# Modeling Student Exam Performance

ALBERTO SALVARESE, CHRIS LAWSON, JEFFREY GORDON, and UTKARSH MUJUMDAR

This study aims to model student exam performance as a function of several variables such as academic indicators, demographics, family background and others, and trying to evaluate which features contribute the most in determining students' academic success, to have a better understanding on where to intervene to guarantee fairness.

**ACM Reference Format:**
Alberto Salvarese, Chris Lawson, Jeffrey Gordon, and Utkarsh Mujumdar. 2023. Modeling Student Exam Performance. 1, 1 (April 2023), 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

Academic performance is a crucial metric to measure student success in any educational system. Over the years, educators have used various techniques to understand student learning and performance, including traditional tests, quizzes, and assignments. However, the advent of data science and machine learning has revolutionized the way we analyze and interpret student performance data. This has led to the development of predictive models that can forecast student exam performance based on various factors such as study habits, attendance, and demographics. This project aims to explore the different approaches to modeling student exam performance and their effectiveness in predicting student success. By doing so, we hope to provide insights into how educators can use these models to better understand student learning and make data-driven decisions to improve academic outcomes.

The use of data science and machine learning techniques to model student performance has gained increasing attention in recent years. The study by Kovacic, Z. [Kovačić and Nz 2010] tried to analyze the prediction of student success using socio-demographic features, i.e., education, work, gender, status, disability, etc., and course features such as course program, course block, etc., for effective prediction. Ethnicity, course program, and course block were identified as the top three features affecting students' success as part of the study. Sotiris B. [Kotsiantis 2012] used four types of features as part of their study: student demographic data, e-learning system logs, academic data, and admission information. Treating the problem as a classification task, they trained five different classifiers: Model Tree (MT), Neural Net, Linear Regression (LR), Locally Weighted Linear Regression, and Support Vector Machine (SVM) with the MT predictor attaining the lowest Mean Absolute Error (MAE).

Aggarwal et al. [Aggarwal et al. 2021] tried to highlight the significance of nonacademic features in predicting student performance, such as demographic information. Similar to the previous study, the academic performance was treated as a categorical class in this study, with Random Forest having the best F-1 score of 93.8% amongst 8 other classifiers. Their findings showed the importance of demographic information for predicting the student's performance. Oyedeji et al. [Oyedeji et al. 2020] applied machine learning-based techniques on a feature-set of past results with the combination of individual attributes such as the student's age, demographic distribution, and family background achieving an MAE of 3.26 using deep learning methods. Cortez et al.[Cortez and Silva 2008] used a dataset of Portuguese secondary school students for two subjects (Math & Portuguese) having academic features such as past performance and study time combined with demographic variables and information about the students' family backgrounds. They were able to achieve RMSE scores of 1.75 and 1.32 for Math and Portuguese, respectively using a Random-Forest regression model.

## 2 DATA

The dataset used in this project contains grades of students from two secondary schools in Portugal (distribution can be seen in Fig 1) in the subjects of Mathematics and Portuguese language, collected during the 2005-2006 school year.

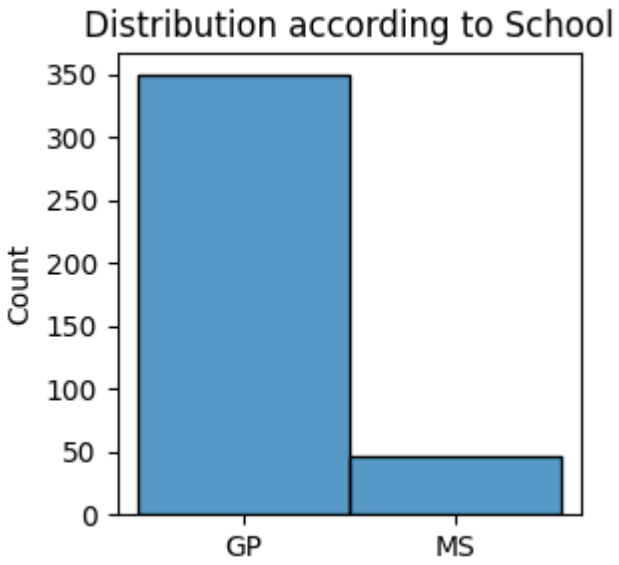


Fig. 1. Distribution of students in the two schools, Gabriel Pereira ('GP') and Mousinho da Silveira ('MS')

In the Portuguese educational system, the students are evaluated in three periods and the last evaluation ('G3') corresponds to the final grade in any subject. 'G3' will be the target variable in our study. The dataset is feature-rich with features representing academic

Authors' address: Alberto Salvarese; Chris Lawson; Jeffrey Gordon; Utkarsh Mujumdar.


XXXX-XXXX/2023/4-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

indicators, demographics, family background and behavioral traits (Table 1)

The dataset consists of 395 datapoints related to Mathematics and 649 related to the Portuguese language subjects, respectively. The database was built from two sources: school reports capturing the academic indicators (i.e. the three period grades and number of school absences) and questionnaires, used to collect several demographic (e.g. mother's education, family income), social/emotional (e.g. alcohol consumption) and school related (e.g. number of past class failures) variables that were expected to affect student performance.

