and Courville 2016, pp. 171, 178.

[23] Efron and Hastie 2016, p. 362.

[24] Chollet and Allaire 2018, pp. 118–119.

[25] Efron and Hastie 2016, p. 361.

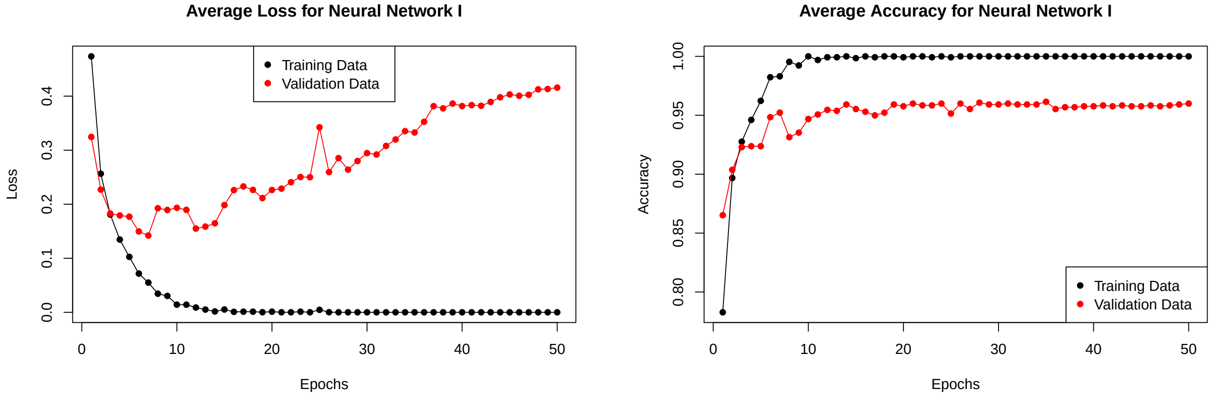
[26] Chollet and Allaire 2018, p. 123.

[27] Chollet and Allaire 2018, pp. 131–133.

[28] Chollet and Allaire 2018, p. 86.

		Validation Loss		
	Architecture	25 Epochs	50 Epochs	Test Error
I	{128} {64} {64} {32} {1}	0.342	0.416	8%
II	{64} {64} {64} {32} {1}	0.178	0.301	7%
III	{32} {32} {1}	0.111	0.117	7%
IV	{16} {16} {1}	0.157	0.120	7%
V	{16} {8} {1}	0.161	0.109	9%

Table 4: Architectures and performance statistics for the five candidate neural networks. Validation loss given to 3 d.p., test misclassification error to the nearest percent.



(a) Loss of neural network I over 50 epochs.

(b) Accuracy of neural network I over 50 epochs.

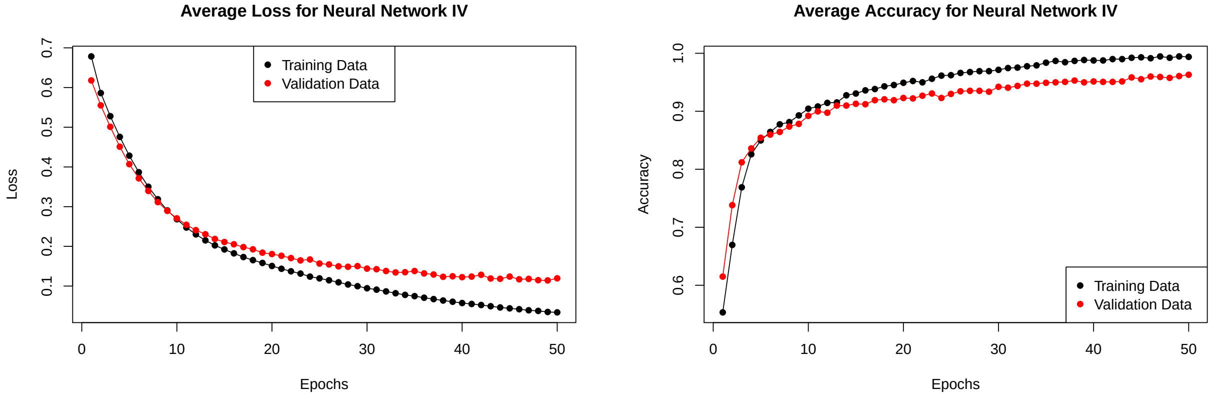
Figure 11

candidate network, which was constructed with a large input layer of 128 neurons and three hidden layers. Figure 11 shows that after only a few epochs, candidate network I had achieved near 100% accuracy on the training data, yet the validation loss function steadily increased after the ninth epoch, indicating that the network was overfitted. The accuracy and loss performances for candidate networks II and III are not plotted, but from Table 4, it can be seen that the validation loss was higher at the fiftieth epoch than at the twenty-fifth, and despite being simpler than network I, still suffer from overfitting.

