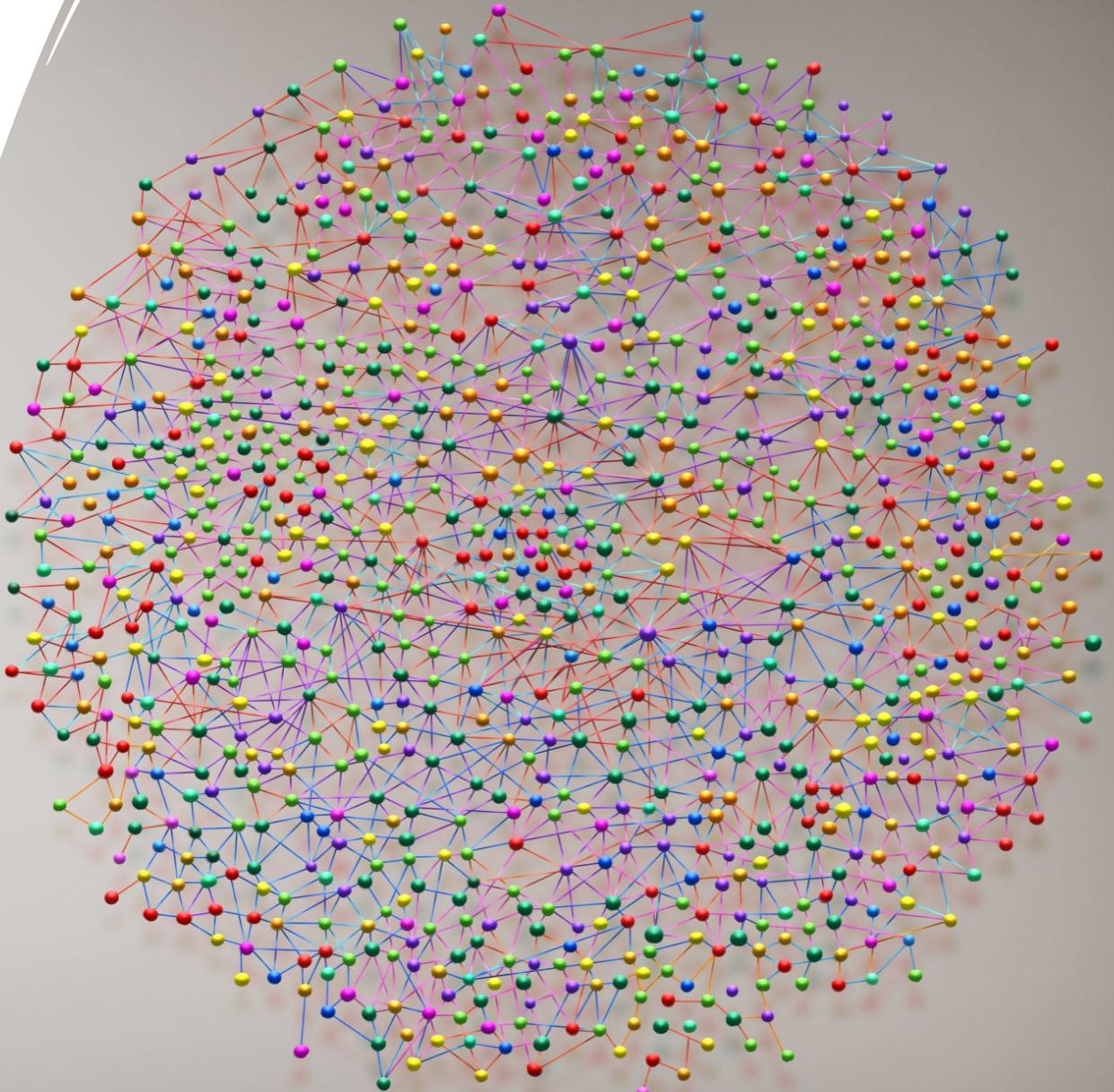


# PREDICTING PORTUGUESE SECONDARY SCHOOL STUDENT PERFORMANCE

[https://github.com/annieptba/data1030\\_project.git](https://github.com/annieptba/data1030_project.git)

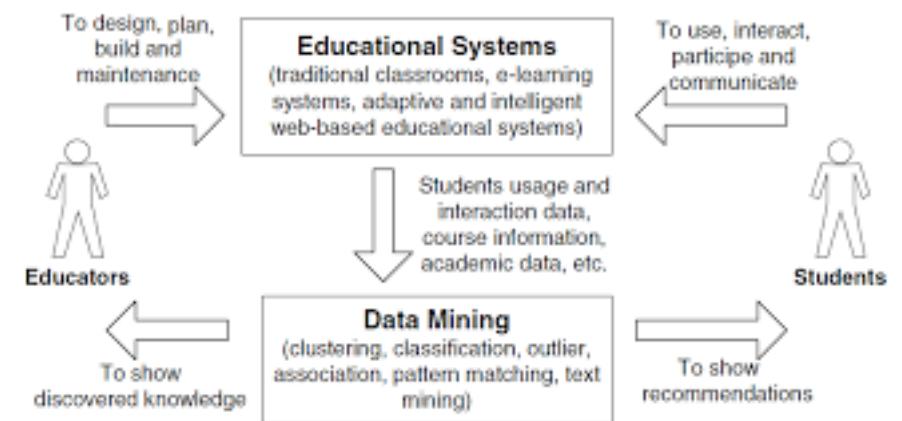
---

ANNIE PHAN (Banner ID: B01309278)



# CONTENT OVERVIEW

- 1. Introduction
- 2. EDA
- 3. Methods
- 4. Results
- 5. Outlook



# 1. INTRODUCTION

Problem:

Regression: predict students' final scores (continuous, integer values from 0 to 20) using first and second period grades and other social, economic, and education factors  
Classification: classify students' final performance (categorical values 'good', 'fair', 'poor') using first and second period grades and other social, economic, and education factors



Data overview and source

<https://archive.ics.uci.edu/ml/datasets/student+performance>

1,044 instances (students) including 395 Mathematics class students and 649 Portuguese language class students

33 features



Implications:

What type of courses can be offered to attract more students? Is it possible to predict student performance?

What are the factors that affect student achievement?

My model explores these similar questions but look further into demographic factors such as family support, romantic relationships, alcohol consumption, and internet access.



Original research:

