# **Methods**

## **Data**

Researchers collected student grades for Portuguese language and mathematics courses during the 2005-2006 school year from two Portuguese schools. Grades were evaluated at three different time points throughout the school year, with the final grade corresponding to the students’ overall course grade. Researchers also surveyed students to gather demographics, social, emotional, economic, and other school related variables. Predictive variables were presented using an ordinal or categorical scale. Students’ grades were presented on an integer scale, with a minimum of zero and a maximum of 20. There were a total of 649 records with complete data. Cortez and Silva (2008) also discarded several variables due to high rates of missing data or a lack of variability among the sample students. These variables were not made available in the current dataset. The following predictor variables were left as integer level variables: age, failures, famrel, freetime, goout, Dalc, Walc, health, absences, G1 (when included as a predictor), and G2 (when included as a predictor).

Grades were predicted using binary classification (i.e., pass/fail), five-class classification (i.e., letter grades A through F), and then on the full integer scale (from 0 to 20). A description of all of the predictor variables and the outcome variable can be found in Cortez & Silva (2008).

Models will be fit using the first period grade and the second period grade plus the remaining predictor variables, the first period grade with the remaining predictor, and neither the first or second period grade to predict the final grade. The best performing model recommendations are based on performance when both the second and first period grades are excluded from the predictors. Table 1 shows the number of observations and variables in each of the training and testing data sets of each version.

Table 1. Summary of training and testing data splits among the three conditions

|  |  | # Observations | # Variables |
| --- | --- | --- | --- |
|  | All data | 649 | 33 |
| Version A:  With G1 and G2 | Train | 454 | 32 |
| Test | 195 | 32 |
| Version B:  With G1 | Train | 454 | 31 |
| Test | 195 | 31 |
| Version C:  Without G1 and G2 | Train | 454 | 30 |
| Test | 195 | 30 |

## **Models**

***Binary Classification***

The binary classification was coded to a pass/fail outcome with passing being 10 or greater school grade. We ran a random forest classifier and a logistic regression with an L2 penalty. Just like in the original paper, we ran the model 3 different times for each outcome classification so there was a Version A, B and C. Version A included G1 and G2, version B included only G1 and version C did not include G1 or G2. For both models, G1 and G2 were the most important predictors of the pass/fail outcome. However, other variables also proved to be useful in the prediction, including age, school attended, romantic relationships, number of absences, and desire for a higher education.

For both logistic regression and random forest, the hyperparameters were selected using 5-fold cross validation on a training set with the best parameters being selected based on the highest average accuracy. Once the best parameters were selected, the model was used on a separate testing set to find the testing accuracy that was reported in the final table.

The logistic regression model tested the value of C, the penalty value, the solver algorithm, and the random state to determine the best model parameters.The Random Forest model tested maximum features, number of estimators, maximum depth of the tree, the minimum samples per split, and the random state used to start the model to determine the best parameters.

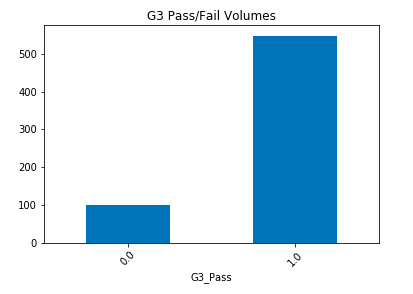


Figure 1. Number of students with passing (0.0) and failing (1.0) grades from Cortez and Silva (2008).

***Five-level Classification***

Several models were evaluated to determine the best model at predicting the five-level classification for the G3 grade. Five-fold cross-validation and a grid search were used for each model to determine the best model parameters before being fit to all of the training data. Accuracy was used to evaluate the model parameters. The Random Forest model tested maximum features, number of estimators, maximum depth of the tree, the minimum samples per split, and the random state used to start the model to determine the best parameters. The Support Vector Classifier (SVC) tested the value of C, gamma, the kernel used, the degree of the kernel for the polynomial model, and the random starting state to determine the best fitting model. The logistic regression model tested the value of C, the penalty value, the solver algorithm, and the random state to determine the best model parameters.

