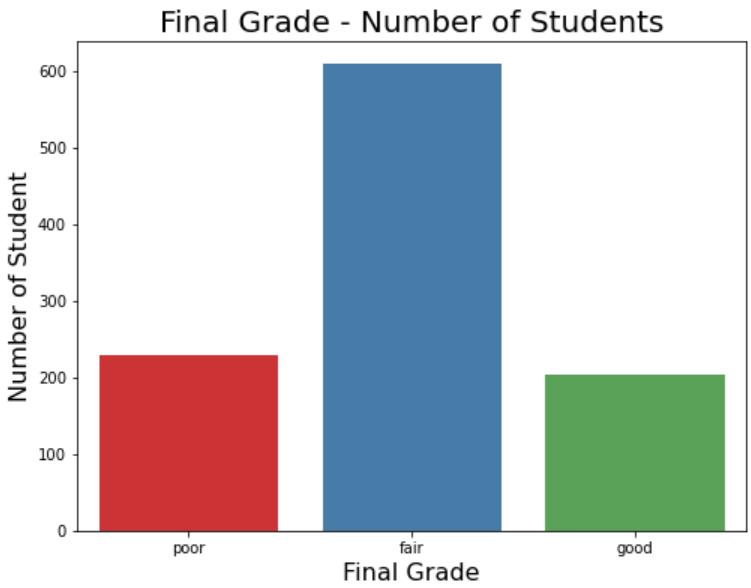
MIDTERM PROJECT REPORT – ANNIE PHAN (Banner ID: B01309278)
PREDICTING PORTUGESE SECONDARY SCHOOL STUDENT PERFORMANCE