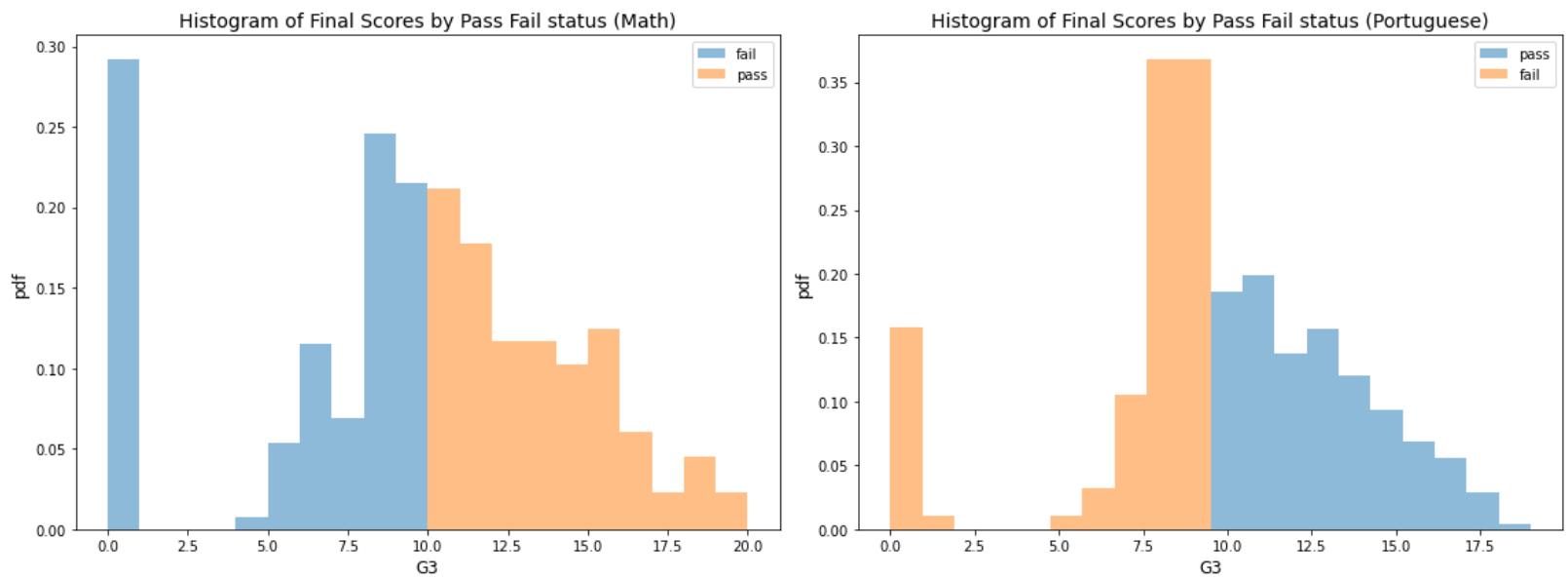
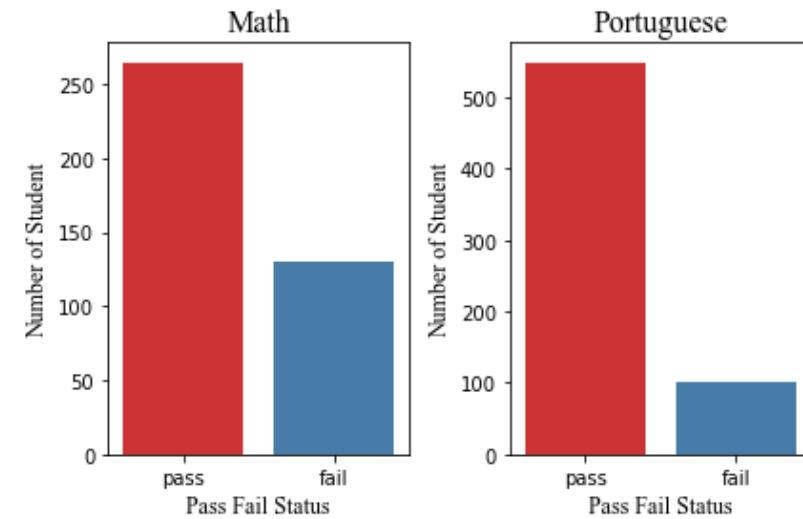
Project: "Using Data Mining To Predict Secondary School Student Performance" by Cortez and Silva in 2008

Method: binary classification (pass/fail), classification with 5 levels (1 very good → V insufficient), regression, with a numeric output that ranges between (0%) and (100%)

Result: students' final grades can be predicted by the first and/or second school period grades and also other relevant features

## 2. EDA - FINAL GRADES AND PASS FAIL OVERVIEW

- More students have a higher Portuguese score than Math scores. This poses questions about Math vs Portuguese teaching level of Portuguese schools
- % of students who fail is quite high: 32.9% for Math and 15.4% for Portuguese

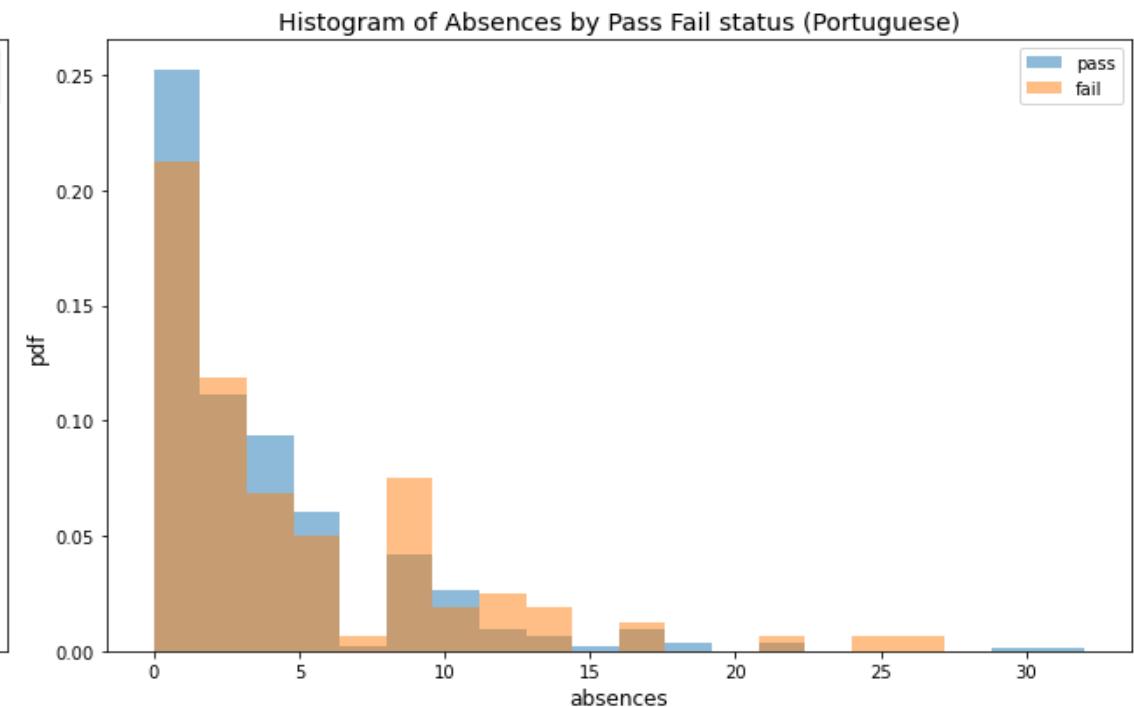
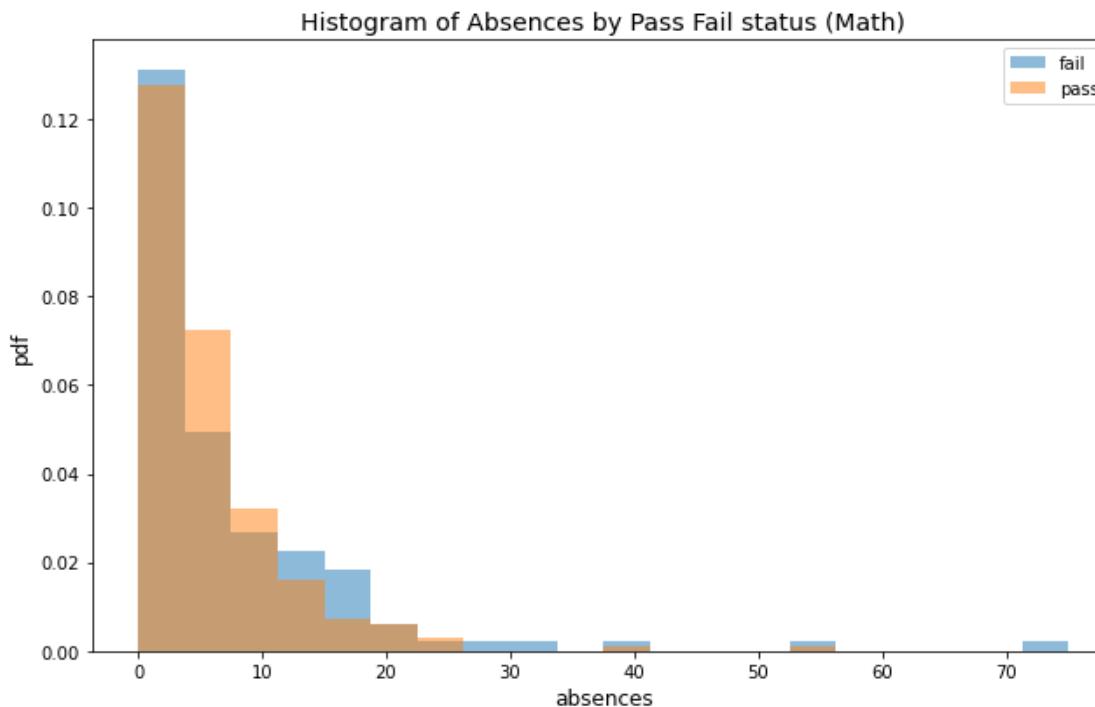