Once the best parameters for each model were determined, and the model fit to the full training data, area under the curve (AUC; calculated using a macro average) and accuracy were used to compare the models.

A logistic regression model, support vector classifier, and random forest classifier were tested to determine the five-level classification’s best performing model. These three models were selected because of their excellent performance in classification problems. All three also adapt very well from binary classification to multi-class classification. Logistic regression models and random forest models are relatively easy to understand. Interpretability is not a strength of the support vector classifier, but the support vector classifier typically performs well with classification problems. Figure 2 shows how the student grades were distributed across the five different levels.

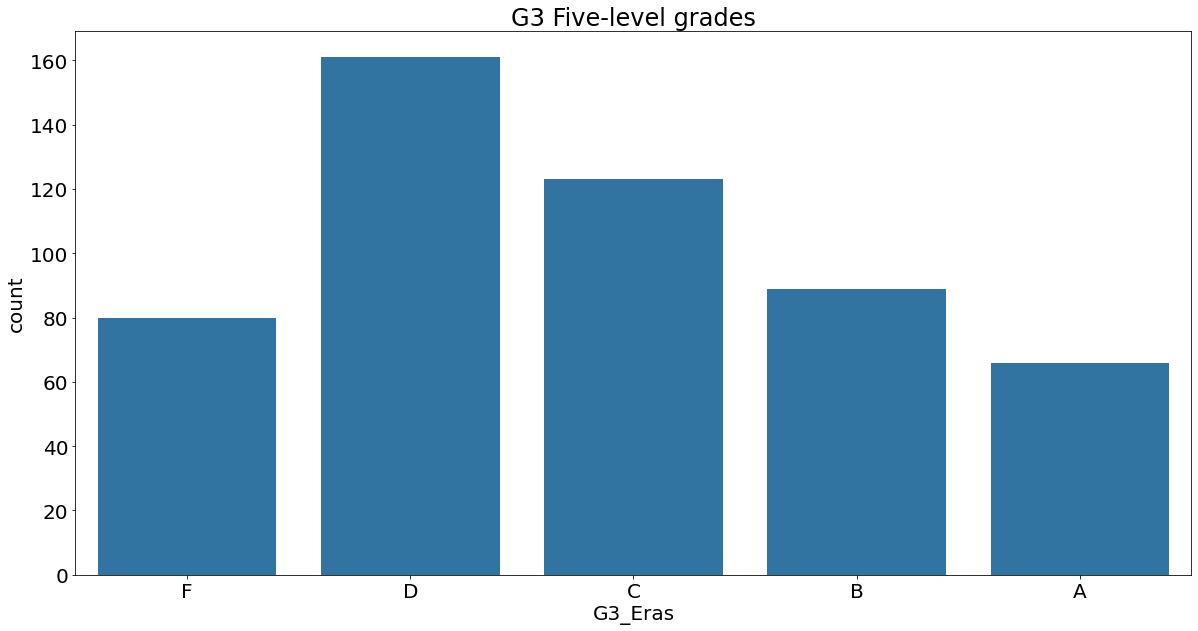


Figure 2. Distribution of student grades using the five-classes from Cortez and Silva (2008).

***Regression***

Several regression models were built and evaluated to determine the best model for predicting a numeric output between 0 and 20 for the G3 variable – the final grade. Linear Regression (using the OLS method), Ridge Regression, Lasso Regression, and ElasticNet Regression. Ridge Regression is an extension of Linear Regression where the loss function is modified by adding a penalty, a value equivalent to the magnitude of the coefficients squared, to minimize model complexity. Lasso Regression is an extension of Linear Regression where the loss function is modified by adding a different penalty, equivalent to the sum of the absolute model coefficients, again to minimize model complexity. ElasticNet Regression is yet another extension of Linear Regression – that is actually a combination of both Ridge and Lasso Regression: it applies both penalties.

Each set of models described above were run with both G1 and G2 included in the array of predictors (Version A, as mentioned above), G1 included (Version B), and neither G1 nor G2 included (Version C) as predictors.

Figure 3 shows how the student grades were distributed across the integer range of 0 to 20.

