STAT 385 Final Project Abstract

By: Vincent Johnson, Joseph Kang, Catherine Sung

Abstract

In this project, we study what academic, personal, and social factors help predict students' final grades in Math and Portuguese. We use a combined dataset of student information and look at things like study time, family background, and school support to see how they affect performance. We start with graphs and summary statistics to better understand the data and choose useful models.

We use many statistical and machine learning methods, including multiple linear regression, lasso, ridge, PCR, PLS, support vector machines (SVM), random forests, and boosting. For each method, we explain how the model works, which variables we used, how we chose the best settings, and how well each model fits the data. We also explain what the results mean and how they answer our questions.

We compare all models using the same training and test sets and measure performance using RMSE and R^2. We set a random seed so results can be repeated. In the end, we suggest the best models based on accuracy and ease of understanding. Our results can help teachers and schools understand what factors matter most for student success.

EDA: Exploratory Data Analysis

We begin our analysis by examining the dataset, which includes detailed information on students from two Portuguese secondary schools. Each row in the data represents one student, and each column is a feature describing that student's academic, personal, or social background.

In total, there are 32 variables. The key variables include:

- **school** – the student's school: Gabriel Pereira (GP) or Mousinho da Silveira (MS)
- **sex** – student's gender: Female (F) or Male (M)
- **age** – student's age (from 15 to 22)
- **address** – type of home address: Urban (U) or Rural (R)
- **famsize** – family size: LE3 (3 or fewer) or GT3 (greater than 3)
- **Pstatus** – parents' cohabitation status: Together (T) or Apart (A)

- **Medu, Fedu** – education level of mother and father (from 0 to 4)
- **Mjob, Fjob** – job type of mother and father
- **reason** – reason for choosing the school
- **guardian** – student's main guardian: mother, father, or other
- **traveltime, studytime** – time spent traveling to school and studying per week
- **failures** – number of past class failures
- **schoolsup, famsup, paid** – access to additional support or classes
- **activities, nursery, higher, internet, romantic** – student's personal background and interests
- **famrel, freetime, goout, Dalc, Walc, health** – lifestyle and well-being factors (rated from 1 to 5)
- **absences** – number of absences from school
- **G1, G2** – grades from the first and second periods (0–20)
- **G3** – **final grade**, which is our target variable and the outcome we aim to predict.

We use summary statistics and visualizations such as histograms, boxplots, and correlation matrices to understand the relationships between variables. This helps us identify which features might be important predictors for the final grade and decide which modeling techniques are most appropriate for the data.

Preparing the Data for Our Use

Before analyzing the data, we performed several steps to combine and clean the datasets so they would be ready for modeling.

We started with two separate datasets: one for students taking Math and another for those taking Portuguese. Each dataset included student information and course-specific grades.We merged them using a set of common background variables. These variables included: school, sex, age, address, family size, parental status, parental education and jobs, reason for choosing the school, whether they attended nursery school, and whether they had internet access at home.

```
math <- read.csv("math.csv")
por <- read.csv("portuguese.csv")

join_cols <- c("school", "sex", "age", "address", "famsize", "Pstatus",
               "Medu", "Fedu", "Mjob", "Fjob", "reason", "nursery", "internet")

merged_data <- merge(math, por, by = join_cols, suffixes = c(".math", ".por"))
```

After merging the data, we renamed the grade columns for clarity:

- `mathp1`, `mathp2`, `mathfinal` for Math grades from periods 1, 2, and the final grade
- `porp1`, `porp2`, `porfinal` for Portuguese grades

We then removed the Math grades when predicting Portuguese outcomes (and vice versa), so that grades from one subject would not be used to predict the other.

All character variables were converted into factors for proper handling in modeling, and we confirmed that the final cleaned dataset had exactly 36 columns.

We also checked for missing values across all variables:

```
cat("\nMissing data summary:\n")
print(sapply(clean_data, function(x) sum(is.na(x))))
```

The result showed that there were **no missing values** in our cleaned dataset. This allowed us to proceed confidently to model fitting without needing imputation or additional cleaning.

After Accomplishing all of this we began the next step which was to begin visualising our data.