## 3 EXPLORATION

### 3.1 Correlation Analysis

Looking at the correlation matrix (Fig. 2), we can make the following observations:

- The features Father's Education and Mother's Education have the largest positive correlation with the target G3
- The features Failures and Goout have the largest negative correlations with the target G3
- The pairs of features Medu & Fedu, Dalc & Walc and G1 & G2 are highly correlated with each other

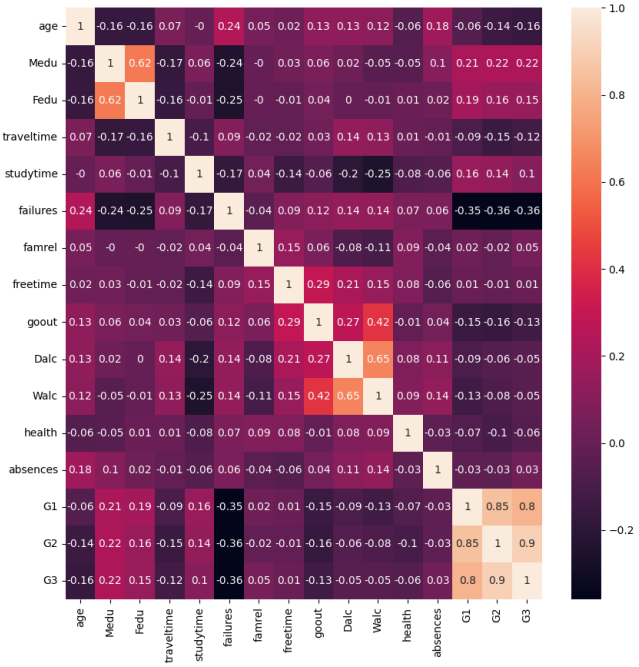


Fig. 2. Correlation Matrix of Features

### 3.2 Features of Interest

After having explored possible correlations between the features, we decided to focus only on the ones with higher correlation. To these, we also added the parents job, as we believed it could effect students' performances. To have a better understanding on how these features influenced students' scores, we plotted the G3 exam results against "Parents Education", "Parents Job", "Age", and "Past Failures" using box and whisker plots (Fig. 3).

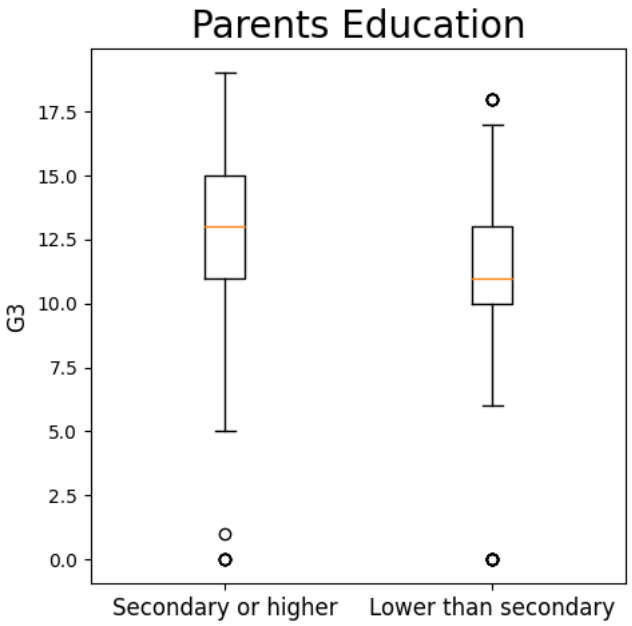


Fig. 3. Parents Education box plot


Fig. 4. Parents Job box plot

| Attribute | Description |
|---|---|
| **sex** | student's sex (binary: female or male) |
| **gender** | 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) |
| **Mjob** | mother's job (nominal) |
| **Fedu** | father's education (numeric: from 0 to 4) |
| **Fjob** | father's job (nominal) |
| **guardian** | student's guardian (nominal: mother, father or other) |
| **famsize** | family size (binary: less than equal to 3 or more than 3) |
| **famrel** | quality of family relationships (numeric: from 1 – very bad to 5 – excellent) |
| **reason** | reason to choose this school (nominal) |
| **traveltime** | home to school travel time (numeric: 0-15 min, 15-30 min, 30-60 min or 60 min+) |
| **studytime** | weekly study time (numeric: 0 to 2 hours, 2 to 5 hours, 5 to 10 hours or 10 hours+) |
| **failures** | number of past class failures (numeric: n if n is less than 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** | in 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) |

Table 1. Features as part of the Dataset

It is possible to see that each of these features seems to have an effect on students' performances. For instance, parents with a higher education favour higher grades (Q1, Q2, and Q3 are higher for higher education). Even by looking at the different statistics for student with at least one parent as a teacher and no teaching parents we see that the former is slightly shifted towards higher grades.

Also the differences in the failures faced by a student seem to strongly influence the results of future exams; the two distributions here are very different, with the one for less failures having much higher Q1, Q2, and Q3. Although not as clearly as in the latter case, age seems to affect the performance of students, as at higher ages the grades distributions seem to converge towards lower values.

### 3.3 KLLR

We then proceeded to search for possible trends with respect to the aforementioned features. We used KLLR with Gaussian kernel on both Math and Portuguese final exam scores, and plotted it against "Parents education", "Age", and "Past Failures". We find the same overall trends for both Portuguese and Math exams. However, due to the size of data being double in the former, the predictions obtained with KLLR are more precise in the former case. Therefore, only those results are shown in the following.

In Fig. 7 we plotted the influence that parents education may have in the students performances. We notice that the mother's and father's education produces the same overall trend, showing us that the higher the parental education, the higher the average score of the students. This is in agreement with what was learnt from the correlation matrix, and will motivate the decision of dropping one of the two feature in the modeling phase.

We investigated then the possible relation between the job of the parents and the students' results. Here, we observed a softly decreasing trend as we move away from teaching parents. Students with teaching parents seem to have a G3 score slightly higher than