# 2. EDA - PASS FAIL STATUS AND ABSENCES

---

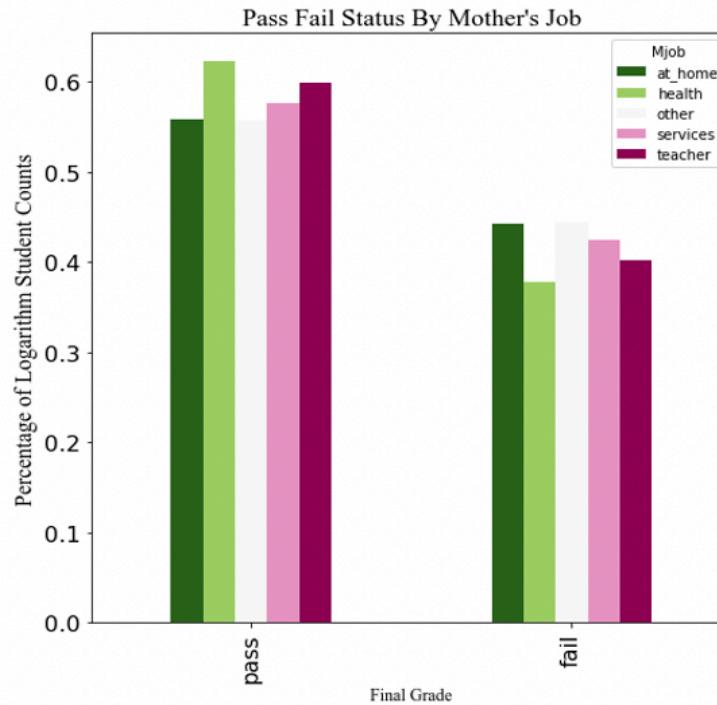
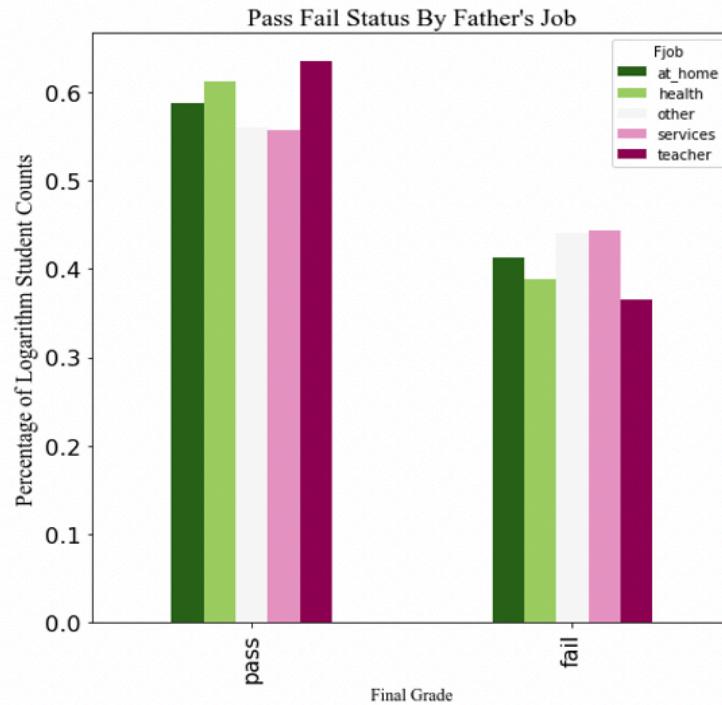
- Students have more absences in Math than Portuguese → students may be less interested or struggle more with this subject
- More students who had higher absences were able to pass Portuguese than Math, which again suggests that Math is harder than Portuguese for the students

- why students are absent more in Math
- how or if they are struggling with the subject,
- what can be done to improve



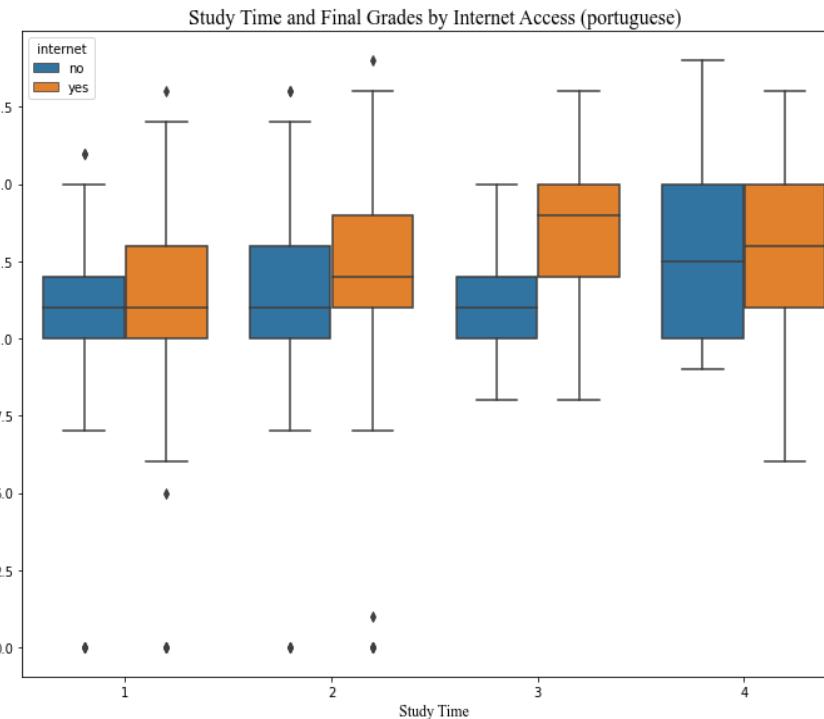
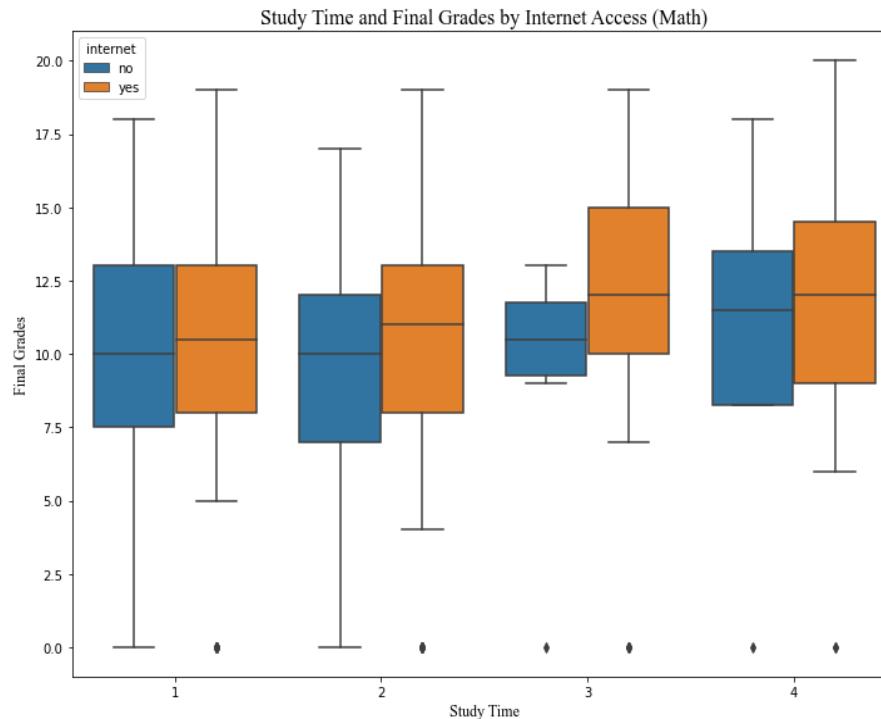
# 2. EDA - PASS FAIL STATUS AND PARENTS' JOBS