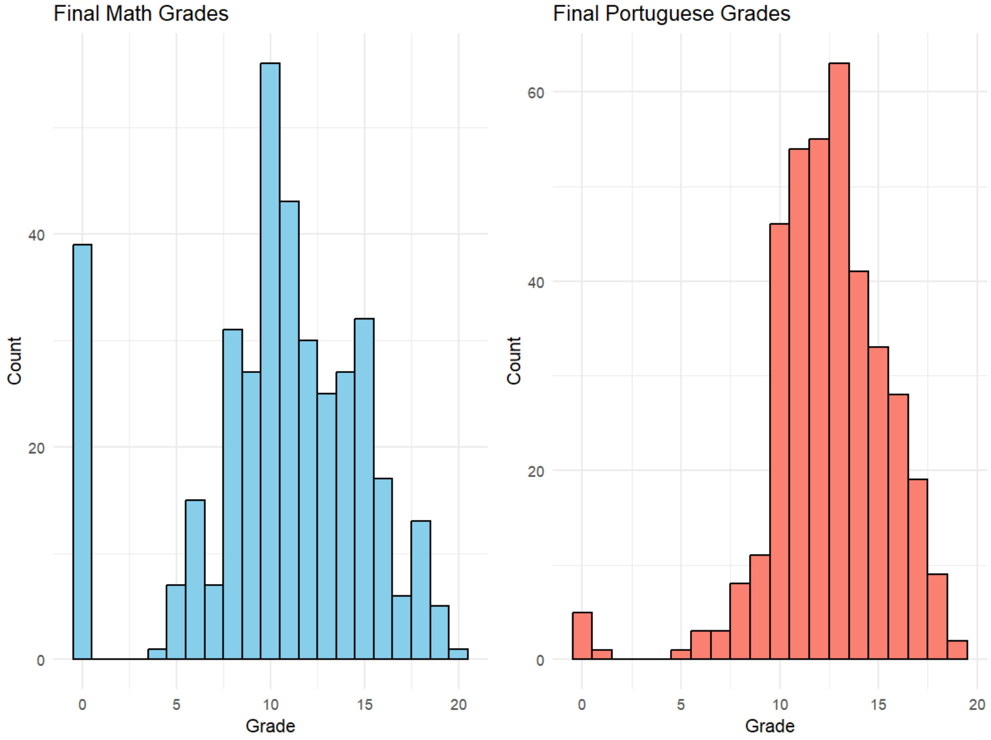

Figure 1: Distribution of Final Grades

The distribution of final grades in Math and Portuguese, as depicted in the histograms, highlights key differences between the two subjects. Most notably, the Math histogram shows a disproportionate number of students receiving a final grade of zero, a pattern that is not

Table 1: Model Performance Summary for Math and Portuguese

Statistic	Math	Portuguese
Min	0.000000	0.000000
1st Quartile	8.000000	11.000000
Median	11.000000	13.000000
Mean	10.387435	12.515707
3rd Quartile	14.000000	14.000000
Max	20.000000	19.000000
SD	4.687242	2.945438

observed in the Portuguese data. This concentration of zeros introduces a strong left skew in the Math distribution and likely reflects a subset of students who failed to engage with the course or dropped out entirely. This anomaly can severely impact the performance of predictive models for Math, as these outliers increase variance and reduce model stability.

In contrast, the Portuguese scores display a more consistent distribution, with a slight shift to the right and a tighter clustering of values around the mean. This is supported by the lower standard deviation in Portuguese (2.95) compared to Math (4.69), as shown in the accompanying table. The Portuguese grades also have a higher mean (12.52) and median (13) than Math (mean = 10.39, median = 11), indicating overall better performance in that subject. Excluding the zeros, the Math scores appear to approximate a normal distribution, but the presence of those extreme values creates challenges in modeling.

These histogram insights help explain why models predicting Portuguese grades generally outperform those predicting Math grades. The greater consistency and fewer extreme values in the Portuguese data provide a more reliable foundation for accurate prediction.

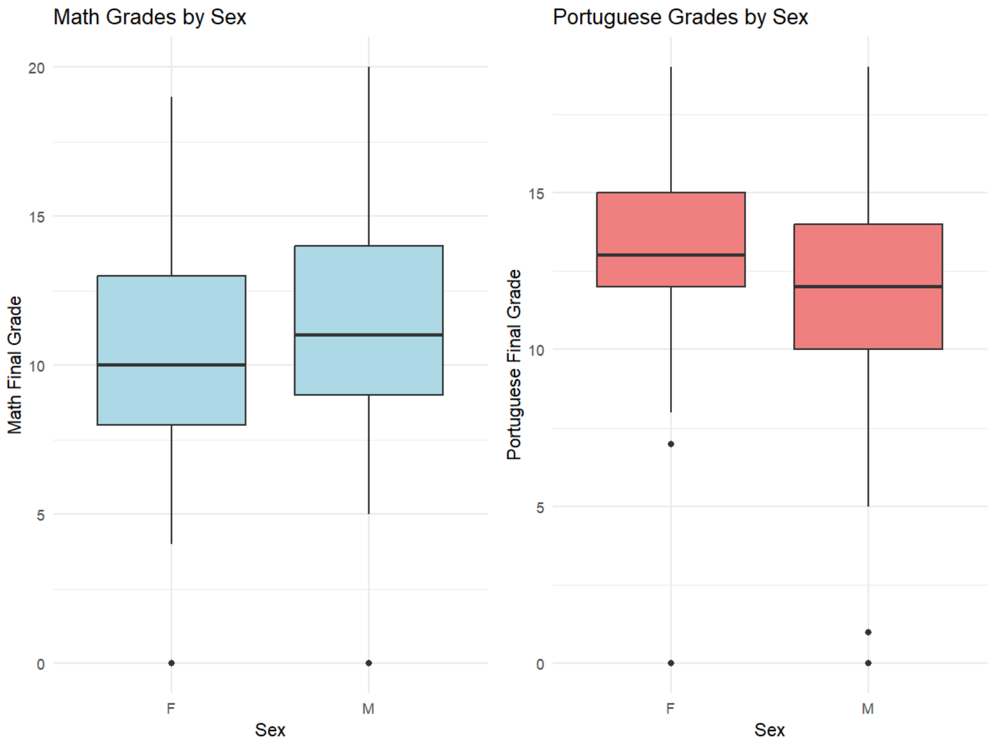

Figure 2: Distribution of Sex by Subject

The distribution of final grades in Math and Portuguese, as illustrated by the histograms, reveals several important differences between the two subjects. Most notably, Math exhibits a disproportionate number of zero grades, likely indicating cases of course failure or withdrawal. These zeros introduce a substantial left skew and inflate the variance, making predictive modeling more difficult and less stable for Math. In contrast, Portuguese grades display a more symmetric, right-shifted distribution with tighter clustering, as confirmed by the lower standard deviation (2.95 vs. 4.69 for Math). When zero values are excluded, the Math distribution resembles a normal curve, but the presence of extreme low values remains a critical factor impacting model accuracy.