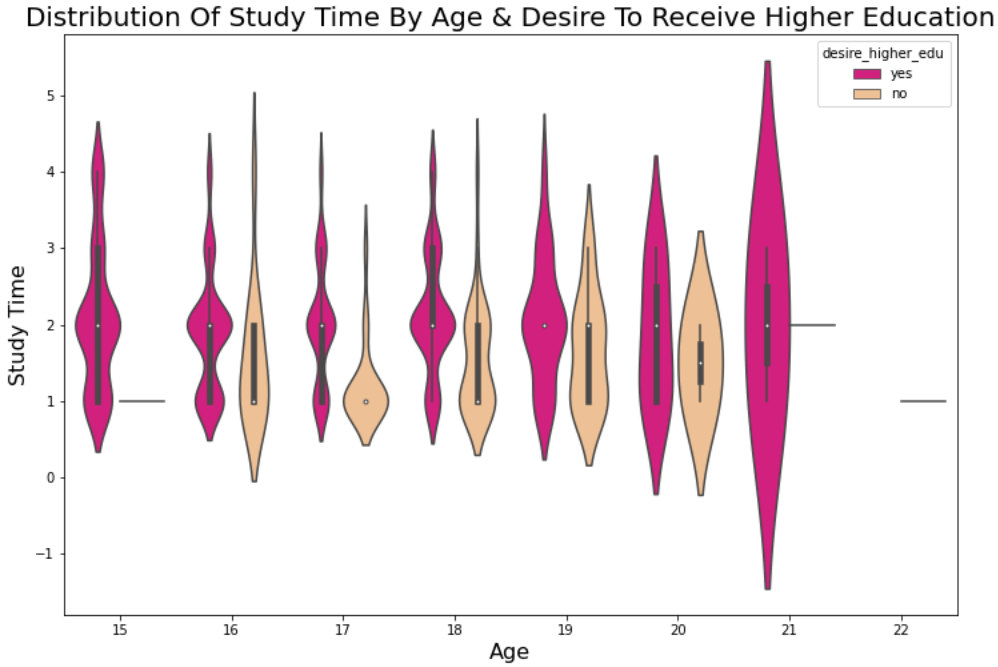
1. Introduction

- For the regression problem, the target variable is the column 'final_score," which is the final period math and Portuguese grades for secondary students. The problem is regression because the target variable is continuous.
- However, the original research paper also carries out some classification modeling. I decide that I could do some classification problem as well to check the accuracy of the model by creating an additional column called "final_grade" which categorizes the students into 'good' performance if they score from 15 to 20, 'fair' if they score from 10 to 14, and 'poor' if they score below 9. This variable is my target variable for the classification problem.
- The dataset is interesting/important because although the educational level of the Portuguese population has increased in last decades, Portugal remains at Europe's tail end due to its high student failure rates. The data not only helps Portuguese educational institutions and government find attributes to best invest in to improve students' performance or identify students that need assistance, but also provides data for Business Intelligence (BI)/Data Mining (DM) to develop automated tools that can improve decision making and optimize success in education. For instance, some interesting questions for this domain that could be answered using BI/DM techniques: 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? In my model, I seek to explore these similar questions but look further into demographic factors such as family support, romantic relationships, alcohol consumption, and internet access.
- This data contains 1,044 instances (students) including 395 Mathematics class students and 649 Portuguese language class students. There are 33 features. The dataset is already well-described and can be found here: https://archive.ics.uci.edu/ml/datasets/student+performance
- In the original research paper named "Using Data Mining To Predict Secondary School Student Performance" by Cortez and Silva in 2008, the two datasets were modeled under binary/five-level classification and regression tasks: "i) binary classification (pass/fail); ii) classification with five levels (from I very good or excellent to V - insufficient); and iii) regression, with a numeric output that ranges between zero (0%) and twenty (100%)." The results show that the students' final grades can be predicted by the first and/or second school period grades and also other relevant features (e.g. number of absences, parent's job and education, alcohol consumption). Another project that tackled this problem was the Michigan State University which also modeled the problem using three classification approaches: "binary: pass/fail; 3-level: low, middle, high; and 9-level: from 1 - lowest grade to 9 - highest score" (Minaei-Bidgoli et al. 2003). The best solution was obtained by a Naive Bayes method with an accuracy of 74%. It was also found that past school grades have a much higher impact than demographic variables (ibid, 2003). There are other researches that have touched on this problem, but I think that the two researches highlighted here are most important. I base my regression and classification problems on the approaches in these researches.

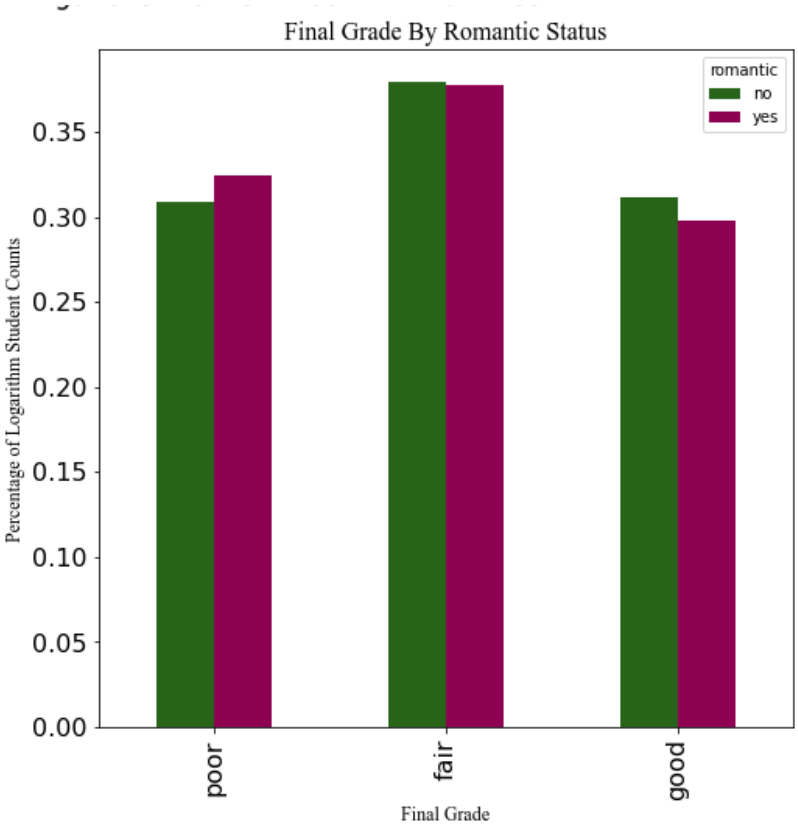
2. Exploratory Data Analysis



- I chose to show the distribution of student's performance to have an idea for the classification problem. The figure demonstrates distribution of students across different final grades performance. Around 60% of the students have a fair performance, but 20% of them have a poor and 20% have good performance. This is a very high % of students have a poor performance and fit with the idea that many Portuguese students were underperforming.

