Data Analytics
Academic Year 2022-2023

Course Assignment N. 10: Educational Level

Davide Casnici - David Alarcon

March 2023

Abstract

This work explores the performance of different machine learning models in predicting the final grade (**G3**) of high school students based on a dataset of student performance in different courses and demographic variables. We first explore the correlations between variables, and find that almost all variables correlated with **G2** are also similarly correlated with **G3**. We then use a probabilistic graphical model and a Bayesian sampler to impute missing data values more accurately. We initially treat the problem as a regression problem but then realize that classification may be a more suitable approach. We evaluate different models using a 10 folds stratified cross-validation, and find that, among the classic machine learning models an SVM model with PCA-transformed data and a carefully tuned hyperparameters achieves the best performance, with an F1-score of 46.95%. We also compare this model to baseline classifiers and show its superiority. Finally, we demonstrate the importance of the **G2** variable in predicting **G3**, but note that other variables are important in cases of low **G2**. Our work highlights the potential of machine learning models while also simpler heuristic of predictions are suitable under certain circumstances.

1 Introduction

In recent decades, Portugal has made significant strides in improving the educational attainment of its populace. However, despite these efforts, Portugal still lags behind other European countries in terms of student performance, primarily due to high rates of academic failure, particularly in core Mathematics courses. This persistent lack of success among students is a cause for concern. To address this issue, we will be analyzing a real-world dataset that includes information on student performance and other personal data. Our ultimate goal is to develop a model that can accurately predict a student's final grade, which

will enable educators to intervene and provide targeted support to those who may be at risk of failure.

2 Data Analysis

We merged the training and test datasets in order to examine their characteristics and for subsequent performance analysis. The resulting dataset contains 395 rows and 33 columns or features, with the **G3** column representing the final grade attained by each student. The features were divided into 17 categorical and 16 numeric variables, with the categorical features modeled as factors to facilitate manipulation. We conducted exploratory data analysis on each variable, and Figure 1 presents the bar charts and histograms for each feature. It is worth noting that the target variable **G3** exhibits a high number of outliers, which make up 9.6% of the dataset.

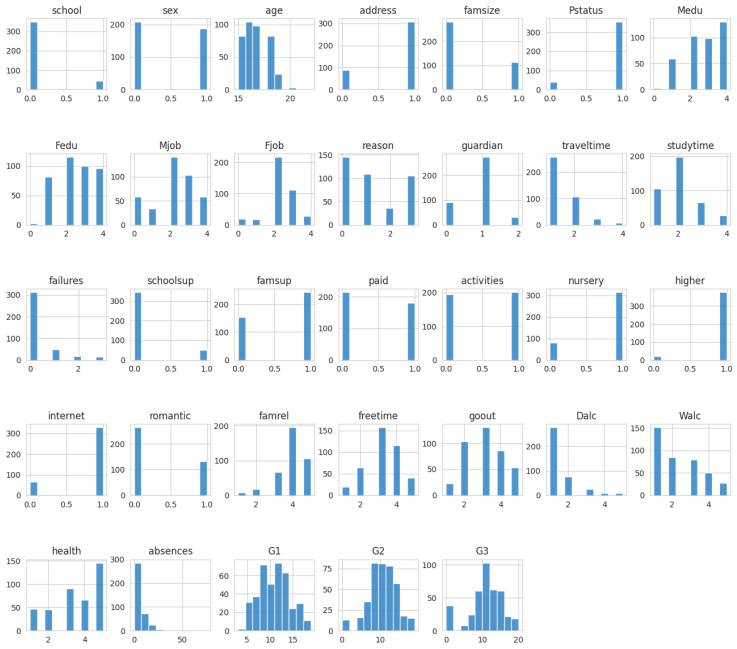

Figure 1: Bar charts and histograms for each feature in the merged dataset, and the **G3** column represents the final grade obtained by each student. The charts provide a visual representation of the distribution of each variable, highlighting the presence of outliers in the target variable **G3**.

Next, we examined the number of outliers in other variables by analyzing boxplots for only the numerical valued variables. We present the boxplots in Figure 2, which show that we obtained 41 datapoints of interest, corresponding to 10.38% of our data. Because the relatively small size of our dataset, we chose not to drop the outliers as they may contain useful information.