Further insights are gained through QQ plots organized by gender, particularly because gender was one of the predictor we beleived would have the biggest impact in our analysis. In Math, both male and female students show wide variability in final grades, with males slightly outperforming females on average. However, in Portuguese, this pattern reverses—female students outperform males, and both genders exhibit relatively tight distributions centered near the mean, with only a few outliers. This suggests gender-based performance differences not only vary by subject but also interact with the overall data spread. The tighter grouping of Portuguese scores across genders again points to greater predictability in that subject, contributing to stronger model performance compared to Math.

To understand how different variables are related to each other and to the final grades, we looked at a correlation plot. This plot helped us see both how the variables interact with one another and which ones are most connected to student performance in Math and Portuguese.

We found some strong positive and negative relationships. For example, students who drink alcohol on the weekends are also more likely to drink during the week. This is important because weekday drinking has a noticeable negative effect on grades. In contrast, higher levels of parental education—especially the mother's—are linked to better student performance. This may be because parents with more education are better able to support their children academically.

Overall, the correlation plot shows that both personal behavior and family background can have a big impact on how well students do in school. This information helps guide which variables we focus on when building and interpreting our prediction models.

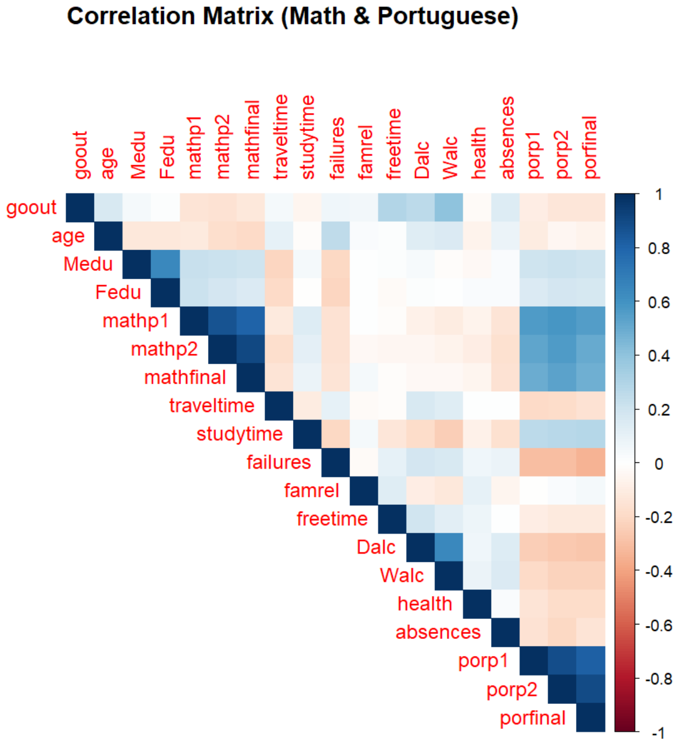

Figure 3: Heat map of Variable Corelation

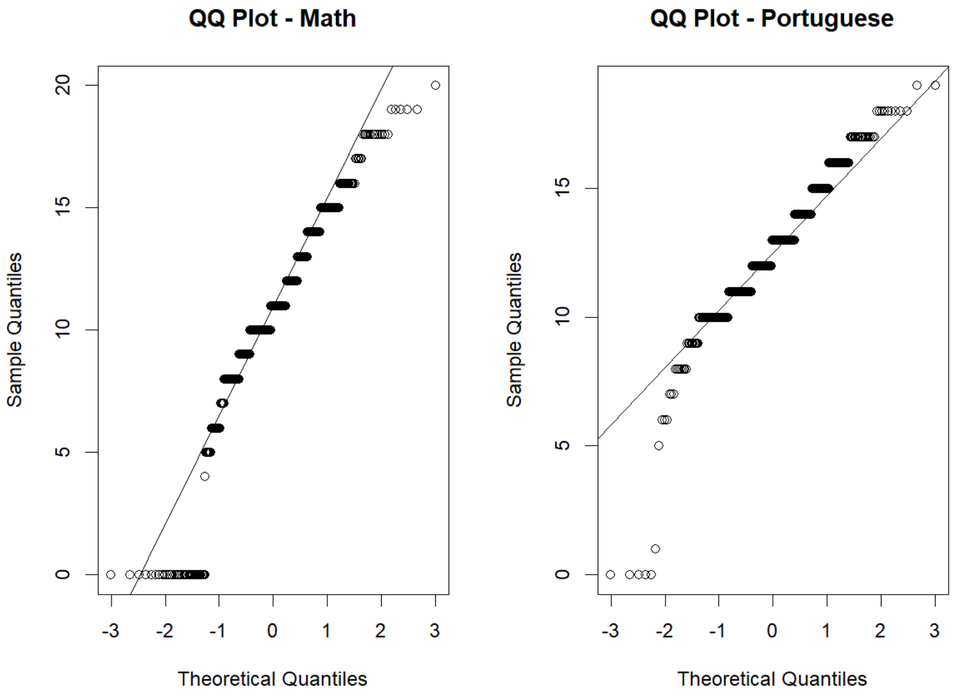

Figure 4: QQ Plot

Looking at the QQ plot, we can immediately see a distinct difference between Math and Portuguese scores. For Math, the slope is much steeper and starts below zero. This suggests that the data for Math is heavily skewed, primarily due to the high number of zeros in the dataset. These zeros, likely representing students who scored very low or had no score at all, have a large influence on the distribution, making the slope steeper in the lower range. This steeper slope at the start indicates that our model has trouble predicting the low scores, possibly because of these zero values that distort the model's accuracy.