- correlation between the parents' job and students' performance
- more students pass if their parents are in a "highly educated" field, such as health or teacher and vice versa.
- there isn't a significant difference between the number of students who pass and fail with respect to the mother or father's job, which suggests that both parents' jobs are equally important.
- suggests some degree of socioeconomic inequality in education



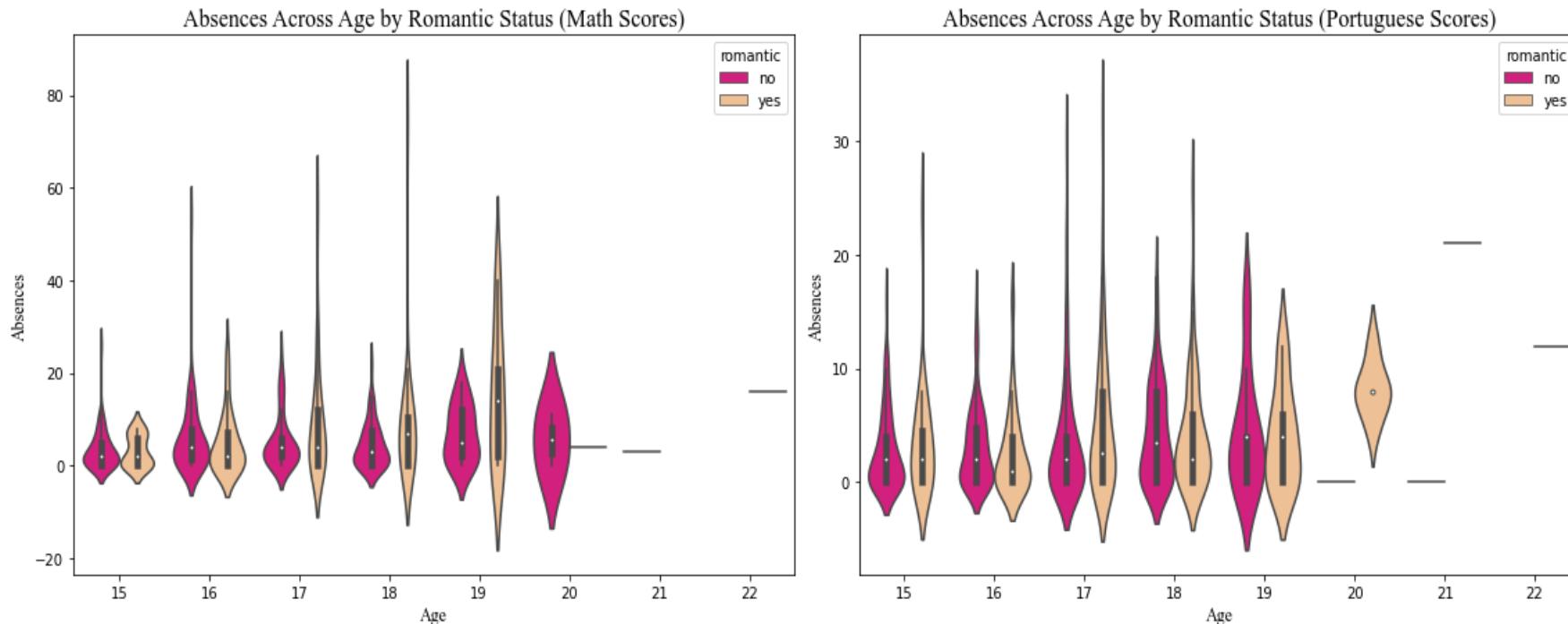
## 2. EDA - STUDYTIME, FINAL GRADE, INTERNET ACCESS

- students who have internet have higher grades than those who don't → internet is quite important for studying.
- for students who study less in both subjects, there is a smaller difference in the final grades between those who don't have internet and those who do compared to students who study more → strong students are more effective at using the internet for studying
- → needs to be more research into what the students do on the Internet, how much time do they dedicate their internet use to studying, and how schools can assist students in using the Internet for more education purposes.