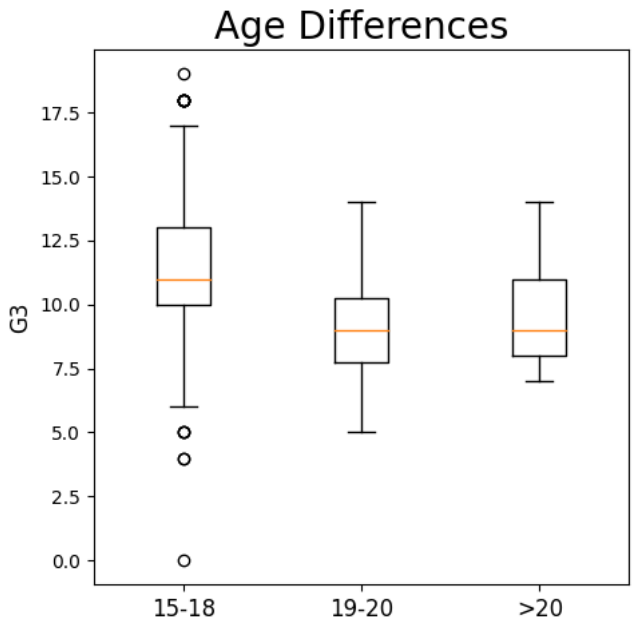


Fig. 5. Student Age box plot

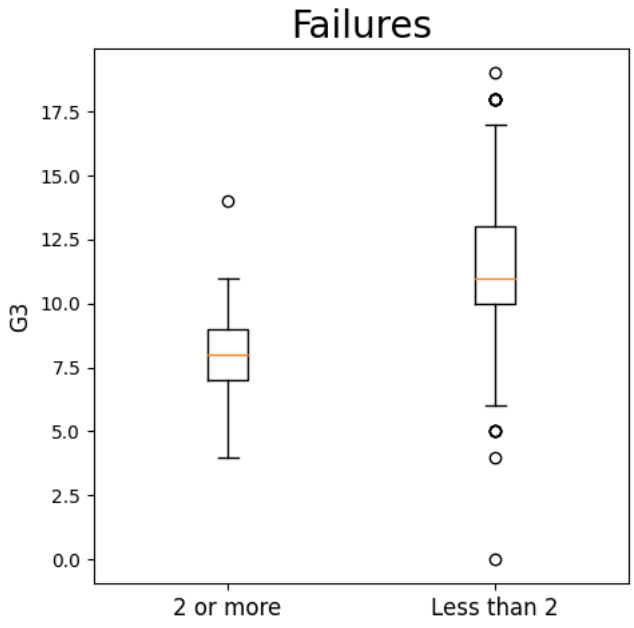


Fig. 6. Number of classes failed box plot

the ones with different parents job.

In Fig. 9, we explored the effect of the age. We can see that, as the age increases, the final exams scores tend to decrease rapidly. We thought that this could be justified by the fact that the data we had referred to students currently enrolled in high school, whose natural extension in Portugal would be from 15 to 18 years old. The Math case, is slightly different as we have a more rapidly decreasing average score even within the 15 – 18 age range. For Portuguese, instead we notice that overall students perform at the same level for the whole duration of their high school path.

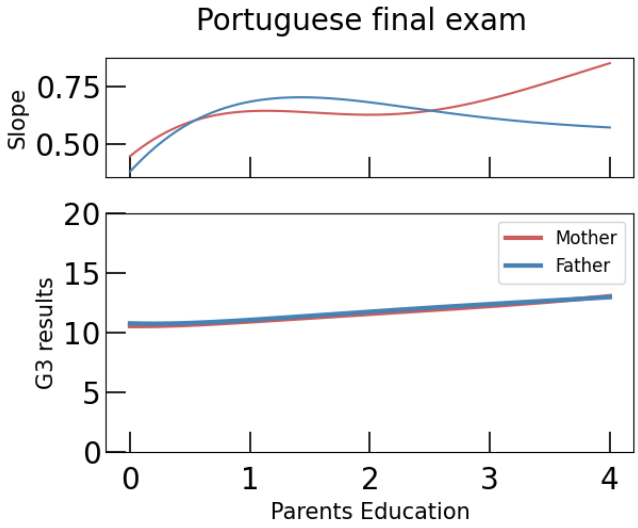


Fig. 7. KLLR for the Parents education

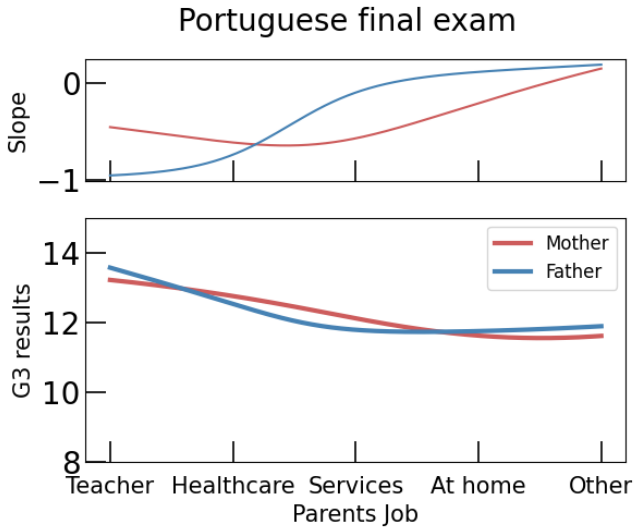


Fig. 8. KLLR for the Parents Job

Finally, in Fig. 10, we find that as the number of failures increases, the final scores decreases, which brings out a possible correlation between past and present students performances, a result already pointed out by Cortez [Cortez and Silva 2008].
As pointed out in the correlation matrix study, also the feature