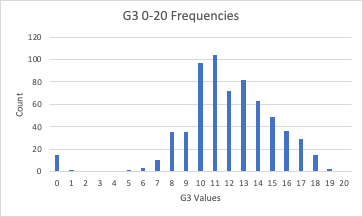


Figure 3. Distribution of student grades using the full grading scale (0 to 20) from Cortez and Silva (2008).

# **Results**

## **Predictive performance**

***Binary Classification***

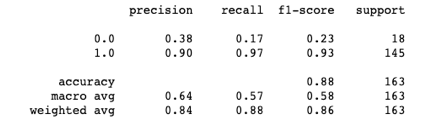
**Random Forest Classifier.** We ran 3 random forest iterations (see above in “Models”). The accuracy rate decreases as the G1 and G2 variables are removed. For the 3 different models, the accuracy varies from 0.91 to 0.88. Outside of G1/G2, the most important variables were: going out with friends, desire for higher education, number of failures, absences, which school the student attended, and weekend alcohol consumption.

See below the 3 iterations where G2 and G1 were removed for each step.

Version A Accuracy (G1 and G2 still in model): 0.91

Version B Accuracy (G2 removed as predictor): 0.91

Version C Accuracy (G1 and G2 removed): 0.88



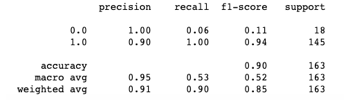
**Logistic Regression with L2 Regularization.**The same iterations where G1 and G2 were removed for each step were performed with the Logistic Regression. Just like with Random Forest, the accuracy was the best with G1 and G2 in the model. The most important variables in this model outside of G1/G2 were age, school, romantic relationship, guardian, higher education seeking, study time, Fedu and Medu. For this model, L2 regularization is performed, so many of the coefficients were shrunk for the final model.

See below the 3 iterations where G2 and G1 were removed for each step.

Version A Accuracy (G1 and G2 still in model): 0.94

Version B Accuracy (G2 removed as predictor): 0.91

Version C Accuracy (G1 and G2 removed): 0.90



Comparing LR to RF for binary classification: The LR model had a higher accuracy, but it had a particularly low recall for predicting failures. A vast majority of the predictions were “passes”, which in the model is the majority outcome. Most of the predictions were “pass”, meaning that the model will inherently have a high accuracy rate.

***Multi-class classification***

Accuracy was used to determine the best model parameters. Accuracy and AUC with the test dataset were used to select the best performing model on the five-level classification. For version C, the logistic regression and SVC models were both equally accurate, while the logistic regression model performed slightly better than the SVC model concerning AUC. Therefore, the Logistic Regression Model is the most preferred model for the Multi-class classification problem. It had an accuracy of .37 and an AUC of .88.

Of note with the models fit to version C, the logistic regression model did not classify any of the students as receiving a grade of “F,” the lowest grade and equivalent to failing in the binary classification problem. The SVC model did classify students as failing, but was not particularly accurate at this, only correctly classifying one out sixteen failing students.

For models fit to version B, the logistic regression model had the greatest accuracy and AUC. The overall accuracy of all of the models is greatly improved when G1 is included as a predictor, with the best performing model having an accuracy of .66 and an AUC of .88. None of the other models performed as accurately, or had as high an AUC, as the logistic regression model in version B.

For models fit to Version A, the logistic regression and random forest models are similarly accurate, but the logistic regression model has a slightly better AUC. Again, when G1 and G2 are included as predictors, the model accuracy improves greatly. The best performing model had an accuracy of .75 and an AUC of .94.

***Regression***

Each of the models were evaluated with two metrics: R-squared (R2) and Root Mean Squared Error (RMSE). R2 is the proportion of variation in the G3 variable that is explained by the other predictors we have. RMSE is the average error resulting from a model in an effort to predict G3 for a single observation. The better models will have a high R2 value and a low RMSE value.

Because three different sets of models were performed (varying in inclusion/exclusion of G1 and G2), let’s analyze the performance of the models with respect to each condition.

