# Predicting Student Academic Performance with Machine Learning

Kostantinos Morfesis

March 11, 2019

## Contents

# 1 Introduction

The question of what causes a countries students to under perform academically is of great concern, not just to individual schools but the country as well. When students under perform, they're less likely to attend college and higher education is associated with: decreased reliance on public assistance, better health choices, and job security [1]. The potential causes of this occurrence can range from personality differences in the students, to different teaching styles, or to government funding. The goal of this capstone is to perform data analysis on intra and inter country data for Portugal. Ultimately determining what factors most influence a students success in education. Given the results of the data analysis, one can then act in such a way that will benefit the students education. This may include changes to school resource allocation, government spending, or altering teaching styles to better accommodate student needs.

The intra data source consists of a questionnaire given to 649 students at a Portuguese secondary school, along with their numeric grades on a scale of 0 to 20 [2]. The questions probe information about the students demographics, family history, and out of school activities. While the inter data source comes from the World Bank [3] and PISA test results [4] [5], with information about different educational outcomes between countries.

The approach can be outlined as follows: First, the data will be cleaned for missing values and outliers such that it is usable for analysis. Next, initial exploratory analysis will be run on the data to determine the relationship between educational outcomes and different features. Lastly, various supervised machine learning techniques will be implemented for grade prediction. Specifically, the analysis will be split into a classification and regression scheme. From there: random forests, logistic regression, support vector machines (SVM), and ridge regression will be utilized.