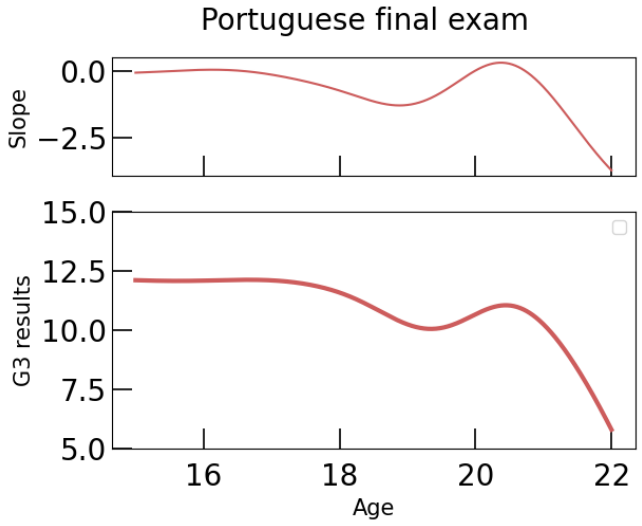


Fig. 9. KLLR for the Student Age

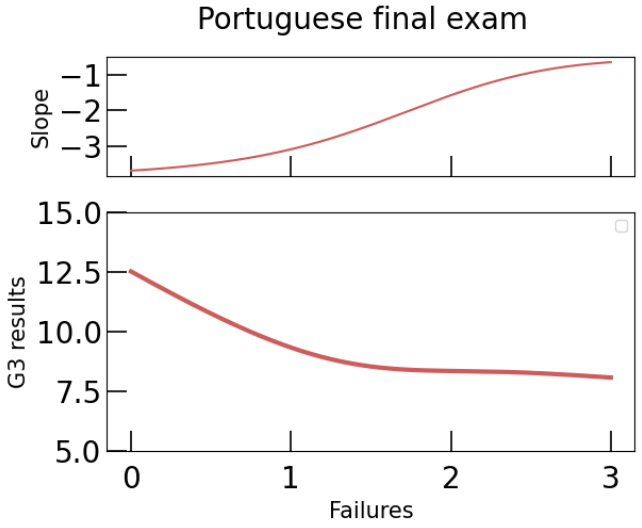


Fig. 10. KLLR for the student number of failures

"Goout" seems to have a high correlation. Therefore, we wanted to search for possible effects of a student's social life in their G3 results. In order to do that, we considered different aspects related to the definition of "social life", and defined a new variable called "Student Quality" (SQ). That was given by summing together the normalized values of "Absences", " (weekly) Hours spent going out", "Daily alcohol", and "Weekend alcohol consumption". This new featured had a maximum value of 3.38 for Portuguese and 3.21 for Math. We then defined a student to be a "Good Student" if SQ falls between 0 and 1, an "Average Student" if $1 < SQ \leq 2.1$, where this upper boundary has been defined by the mean of SQ plus its standard deviation. Finally, a "Bad Student" was defined such that $SQ > 2.1$.

We then proceeded by stratifying the trends for parents education vs G3 results with respect to SQ (Fig. 11).

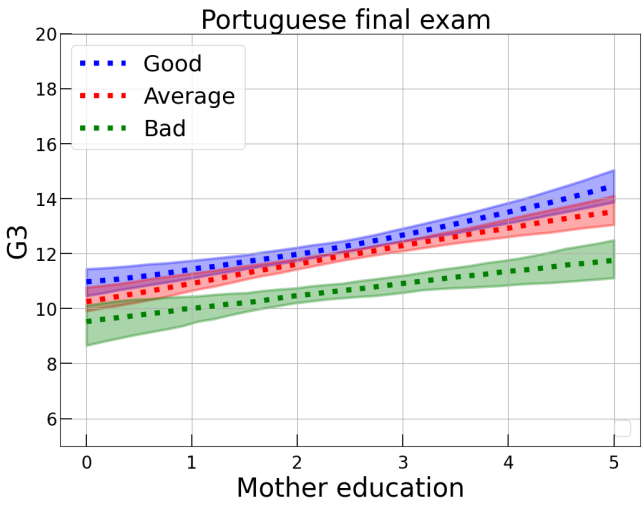


Fig. 11. New Student Quality KLLR

We can see from here that the quality of a student seems to have a direct impact on their average performances. This effect is the same no matter the level of education of the parents. Good students in general perform better than bad students with an average score of $2 - 3$ points higher.
Then, we did the same for the school difference, stratifying the students with respect to their school: Gabriel Pereira and Mousinho da Silveira.

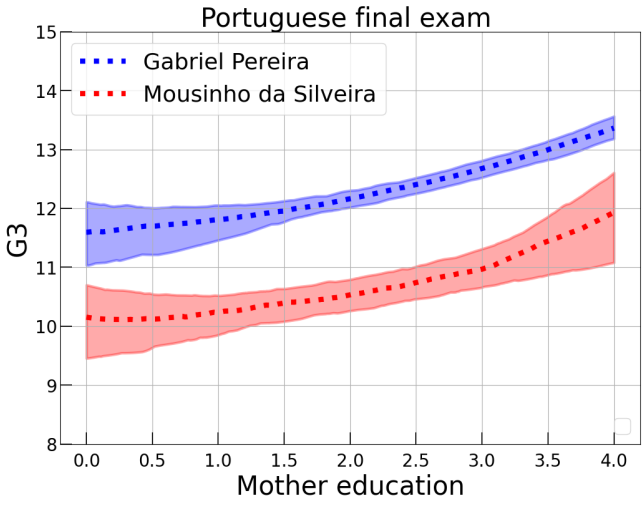


Fig. 12. KLLR on mother education for Portuguese

In Fig. 12 we observe the same trends for both schools, which confirms the importance of parents education as a positive correlation

factor for students performances. However, it seems that a different school favours higher average grades at each level of mother education. On the other hand, as noticed in the data section (Fig 1, the distribution of students in the two schools is extremely different. Students coming from Mousinho da Silveira, are one seventh of the ones form Gabriel Pereira. This fact motivated us to doubt the precision of this prediction and to discard the hypothesis that school choice could effect the students performance.

### 3.4 Hypotheses

All the results obtained with KLLR analysis suggests several hypotheses we decided to follow:

(1) Parents education has a positive correlation with students performances
(2) Students with one or more parents that are teachers have higher grades
(3) Features related to student quality have a negative correlation with students performances
(4) Past failures have a negative correlation with students performances
(5) Age has a negative correlation with students performances

## 4 MODELING

### 4.1 Modeling Pipeline

There were three steps to our modeling pipeline to get to the results that we got. Those steps are:

(1) Feature Manipulation
(2) Data Manipulation
(3) Model Hypertuning

*4.1.1 Feature Manipulation.* With all the information we obtained within the feature exploration in the sections above, we needed to decide which features we wanted to keep and which ones we wanted to discard.