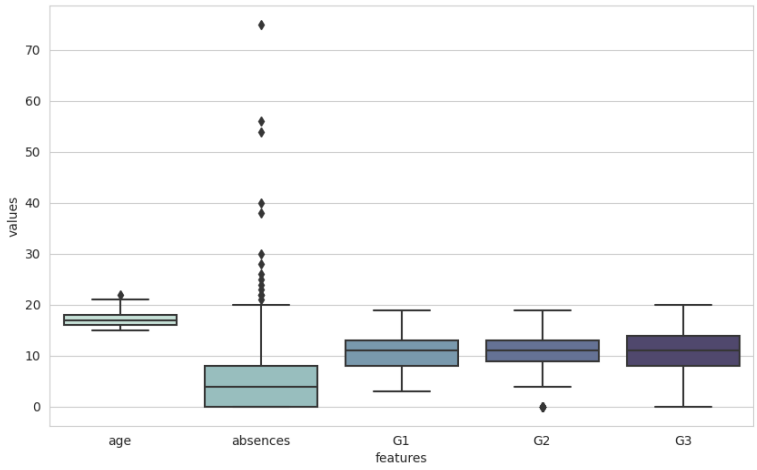

Figure 2: box plots for the numerical valued variables in the datasets, with outliers depicted as dots above the whiskers. The variable **absence** has the highest number of outliers. The box plots provide a visual representation of the distribution of each variable, highlighting the presence of outliers and the spread of the data.

We conducted an analysis of the relationship between the variables by computing both linear (Pearson method) and non-linear (Kendall method) correlations among them. Figure 3 shows that the values of the linear and non-linear correlations are almost identical. Upon examining the correlation matrices, we observed that most of the variables are poorly correlated with each other, with a few exceptions such as the variables **Medu** (Mother education) and **Fedu** (Father education), which display a clear positive correlation between the education levels of a student's parents. We also found that almost every variable is correlated to the feature **G2** in the same way as with the target variable **G3**, indicating that **G2** is the most relevant feature with respect to the target. This is confirmed by the fact that the correlation value between **G3** and **G2** is the highest. Overall, while most of the features are not highly correlated with the target variable **G3**, the first and second period grades (**G1** and **G2** are strongly correlated with it. This makes sense, as a student who did not perform well in the previous exams is less likely to perform well in the final one (perhaps due to lack of study). However, even though the other features are not strongly correlated with **G3**, they may still provide useful information for a predictive model.

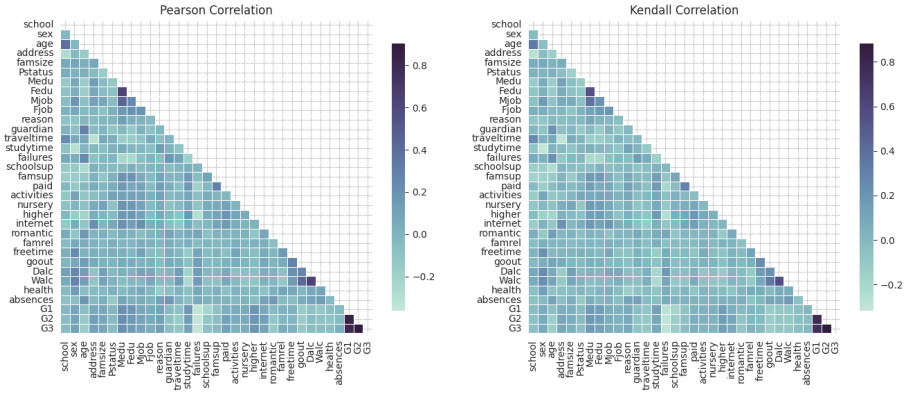

Figure 3: Linear and non-linear correlations among the variables in the dataset. The plot highlights how the values of the linear and non-linear correlations are nearly identical. The correlation matrices reveal that most of the variables are poorly correlated with one another, with some exceptions such as the positive correlation between the education levels of a student's parents. Furthermore, almost all the features have a similar correlation with both **G3** and **G2**, with **G2** being the most informative feature. The correlation between **G3** and **G2** is the highest, while the first and second period grades (**G1** and **G2**) are strongly correlated with the target variable **G3**. Despite the weak correlations among many of the variables and the target variable, these features may still provide useful information for developing a predictive model.