As we move through the middle of the data, the accuracy improves, reflecting that the model can better predict the students who are performing at average or higher levels. However, the accuracy again drops as we approach the higher scores. This drop is likely due to the difficulty in predicting the performance of students who score high, as the zeros in the dataset may make it harder to capture the more extreme ends of the distribution properly. Essentially, the zeros drag down the model's ability to generalize well at the higher end of the scale.

On the other hand, the QQ plot for Portuguese shows a much less steep slope. Although the accuracy starts low (as is common with many models at the lower end), the slope is more gradual, and the accuracy remains relatively stable throughout. This suggests that the data for Portuguese is more normally distributed, with fewer extreme values or outliers (such as the zeros seen in Math). The more consistent distribution allows for a more reliable and stable model, especially for higher scores, where the model can maintain higher accuracy without being distorted by outliers or extreme low values.

In summary, the QQ plots for both Math and Portuguese highlight key differences in the data distributions. The steepness of Math's plot indicates that the presence of zeros is a significant challenge for predicting scores, especially at the extremes. Portuguese, on the other hand, appears to have a more evenly spread dataset, making it easier for the model to predict scores with more consistency across the entire range.

Linear Regression

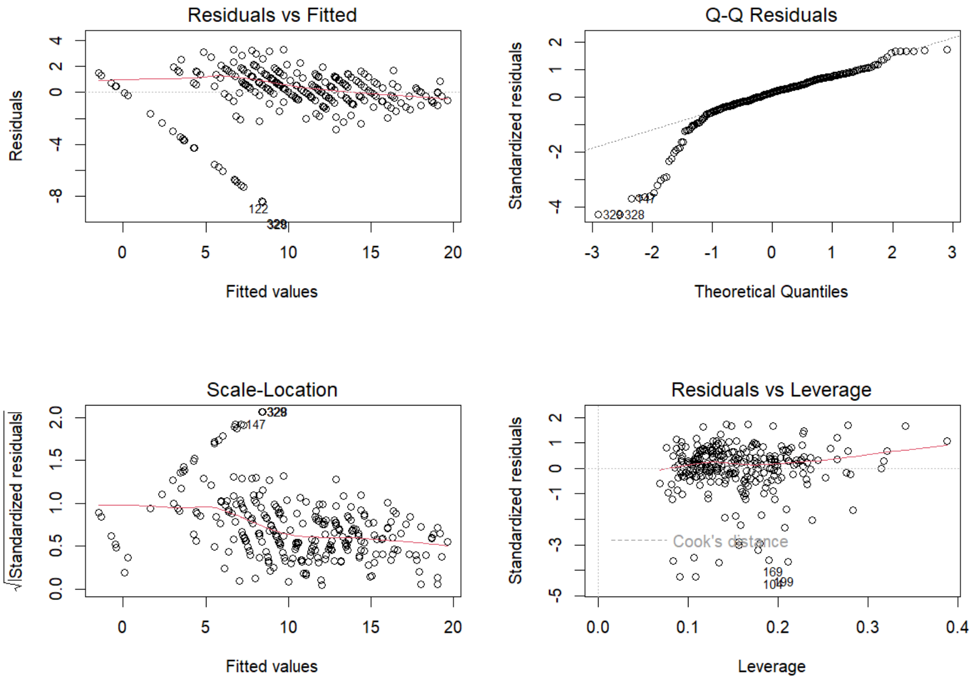

Figure 5: Assumptions Check for LR

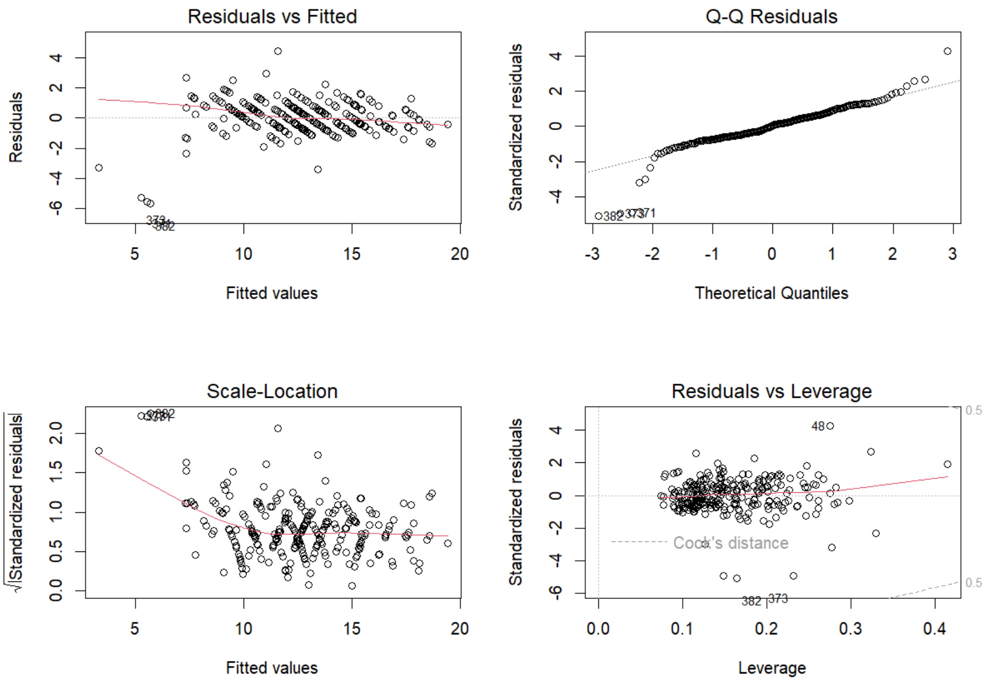

Figure 6: Assumptions Check for LR

Here we don't have much to worry about in regards to Multicellularity. There are some worrisome points that go just over 5 witch is considered relatively high, however, none go over ten witch is our cutoff point. We also see some worrisome points for the Assumptions checks, however, it is not to bad overall and we believe all of our requirements are met.

