Influence of demographic, social and school factors on Mathematics performance

Erick Calderon-Morales

Fall 2021

Introduction

Education is one of the main key areas of investment that dominate the political agenda in the European union (EU). Despite the recent global economic crisis, policy makers has set ambitious goals like reducing school dropouts, and increase the percentage of the population with an educational degree. With this kind of strategies policy makers are trying to address the increasing social inequalities that affect the citizens of the EU (Busemeyer et al, 2018).

Achieving equality in education is recognized by the politicians as key, but among the EU members there is variation in the socioeconomic factors that influence the citizens of each country. This kind of variation among countries makes difficult to pass general bills that reduce education inequality in the EU. Because of this, it is important to investigate each country separately in order to draw conclusions that can potentially be generalized to the the whole EU. (Hippe, Araujo and da Costa, 2016)

Portugal educational system is characterized for providing universal access to its citizens. This has lead to reducing the number of schools dropouts, (having the lowest out of school rate at the high school level among OCDE members). Despite of this, Portugal has one of the highest unemployment rates of recent graduates with upper secondary education and one of the lowest medium and high educational attainment (O.E.C.D, 2021)

In this report, I investigate the demographic, social and school factors that influence the performance of students in the mathematics test. Specifically I address the following questions, which factors determine the score in mathematics (A1) and which factors can be use to predict if a student fail or pass the test (A2). The main goal of this report is to develop predictive models for the final year grade (G3) using statistical learning methods.

Methods

For addressing which factors determine the score in mathematics (A1) I used multiple regression analysis, Partial least squares regression, Principal components regression, LASSO regression an Ridge regression, while for addressing which factors can be use to predict if a student fail or pass the test (A2) I used decision trees, Bagging, Random forests, Random forest with gradient boosting and k-Nearest Neighbor.

The data set was split into a training set containing the 75% of the observations and testing set containing the 25% of the observation. For the evaluation of all the models I used the testing set. In the case of the linear models I used the mean squared error on the testing data set for evaluating the fit of the models and for the classification models I estimated the accuracy, the sensitivity and specificity of each model using the test data set.

Results

Exploratory study of the variables.

The data set used contains a total of 395 observations (with no missing values) of students from two schools Gabriela Pereira and Mousinho da Silveira. In total 34 variables were measured among students which 31 were considered as factors (Figure 1) and 3 were considered as numeric (g1, g2, g3).

In general, most of the students, come from the school called Gabriela Pereira (88%), live in most of the cases in Urban areas (78%), travel approximately 1h to get to school (65%), dedicated 2h to studying (50%), failed almost no class (78%), do not received extra class support from the school (87%), the parents support them to stay in school (61%), attended to nursery school (70%), want to take higher education (95%), have internet access (83%), have no romantic relationships (67) and have a low alcoholic consumption during the workday. In terms of family, most of the students have families with size greater than 3 (71%) and have good family relationships (40% and 27%) and most of the parents have at least some degree of education. Also in terms of absences, there was 6 students with more than 30 absences (Figure 1).

All students took three tests, the first period grade (g1), the second period grade (g2) and the final grade (g3). All of these have a score that varies from 0 to 20 being 0 the lowest score possible. Overall among students, g1 had a mean value of 10.90 with a standard deviation of 3.31, g2 had a mean value of 10.71 and a standard deviation of 3.76 and g3 had a mean of 10.41 with a standard deviation of 4.5 (Table 1). It is worth mention that each variable has a high correlation between each other which get less stronger across time (Table 2). Finally is important to mention that the proportion of students that got a score of 0 increased across tests. In the g1 test a total of 0 students had a score of 0 while in the g2 and g3 a total of 13 and 38 got a score of 0 respectively. In this report I used g3 as response variable and g1 and g2 as predictors. No outliers or skewed distributions were found in any of these variables (Figure 2).

When each categorical variable was evaluated I couldn't find any dominating trend. This makes me think that the determinants of students success in the g3 test is a combination of different socioeconomic factors (Figure 3)

The zero (zeroVar) and near zero variance (nzv) predictors are those categorical predictors that only have a unique or a highly extreme frequency of values. These kind of predictors are characterized by having a fraction of unique values over the sample size is low (say 10 %) and having a ratio of the frequency of the most prevalent value to the frequency of the second most prevalent value is large (say around 20). I evaluated this in all the categorical and integer predictors and none of them a have this problem (Kuhn and Johnson, 2013) (Table 2).

Linear regression selection/dimension reduction methods to address (A1)

For answering which factors determine the score in mathematics (A1) a total of five models were used.

Classification analysis (A2)

References

Busemeyer, M. R., Garritzmann, J. L., Neimanns, E., & Nezi, R. (2018). Investing in education in Europe: Evidence from a new survey of public opinion. Journal of European Social Policy, 28(1), 34-54.

Hippe, R., Araujo, L., & da Costa, P. D. (2016). Equity in education in Europe. Luxembourg: Publications Office of the European Union.

Indicators, O.E.C.D. (2021). Education at a Glance 2016. Editions OECD, 90.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26, p. 13). New York: Springer.