Network IV was the first network that did not show symptoms of overfitting, while still retaining comparable accuracy with the previous networks. Comparing the plots of the accuracy of networks I and IV (Figure 12) it can be seen that it takes network IV the full fifty epochs to achieve a similar fit to that which network I achieved after only ten epochs. Network V, however, despite having an even lower loss value for the validation data, struggled to achieve the same fit and accuracy within fifty epochs. Here the network is too generalised and under-fitted to the validation data. Network IV was judged to be a suitable compromise between optimisation and generalisation, and was the network ultimately chosen.

Finally, it was noted as that the variability in the data and the relatively small number of observations (in the context of a deep learning setting) produced some variation in a network's performance if only a single three-way train-validate-test fold was created in the data. As shown in section 1 and noted in the discussion of the tree-based methods, the observations of certain variables, particularly ones shown to be significant such as `failures`, `school` and `sex`, are not normally distributed, and as such a model can be trained on unrepresentative data. Therefore, for each of the five candidate models, k-fold cross validation was used in computing the loss, accuracy and test misclassification error statistics.[29] The data was split into thirds and the network fitted three times, with each third serving as the test,

[29] Chollet and Allaire 2018, p. 114.



(a) Loss of neural network IV over 50 epochs.

(b) Accuracy of neural network IV over 50 epochs.

Figure 12

Method	Error Rate
Unpruned Single Tree	13%
Pruned Single Tree	20%
Random Forest	16%
Neural Network	7%

Table 5: Error misclassification rate of the modelling methods compared. Note, for the random forest the rate is the mean of ten forests, for the neural network the mean of three networks.

validation or training fold in turn. The mean loss, accuracy and misclassification error rate was calculated for each model and these were the values used to assess each network's performance.

3 Discussion

3.1 Different Methods Compared

Until now, the wider issues of performance, utility and interpretability for each method have not been discussed in detail, and will be treated here. As noted in section 1, while the Student Performance data set does not contain a particularly small number of observations, the inverse log distribution of variables such as `failures`, `absences` and `Dalc` can lead to the fold of the training data being highly unrepresentative, particularly as these variables were discovered to be very significant predictors (Figure 10). This has been one of the main issues in preparing a classification model of this data; managing unrepresentative folds of the training data and the resulting instability in model structures. This discounts the use of the single trees of section 2.1.1: the main advantage of single trees, namely their interpretability by non-experts, is negated by their variability in this instance, which also disqualifies them for useful prediction.[30]

As outlined in section 2.1.2, random forests overcome the shortages of single trees by using random selection of variables. In this particular case, however, it was found that random forest methods did not consistently produce the same results, and so the forest was generated multiple times and the mean variable importance and error rates calculated. This serves to highlight again the nature of the distribution of the variables in this data set, and how

[30] "Often a small change in the data can result in a very different series of splits, making interpretation somewhat precarious." Friedman, Hastie, and Tibshirani 2009, p. 312; "trees can be very non-robust." James et al. 2013, p. 316; "Discontinuous regression surfaces $\hat{S}$ . . . disqualify them for . . . estimation." Efron and Hastie 2016, p. 127.

extreme care must be taken to mitigate against the model being trained on unrepresentative data. Nevertheless, despite being a "black box"-type model, the results can be very easily displayed and interpreted, even by non-experts (Figure 10). Moreover, this model can be effectively used for prediction, having a lower mean misclassification rate and a much smaller range than the single trees (Table 5).[31] Ultimately, I judge the random forest to be the optimum model in this scenario in terms of the trade-off between fit, interpretability and prediction accuracy. In addition, the performance of this model could be further improved by incorporating boosting methods, which were developed specifically for binary classification problems of this type, though the constraints of time and brevity have prevented them from being explored in this enquiry.[32]

Finally, as can be seen from Table 5, the neural network performed the best in terms of prediction accuracy. Yet again, there was an issue of data variability and so three-way *k*-fold cross-validation was used to calculate the mean misclassification rate. As was noted in section 2.2, regularisation or drop-out techniques were not used to mitigate over-fitting (again due to the constraints of brevity), but by constructing a reasonably simple network such issues should not be present.[33] Despite the much better predicton accuracy, however, the neural network has very limited utility in anaylsing the data in sociological or political contexts due to its "black box" nature and inability for the parameters to be interpreted in any meaningful way. It is notable that Cortez and Silva use both neural networks and support vector machines in their analysis for exactly these reasons: one set of methods is used for prediction, the other for analysis of the variables.[34]