We decided to remove the following variables:

- Reason
- School
- Guardian
- Father Education
- Weekday Alcohol Consumption

The first 3 features were discarded due to their low variability and the last two features were discarded due them being highly correlated with another feature. Father Education being correlated with Mother Education and weekday alcohol consumption being correlated with weekend alcohol consumption.

Next, we needed to one hot encode the remaining categorical variables into their individual values. Prior to the one hot encoding we had 34 features and after encoding we would expand to 48 features so this would not impact the feature space drastically.

Lastly, we needed to drop the G1 and G2 scores given how correlated the values are with the G3 score and this would be unfair to give to the model and would not lead to any important information gained.

*4.1.2 Data Manipulation.* Next, we have three different manipulations to the data. These steps are in place to add more variety into the dataset itself.

(1) Base Models with Hyper-tuning
(2) Scaling Features
(3) Feature Reduction

The way we use all the different steps is by progressively attaching each new modification on top of the last step. This means that for step 2, we will be scaling the features along with hyper-tuning and for step 3 we will be including everything.

The first step includes just using our base models with various values to hyper-tune (visit section 4.2.2 for specifics). This is to ensure that we have allowed the model the best opportunity to conform to the data.

The next modification we introduced was to scale our features before running another set of models with hyper-tuning. We chose to include this step due to us including models that are very specific to the dimensionality of the data such as SVM and Linear models.

The last step that we chose to include was feature reduction through Principle Component Analysis (PCA). This was to try to further reduce the features that we had as it could help the models even further with trying to fit with the data itself.

### 4.2 Basic Models

*4.2.1 Chosen Models.* We wanted a wide range of models to try to capture the problem as much as possible. This includes linear, kernel and tree type models. The models we chose are the following:

- Gaussian Process
- Kernel Ridge
- SVM
- Decision Tree
- Random Forest

These models help to encapsulate all aspects of the machine learning concepts that we learned over the course of the semester.

*4.2.2 Hypertuning Specifics.* The following are the features that were hypertuned for each model and values used for each model.

- Gaussian Process
  - kernel: [Constant * RBF, Dot + White, RBF + White]
- Kernel Ridge
  - kernel: [Linear, Polynomial, RBF, Sigmoid]
  - alpha: 0 to 1 in increments of 0.1
- SVM
  - kernel: [Linear, Polynomial, RBF, Sigmoid]
  - degree: 0 to 12
- Decision Tree
  - Max Depth: 5 to 20 in increments of 5
  - Max Features: 5 to 15 in increments of 5
  - Minimum Leaf Samples: 5 to 20 in increments of 5
- Random Forest
  - Max Depth: 5 to 20 in increments of 5
  - Max Features: 5 to 15 in increments of 5
  - Minimum Leaf Samples: 5 to 20 in increments of 5

### 4.3 Predictor + GAN Neural Network

We also constructed a deep learning architecture for our last model, comprising of two neural networks: PredictorNet and a GAN. PredictorNet utilized 30 student performance features (excluding G1 & G2) to predict the G3 score. The GAN had two networks - the generator and the discriminator. The generator generated data that appeared to be from the same distribution as the actual data, while the discriminator evaluated whether the sample was real or from the generator. Both models were trained concurrently, learning from one another.

Our primary objective was twofold. Firstly, to develop an accurate neural network capable of predicting the G3 score, and secondly, to determine if this network could predict the score of unseen, simulated data. If successful, this would provide valuable insight into the network's ability to predict test scores in future datasets or even different exams.

### 4.4 Validation

For all models that we created, we performed 3 different kinds of validation strategies.

(1) 10 fold validation
(2) Quartile-Quartile (QQ) plots
(3) SHAP values

We chose the 10 fold validation to account for a fair split of data for the model to both be able to run and train with. Next we wanted to look at the QQ plots in order to validate that our models were truly learning or if any kind of issues were occurring within the model learning. Lastly we chose SHAP values as a validation technique given that we are interested in which features are most predictive in the students test score and this would be a good way of interpreting which features helped correlate with the test score.

## 5 RESULTS

### 5.1 Basic Models Results

*5.1.1 Mean Absolute Error Values.* The following table contains the Mean Absolute Error (MAE) values and their standard deviation for the top 12 models. The units are in grade points.

| Model | Step | MAE | MAE-STD |
|---|---|---|---|
| RF | 2 | 3.167 | 0.515 |
| GP | 1 | 3.298 | 0.617 |
| SVM | 3 | 3.301 | 0.610 |
| GP | 2 | 3.301 | 0.610 |
| GP | 3 | 3.335 | 0.603 |
| L | 1 | 3.335 | 0.603 |
| L | 2 | 3.356 | 0.598 |
| RF | 3 | 3.370 | 0.598 |
| L | 3 | 3.383 | 0.600 |
| SVM | 1 | 3.433 | 0.667 |
| D | 2 | 3.533 | 0.585 |
| KR | 1 | 3.562 | 0.544 |

The "Model" column refers to the Model that was run where GP stands for Gaussian Process, RF stands for Random Forest, L stands for Linear, SVM stands for Support Vector Machine, KR stands for Kernel Ridge and D stands for Decision Tree. "Step" column represents which data manipulation step the model was trained with.

*5.1.2 Best Hyper Parameters.* For the best 3 models, these are the best parameters.