When both G1 and G2 are included as predictors, OLS, Ridge, and Lasso regressions all perform similarly, with an RMSE value of approximately 1.46 and an R^2 value of 0.82. The ElasticNet regression performs slightly worse, with an RMSE value of 1.60 and an R^2 value of 0.79. G1 and G2 are the best predictors of G3, so this set of models performed the best.

When G2 was dropped from the array of predictors, OLS, Ridge, Lasso, and ElasticNet regressions all performed similarly, with an RMSE value of about 2.21 and an R^2 value of 0.59.

When both G1 and G2 were excluded as predictors, it became very clear that the predictive ability of the models suffered, with all of the models having an RMSE value of about 2.88 and an R^2 close to 0.31.

When either G1, G2, or both were dropped as predictors, it can be seen that the models had a much more difficult time predicting G3, than when both G1 and G2 were included.

Table 2. Model performance for the different classification/regression problems

| Prediction Problem | Linear Models | Non-linear models |
| --- | --- | --- |
| Binary Classification (I.e., Pass/Fail)\* | Logistic Regression (L2 regularization) -  A: 0.94  B: 0.91  C: 0.90 | Random Forest -  A : 0.91  B : 0.91  C : 0.88 |
| 5-Level Classification (e.g., Letter Grade)\* | Logistic Regression:  A: 0.75  B: 0.66  C: 0.37^ | Random Forest -  A: 0.75  B: 0.60  C: 0.36  Support Vector Machine -  A: 0.66  B: 0.48  C: 0.37 |
| Regression (i.e. 20 point scale)& | Please consult Table 3 (below) | N/A |

Table 3. Model performance for the various types of regression on the numeric output

| Condition | Regression Type | RMSE | R^2 |
| --- | --- | --- | --- |
| Version A  With G1 and G2 | OLS | 1.467682918 | 0.821520947 |
| Ridge | 1.468160855 | 0.821404688 |
| Lasso | 1.455151962 | 0.824555615 |
| ElasticNet | 1.596542643 | 0.788804908 |
| Version B  With G1 | OLS | 2.220143366 | 0.591600864 |
| Ridge | 2.219663861 | 0.591777257 |
| Lasso | 2.219122334 | 0.591976419 |
| ElasticNet | 2.213953112 | 0.593875105 |
| Version C  Without G1 and G2 | OLS | 2.882929822 | 0.311362108 |
| Ridge | 2.882671576 | 0.311485476 |
| Lasso | 2.888165231 | 0.3088587 |
| ElasticNet | 2.880599309 | 0.312475025 |

\*Accuracy was used to assess these models.

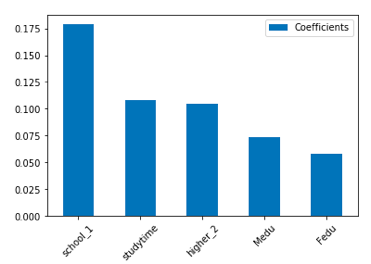
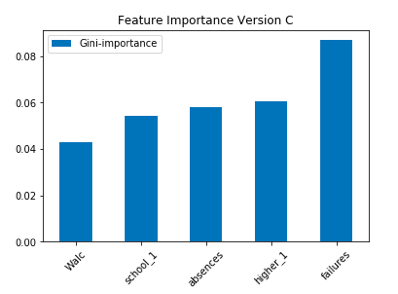
&R2 and RMSE were used to assess these models.

^Area Under the Curve was used as a tiebreaker to determine the better model for predicting the 5-level classification.

**Descriptive Knowledge**

The various models used have different ways that feature importance can be assessed. Random Forest models can be used to determine Gini impurity, which can be used for feature selection. Regression coefficients can also be used to assess feature importance. The feature importance for the various predictive problems is compared below.