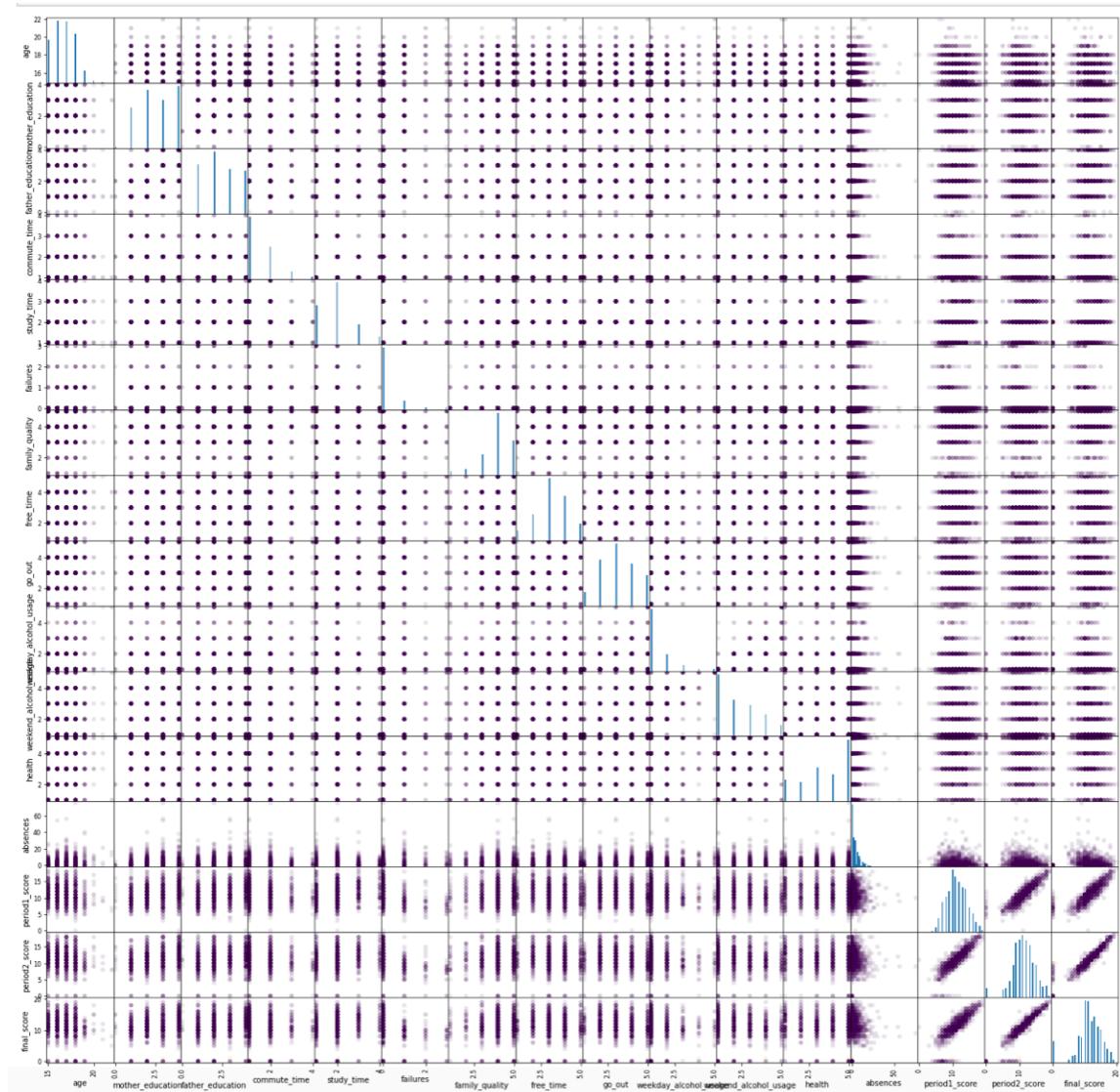
## 2. EDA - ABSENCES, AGE, ROMANCE

- students who are in a relationship have more absences than those who don't. → suggests that relationships can distract students from school.
- In math, students between 17 - 19 years old have the most absences and are in a relationship more.
- In Portuguese, there are more students who are in a relationship than Math, and these students also have more absences → suggests that either students in Math are more "disciplined" or math is harder and require students to study more.



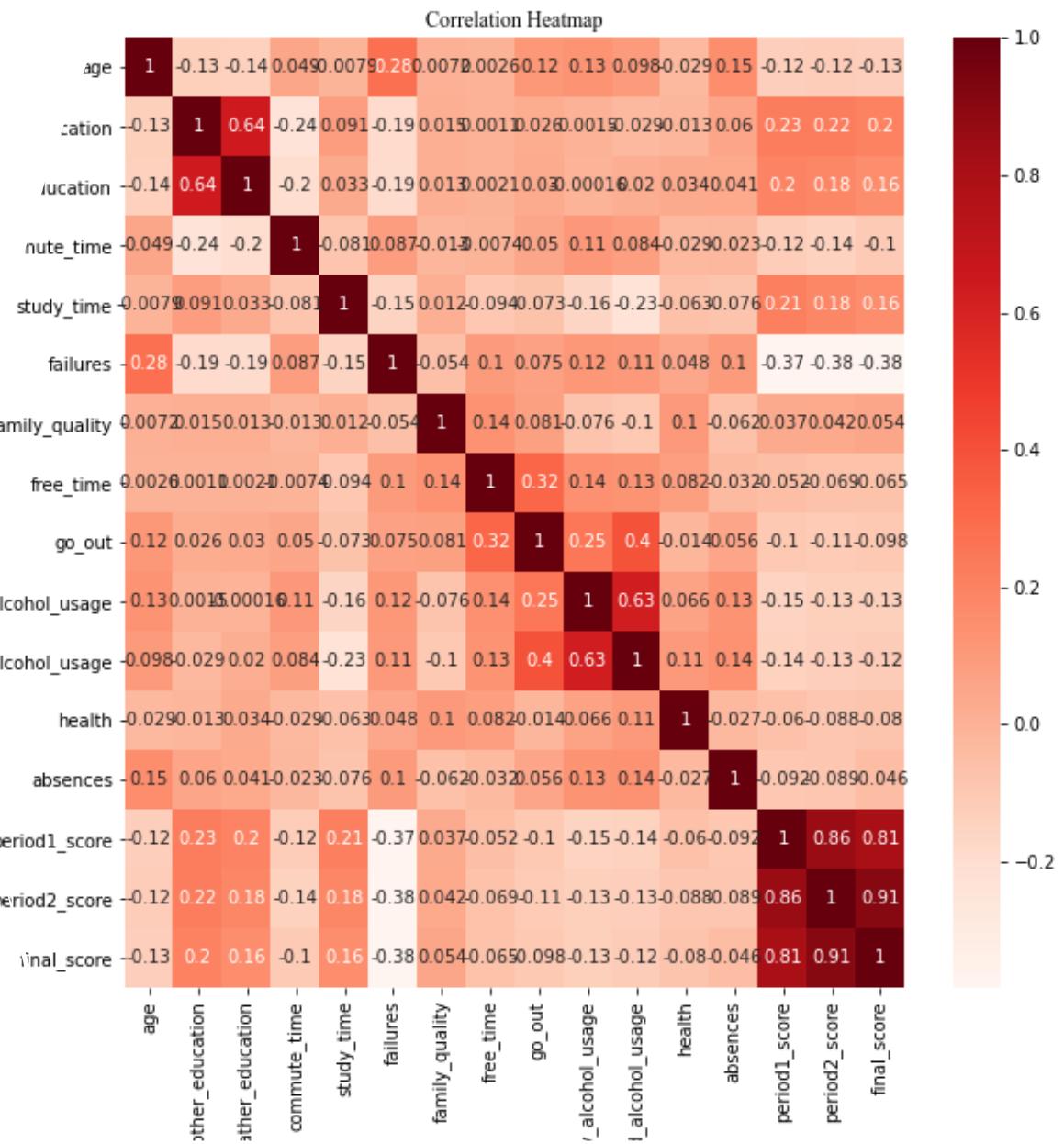
## 2. EDA - SCATTER MATRIX

- a strong positive relationship between the 1st and 2nd period scores and final scores, signaling that score prediction can be quite accurate.
- relationships between final scores and other variables aren't so obvious.



# 2. EDA - CORRELATION HEATMAP

- a strong positive relationship between the 1st and 2nd period scores and final scores, signaling that score prediction can be quite accurate
- final scores also seem to have quite strong positive relationships with the mother and father's education and study time
- other positive but weaker relationships between finals cores and other variables include family quality and free time
- a negative relationship between final age and scores, which can be because the materials get harder
- a weak negative relationship between health and final scores, raising concerns about mental health.



# 3. METHODS - MACHINE LEARNING MODELS & FEATURE IMPORTANCES

---

- Splitting: Both data sets are IID, relatively small, and don't have group structure, time series, and missing values → basic train\_test\_split and a KFold split
- Preprocessing: ordinal features already encoded, OneHotEncoder on the remaining non-bounded/ranked, categorical features and the MixMaxEncoder on the bounded continuous features, LabelEncoder on the target variable
- choose accuracy score because my datasets are both balanced (0.671 Class 1 for Math and 0.846 Class 1 for Portuguese)
- measure uncertainties due to splitting and non-deterministic ML models, I loop through 10 random states and calculate the mean and standard deviation of the test scores
- additional cross-validation confusion matrix and ROC curve
- developed an ML pipeline using K-Fold Cross Validation (GridSearchCV),

create an additional feature called "pass\_fail" which denotes that a student fail if their score is below 10 and pass if otherwise.