# 2 Data Cleaning and Preprocessing

## 2.1 Data Sources

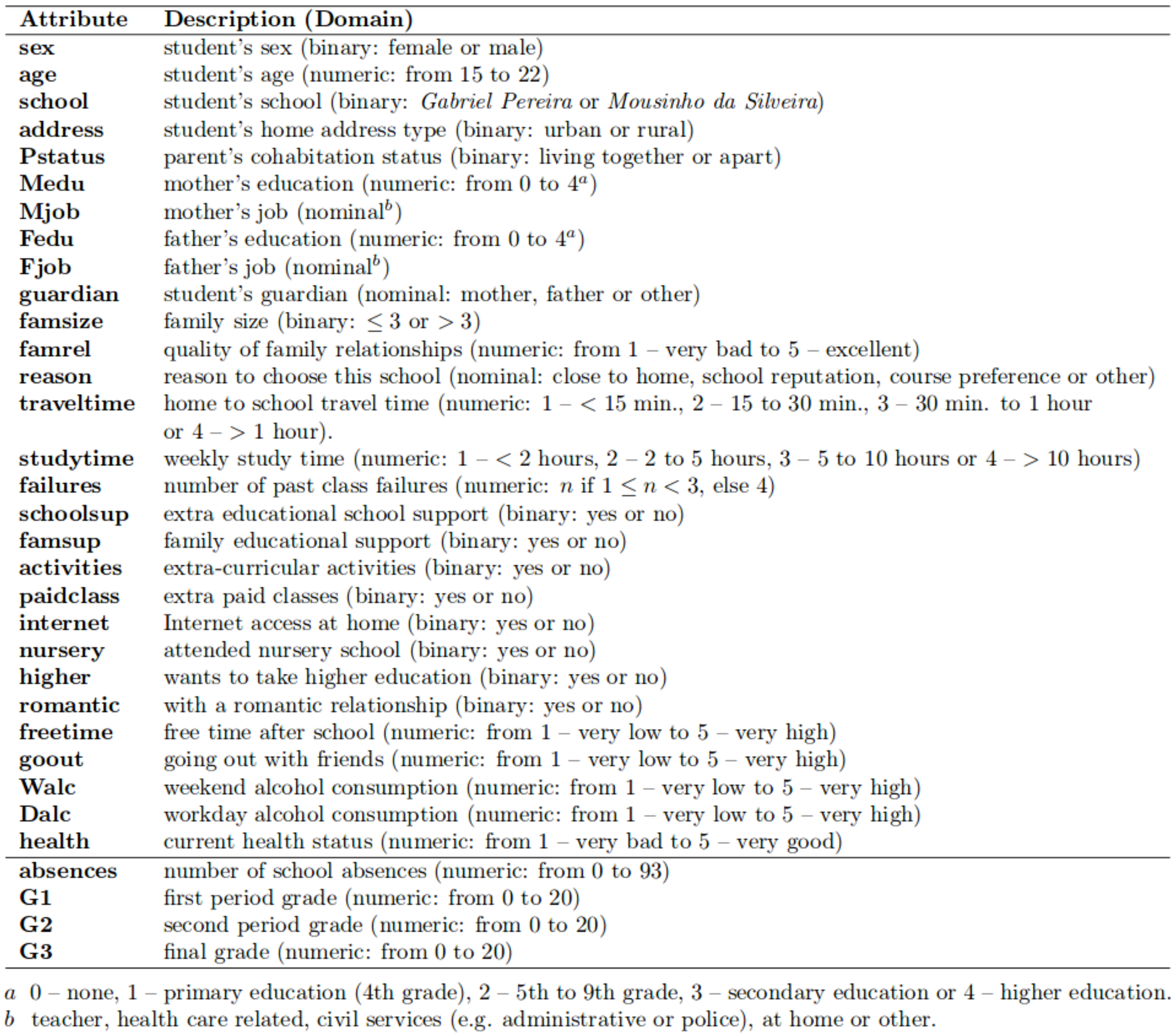
**Intracountry data for Portugal:**

The intracountry data consists of only a questionnaire given to 649 students at two different schools, along with their grades in Portuguese and math for the three periods of the school year. The different attributes and their descriptions are available in table 1. Features such as the time spent studying (studytime), mothers education (Medu), and going out with friends (goout), first and second period grades (G1 and G2) are all accounted for along with many others. As will be observed during the analysis, these features can explain a significant portion of the variation in final grades (G3).

**Intercountry data:**

The intrercountry data consists of the following data sources:
- The Programme for International Student Assessment (PISA) test results according to each country. This is such that we can compare the learning outcomes of each individual country. The PISA test results we look at are only for reading and math, to remain consistent with the intracountry source. [4] [5].

**Table 1:** Questions in questionnaire and the corresponding attributes, along with individual grades.



| Attribute | Description (Domain) |
|---|---|
| **sex** | student's sex (binary: female or male) |
| **age** | student's age (numeric: from 15 to 22) |
| **school** | student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*) |
| **address** | student's home address type (binary: urban or rural) |
| **Pstatus** | parent's cohabitation status (binary: living together or apart) |
| **Medu** | mother's education (numeric: from 0 to $4^a$) |
| **Mjob** | mother's job (nominal$^b$) |
| **Fedu** | father's education (numeric: from 0 to $4^a$) |
| **Fjob** | father's job (nominal$^b$) |
| **guardian** | student's guardian (nominal: mother, father or other) |
| **famsize** | family size (binary: $\leq 3$ or $> 3$) |
| **famrel** | quality of family relationships (numeric: from 1 – very bad to 5 – excellent) |
| **reason** | reason to choose this school (nominal: close to home, school reputation, course preference or other) |
| **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). |
| **studytime** | weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours) |
| **failures** | number of past class failures (numeric: $n$ if $1 \leq n < 3$, else 4) |
| **schoolsup** | extra educational school support (binary: yes or no) |
| **famsup** | family educational support (binary: yes or no) |
| **activities** | extra-curricular activities (binary: yes or no) |
| **paidclass** | extra paid classes (binary: yes or no) |
| **internet** | Internet access at home (binary: yes or no) |
| **nursery** | attended nursery school (binary: yes or no) |
| **higher** | wants to take higher education (binary: yes or no) |
| **romantic** | with a romantic relationship (binary: yes or no) |
| **freetime** | free time after school (numeric: from 1 – very low to 5 – very high) |
| **goout** | going out with friends (numeric: from 1 – very low to 5 – very high) |
| **Walc** | weekend alcohol consumption (numeric: from 1 – very low to 5 – very high) |
| **Dalc** | workday alcohol consumption (numeric: from 1 – very low to 5 – very high) |
| **health** | current health status (numeric: from 1 – very bad to 5 – very good) |
| **absences** | number of school absences (numeric: from 0 to 93) |
| **G1** | first period grade (numeric: from 0 to 20) |
| **G2** | second period grade (numeric: from 0 to 20) |
| **G3** | final grade (numeric: from 0 to 20) |