The binary classification random forest model selected the number past class failures, whether the student wanted to seek higher education, how many absences the student had, the school the student attended, and the weekly alcohol consumption as the most important features (figure 4). The logistic regression model for binary classification selected the school the student attended, the amount of time spent studying every week, whether the student intended to pursue higher education, and the mothers and fathers education as the most important variables (figure 5). The two models gave importance to slightly different features, although both selected the school and whether the student intended to select higher education.



| Figure 4. Random Forest Gini impurity for the top five features in the binary classification model | Figure 5. Logistic Regression (L2) Coefficients for the top five features in the binary classification model. |
| --- | --- |

The random forest model was used to assess what features may be most important in the five class classification problem. This model performed moderately well in the classification problem, even though it was not the most accurate model, or had the highest AUC of the models tested. The random forest is also uniquely suited to determining the most important features for a predictive model.

Figure 6 shows that the top five features for the five class classification problem are the number of absences, the number of previous failures, how frequently students go out with friends, how much free time students have after school, and the students' self-assessment of their health.

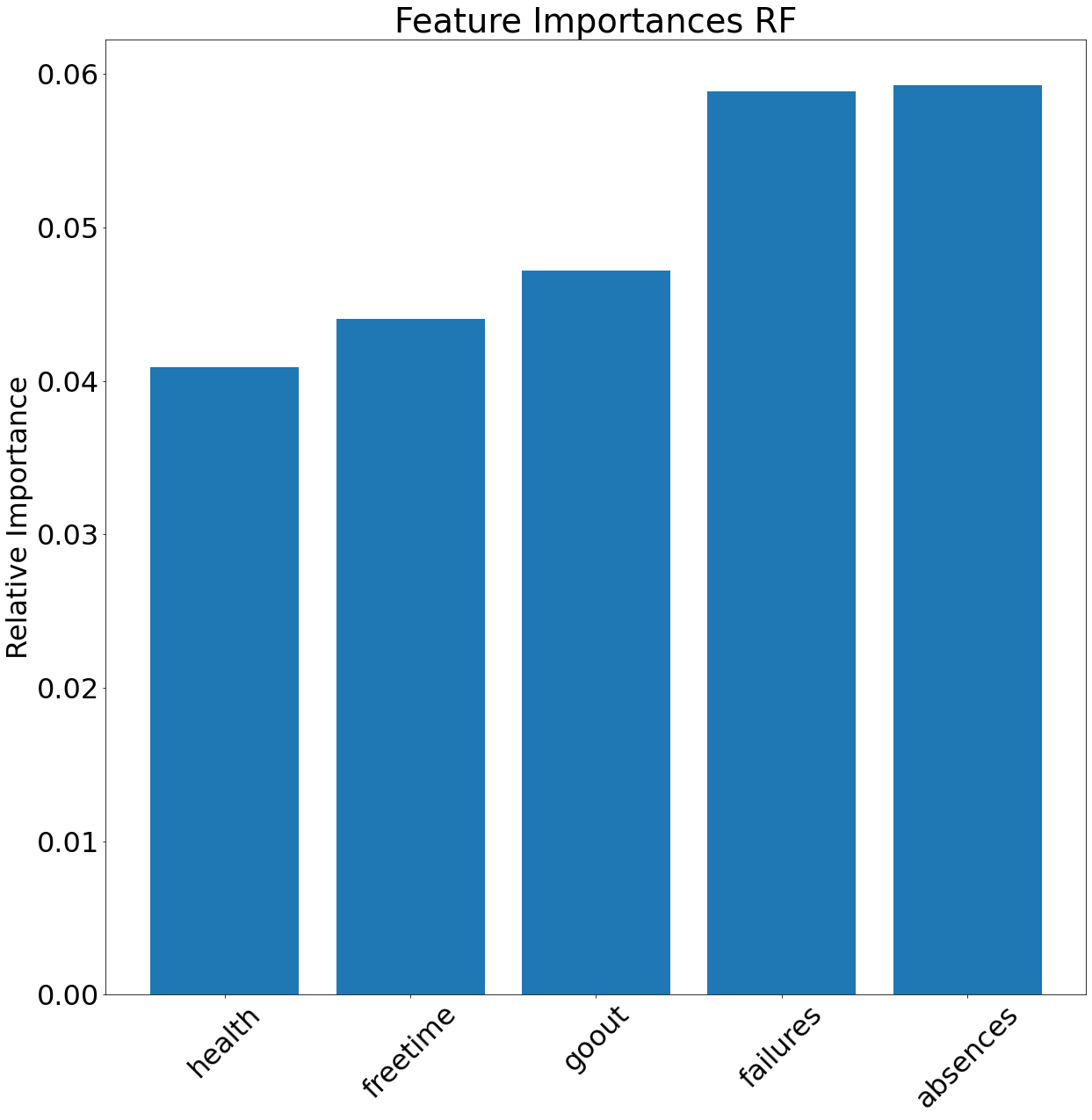


Figure 6. Random forest Gini-impurity index for the five-class classification model.