(1) Random Forest - Scaled:
- Max Depth: 5
- Max Feature: 15
- Min Samples Leaf: 5

(2) Gaussian Process - Base:
- kernel: DotProduct + WhiteKernel
- alpha: 1.0

(3) SVM - All:
- degree: 2
- kernel: Sigmoid

*5.1.3 QQ Plots.* The figures 13 14 and 15 contain the QQ plots for the top 3 models

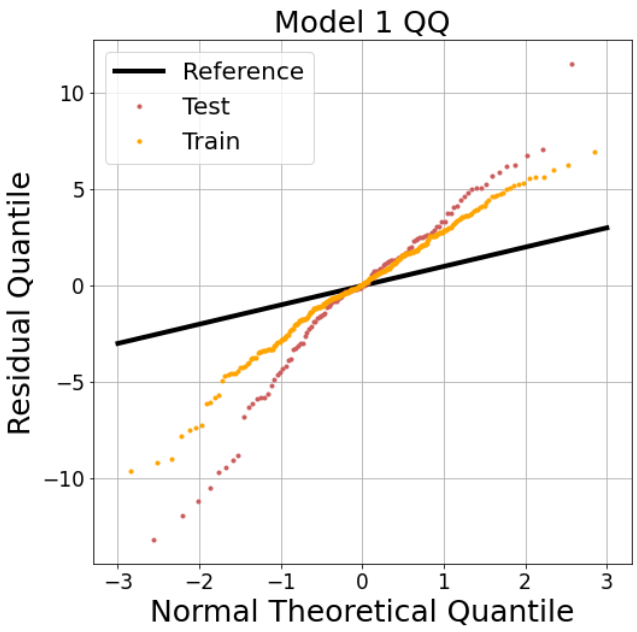


Fig. 13. QQ plot for RF - scaled model

*5.1.4 SHAP Value.* Figure 16 is the SHAP beeswarm plot of the best model to look at the most influential features in predicting the target feature.

### 5.2 Predictor Network Tests & Results

We began by preprocessing the data, which involved isolating the G3 feature (did not scale or normalize these values) and dropping G1 & G2. Next, we normalized the input features to scale them between 0 to 1 to expedite the neural network's training and convergence. Then, we created six unique Predictor networks, each with a different configuration. Specifically, we had the following networks:

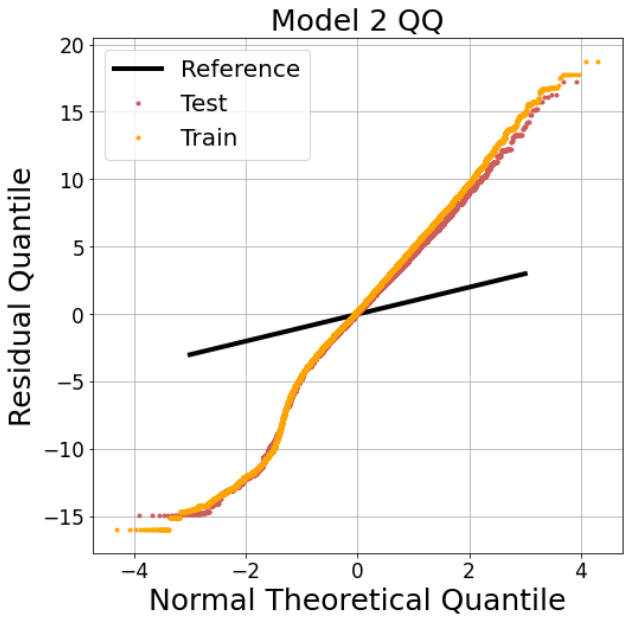


Fig. 14. QQ plot for GP - base model

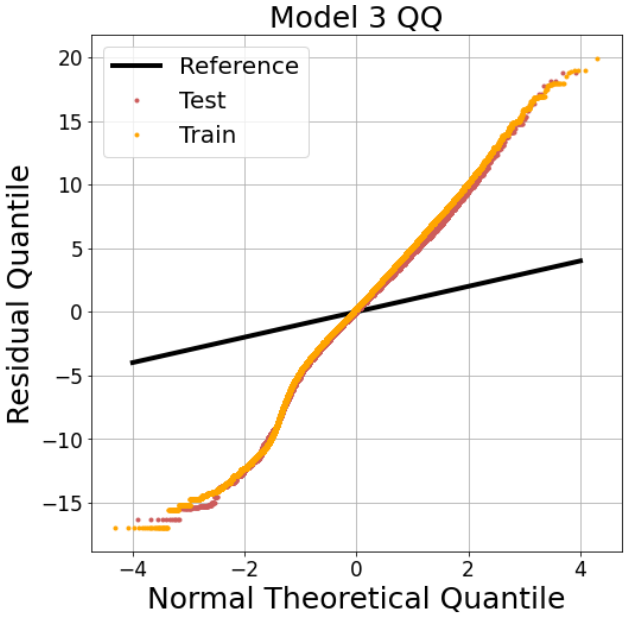


Fig. 15. QQ plot for SVM - all model

- PredictorNet-1-32: 1 Hidden Layer, 32 Nodes per Layer
- PredictorNet-2-32: 2 Hidden Layers, 32 Nodes per Layer
- PredictorNet-1-64: 1 Hidden Layer, 64 Nodes per Layer
- PredictorNet-2-64: 2 Hidden Layers, 64 Nodes per Layer
- PredictorNet-4-32: 4 Hidden Layers, 32 Nodes per Layer
- PredictorNet-4-64: 4 Hidden Layers, 64 Nodes per Layer

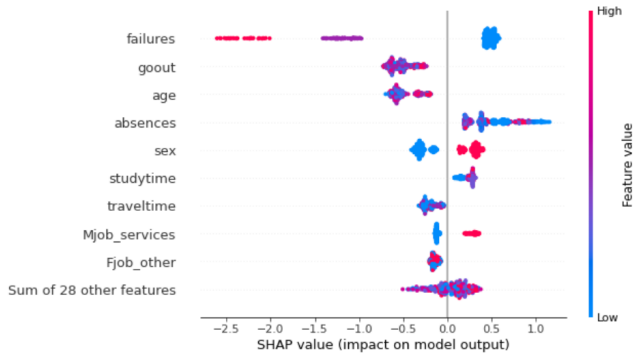


Fig. 16. SHAP beeswarm for RF - scaled model

All the layers of the neural networks utilized a ReLU activation function. We trained the models with a batch size of 32 for 100 epochs. The next figure shows the loss of each model per epoch. MSE = $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, $Y_i$ being the actual score, $\hat{Y}_i$ being the predicted. The results shown in Figure 17