2.1 NaN values

There are 33 datapoints in the dataset that contain **NaN** values in at least one variable. Rather than dropping these datapoints, we opted to sample from a stochastic process that takes into account the distribution of the variable given the rest of the observed variables in the datapoint. The variables that contain **NaN** values are **Medu**, **Fedu**, **famrel**, **Dalc**, and **absences**. To capture the most relevant dependencies between these variables, we constructed a directed probabilistic graphical model using the Peter and Clark (PC) algorithm. The PC algorithm revealed that the distribution of the **Dalc** variable is dependent on the distribution of the **Walc** variable, while the **Fedu** variable has a dependency relationship with the **Medu** variable. Furthermore, when other variables are observed, the distributions of the **famrel** and **absences** variables are independent.

To generate synthetic values from a posterior distribution, conditioned on the observed values of the dependencies for each variable, we created a Bayesian sampler. The parameters for the prior were chosen based on the statistics of the variable of interest in the training set. The evidence data used in the Bayesian sampler was obtained by filtering the datapoints based on the corresponding observed value of the dependent variable. This approach allowed us to preserve the overall structure and patterns in the data, while still accounting for the missing values in the dataset.

For example, the conditional variable **Dalc** | **Walc** (workday alcohol consumption given the weekend alcohol consumption) is modeled as a Poisson variable with a parameter $\lambda_{k k=1,\ldots,5}$, where k is the observed value of **Walc**. The conjugate prior for the Poisson variable is a gamma distribution with parameters 1 and 1/mean(**Dalc**), ensuring that the expected value and variance of the distribution coincide with those of the interest variable. The evidence used to compute the posterior is taken from the **Dalc** values in the training set, filtered based on the specific value of **Walc** observed in our datapoint with **Dalc** = **NaN**. Finally, we sample from the resulting gamma distribution to obtain a value for λ_k, and then the **NaN** value is filled with a sample from Poisson(λ_k).

This heuristic is repeated for all variables with **NaN** values in the dataset, the other models used for the rest of the variables of interest are detailed in the working notebook. This procedure allowed us to impute missing values in a way that preserves the underlying structure and patterns in the data.

3 Experiments

After filling in the null values, we investigated the relationship between **G2** and **G3** by plotting their distributions using a seaborn jointplot, as shown in Figure 4. This gave us an intuitive insight into their relationship and suggested that the problem could perhaps be approached as a regression problem.

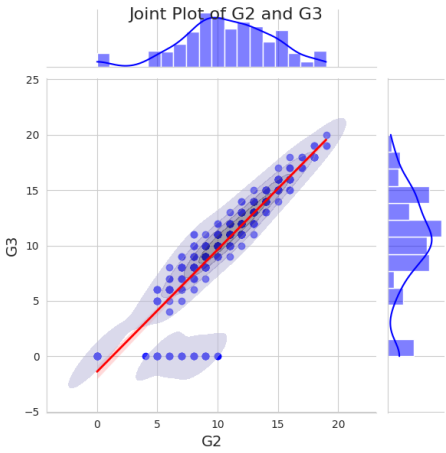

Figure 4: Joint distribution plot of **G2** and **G3** variables, showing a strong positive correlation between the two, except for a few outliers corresponding to students who obtained a final grade of 0. The plot suggests that the problem could be approached as a regression problem.

Therefore, we used three regression models, a linear regression, a random forest regressor, and an MLP regressor, all from the sklearn library, to predict

the value of **G3**. As previously mentioned, we kept all features to predict **G3** as they may still contain useful information for the model. Figure 6 shows the results obtained by the three regression models, with the performance of a 'perfect' regression shown in red. We can observe that the predictions of the regression models are affected by the outliers, which have a greater impact on regression models in general. The resulting pattern is similar to the correlation pattern shown in Figure 4.

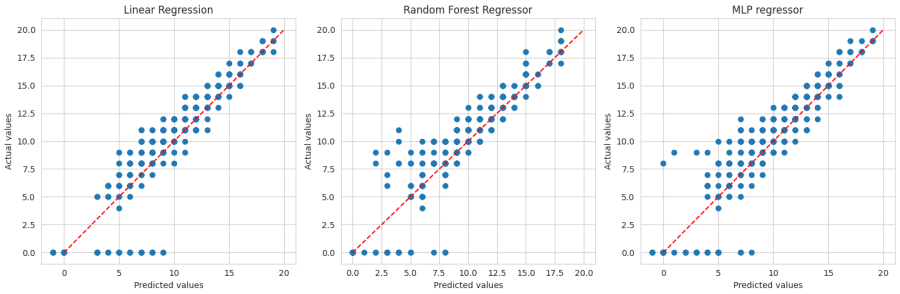

Figure 5: Predictions of the linear regression, random forest regressor, and MLP regressor models for **G3** variable. The figure shows the impact of outliers on the predictions, which closely follow the correlation pattern seen in Figure 4.