3.2 Student Performance

Using the mean variable importance plot from the random forest model (Figure 10), it can be seen that `failures` is the most significant predictor as to whether a student will pass or fail. As was noted in the discussion of the pruned tree in section 2.1.2, it is not too suprising that the minority of students who are taking the exam again would be expected to pass. The next four most important variables, `absences`, `higher`, `goout` and `famrel`, might be grouped together as indicating a student's individual behaviour: whether they skip school, go out often, have good familial relations and intend to pursue higher education. This is also the case for alcohol consumption `Walc` and `Dalc`, though to a lesser extent.[35]

From a sociological point of view, these predictors might themselves be interpreted as the results of a student's wider background. Yet they could also indicate a student's personal attitude and commitment to study, particularly as `reason` is the sixth most important variable, though it is unclear to what extent the students' parents influenced their choice of school, or indeed how much choice in which secondary school to attend is available in that particular region. The significance of `school` should be compared to discussions in section 1 of the imbalance of responses from the two schools surveyed, but it may also legitimately be the case that one school has better teaching than the other. Note that here unlike in a single tree, it cannot be seen attending which of the two schools predicts a better exam result.

Nevertheless, it is notable that academic predictors have very little importance: `paid` and `schoolsup` are third and second least important predictors, while `nursery` and `famsup` likewise are low down on the list. Likewise, other predictors that might indicate wealth, in particular parents' education (which are almost as important as each other) and career, do not seem especially important. The excpetion is `Mjob`, which has a Gini index reduction value of 4.467, higher than `Fjob` 3.306. Again, it cannot be seen here which jobs a mother has predicts better exam performance, but taking all these observations together, one might hypothesise that a student's home life (note too the importance of `freetime` and familial relationships may be just as important as their motivation to study, and more important than any formal academic support, whether offered by their parents or their school.

[31] "Random forests are an effective tool in prediction." Breiman 2001, pp. 6, 29.

[32] Efron and Hastie 2016, p. 333.

[33] "The simplest way to prevent overfitting is to reduce the size of the model." Chollet and Allaire 2018, p. 119.

[34] Cortez and Silva 2008, pp. 4, 7.

[35] Despite being significant in the single trees, `romantic` has only a Gini index reduction value of 1.564, and so this particular behaviour does not seem to influence students' results as much as the single trees suggest.

References

Breiman, L., Cutler, A., Liaw, A., and Wiener, M. (2018). *Package 'randomForest': Breiman and Cutler's Random Forests for Classification and Regression*. Version 4.6–14. URL: cran.r-project.org/web/packages/randomForest.

Breiman, R. (2001). "Random Forests". In: *Machine Learning* 45, pp. 5–32.

Chollet, F. and Allaire, J. (2018). *Deep Learning with R*. New York: Manning.

Cortez, P. (2014). *Student Performance Data Set*. University of Minho, Guimarães, Portugal. URL: archive.ics.uci.edu/ml/datasets/Student+Performance.

Cortez, P. and Silva, A. (2008). "Using Data Mining to Predict Secondary School Student Performance". In: *Proceedings of the 5th Future Business Technology Conference*. Ed. by A. Brito and J. Teixeira. Porto, Portugal: EUROSIS, pp. 5–12. URL: www3.dsi.uminho.pt/pcortez/student.pdf.

Dua, D. and Graff, C. (2017). UCI Machine Learning Repository. URL: archive.ics.uci.edu/ml.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Institute of Mathematical Statistics Monographs. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781316576533.

Falbel, D. (2020). *Package 'keras': R Interface to Keras*. Version 2.3.0.0. URL: keras.rstudio.com.

Friedman, K., Hastie, T., and Tibshirani, R. (2009). *The Elements of Statistical Learning*. 2nd ed. Springer Texts in Statistics. New York: Springer.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, USA: MIT Press. URL: deeplearningbook.org.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. New York: Springer.

Kuhn, M. and Wickham, H. (2020). *Package 'recipes': Preprocessing Tools to Create Design Matrices*. Version 0.1.15. URL: cran.r-project.org/web/packages/recipes.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Version 4.0.3. R Foundation for Statistical Computing. Vienna.

Ripley, B. (2019). *Package 'tree': Classification and Regression Trees*. Version 1.0–40. URL: cran.r-project.org/web/packages/tree.