*a* 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.
*b* teacher, health care related, civil services (e.g. administrative or police), at home or other.

- Percent of the government expenditure which is spent on secondary education by country. Such that we may be able to determine if spending on education impacts educational outcomes [3].

- Average class size by country. To determine the effect of class size on educational outcomes [6]

To remain as consistent as possible, all of the data is from years as close to 2008 as possible. The PISA test data is from 2009, average class size by country is from 2010 and the government expenditure information comes from 2009.

## 2.2 Data Cleaning Steps

The data cleaning steps begin with importing the data sets one at a time, making sure to clean them before moving on to the next. This works to modularize the code/procedure and increases understandable as well as allowing for easy updates.

For the questionnaire, the following steps were implemented: The data was downloaded in two csv files, one for each class (Math or Portuguese). To simplify the data into a single dataframe, the data was checked to see if the same students were prevalent in each dataframe. However, since their was no unique identifier (student ID) the two dataframes could not be joined inner or outer joined. Consequently, the data sets were concatenated for simplification. This was done because the purpose of this study is to understand how certain features affect a students general academic success. The purpose is not to understand the individual differences which arise when looking at grades for different courses (math or Portuguese). Next outliers were checked using the mean, min, and max of each column. Since the questionnaire was multiple choice there were few outliers. However, the absences recorded were not multiple choice. To make the absences more tenable, they were grouped by how much of the data set was in each range. The range was chosen as 40, 20, 20, and 20 percent of the data for the 4 groups. This was chosen to appropriately spread out the data. This resulted in groups of absences of: $x < 2, 2 \leq x < 4, 3 \leq x < 7, x > 6$

For the intercountry data, the following was done for all: The data sets were brought into python and made into dataframes. Columns not needed for analysis were removed for simplification. Any NaN values were removed from the dataframes. This is because an NaN for these data sets implies unknown values. Consequently they should not be used for analysis. The data was then grouped to obtain the desired columns. For instance, in the PISA test the results were split up by gender and by country. For this study were looking at differences between countries only and so the different gender values were averaged for each country. Last, the data frames were merged. Specifically, the PISA test data was merge with the education spending dataframe and the average class size data frame each separately. Care was required when merging since the same countries were not all present in each dataframe.

# 3 Inferential Statistics and Hypothesis Testing

## 3.1 Testing Framework

To test the strength of correlation between different independent and dependent variables for the intra country data source, a hypothesis test is performed below. The hypothesis is: if there is a correlation between two random variables $(X, Y)$, a non-zero Pearson correlation coefficient will be observed. Where the Pearson correlation coefficient is defined as:

$$\rho_{X,Y} = \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y} \tag{1}$$

Hence the following hypotheses can be postulated:

$H_0$: $\rho_{X,Y} = 0$, there is no linear correlation between a pair of random variables $(X, Y)$
$H_A$: $\rho_{X,Y} \neq 0$, there is a linear correlation between a pair of random variables $(X, Y)$

To test whether the categorical variables have statistical significance a two-sample bootstrap test is used. This is achieved such that the highest and lowest average values for a category (as determined by the respective bar graphs) are compared to one another. It should be noted that only some variables have significance which can be easily determined from the plots. That is, when two confidence intervals overlap for the categories, the correlation of the variable with G3 should be statistically insignificant. Consequently one can choose to only focus on those which have no overlap. Therefore, our hypothesis test is as follows:

$H_0$: the difference between the highest and lowest mean values in a category is zero
$H_A$: the difference between the highest and lowest mean values in a category is non-zero