Since **G3** is a discrete numerical value, we rounded the regression predictions to the closest integer value. This was also done to allow for a fair comparison of performance metrics with the subsequently used classification models.

After rounding the regression predictions to the closest integer value, we attempted to perform dimensionality reduction on our dataset by applying Principal Component Analysis (PCA) while retaining 95% of the variance. This resulted in a reduced set of 12 features, which is less than half of the original number of features, and led to a denser dataset. By plotting the data with respect to the first three principal components, as shown in Figure 6, we noticed a pattern that suggested a clustering approach and indicated that the problem could be treated as a classification problem.

We employed three classification models for our analysis: the Random Forest Classifier, Multi Layer Perceptron, and Support Vector Machine. These models were trained using both the original data and the data processed with the PCA library of scikit-learn.

All the models (regression and classification) were trained using a stratified 10-fold cross-validation approach. This method preserves the distribution of the target variable in each test fold (10% of the entire dataset), and provides a more accurate indication of the performance of the models, since each model is tested with different training and test data.

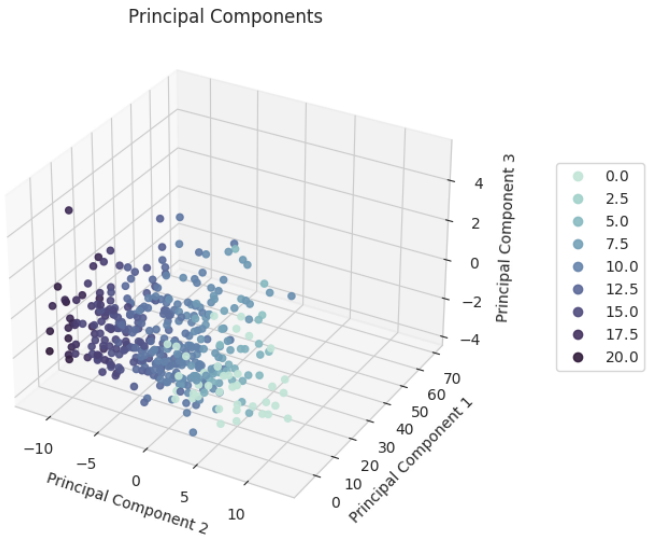

Figure 6: Data projection onto the first three principal components, showing a pattern that suggests a clustering approach and indicates that the problem could be treated as a classification problem.

4 Results

In Figure 7, we observe the results obtained by each model on the 10 validation folds (averaged mean and standard deviation of the 10 trials). The best-performing model is the SVM, which is typically robust against outliers due to its margin, and used with PCA-transformed data. This model achieves an F1-Score performance of 44.75% without tuning (averaged on the 10 folds). As the target variable is imbalanced, accuracy was not used as a comparison metric, and instead, the F1-score was used as the primary metric for model evaluation. Recall that the F1-score is the harmonic sum of precision and recall.

After comparing the models, we selected the best one and proceeded with fine-tuning its hyperparameters. Specifically, we searched for the optimal combination of SVM's kernel type, gamma (kernel coefficient), c (regularization parameter), and degree (degree of the polynomial kernel function). Following the hyperparameter tuning, we evaluated the best model's performance and compared it to that of the two baseline classifiers (random and biased). As depicted in Figure 8, the best model achieved an average F1-score of almost 47% (precisely, 46.95%) using 10-fold stratified cross-validation, exhibiting a markedly superior performance compared to the random and biased models. We can notice that, using **G2** as a direct predictor led to a similar performance as our best machine learning model, but such an approach does not leverage information from other features, making it more susceptible to outliers and yielding a higher standard deviation.