Table 2: Side-by-Side VIF Comparison for Math and Portuguese Linear Models

Variable	M_VIF	M_Df	M_VIF^(1/(2*Df))	P_VIF	P_Df	P_VIF^(1/(2*Df))
guardian	1.480193	2	1.103010	1.571115	2	1.119572
goout	1.608989	1	1.268459	1.681249	1	1.296630
school	1.641374	1	1.281161	1.614344	1	1.270568
sex	1.480817	1	1.216888	1.496284	1	1.223227
age	1.683393	1	1.297456	1.633526	1	1.278095
address	1.434334	1	1.197637	1.520429	1	1.233057
famsize	1.155246	1	1.074824	1.248482	1	1.117355
Pstatus	1.189486	1	1.090636	1.220292	1	1.104668
Medu	3.225022	1	1.795835	3.197912	1	1.788271
Fedu	2.256879	1	1.502291	2.371832	1	1.540075
Mjob	4.711666	4	1.213799	4.768852	4	1.215631
Fjob	2.917813	4	1.143226	3.076145	4	1.150803
reason	1.812124	3	1.104158	1.901376	3	1.113041
nursery	1.247515	1	1.116922	1.200685	1	1.095758
internet	1.282259	1	1.132369	1.337381	1	1.156452
mathp1	5.567227	1	2.359497	NA	NA	NA
mathp2	5.443060	1	2.333037	NA	NA	NA
traveltime	1.374866	1	1.172547	1.388562	1	1.178373
studytime	1.453808	1	1.205739	1.601372	1	1.265453
failures	1.475491	1	1.214698	1.495908	1	1.223073
schoolsup	1.395075	1	1.181133	1.401559	1	1.183875
famsup	1.250634	1	1.118317	1.199915	1	1.095406
paid	1.233115	1	1.110457	1.251845	1	1.118859
activities	1.209743	1	1.099883	1.303474	1	1.141698
higher	1.367377	1	1.169349	1.432669	1	1.196942
romantic	1.278935	1	1.130900	1.262127	1	1.123444
famrel	1.220360	1	1.104699	1.226586	1	1.107514
freetime	1.448505	1	1.203539	1.338110	1	1.156767
Dalc	1.941432	1	1.393353	2.008371	1	1.417170
Walc	2.239369	1	1.496452	2.403011	1	1.550165
health	1.295174	1	1.138057	1.351867	1	1.162698
absences	1.260738	1	1.122826	1.312955	1	1.145842
porp1	NA	NA	NA	5.218640	1	2.284434
porp2	NA	NA	NA	5.504888	1	2.346250

Stepwise Selection

Math Model Analysis

Table 3: Stepwise Model Comparison for Math

Method	Math_Vars	Math_R2	Math_AdjR2	Math_AIC	Math_BIC
AIC	age mathp1 mathp2 activitiesyes famrel Walc	0.8230933	0.8190420	1147.288	1176.046
BIC	mathp2 famrel	0.8144676	0.8130726	1152.094	1166.473
Forward	mathp2 famrel age activitiesyes mathp1 Walc	0.8230933	0.8190420	1147.288	1176.046
Backward	age mathp1 mathp2 activitiesyes famrel	0.8213539	0.8179576	1147.920	1173.083

In the table, we observe the performance of four different stepwise model selection methods: AIC, BIC, Forward, and Backward. Each method uses a set of variables to predict the Math final score (dependent variable), and we evaluate the models based on several metrics such as R^2, adjusted R^2, AIC (Akaike Information Criterion), and BIC (Bayesian Information Criterion).

- **AIC Model:**
  - **Variables:** age, mathp1, mathp2, activitiesyes, famrel, Walc
  - **R^2:** 0.8231
  - **Adjusted R^2:** 0.8190
  - **AIC:** 1147.288
  - **BIC:** 1176.046

  The AIC model includes a variety of variables, with a fairly high R^2 value indicating a good fit.

- **BIC Model:**
  - **Variables:** mathp2, famrel
  - **$\mathbf{R^2}$:** 0.8145
  - **Adjusted $\mathbf{R^2}$:** 0.8131
  - **AIC:** 1152.094
  - **BIC:** 1166.473

  The BIC model is more restrictive, focusing on fewer variables. While its performance is slightly lower than the AIC model, it still provides a solid fit.

- **Forward Model:**
  - **Variables:** mathp2, famrel, age, activitiesyes, mathp1, Walc
  - **$\mathbf{R^2}$:** 0.8231
  - **Adjusted $\mathbf{R^2}$:** 0.8190
  - **AIC:** 1147.288
  - **BIC:** 1176.046

  This model uses a forward selection approach, incorporating a slightly larger set of variables. It performs identically to the AIC model, indicating that this set of predictors is likely optimal for predicting Math scores.

- **Backward Model:**
  - **Variables:** age, mathp1, mathp2, activitiesyes, famrel
  - **$\mathbf{R^2}$:** 0.8214
  - **Adjusted $\mathbf{R^2}$:** 0.8180
  - **AIC:** 1147.920
  - **BIC:** 1173.083

  The backward model removes certain variables, but it still performs similarly to the AIC and Forward models, showing that the exclusion of variables has little impact on performance.