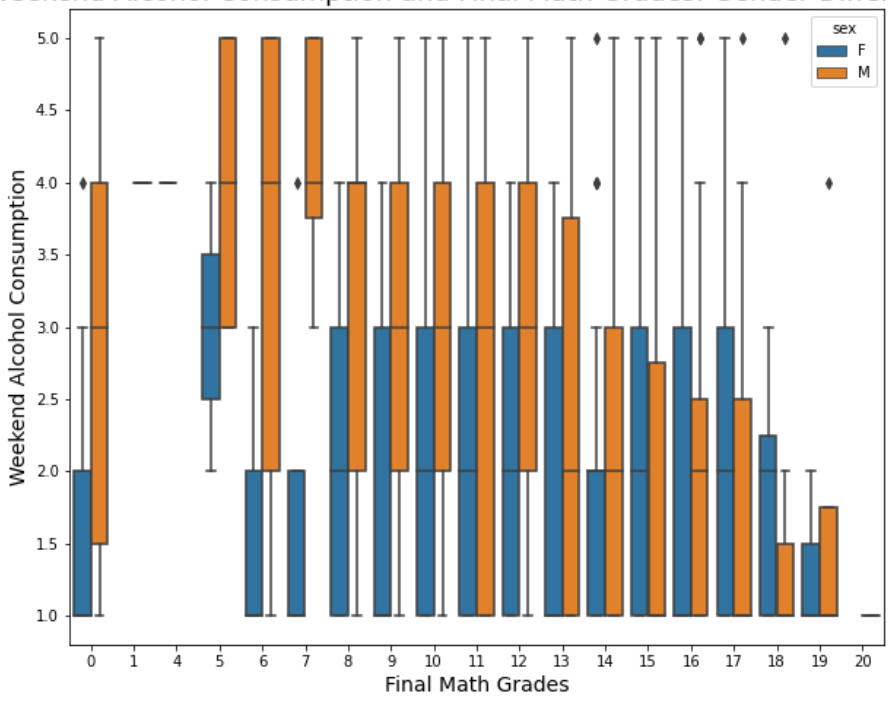


-I chose to explore the relationship between study time and desire for higher education of children across gender to see study patterns and desires across gender. The figure demonstrates distribution of students' study time across different age, classified into whether the student wants to achieve higher education or not. It seems that the older the students get, the more they want to achieve higher education, but also overall, however, the distribution of students who don't want to achieve higher education is also quite close to those who want to, which raises some concerns about how the Portuguese school system and other social factors influence students' choice to achieve higher education. There is also a positive relationship between the students' study time and whether they want to achieve higher education. This makes sense because more motivated students will want to achieve higher education. However, this alerts that more measures should be taken to incentives students to study more and year higher education in a way that balance their well-being and happiness as well.