The LASSO regression model applies a penalty to the coefficients, forcing some of them to reach zero and others to take non-zero weights. The LASSO regression coefficients can be useful to also determine feature importance. Figure 7 shows that the top five features for the numeric output prediction case were if the students had access to and taken extra paid classes within the course subject, the size of the student’s family, if the student wanted to pursue higher education, the quality of the student’s family relationships, and the student’s workday alcohol consumption.

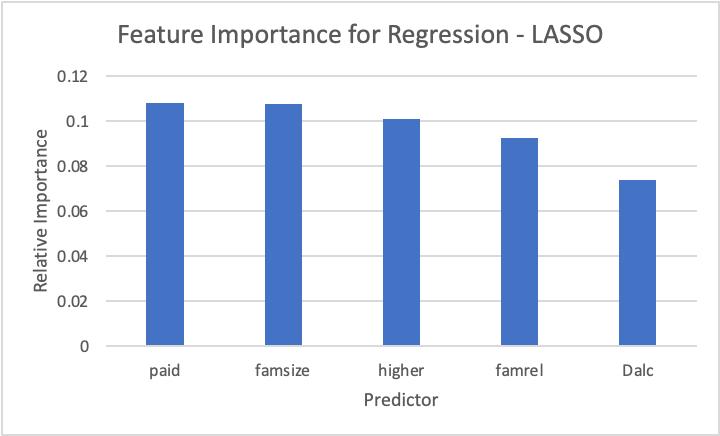
****

Figure 7. Coefficients for the top five features in the LASSO regression model.

# **Discussion**

For this analysis, linear models tended to have higher performance than non-linear models by a slight margin. In addition, the model performance varied significantly based on binning of the outcome variable. For the two classification problems, the accuracy was unsurprisingly much higher for the binary outcome where a majority of the outcomes were “pass”, whereas the multi-class classification saw significantly lower accuracy. Additionally, while not presented, the logistic regression model that was the highest performing model when G1 and G2 were excluded did not classify any students as failing (receiving a grade of “F”). This indicates that the logistic regression model is not equivalent in performance to the binary model with respect to predicting pass/fail grades. If letter grades need to be predicted, perhaps some combination of the pass/fail classification and the letter grade predictor could be most accurate, with the pass/fail model used to predict students expected to receive a grade of “F,” and the multilevel model applied to the remaining students across the remaining 4 letter grades.

None of the models constructed for any of the prediction problems used engineered features beyond one hot encoding. Many of the variables were treated as categorical, even though they were encoded using numerical codes. The variables in the dataset did consist of many that were on an ordinal scale. Ordinal variables can be treated as categorical or integer variables. Treating the predictor variables as integer level variables could allow for greater feature engineering resulting in better performing models. This direction could be taken in future experiments.

Neural networks were not tested in the current paper. Neural networks are very powerful machine learning algorithms that can be applied to both classification and regression problems. Neural networks are also able to learn complex relationships and features and use them to accurately make predictions. Cortez and Silva (2008) did use Neural Networks in their original work, although the neural network was not the best performing model. Using neural networks, especially in conjunction with feature engineering as discussed above, could also result in better performing models. It should be cautioned that neural networks can be quite computationally expensive making them less desirable for algorithms intended to be put into production. Neural networks are also very difficult to understand and interpret. If model understanding is imperative, either for research purposes or to advocate for model implementation, then neural networks would not be recommended.