Portuguese Model Comparison

Table 4: Stepwise Model Comparison for Portuguese

Method	Port_Vars	Port_R2	Port_AdjR2	Port_AIC	Port_BIC
AIC	schoolMS reasonhome reasonother reasonreputation internetyes porp1 porp2	0.8441477	0.8399678	863.9884	896.3408
BIC	schoolMS porp1 porp2	0.8338638	0.8319830	873.1773	891.1509
Forward	porp2 schoolMS porp1 reasonhome reasonother reasonreputation internetyes	0.8441477	0.8399678	863.9884	896.3408
Backward	schoolMS reasonhome reasonother reasonreputation internetyes porp1 porp2	0.8441477	0.8399678	863.9884	896.3408

In this table, the same methods (AIC, BIC, Forward, and Backward) are used to select variables for predicting the final grade in Portuguese.

- **AIC Model:**
  - **Variables:** schoolMS, reasonhome, reasonother, reasonreputation, internetyes, porp1, porp2
  - **R^2:** 0.8441
  - **Adjusted R^2:** 0.8400
  - **AIC:** 863.9884
  - **BIC:** 896.3408

The AIC model for Portuguese includes several variables and yields the highest $\mathbf{R^2}$, suggesting a strong fit for the model. It is the best performing model in terms of explained variance.

- **BIC Model:**
  - **Variables:** schoolMS, porp1, porp2
  - **$\mathbf{R^2}$:** 0.8339
  - **Adjusted $\mathbf{R^2}$:** 0.8320
  - **AIC:** 873.1773
  - **BIC:** 891.1509

  The BIC model, similar to the AIC model, includes fewer variables but still provides a strong fit with a slightly lower $\mathbf{R^2}$.

- **Forward Model:**
  - **Variables:** porp2, schoolMS, porp1, reasonhome, reasonother, reasonreputation, internetyes
  - **$\mathbf{R^2}$:** 0.8441
  - **Adjusted $\mathbf{R^2}$:** 0.8400
  - **AIC:** 863.9884
  - **BIC:** 896.3408

  The Forward model's performance matches the AIC model, suggesting that the forward selection process leads to a good selection of variables.

- **Backward Model:**
  - **Variables:** schoolMS, reasonhome, reasonother, reasonreputation, internetyes, porp1, porp2
  - **$\mathbf{R^2}$:** 0.8441
  - **Adjusted $\mathbf{R^2}$:** 0.8400
  - **AIC:** 863.9884
  - **BIC:** 896.3408

  The Backward model produces identical results to the AIC and Forward models, confirming that the removal of certain variables doesn't negatively impact the model's performance.

Comparison of Overall Performance

Best Performing Model for Math:

The **AIC model** is the best performing for Math, with an R^2 of 0.8231 and an adjusted R^2 of 0.8190. This model incorporates six variables and provides a good balance between model complexity and performance. The AIC, Forward, and Backward models all perform similarly, indicating that the inclusion of more variables improves the prediction of final Math scores.

Best Performing Model for Portuguese:

The **AIC model** for Portuguese stands out with the highest R^2 (0.8441) and adjusted R^2 (0.8400). This model also shows the best performance in terms of AIC and BIC values, indicating that it offers the best trade-off between fit and complexity. The Forward and Backward models yield identical results to the AIC model, confirming that variable selection is robust.

Conclusion

- For **Math**, the AIC model is the top performer, closely followed by the Forward and Backward models. The slight differences in performance suggest that the inclusion of more variables in the model improves the prediction of final Math scores.

- For **Portuguese**, the AIC model is also the best performer, providing the highest R^2, adjusted R^2, and the lowest AIC and BIC. This indicates that the model is well-suited for predicting Portuguese final grades.

- In both subjects, the Forward and Backward models yield very similar results to the AIC model, suggesting that variable selection methods are not overly influential for these data sets.

In conclusion, the AIC method provides the best balance for both subjects, and while there are slight variations in the models, the choice of variable selection method does not drastically alter the performance for either Math or Portuguese.

Lasso Regression

Table 5: Lasso Coefficients for Math and Portuguese Final Grades

Variable	Math_Coefficient	Portuguese_Coefficient
(Intercept)	-1.0192678	0.1989165
schoolMS	0.0000000	-0.5722101
age	-0.0449502	0.0000000
Mjobother	0.0000000	-0.0442212
reasonother	0.0000000	-0.5770312
reasonreputation	0.0213256	0.0000000
internetyes	0.0000000	0.1053971
mathp1	0.0937878	NA
mathp2	0.9849850	NA
famrel	0.1469856	0.0000000
porp1	NA	0.1701065
porp2	NA	0.8470918

In the Lasso regression model, we examine the coefficients for various predictors of final grades in both Math and Portuguese. The table presents the coefficients for each variable, and we can analyze their impact on the final grades for each subject.

Math Analysis

For Math, several variables have non-zero coefficients, including:

- **Intercept:** The intercept coefficient is -1.0193, indicating the baseline level of Math grades when all predictors are zero.
- **Mathp2:** The coefficient of 0.985 for mathp2 suggests a strong positive relationship with the Math final grade.
- **Mathp1:** With a coefficient of 0.094, mathp1 also shows a positive relationship, though its effect is smaller compared to mathp2.
- **Famrel:** The coefficient of 0.147 for famrel shows a positive association with the Math final grade.
- **Age:** The negative coefficient of -0.045 for age implies a slight decrease in Math final grades as age increases, though the effect is small.

Portuguese Analysis

For Portuguese, we see different variables have a significant impact:

- **Intercept:** The intercept is 0.1989, indicating the baseline level of Portuguese grades.
- **SchoolMS:** With a negative coefficient of -0.5722, schoolMS seems to have a detrimental effect on Portuguese final grades.