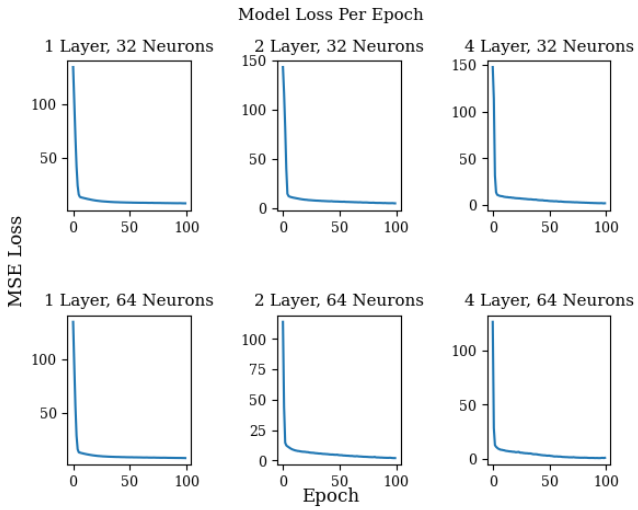


Fig. 17. Predictor MSE

The PredictorNet-4-64 model delivered the best performance, with a mean squared error (MSE) loss of 0.76, which is reasonable given the output range of 0-20. To avoid overfitting, we conducted a 10-fold cross-validation, which yielded an impressive average MSE loss of 0.15.

To determine the effect of the model's complexity on its performance, we plotted the model parameter size against the loss. The next figure presents the model parameter size and the corresponding average MSE loss of the 10-Fold Cross Validation. This being shown in Figure 18

As expected, increasing the model size greatly improved the performance. We were particularly pleased with the results of PredictorNet-4-64 and decided to use it for further analysis.

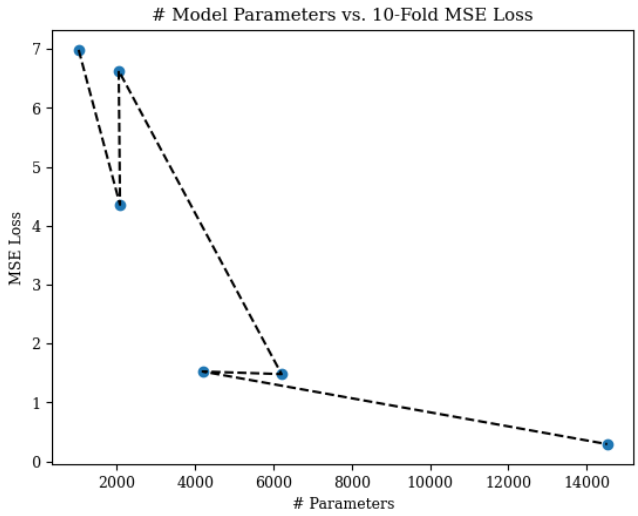


Fig. 18. Predictor Loss training

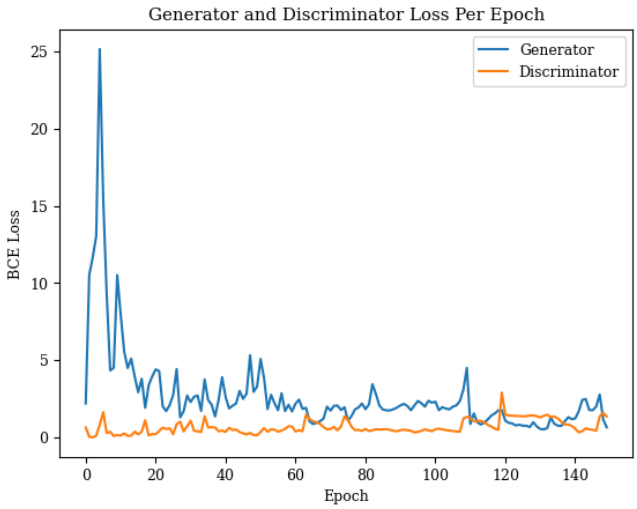


Fig. 19. GAN Discriminator and Generator Loss

### 5.3 GAN Tests & Results

We implemented a large GAN network despite having a relatively small dataset. The generator and discriminator both had eight hidden layers, with 64 nodes in each layer. We opted for a large model because with a smaller one, we observed that the models never reached equilibrium in loss, which prevented them from training off each other.

The generator received input in the form of a batch size x 128-dimensional random vector and produced a batch size x 30 matrix of simulated samples. The discriminator, on the other hand, took in a batch size x 30 matrix of samples (real or fake) and predicted whether they were genuine or fake. During training, the discriminator was first trained by comparing its predictions against a tensor of all 1's or all 0's, depending on whether we input fake or real data. Then, the generator was trained by comparing the discriminator's predictions against a tensor of all 1's, aiming to make the discriminator perceive all the samples it receives as real.

All of the generator layers employed the ReLU activation function, and the output layer also had ReLU, preventing simulated data from having negative feature values. The input data was normalized similarly to before.

The discriminator layers used ReLU activation, but the output layer employed a sigmoid activation to obtain a probability of 'real'.

We trained the GAN model for 150 epochs, with a batch size of 32, and the Binary Cross-Entropy (BCE) Loss per epoch is displayed in the next figure. The BCE loss compares the actual real/fake label to the predicted probability of being real from the network.

Figure 19 illustrates that the discriminator was challenging the generator significantly until around epoch 115, after which an equilibrium was nearly achieved. Given the already large model and the number of epochs, we concluded that this model would suffice for our purposes.

We utilized the generator to produce 500 simulated student performance samples and then ran them through our predictor network. The lowest predicted G3 score was 9.40, with the lowest actual score being 0, while the highest predicted G3 score was 15.54, with the highest actual score being 19. We were satisfied with these results, as the predicted score range was within the actual range. This gave us confidence that our model could perform well on unseen data. Hence, if we were to receive more student performance data or use the model to predict scores for new students, we could rely on it.