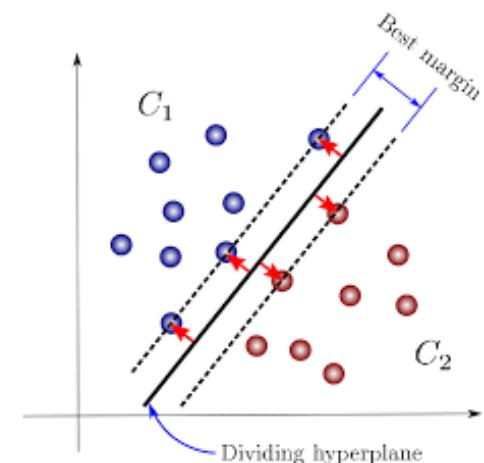
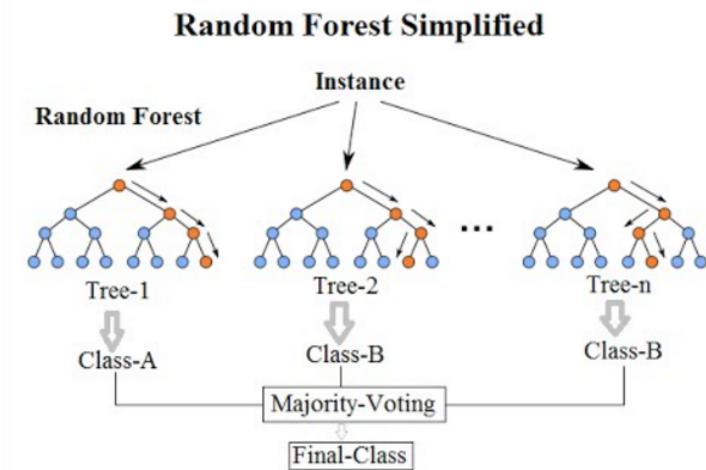
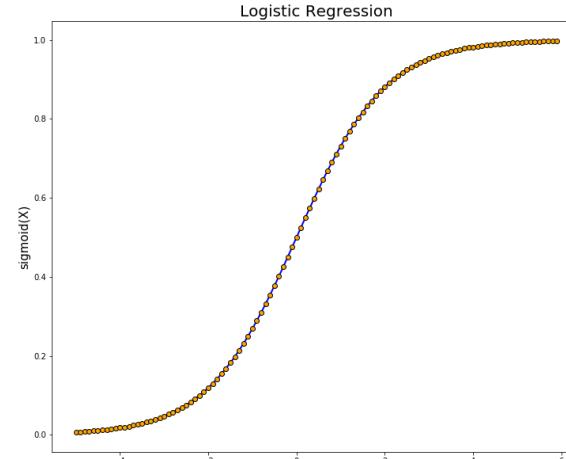
G3	pass_fail
6	fail
6	fail
10	pass
15	pass
10	pass

3 sub-models: Model 1 consists of all midterm and final scores, Model 2 consists of only the first midterm scores and final scores, and Model 3 consists of no midterm and only final scores

G1	pass_fail	G1	G2	pass_fail	pass_fail
5	fail	0	11	pass	fail
5	fail	9	11	pass	fail
7	pass	12	13	pass	pass
15	pass	14	14	pass	pass
6	pass	11	13	pass	pass

# 3. METHODS - MACHINE LEARNING MODELS & FEATURE IMPORTANCES

- 3 ML algorithms: logistic regression, random forest classifier, support vector classifier.
- Hyperparameter tuning
  - `LogisticRegression(penalty='l1', solver='saga', max_iter=10000, random_state = 20)`
    - Tune: 'logisticregression\_\_C': `np.logspace(-2,2, num=8)`
  - `RandomForestClassifier(n_estimators = 100,random_state=random_state)`
    - Tune: 'randomforestclassifier\_\_max\_depth': `np.linspace(2,30,num=9,dtype=int)`
    - Tune: 'randomforestclassifier\_\_max\_features': `np.linspace(0.25,1,5)`
  - `SVC(random_state = 20)`
    - Tune 'svc\_\_C': `np.logspace(-3,4,num=8)`
    - Tune: 'svc\_\_gamma': `np.logspace(-3,4,num=8)`
- Feature importance
  - LR: perturbation, linear coefficients, SHAP (LinearExplainer).
  - RFC: perturbation, random forest metrics, SHAP (TreeExplainer).
  - SVC: perturbation and SHAP (KernelExplainer).



# 4. RESULTS -MODEL PERFORMANCE

- highest accuracy from Model 1 because the algo have more score information to leverage from to predict the performance, accuracy score erodes over Model 2 and Model 3 because the algo have less score information to predict from.
- Model 2 performs pretty well, and its results fit into an academic context because it allows schools and teachers to identify weak students early on to take necessary actions

Sub-model (2-class)	ML Mode	Baseline	Mean Accurac	Std Accurac	Std Above Baseline	Best Parameters	Best Model
Math (all score)	LR	0.671	0.919	0.019	13.059	C = 0.5179	Logistic Regression
	RFC		0.916	0.02	12.256	max_depth = 5, max_features = 0.625	
	SVC		0.905	0.024	9.755	C = 10000, gamma = 0.001	
Math (no midterm 2 scores)	LR	0.671	0.838	0.036	4.642	C = 26.827	Random Forest Classification
	RFC		0.849	0.037	4.814	max_depth = 9, max_features = 0.625	
	SVC		0.841	0.04	4.253	C = 10000, gamma = 0.001	
Math (no midterm 1&2 scores)	LR	0.846	0.67	0.062	-0.014	C = 0.01	Support Vector Classification (but All 3 Models Perform Quite Similarly)
	RFC		0.676	0.056	0.091	max_depth = 2, max_features = 0.25	
	SVC		0.68	0.055	0.166	C = 100, gamma = 1	
Portuguese (all score)	LR	0.846	0.931	0.02	4.254	C = 1.93	Logistic Regression
	RFC		0.932	0.018	4.782	max_depth = 2, max_features = 0.4375	
	SVC		0.925	0.023	3.438	C = 10000, gamma = 0.001	
Portuguese (no midterm 2 scores)	LR	0.846	0.889	0.02	2.154	C = 100	Random Forest Classification
	RFC		0.915	0.022	3.14	max_depth = 2, max_features = 0.625	
	SVC		0.875	0.024	1.212	C = 1000, gamma = 0.001	
Portuguese (no midterm 1&2 scores)	LR	0.846	0.84	0.031	-0.191	C = 0.01	All 3 Models Perform Quite Similarly
	RFC		0.841	0.032	-0.154	max_depth = 2, max_features = 0.25	
	SVC		0.84	0.031	-0.191	C = 1, gamma = 10	

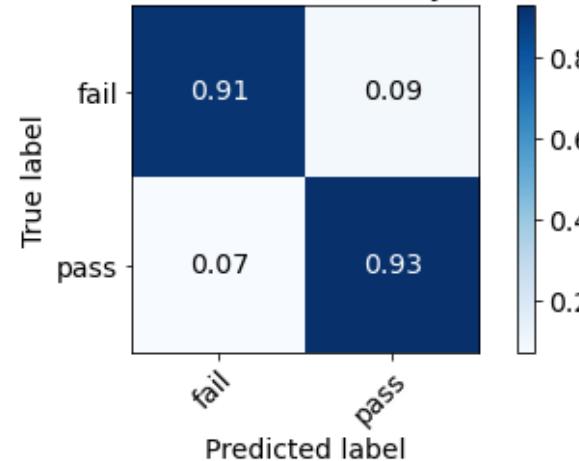
# 4. RESULTS - CONFUSION MATRICES AND ROC CURVES

- Model 2 is quite accurate and allows the most actionable insights for educators to help teachers improve students performance before it's too close to the end of the semester
- Display Random Forest Classification's Confusion Matrix and ROC Curves for Model 2 for both the Math and Portuguese dataset because these were the best models

Math, Model 2:

- Random Forest Classification (`n_estimators = 100, random_state = 42, max_depth = 9, max_features = 0.625`)
- Test accuracy: 0.924