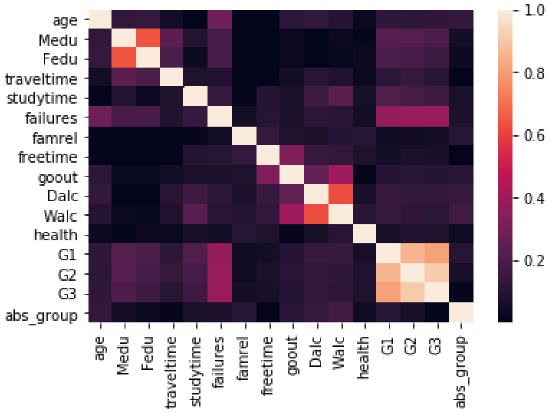


**Figure 1:** heat map indicating the $|\rho|$ values, or correlation strength.

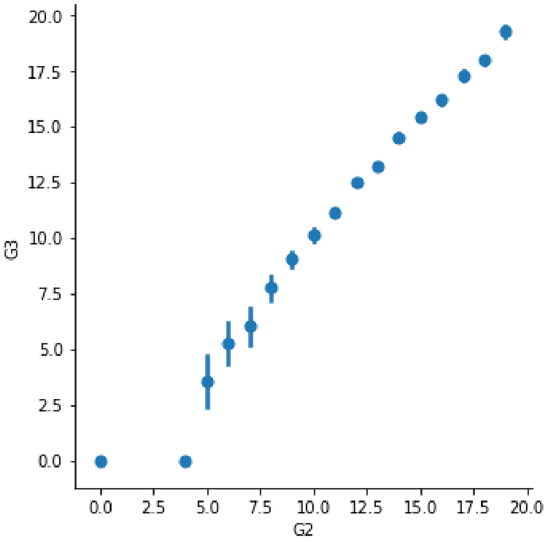


**Figure 2:** Strong correlation between final grades and second third grades. This is the strongest correlation between any variables.

## 3.2 Numerical Feature Analysis

using scipy.stats.pearsonr, the p-value along with $\rho$ is calculated. Any p-value less than .05 is discarded as statistically insignificant. All of the remaining p-values are kept and stored. There are 83 pairs of variables which have a statistically significant correlation between them. Of course, some are more significant than others and so the data frame is put in descending order with the first value being the most statistically significant. Further, a heat map of $\|\rho|$ can be seen in fig. 1, giving an idea of the correlation strength of different features.

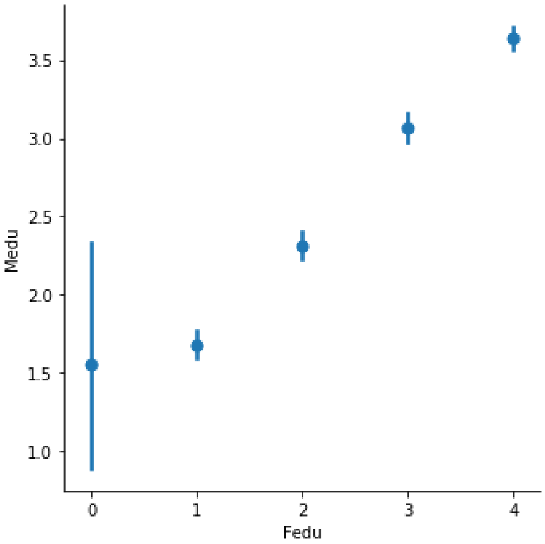


**Figure 3:** Correlation between mother and father education levels. This is the strongest correlation outside of correlation between grades.

As expected, the variables with the most significance to one another are the grades (G1, G2, G3) with one another (fig.2). Interestingly however, the next most significant correlation exists between mothers education and fathers education (fig. 3). Then following with weekend alcohol consumption correlating with M-F alcohol consumption. These correlations are important for our analysis.