## 6 DISCUSSION

### 6.1 Basic Model Interpretation

Here, it seems that we are able to get some good results with our best model with an mean absolute error of around 3. Here given the fact that our test scores are measured on the scale from 0 to 20, an MAE of 3 suggests that there is still quite a bit of error in our model to be able to predict accurately. However, due to this low variablity in the score, we do have to be flexible with our model.

### 6.2 QQ plot Interpretation

The next method we used to try to interpret our model's performance and accuracy was through constructing their respective Quartile-Quartile plots (QQ plots) in Figures 13, 14, and 15. For all 3 models, but especially the last 2, we see that with these plots are very abnormal with respect to the identity line that represents a perfect model. We notice that for models 2 and 3, we have some very interesting behavior in the negative region of the graphs.

The QQ plot for model 1 is slightly better in that it is closer to the identity line, but not to the point where it can be considered acceptable.

### 6.3 SHAP Plot Interpretation

We only calculated the beeswarm plot for the best model (Figure 16) to interpret what the features best help to influence the model prediction. Here we see that, in accordance with a few of our hypothesis, the top features in grade prediction are: absences, failures, how

much they go out, and amount of time spent studying. All of these features fall under the "academic" variables that was mentioned in our hypothesis.

There are two other features that were mentioned in the plot that stood out as interesting, those two being the students age and sex. When discussed further, the fact that age had a negative correlation on their test score made sense. In that, the higher the age someone is in their last year, the more likely the chance that they have a low grade.

What was very interesting was that there was a very clear separation between a students sex and the impact on their test score. While the feature itself didn't have that much effect on the test score the more interesting aspect was that it was incredibly homogeneous in the distribution.

### 6.4 Limitations

*6.4.1 Data Limitations.* There were two main limitations with the data. The first one being that there were multiple disproportionate group sizes within the data. This was mainly explored in the Data Exploration section, but to reiterate, there were multiple variables that had a very low variance, which school a student went to, whether they lived in a rural or urban area, whether they wanted to go to higher education, and various other instances. This in addition to the low amount of data that we have causes us to be worried about how well our models can do to extrapolate the underlying manifold the data lies on.

The next issue we encountered with the data was the way we interpreted the data. For a majority of the data, the numerical data that they contain actually a categorical tag. For example, with parents education, the values 0 through 4 represent specific education levels. Where 0 represents no education, and 4 represents them higher education. This means that these numbers should not really be treated as numerical values but as the categories for which they describe. However, the reason why we did continue to go with this plan was because we believed that the patterns that the numbers picked were representative in how impactful we believed the feature was on the test score.

## 7 CONCLUSION

Throughout the course of our project we were able to explore, clean and eventually model the students in Portugal to see if we could get a better understanding as to what was most important in determining their final grade from school. We saw through feature exploration that there were various different groups that we could identify and view and saw meaningful differences in scores. From this we came up with 4 different hypothesis about our data.

(1) A Parent's Education level had an effect on the student's performance.
(2) A Parent's job (specifically teachers) had an effect on a student's performance.
(3) A student's after-school life had an impact on their grade.
(4) Past failures affects the students' performances.
(5) A student's age had an impact on their grade.

Through modeling, we were able to identify our best model along with the best data and model pipeline to get the best results. This being a Random Forest Model (parameters used: max depth being 5, max feature being 15, minimum samples leaf being 5) with a scaled dataset to get a mean absolute error of 3.167 grade points with a standard deviation of 0.515.

From this model we saw that while it did not have very representative Quartile-Quartile plots, it's SHAP values were much more interesting. Show casing that the most important features in predicting a students final score being the number of their absences and the number of failed classes. From these results we can make the following conclusions about our hypothesis:

- Hypothesis 1: **True**. We see this from the important features from from the SHAP of the top model.
- Hypothesis 2: **Plausible**, Similar to hypothesis 1, we did not have a very large sample of students that had parents as teachers and thus we were not able to conclude with certainty if it was true or not.
- Hypothesis 3: **Plausible**, Here we see a few variables shown in the SHAP features such as "going out" to have an impact, but not all the features we described. Thus we leave this as "plausible."
- Hypothesis 4: **True**, from the SHAP plot, we see that failures negatively influence the students' performances.
- Hypothesis 5: **True**, We see this from the important features from the SHAP of the top model.

These ultimately these results are somewhat conclusive with our hypotheses. However, it seemed like the most impactful features were those related to a students academic career. These results were very conclusive with the work done in work done by Cortez and Silva.[Cortez and Silva 2008].

## 8 ACKNOWLEDGEMENTS

The contribution of the team members can be referred to in Table 2

| Name | Contribution |
|---|---|
| Alberto Salvarese | 100 |
| Chris Lawson | 100 |
| Jeffrey Gordon | 100 |
| Utkarsh Mujumdar | 100 |

Table 2. Contribution of each team member

## REFERENCES

Deepti Aggarwal, Sonu Mittal, and Vikram Bali. 2021. Significance of Non-Academic Parameters for Predicting Student Performance Using Ensemble Learning Techniques. *International Journal of System Dynamics Applications (IJSDA)* 10, 3 (2021), 38–49. https://ideas.repec.org//a/igg/jsda00/v10y2021i3p38-49.html Publisher: IGI Global.

Paulo Cortez and Alice Silva. 2008. Using data mining to predict secondary school student performance. *EUROSIS* (Jan. 2008).

S. B. Kotsiantis. 2012. Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. *Artificial Intelligence Review* 37, 4 (April 2012), 331–344. https://doi.org/10.1007/s10462-011-9234-x

Zlatko Kovačić and Nz. 2010. Early Prediction of Student Success: Mining Students Enrolment Data. (Jan. 2010).

Ajibola Oluwafemi Oyedeji, Abdulrazaq M. Salami, Olaolu Folorunsho, and Olatilewa R. Abolade. 2020. Analysis and Prediction of Student Academic Performance Using Machine Learning. *JITCE (Journal of Information Technology and Computer Engineering)* 4, 01 (March 2020), 10–15. https://doi.org/10.25077/jitce.4.01.10-15.2020 Number: 01.