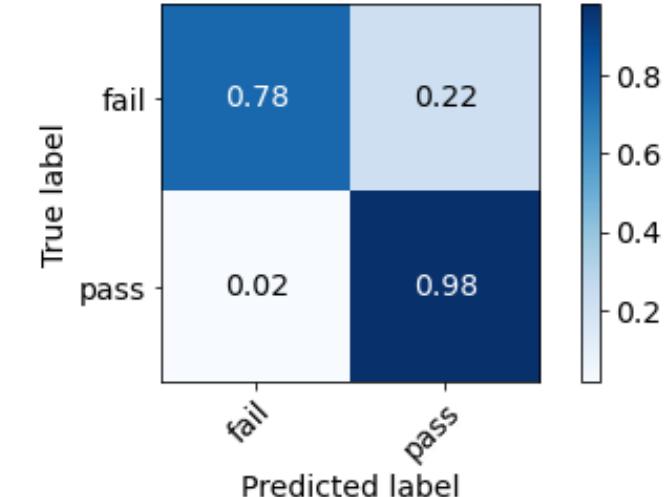
Confusion matrix, norm (onlyG1math, rf)



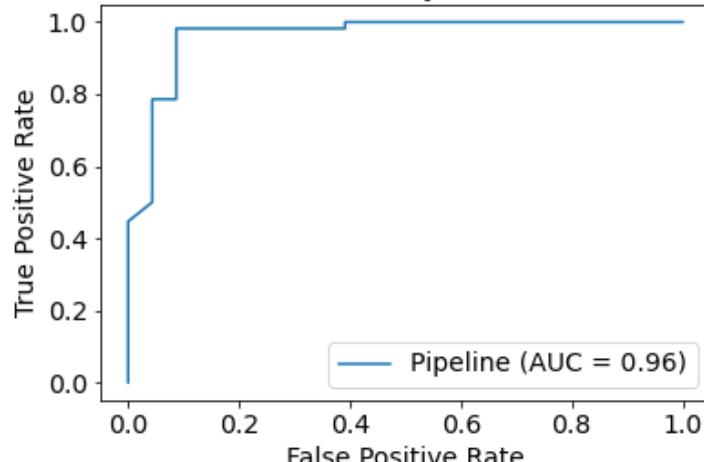
Portuguese Model 2:

- Random Forest Classification (`n_estimators = 100, random_state = 42, max_depth = 2 max_features = 0.625`)
- Test accuracy: 0.954

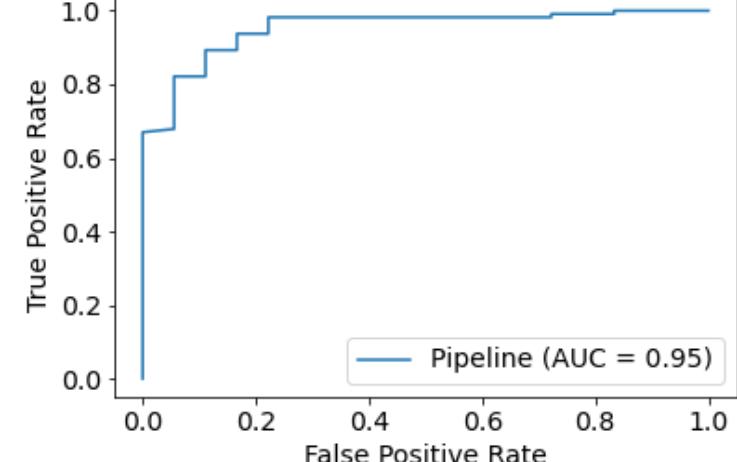
Confusion matrix, norm (onlyG1por, rf)



ROC curve (onlyG1math, rf)

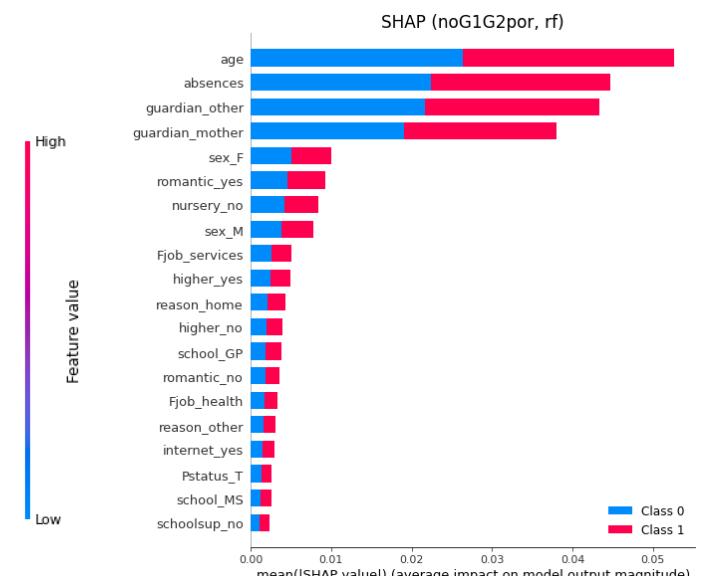
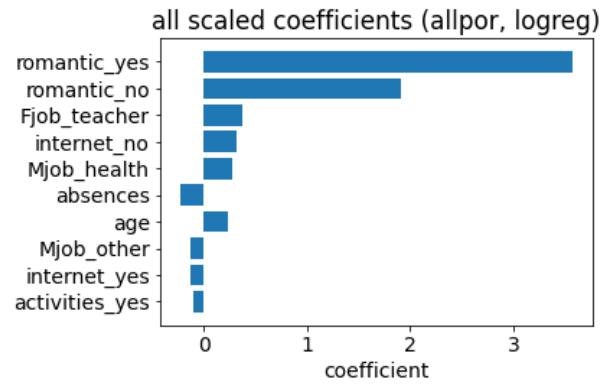
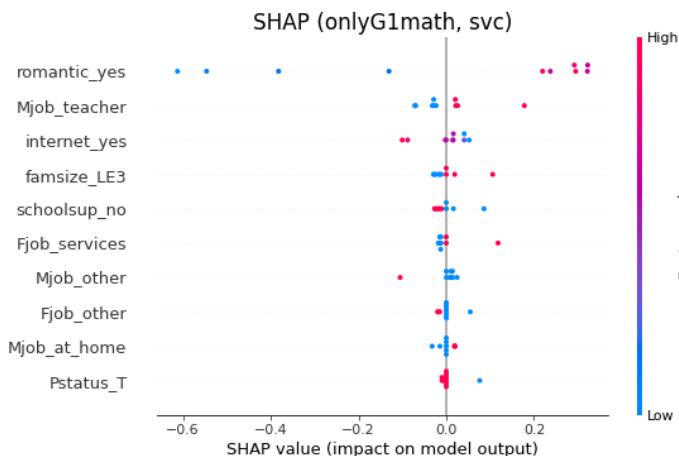
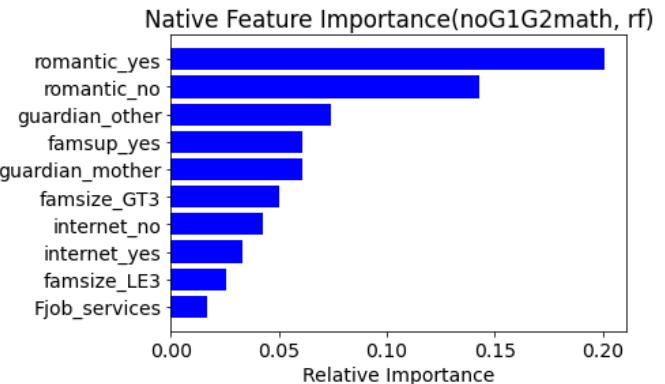
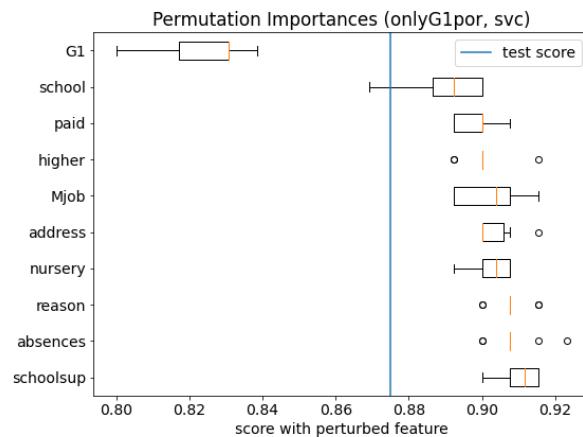
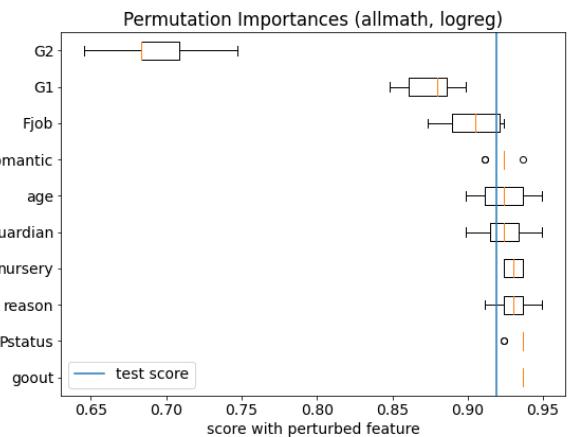