The end goal is to determine what most affects G3, and as such we can look at the $\rho$ values to see which variables are most important. However, the relationship between independent variables is of importance too when utilizing machine learning models. That is, pairs of independent variables with high correlation can have one of the variables neglected. In example, since mothers education and fathers education are strongly correlated, one can be neglected in certain machine learning models. This will help to increase the accuracy of certain models.

For the intercountry data sources, the correlation tests indicate a a p-value of greater than 0.5. Consequently, the null hypothesis cannot be rejected and for the remainder of the analysis only the questionaire will be explored. In further analysis better features and segmentation should be utilized. That is, perhaps a better way to observe the correlation of average classroom size is to look at purely first world countries for comparison. In this way an extraneous feature is accounted for.

### 3.3 Categorical Feature Analysis

From the two-sample bootstrap method one can see all of the variables which were deemed to be significant. The eight which were chose (address, mjob, reason, guardian, schoolsup, higher, internet, romantic) were chosen because they had no overlap in the confidence intervals for the two outermost average values, as can be seen in the above bar graphs. Further, from the p-values one can see they are all indeed statistically significant, and the two most significant are: mothers job (mjob) and whether the student wishes to pursue higher education(fig.4). While the least significant is the difference of guardian value, which makes sense from the guardian bar graph(fig. 5).

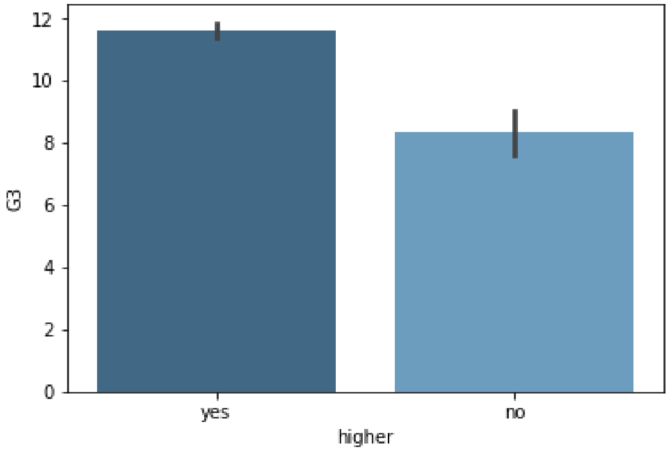


**Figure 4:** lowest p-value came from whether the students planned on going into higher education or not. This is the most statistically significant categorical variable relationship to G3.

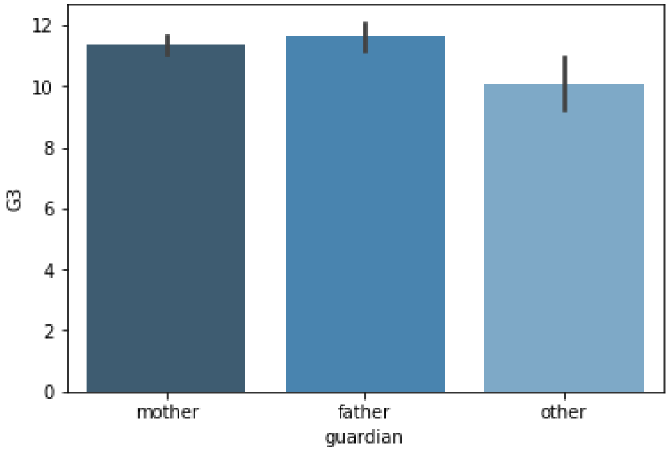


**Figure 5:** The relationship between guardian and G3 had the largest p-value which was still under .05.

# 4 Machine Learning Training and Predictions

## 4.1 Preprocessing

To perform more in-depth analysis, a modeling of the final grades (G3) will be done. The modeling will be split up and trained in the following way: With G2 and G1 included in prediction of G3 and without. Clearly those models without G2 and G1 will have the least predictive accuracy. But their importance is greater because G2 and G1 are variables which occur after part of the year is already over. This is not as useful for determining at risk students for intervention before the year begins.

