Socio-economical and Educational Impacts on Student Performance

By: Suzanne Chung

Introduction

Students academic performance can be affected by multiple variables, from family, health, lack of interest, social activities etc. the list can go on. Are we able to narrow down what actually affects the students' performance in school?

A survey was conducted on students who are enrolled in Math and Portuguese class who were questioned on numerous social and educational factors. The dataset used included the students' grades throughout their enrollment in these courses. Based on the data given, what are the significant social and educational impacts that negatively affect performance?

I will observe the students who are performing lower than average based on the students who participated in the survey. Using exploratory analysis, we will able to visualize which variables have the biggest impact on performance. Following this analysis, I will perform regression and classification analysis to find the significance in these variables.

Literature Review

Is Alcohol Consumption Associated with Poor Academic Achievement in University Students?

This report is observing students at the University of Gloucestershire, UK and alcohol consumption specifically. The data collected had five alcohol consumption measures: length of time and amount consumed during recent drinking occasion, frequency of alcohol consumption, heavy episodic drinking and drinking problems. Three educational measures were used: the importance of achieving good grades, students' appraisal on their performance against their peers and students' actual mark. The research was broken down to demographics specifically sex and age and to see which measure was strongly associated with academic outcomes and how academic achievements are associated with alcohol consumption. The results from regression analysis demonstrated that males were positively associated with all alcohol consumption measures. The students' actual mark was not associated with any alcohol consumption measure. Overall, Alcohol consumption showed negative associations with academic performance.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3843305/

Prediction Accuracy of Academic Performance of Students using Different Datasets with High Influencing Factors by Jai Ruby and Dr. K. David

This research focuses on two different datasets (Arts College and UCI) that analyzes the prediction accuracy of academic performance of students while using Multi Layer Perceptron (MLP) classification algorithm. The main attributes used in both datasets were fathers' job, mothers' job, travel time, first period grade, extracurricular activities, second period grade and previous class performance. The datasets were split into two-thirds training sets and one-third testing set and found that Arts College had a prediction accuracy of 64.5% while UCI was 91.42%. It was concluded that the attributes chosen under the UCI dataset are high influences on students' performance.

http://www.ijarcce.com/upload/2016/february-16/IJARCCE%2017.pdf

The Effects of Socioeconomic Characteristics of Students on Their Academic Achievement in Higher Education by Ekber Tomul and Gokhan Polat

A study was conducted with 691 undergraduate students to see if their academic performance was heavily influenced by school-related and socioeconomically factors. The focus was on families socioeconomic status such as: parents' educational status, family income, the settlement where the family lives, the status of the fathers' workplace, number of siblings and the educational background of the student. The study uses correlation and regression analysis in order to determine whether socioeconomic characteristics are related and have an effect on academic achievement. The study found that there were no significant relationship between family's socioeconomic status despite numerous research and articles that state otherwise. The results show that the type of high school that the student was enrolled in has a strong relationship with the students' performance.

http://pubs.sciepub.com/education/1/10/7/

Dataset

In this research project, I will be using two datasets provided by UCI Machine Learning Repository, which can be found here: https://archive.ics.uci.edu/ml/datasets/Student+Performance. Both datasets represent the students' performance in two different classes, Math and Portuguese. The data collected has 32 attributes that reflect family socioeconomic status, demographics and school-related features.

The attributes are listed below, provided by UCI Machine Learning Repository:

- school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
- sex - student's sex (binary: "F" - female or "M" - male)
- age - student's age (numeric: from 15 to 22)
- address - student's home address type (binary: "U" - urban or "R" - rural)
- famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
- Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

- Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
- guardian - student's guardian (nominal: "mother", "father" or "other")
- traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- failures - number of past class failures (numeric: n if 1<=n<3, else 4)
- schoolsup - extra educational support (binary: yes or no)
- famsup - family educational support (binary: yes or no)
- paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- activities - extra-curricular activities (binary: yes or no)
- nursery - attended nursery school (binary: yes or no)
- higher - wants to take higher education (binary: yes or no)
- internet - Internet access at home (binary: yes or no)
- romantic - with a romantic relationship (binary: yes or no)
- famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- 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)
- Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- Walc - weekend 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, output target)