- **Reasonother:** This variable has a negative coefficient of -0.5770, suggesting that the reason "other" for choosing a school reduces Portuguese grades.

- **Porp2:** The coefficient of 0.8471 indicates a strong positive association with the final grade in Portuguese.

- **Porp1:** Similarly, porp1 shows a positive effect on the Portuguese grade with a coefficient of 0.1701.

- **Internet:** Internet access has a positive coefficient of 0.1054, suggesting a slight increase in the final grade when internet access is available.

Comparison of Math and Portuguese Models

The Lasso models for both Math and Portuguese show that different variables are significant in predicting final grades:

- For **Math**, the strongest predictors are `mathp2` and `mathp1`, with high positive coefficients. `famrel` and `age` also contribute, though to a lesser extent.

- For **Portuguese**, the strongest predictor is `porp2`, followed by `schoolMS`, `reasonother`, and `porp1`.

- **Age and internet access** appear in the Math model with small coefficients, while in Portuguese, `internet` has a slightly positive effect, and age is not a significant factor.

- `schoolMS` has a negative impact on Portuguese grades but does not appear in the Math model.

In summary, the Lasso regression for Math and Portuguese highlights different key variables, with Math relying more on past performance measures (mathp1, mathp2) and Portuguese being influenced by factors like `schoolMS`, `reasonother`, and `porp2`. This reflects the distinct nature of predictors for academic performance in each subject, offering insights into the factors that contribute to final grade outcomes.

Principal Component Regression (PCR) and Partial Least Squares (PLS)

Table 6: PCR and PLS Model Comparison Using Optimal Number of Components

Model	Optimal_Components	RMSE
PCR (Math)	41	1.892848
PCR (Portuguese)	41	1.537861
PLS (Math)	15	1.893054
PLS (Portuguese)	14	1.538132

Summary of PCR and PLS Model Comparison

The Principal Component Regression (PCR) and Partial Least Squares (PLS) models were evaluated using the optimal number of components to predict the final grades for both Math and Portuguese. The performance of each model is assessed based on the Root Mean Squared Error (RMSE).

PCR Results

For **Math**, the PCR model uses 41 components, resulting in an RMSE of 1.8928. The high number of components indicates that the model captures a considerable amount of variance in the data but also suggests potential overfitting, which is typical for PCR models with a large number of components.

For **Portuguese**, the PCR model also uses 41 components but performs slightly better, with an RMSE of 1.5379. The similarity in the number of components between Math and Portuguese indicates that both datasets are similarly complex, but the Portuguese model achieves a lower error, suggesting a better fit to the data. This is supported by earleir analysis!

PLS Results

For **Math**, the PLS model uses 15 components and results in an RMSE of 1.8931, which is nearly identical to the PCR model's RMSE for Math. However, the PLS model uses fewer components, suggesting that it is able to capture the most important variance in the data with less complexity compared to PCR.

For **Portuguese**, the PLS model uses 14 components and produces an RMSE of 1.5381. This is very similar to the PCR result for Portuguese, but again, the PLS model requires fewer components to achieve a comparable error. This indicates that PLS is more efficient in reducing dimensionality while still providing a good fit to the data.

Comparison Between PCR and PLS

When comparing PCR and PLS, we can observe the following:

- The **RMSE values** for both models are very similar for both Math and Portuguese, with slight differences. The PCR model for both subjects has a marginally lower RMSE, but the difference is minimal.
- The **number of components** used by PLS is significantly lower than that of PCR. For Math, PLS uses 15 components compared to 41 for PCR, and for Portuguese, PLS uses 14 components compared to 41 for PCR. This suggests that PLS is more efficient in selecting relevant components for prediction.
- Both models perform similarly in terms of predictive accuracy, but the PLS model is preferred when aiming for a more parsimonious model with fewer components.

In conclusion, while both PCR and PLS provide similar performance in predicting final grades, PLS tends to be a more efficient method, requiring fewer components to achieve comparable accuracy. This makes PLS a better choice when seeking to reduce model complexity while maintaining performance.

Bagging and Boosting

Table 7: Model Comparison: Bagging and Boosting (Math and Portuguese)

Model	RMSE
Bagging (Math)	1.883409
Boosting (Math)	2.201850
Bagging (Portuguese)	1.570195
Boosting (Portuguese)	1.811127

Summary of Bagging and Boosting Results

Math Performance

For predicting Math final grades, the bagging model achieved a lower RMSE (1.8834) compared to boosting (2.2019). This suggests that bagging was more effective in reducing variance and produced more stable predictions in this case. The higher error in boosting may be due to overfitting or the model placing too much weight on difficult-to-predict observations.

Portuguese Performance

For Portuguese grades, bagging again outperformed boosting, with an RMSE of 1.5702 compared to 1.8111 for boosting. Like in Math, bagging seems to generalize better for this dataset. However, both models performed reasonably well overall.

Comparison Between Methods

- **Bagging consistently outperformed boosting** for both subjects, achieving lower RMSE values.
- The difference in performance is more pronounced in the Math models, possibly indicating that the boosting algorithm struggled with the structure or outliers in the Math dataset.
- While boosting can excel in many settings by reducing bias, in this case, it may have overfit or failed to capture patterns that bagging managed more effectively.

Overall, bagging proved to be the better ensemble method for both Math and Portuguese grade prediction in this analysis, providing lower error and more reliable generalization across both subjects.