Further, G3 will be split up into three different classifications to see how powerful the prediction can become. Specifically G3 will be predicted in the three following classification schemes:

1.) Binary: pass (if $G3 \geq 10$), fail (if $G3 < 10$)

2.) 5-level classification: using the following scheme: A:(16 - 20), B:(14 - 15), C:(12 - 13), D:(10 - 11), F(0 - 9)

3.) Regression: using G3 as the value

In total this yields six combinations to be tested. One would expect the binary classification to have the highest accuracy, and the regression to have the lowest. To perform this analysis, some preprocessing of the data is required Columns associated with the binary and 5-level classification schemes must be added. Further, it is necessary to create dummy variables for all categorical variables in the data set. This creates features for the categorical variables corresponding to binary classification. As an additional preprocessing step, feature scaling is implemented. This will normalize all features of the data set such that they have mean zero and standard deviation of one. This is relevant since the distance between two points is used in SVM. If one feature has a higher scale than another, it will be the dominant feature in the model. Additionally, the accuracy of each model and classification scheme is shown in in the following tables (tables 2, 3, and 4)

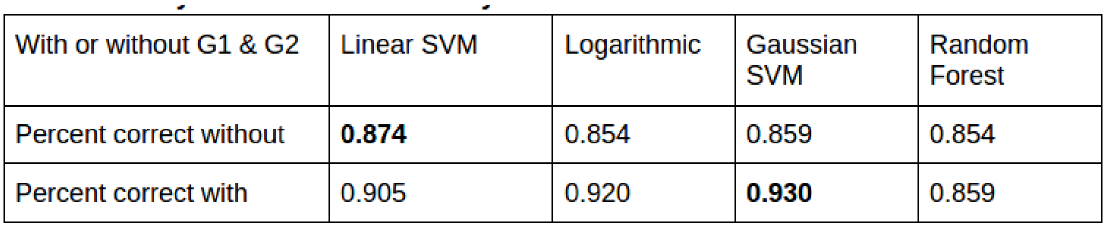
## 4.2 Binary Classification

Since the data set is not too large (less than 10k), a stochastic gradient descent (SGD) classifier was used along with random forest and a Gaussian SVM. For the loss function of the SGD classifier, both linear SVM and log loss functions were used and their accuracy was compared. For the binary classification the accuracy was assessed using a percentage of correct classifications (accuracy score). The accuracy for each model can be seen in table 1. As expected, the accuracy score with G1 and G2 was greater than without. Further, it seems in both cases the SVM classifiers performed better than either the logarithmic or random forest classifiers. For the training, the data was split randomly into 80

## 4.3 5-Level Classification

The 5-level classification was done similar to the binary classification. The training data was split in the same way, along with performing 5-fold cross validation. From the accuracy scores (table 2), the random forest classifier performed the best. For the random forest classifier, extra care was needed to achieve maximum accuracy. The maximum number of features for a given tree was set

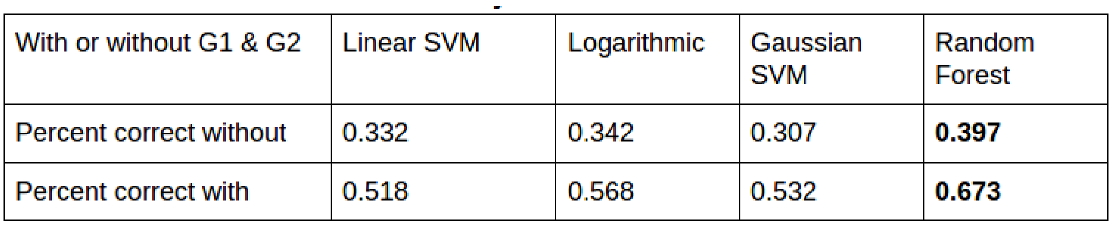
**Table 2: Binary classification percent correct ratio:** bold indicates highest accuracy



| With or without G1 & G2 | Linear SVM | Logarithmic | Gaussian SVM | Random Forest |
|---|---|---|---|---|
| Percent correct without | **0.874** | 0.854 | 0.859 | 0.854 |
| Percent correct with | 0.905 | 0.920 | **0.930** | 0.859 |