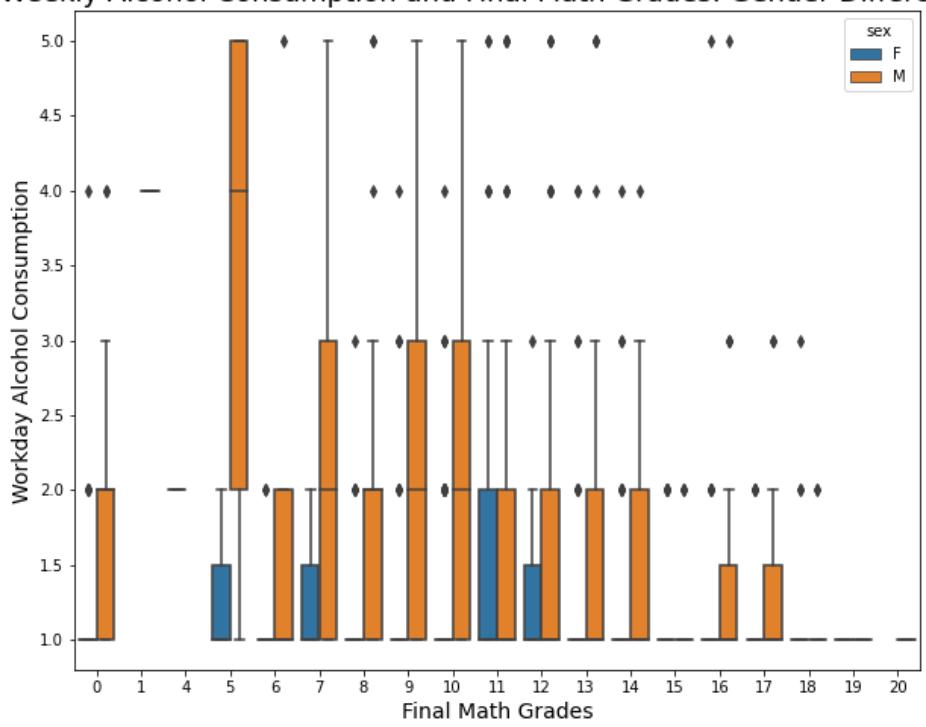


-I chose to explore the relationship between romantic status and students' performance since the dataset covers students of adolescent ages, a group that tend to be "distracted" from school due to desires for romantic relationships. The figure demonstrates distribution of students' relationship status across different final grade performance. It seems that more students who don't have a relationship have better final grade performance and vice versa, though the disparity isn't significant. This signals that students should not be in a relationship if they want to do well in school, but since the difference is not big, intervention measures need to be sensible and cognizant of the students' social development.

Weekend Alcohol Consumption and Final Math Grades: Gender Differences



Weekly Alcohol Consumption and Final Math Grades: Gender Differences



-I chose to explore the relationship between alcohol consumption and students' performance since the dataset covers students of adolescent ages, a group that tend to be more frivolous and excited to drink since they might just be approaching the drinking age. The figures demonstrate final grades across measures of distributions of students' weekend and weekday alcohol consumption, classified by gender. There are some expected observations here. Students who have less alcohol consumption tend to get higher grades and vice versa, and more male students consume alcohol than female students. Students drink significantly more on the weekend than weekday. This suggests that some intervention and educational efforts should be executed to teach students about appropriate drink habits and how it can affect their study.

3. Data Preprocessing

- For the regression question, I split the dataset using the basic approach of 60% of the data in train, 40% validation, and 40% set because the dataset is IID. I know this because it all samples stem from the same generative process and the generative process is assumed to have no memory of past generated samples and is a small dataset. It doesn't have group structure and time series.
- For the classification question, I use K-fold split to shuffle the data a little bit to ensure enough randomness and reduce errors. This is a good split because the dataset is small. I have 20% of the data in test and split the other set into 5. I drop the "final_grade" column because I don't need it for classification and keeping it could skew my results. In the processed data, for the regression problem, I have 626 points in the train set, 209 in the validation and 209 in the test set. For the classification problem, I have 668 points in the train set, 167 in validation and 209 points in the test set.
- For both problems I only apply encoders to the columns that haven't been processed yet. I apply the MixMaxEncoder to the 'age' and 'absences' columns because they both have bounded features ('age' ranges from 15 to 22 and 'absences' ranges from 0 to 93). I apply the One Hot encoders to the categorical columns that have 'yes'-'no' features because they haven't been encoded yet (the 'school', 'sex', 'age',' address', 'family_size',' parents_status', 'mother_job', 'father_job',' reason','guardian', 'school_support', 'family_support', 'paid_classes',' activities', 'nursery', 'desire_higher_edu', 'internet', 'romantic' columns). For the classification problems, I apply a Label Encoder to the target variable 'final_grade' because it is categorical but hasn't been encoded

4. References

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Minaei-Bidgoli B.; Kashy D.; Kortemeyer G.; and Punch W., 2003. Predicting student performance: an application of data mining methods with an educa- tional web-based system. In Proc. of IEEE Frontiers in Education. Colorado, USA, 13–18.

5. Github Repository:

- Link: https://github.com/annieptba/data1030_project_-portugese-secondary-student-performance.git

- Python filename: midterm-annieptba.ipynb