Random Forest and Support Vector Machine (SVM)

Random Forest and Support Vector Machine (SVM) Performance Analysis

Table 8: Model Comparison: Random Forest and SVM for Math and Portuguese

Model	RMSE
Random Forest (Math)	2.078963
Random Forest (Portuguese)	1.505564
SVM (Math)	2.197250
SVM (Portuguese)	1.528807

Math Grade Prediction

In predicting final grades for Math, the Random Forest model outperformed the Support Vector Machine, yielding an RMSE of approximately 2.08 compared to 2.20 for SVM. While both values indicate reasonably strong predictive performance, Random Forest's advantage may stem from its ability to effectively handle non-linear interactions and high-dimensional feature spaces without the need for heavy preprocessing. Given the complex relationships likely present between academic, personal, and social variables and math performance, Random Forest's ensemble nature offers robustness to noise and feature importance ranking, aiding generalization.

In contrast, SVMs—though powerful for classification—can be sensitive to parameter tuning (such as kernel choice and cost settings) and may struggle to generalize well with small-to-medium numeric datasets if hyperparameters are not extensively optimized. Additionally, the potentially limited linear separability of student performance data may constrain SVM regression effectiveness.

Portuguese Grade Prediction

Similar patterns were observed for Portuguese grades. Random Forest again outperformed SVM, though the margin was narrower: RMSE of 1.51 versus 1.53. This suggests both models captured the relevant patterns reasonably well, but Random Forest still provided a slight edge. The smaller error range may indicate that Portuguese grades are somewhat more linearly predictable or less influenced by complex nonlinear interactions than Math, making SVM relatively more competitive in this context.

Model Characteristics and Interpretation

Random Forest benefits from being an ensemble of decision trees, reducing variance through averaging and capturing variable interactions naturally. It also provides variable importance measures, which are valuable for interpreting which student attributes most strongly influence final grade outcomes. Its relatively strong performance across both subjects suggests the dataset contains nontrivial interactions and possibly non-linear relationships among predictors that tree-based models can exploit.

SVM, while effective in many machine learning applications, tends to require more feature scaling and fine-tuned kernel selection to excel. The performance gap here—especially in Math—highlights the challenges of applying SVM regression without deep kernel optimization, particularly when data complexity is high.

Conclusion and Recommendations

- **Math:** Random Forest is clearly the superior model in this case, with a lower RMSE and greater capacity to model complex student performance patterns.

- **Portuguese:** Although the difference is smaller, Random Forest still outperforms SVM and provides a more interpretable and robust solution.

Given these results, Random Forest emerges as the more reliable and effective choice across both subjects, especially in settings where the relationship between predictors and outcomes is nonlinear or involves interactions that are hard to capture with margin-based methods like SVM.

Final Model Comparisons and Overall Performance Analysis

Table 9: Model Performance Comparison: RMSE and R^2 for Math and Portuguese Final Grades

Model	RMSE_Math	R^2_Math	RMSE_Portuguese	R^2_Portuguese
Linear	1.893	0.834	1.538	0.732
StepAIC	1.836	0.842	1.538	0.729
StepBIC	1.867	0.837	1.526	0.733
Forward	1.836	0.842	1.538	0.729
Backward	1.843	0.840	1.538	0.729
Lasso	1.847	0.840	1.503	0.742
PCR	2.987	0.834	1.721	0.732
PLS	1.946	0.834	1.575	0.732
Bagging	1.883	0.834	1.570	0.721
Boosting	2.202	0.774	1.811	0.641
RF	2.079	0.815	1.506	0.746
SVM	2.197	0.803	1.529	0.736

Math Grade Prediction

Among all models tested for predicting final Math grades, the best performance in terms of RMSE was achieved by the **StepAIC** and **Forward selection** models, both with an RMSE of 1.836 and R^2 of 0.842. These models indicate strong linear relationships and effective variable selection under regular linear modeling. Interestingly, Lasso regression and Backward selection followed closely with RMSEs around 1.84, suggesting minimal trade-off between model simplicity and prediction accuracy.

Tree-based and non-linear methods like Random Forest (RMSE = 2.08, $R^2 = 0.815$) and Bagging (RMSE = 1.88, $R^2 = 0.834$) were slightly less accurate than the best subset-based methods, but still competitive. Boosting and SVM underperformed slightly, likely due to either overfitting (in the case of boosting) or lack of optimal hyperparameter tuning (SVM).

PCR significantly underperformed in RMSE (2.99), despite having a similar R^2, suggesting that the high number of components used may have led to overfitting without improving predictive error.

Portuguese Grade Prediction

For Portuguese grades, the best RMSE was achieved by **Lasso regression** (1.503), with a strong R^2 of 0.742. This implies that Lasso's regularization improved generalization by excluding less informative variables. Random Forest also performed very well (RMSE = 1.506, $R^2 = 0.746$), offering strong performance with interpretability through feature importance.

StepBIC (RMSE = 1.526) and SVM (RMSE = 1.529) followed closely, again confirming that both linear and non-linear methods can perform comparably when tuned well. Notably, PLS and Bagging had moderate errors, while Boosting exhibited the poorest performance (RMSE = 1.811, $R^2 = 0.641$), suggesting potential overfitting or inappropriate depth in tree construction.

Overall Best Models

- **Math:** StepAIC and Forward Selection models performed best, showing that linear models with careful variable selection offer superior predictive power in this case.
- **Portuguese:** Lasso regression emerged as the top model, balancing low error with strong interpretability. Random Forest was a close second and may be preferred when variable interactions are critical.

Conclusion

Model performance varied slightly between the two subjects, possibly due to differences in variable relevance, underlying complexity, and noise. Linear models performed better for Math, suggesting a more structured and additive relationship between predictors and outcome. In contrast, Portuguese benefited from regularized and ensemble techniques, indicating more subtle interactions or noise that simpler models might not capture.