Figure 9 displays the confusion matrices of the model's predictions, as well

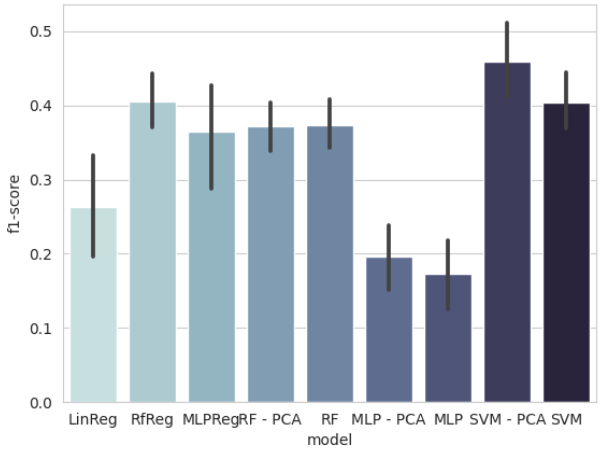

Figure 7: Results obtained by the classification models on the 10-fold cross-validation, displayed as the mean and standard deviation of the F1-score for each model and data configuration. The SVM model used with PCA-transformed data achieved the highest performance.

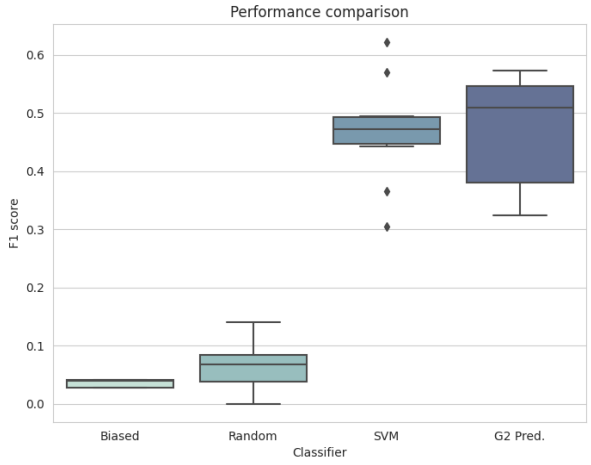

Figure 8: comparison of the final tuned SVM model's performance with the baseline classifiers (random and biased). The models were evaluated using F1-score achieved through 10-folds stratified cross validation. The final model is clearly superior to the baseline models.

as using **G2** as a predictor. The results show that while **G2** is a better predictor for non-outlier datapoints, it struggles to predict outlier datapoints accurately. Conversely, the SVM model consistently outperforms **G2** in predicting outlier datapoints, while being slightly less precise for non-outlier datapoints. However,

the model only misclassifies the grades of just one value in most cases, which is still acceptable given the nature of the problem.

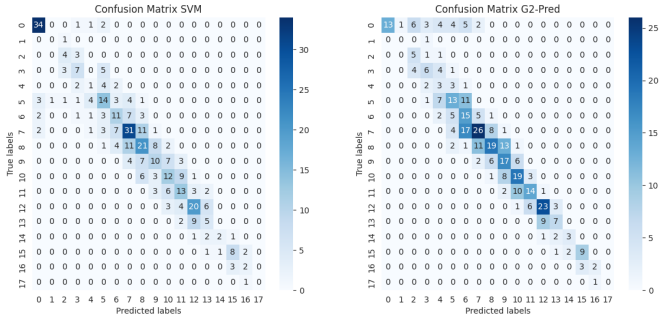

Figure 9: The confusion matrices for the final SVM model and the model using **G2** as the predictor are displayed in Figure 9. The results illustrate how **G2** performs better as a predictor for the non-outlier datapoints, while the SVM model is more effective in handling the outliers. Although the SVM model may misclassify the grades of a few students by just one value, the overall performance of the model is still satisfactory for the given problem.

5 Conclusions

Upon exploring the correlations among variables, we observed that the variables correlated with **G2** were also almost identically correlated with **G3** (linearly and not linearly). This suggests that by observing **G2**, the statistical influence of the remaining variables on **G3** is mitigated. Additionally, we found that the grade before the final test (**G2**) carries the highest amount of information regarding the performance of the students on the final test.

Initially, we approached the problem as a regression task, but we obtained poor performance. Therefore, we rephrased the problem as a classification task. Among the classifiers tested, the SVM achieved the best performance. Even in cases of misclassification, the predicted values were close to the actual values, making the model an effective indicator of the student's final grade.

Based on the available data, a direct predictor model for the student's performance in **G3** could be achieved by assessing their performance in the **G2** test. However, for cases of low **G2**, the SVM model could be used to incorporate information from other variables and improve the prediction performance.

For future work, it would be beneficial to collect more data to create a better dataset that could allow the model to achieve better performance. Additionally, collecting more evidence on the correlations between the analyzed features could also help to improve the model.