to the square root of the number of features, the number of trees was set to 100, and the minimum samples per leaf was found to maximize the accuracy score at 50. The relative feature importance for the top 8 features is shown in figure 1. In table 3 one can see the accuracy of different models. The random forest performed the best under both circumstances. This seems to be due to the low effective dimensionality of the feature space. Random forest performs better in these cases, since it can more easily ignore unhelpful features. This was further validated by removing all features and performing Gaussian SVM by adding one additional feature each time til maximum accuracy was reached. With G1 and G2 this occured at 4 features and the accuracy was close to that of the random forest.

**Table 3: 5-level classification percent correct ratio:** bold indicates highest accuracy



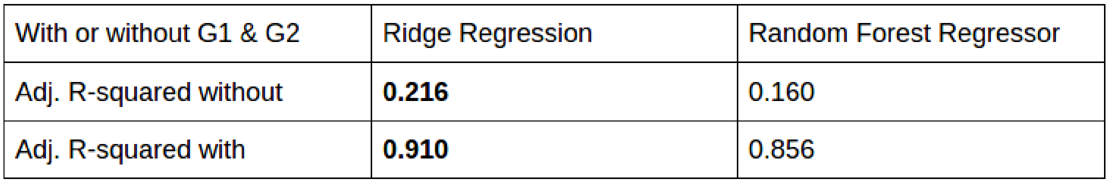
| With or without G1 & G2 | Linear SVM | Logarithmic | Gaussian SVM | Random Forest |
|---|---|---|---|---|
| Percent correct without | 0.332 | 0.342 | 0.307 | **0.397** |
| Percent correct with | 0.518 | 0.568 | 0.532 | **0.673** |

## 4.4 Regression Analysis

Regression was performed with a random forest regressor and ridge regression. Ridge regression was done because of the high collinearity of some of the variables. Again a regularization constant, $/alpha$, for the Ridge regression was grid searched to find the most optimal value. The ridge regression was far superior to the random forest for accuracy. Additionally, the predicted values needed to be rounded to obtain true predictions. This is because the actual G3 values can only take on integer values, but the regression will fit to any value. In fig. 7 one can see the predicted G3 vs. actual G3. This gives a good indication of how well G3 is predicted with the ridge regression. Further, in table 4 one can see the adjusted $R^2$ values for the different models under different circumstances. The Ridge Regression performed the best in both cases. To determine the optimal adjusted $R^2$ value for the regression models, features were added one at a time according to their correlation coefficient values. This is such that any additional feature added has a statistically significant contribution which can be utilized for future data. The models which yielded the best adjusted $R^2$ values had the following features:

**Without G1 and G2:** failures, Medu, studytime, higher, goout, internet
**With G1 and G2:** G2, G1, failures, Medu

**Table 4: Regression adjusted $R^2$ values:** bold indicates highest accuracy



| With or without G1 & G2 | Ridge Regression | Random Forest Regressor |
|---|---|---|
| Adj. R-squared without | **0.216** | 0.160 |
| Adj. R-squared with | **0.910** | 0.856 |