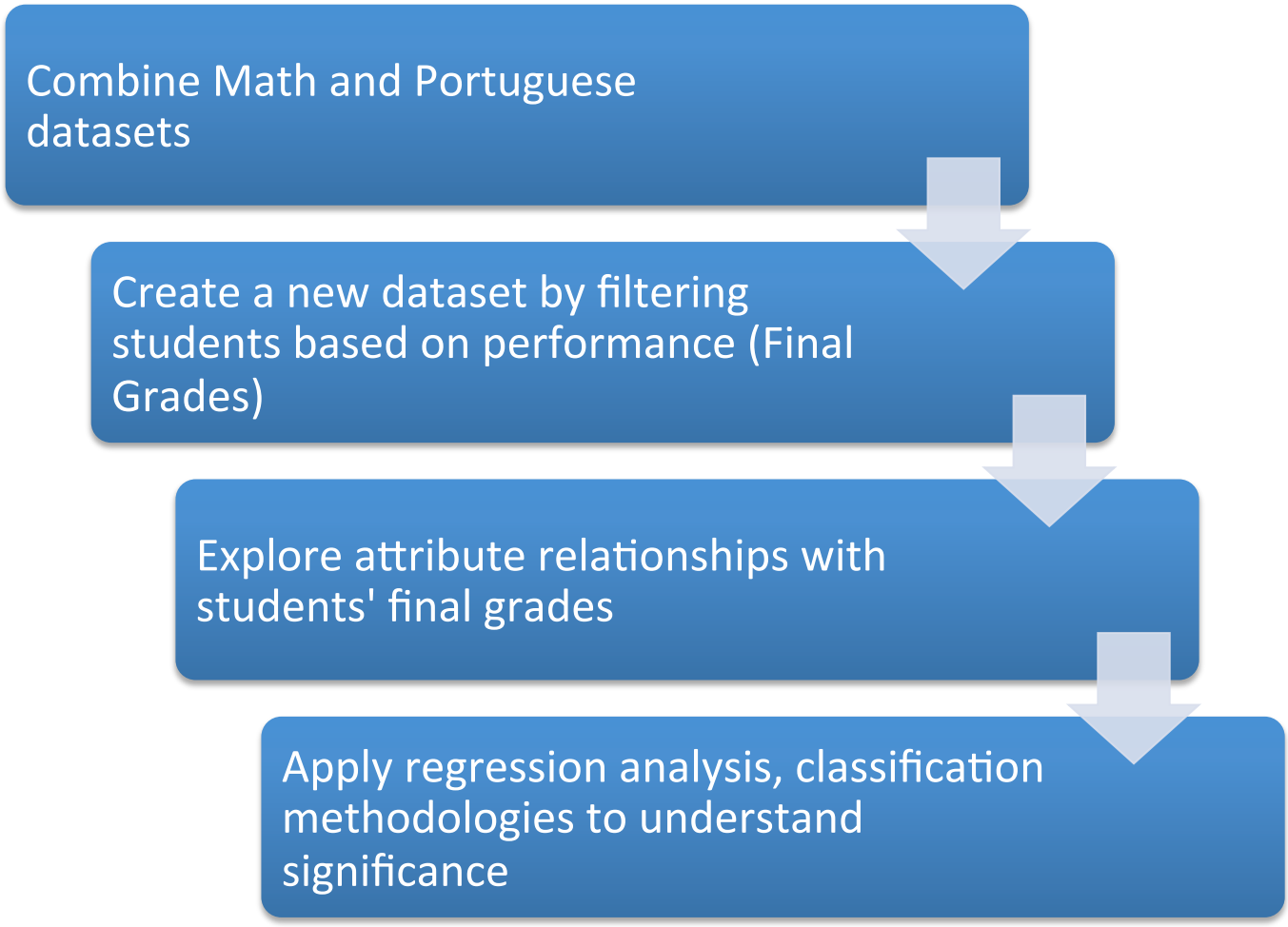
The attributes I will be using are:

- Sex
- Mother’s education
- Father’s education
- Quality family relationship
- Study time
- Failures
- Higher (seeking higher education)
- Extra school support
- Family Educational Support
- Weekday alcohol consumption
- Weekend alcohol consumption

- Absences
- G3 – Final Grades

Based on the literature reviews, I found that my focus would be on family socioeconomics and attributes that are believed to be the causes and significant impacts on a student's grade. The students' final grades will only be used in this analysis to avoid any ambiguity when comparing to the attributes I will be using (ie. did the study time increase/decrease after seeing their first grade, did school support happen the entire enrollment or midway into it etc).

Approach



Step 1: Combine Math and Portuguese datasets

There are two datasets that represents students enrolled in Math and /or Portuguese courses. I will be filtering data to students that are enrolled in both courses. To ensure that that I've captured the students corrected, we merge the datasets based on these attributes below:

- School
- Sex

- Age
- Address
- Family size
- Parents cohabitation status
- Mother's education
- Father's education
- Mother's job
- Father's job
- Reason to choosing specific school
- Nursery attendance
- Internet availability

Step 2: Create a new dataset by filtering students based on performance

I will be creating a new dataset based on Step 1, where I will only be analyzing a group of students who are performing below the average. To do this, the following steps are:

- Calculate the average final grade (G3) of all the students
- Students that are below average will be added into the new dataset called lowperformers.

Step 3: Explore attribute relationships with students' final grades

I will be using various graphs to analyze the relationship between the student grades against the attributes that I will be using in this project. In addition to plots, I will be using correlation analysis. Some questions that I will be exploring are:

- How many of the students' mothers attend schooling?
    - Is this related to the students' performance?
    - Does this affect students' failure rate?
- How many of the students' fathers attend schooling?
    - Is this related to the students' performance?
    - Does this affect students' failure rate?
- Is study time, absences, alcohol consumption related to each other?

More analysis will be conducted once I review results and eliminate attributes that don't significantly impact student performance.

Step 4: Apply regression analysis and classification methodologies to understand significance

This step is still in progress. As stated, regression analysis and classification will be conducted after reviewing all attributes and filtering the ones that have a significant impact on students' final grade.