The models in the paper demonstrate that final grades for students can be predicted with the highest accuracy using previous test grades for the course. However, that information is not always available and changes on a semester basis. Oftentimes, a prediction needs to be made before any grades are yet available, so in the absence of previous grades social factors such as number of absences, previous failures, age and health can also have a significant impact on grades. These social factors need to be further studied in order to determine their isolated impacts on grades so that the proper intervention strategies can be designed for students to help future academic performance in schools.

**Supplementary Figures**

Figure 1: Full feature importance graph - LASSO regression

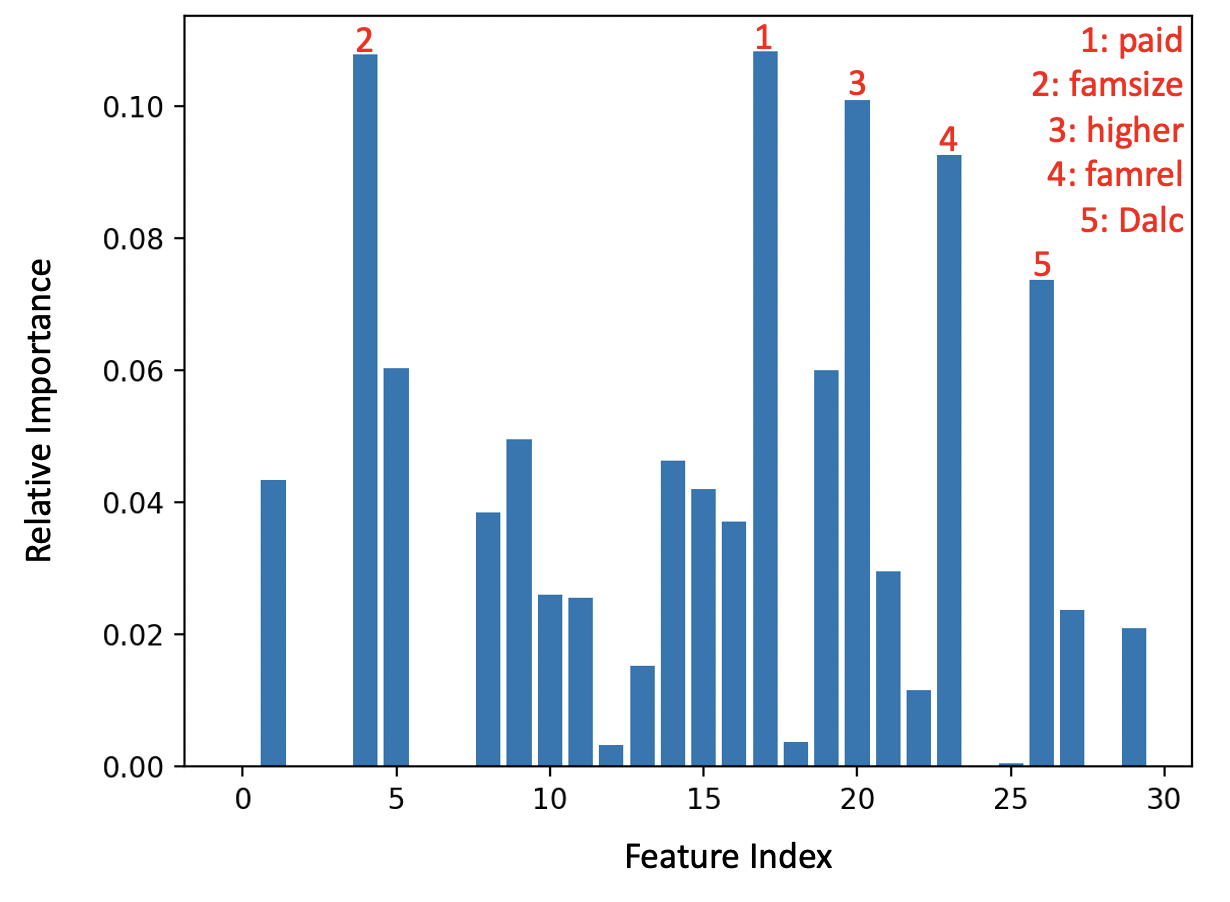


Figure 2: Full feature importance graph - Random Forest for five-classifier model