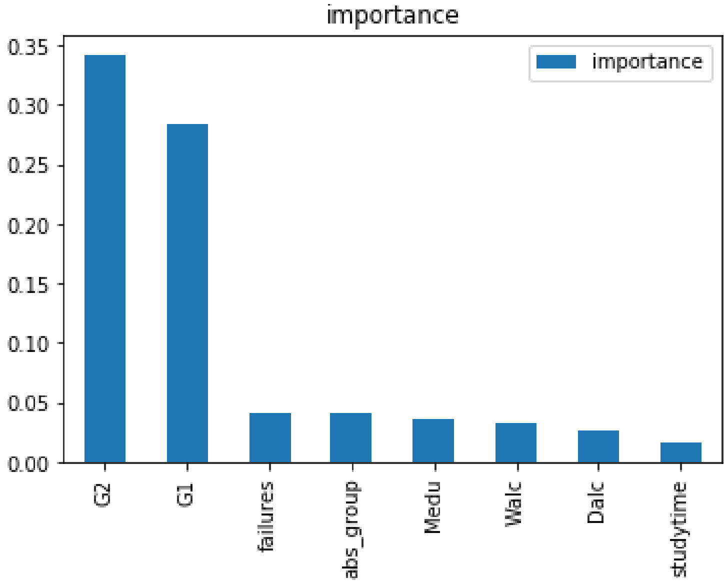


**Figure 6:** relative importance of each feature in the 5-level classification using random forest algorithm.

## 5 Results and Conclusions

Determining the antecedents of academic performance is of great utility of any nation. Using data analysis, one can extract this information and use it for the advantage of the country in a number of ways. The method utilized in this paper is to take information from students obtained through a questionnaire. From there, different machine learning algorithms yielded predictive capability for three different instances: binary (pass fail) classification, 5-level classification, and regression.

From the results in the tables, it is clear that prediction is greatest with binary pass fail classification. Further, this is probably the more useful of the results, as it would allow schools to determine students which are at risk of failure before it's too late. As indicated in table 1, even without G1 and G2, one can predict failure with 87 percent accuracy. From the models, it seems the five features (aside from G1 and G2) which most reflect a students failure possibility are:

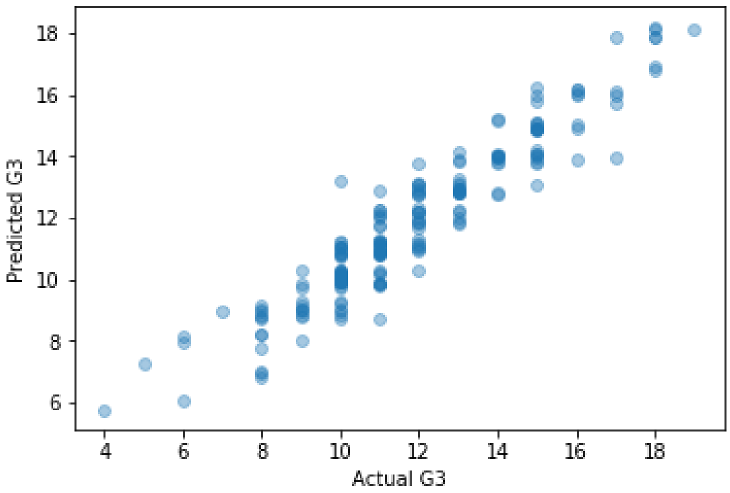


**Figure 7:** Predicted G3 vs. Actual G3 shows how well the ridge regression performed in predicted G3 with G1 and G2

- Previous failures
- Absences
- Study time
- Frequency of going out with friends (Goout)
- Father and mothers education levels

There are multiple solutions which can be implemented given this information. The final decision will have to rest within the schools hands. One possible solution would be to send out an optional/mandatory questionnaire about relevant features. From there, counseling can be provided to at risk students with parent approval. From there, additional analysis can be done to determine the efficacy of the counseling in deterring failure. In this particular analysis it seems the inter-country information had no effect in determining students grades. In additional analysis, more inter-country features should be utilized to determine what Portugal's students would most benefit from.

## References

[1] Jennifer Ma; Matea Pender; Meredith Welch. Education pays 2016, the benefits of higher education for individuals and society. *CollegeBoard*, 2016.

[2] Paulo Cortez and Alice Silva. Using data mining to predict secondary school student performance. *Department Information Systems/Algoritmi R and D Centre University of Minho*, 2008.

[3] The World Bank. percent of gdp spent on education. 2009.

[4] OECD (2019). Mathematics performance (pisa) (indicator). *doi:10.1787/04711c74-en*, 2009.

[5] OECD (2019). Reading performance (pisa) (indicator). *doi:10.1787/04711c74-en*, 2009.

[6] OECD (2019). Average class size. 2010.