ROC curve (onlyG1por,rf)

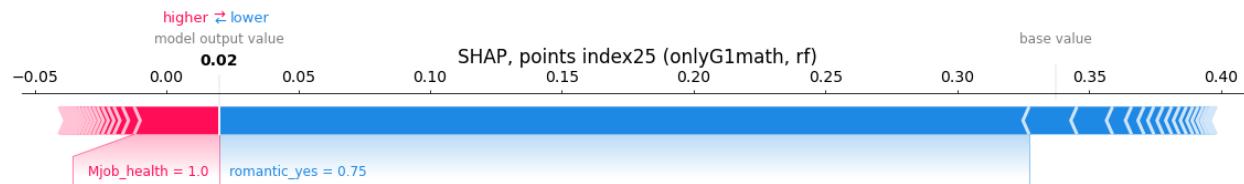
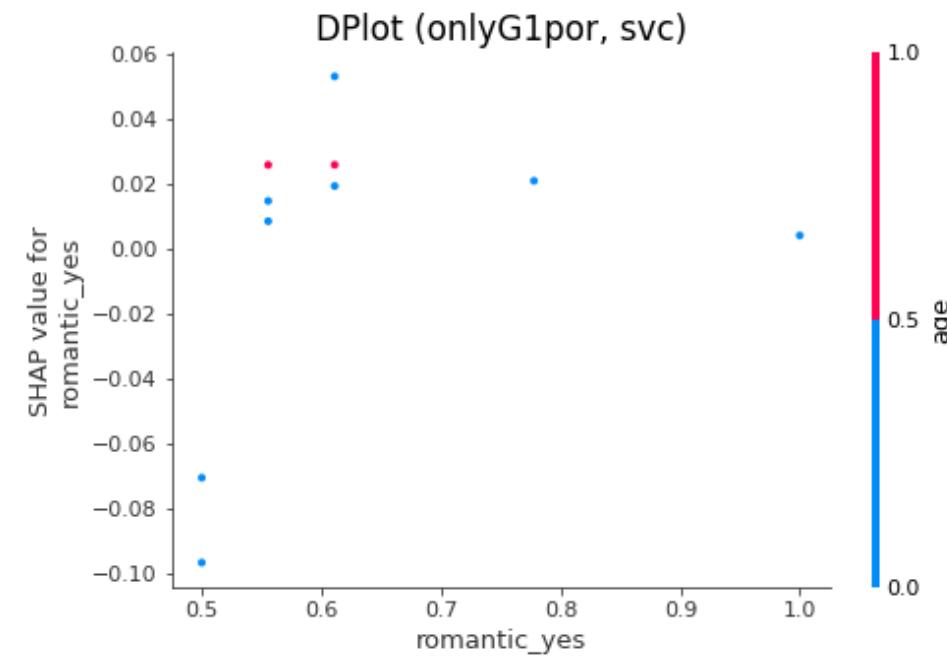
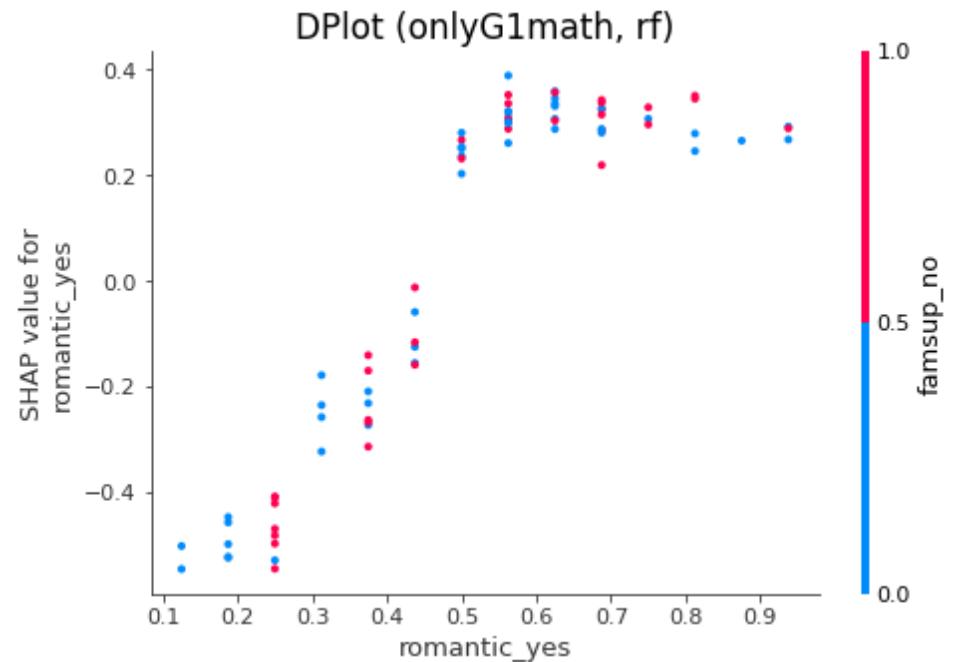


# 4. RESULTS - FEATURE IMPORTANCES (GLBOAL)

- Features importance varies by model, but there are some commonality.
- For Models 1 and 2, the midterm scores are the most important features.
- Other important features are romantic status, internet access, the jobs of the parents, absences, and school supplement and fam supplement.
- Some of the least important features are health, family size & relationship, and travel time.
- Educators should focus on helping students navigate adolescent relationships in a way that could positively impact their performance, work closely with parents of students who have weak performance, and provide more supplement when necessary.



# 4. RESULTS - FEATURE IMPORTANCES (LOCAL)



## 4. OUTLOOK

- Weak spots: used all features for prediction
- Improvement suggestions: feature selection, feature engineering, multi-class classification, more complex ML models such as XGBoost, PCA, KNN, etc.
- Additional info: homework scores, school attitude, prior exposure to the subject, enjoyment of the subjects, more continuous or ordinal features, students' performance in more countries, more holistic measures of academic performance, such as average grades across the school year or school attitude should be considered.

