EDM Analysis on Student Achievement in Secondary Education of Two Portuguese Schools

HUDK4050

Group 13

Chendanni Liu, Jianan Dingqian,

Simon Chen, Tianyi Dai, Zecheng Chang

INTRODUCTION

Education is a foundation in long-term economic progress. The educational level of Portuguese has been gradually improving over the last several decades with a more mature education system and structures (2021). However, even with those improvements, Portugal still stands at the tail end of Europe by analyzing educational statistics, which showed that the high student failure rates and dropping out rates are the two of the main reasons. For example, in 2006 the early school dropping rates in Portugal was 40% among 18 to 24 year-old student groups, while the European Union was just 15%, also 14% in Portugal versus 10% targeted dropping rates in Europe (Luan J., 2002. Minaei - Bidgoli et al). From the reports, the failure in the core classes: Mathematics and Portugueses, is extremely severe, which were supposed to provide fundamental knowledge for the success in the other school subjects like Physics or History class.

Another important circumstance in Portugal is the advances of Information Technology. More and more dedications on the data collection lead to the exponential growth of organizational databases that contain invaluable information and prefigurative relations. With proper manipulations by data analysts, it could grasp certain trends or patterns of a series of problems. After specific analysis, the results could benefit the decision making process and optimize success. Many influential research of similar topics were already conducted. For example, Who are the students taking the most credit hours? What are the factors that affect student achievement? (P. Cortez and A. Silva., 2008) The paper focuses on those practical problems and could benefit the following related research and increase the accuracy of predictions. Then the administrative and professors in higher education facilities could perform corresponding help for targeted students. As an analyst, we could absorb the techniques and experiences from those excellent research.

In this work, the team will analyze the real-world data from two Portuguese secondary schools, and the data sources are from mark reports and questionnaires. Educational success could not only be determined by a student's efforts, but also other details - parents' influences. The main problem needed to be analyzed is the correlation between parents' education background/ occupations and students' decision on taking higher education and their performance.

Breaking into details of the main purpose, several practical approaches were implemented in this research:

i) First problem to explore is whether students' willingness to participate in higher education is influenced by their parents' education backgrounds. We assumed that students who have parents with higher education backgrounds are more likely to seek continued study. When applied, it serves as a motivation for students to perform well in those fundamental classes which are supportive for more fields of study. The problem needs to be addressed because if the situation really occurs, it leaves more thoughts and works for educators to motivate students in pursuing higher education.

ii) Second question, are students' study trends affected by their parents' occupations? Representing this dataset is finding the correlation between students' Mathematics and Portuguese classes performance and their parents' jobs. Being independable on parents' education background, parents' occupations may have a decisive influence on students' preference in learning specific courses. For instance, a student with one of the parents being an engineer will show higher interest in mathematics, and a student with a parent being a teacher or politician will be more engaged in Portuguese class. These types of preference may cause a significant difference in study outcomes.

iii) The third problem is an expansion of problem two. What we would like to define is, what are factors that significantly support students’ achievement in one or more classes (Mathematics and Portuguese). Based on the solid dataset we have, there are variables with demographic and school attributes to analyze. Factors include but are not limited to the student’s family support, extra educational support, extra-curricular activities, Internet access at home, alcohol consumption, and quality of family relationships. A model is desired to explain students’ performances based on these factors.

iv) Fourth, we would like to generalize methods to predict student performance, from these two specific courses to general. Main idea of this approach is not merely to evaluate student performance for all courses . Instead, it would be appropriate to use the prediction for students’ future study plan. Impossible in requiring every student to develop well in all subjects, the predicting model can provide a guide for students in finding their suitable fields of study and courses, with savings in time and effort. It can also be viewed as a tool for educators and parents in helping students from an educational perspective.

v) For the last point, hopefully we can find a way to discard negative educational influences from parents for both students and educators. As educators are devoting themselves to provide an equity in education for all students. It is regrettable to see someone giving up the opportunity of higher education merely due to lack of right guidance.

LITERATURE REVIEW

There is lots of research focused on finding factors that influence students' performance, and one relevant paper that uses the same dataset is the paper called "using data mining to predict secondary school students' performance". This research is concentrated on predicting the final grades of students in Mathematics and the Portuguese language classes using the first and second period of grades along with other factors such as their demographic social and other school related data. The author modeled under these three goals: binary classification and classification with five levels and regression with numeric output between zero and twenty. The data mining result is based on e Percentage of Correct Classifications (PCC) values in both binary classification and Five-level classification in percentage between 78.5%-91.9% presents the relative importance of the input variables; and the Root Mean Squared Error (RMSE) value from regression analysis represents the low global error. Then further tested the prediction's accuracy by using decision trees, random forests, neural networks and support vector machines. The research has shown graphically on these tests and all suggest that the prediction model is highly accurate and can be used in the future prediction practice.

The paper then concluded that it is possible to predict students' grades when 1st and 2nd period grades are known, which confirmed the previous research of "Predicting Students' Performance in Distance Learning Using Machine Learning Techniques" by Kotsiantis in 2004 that students' achievement is highly affected by previous performances. What the paper did not specifically mention is how family alone has affected and influenced their performance, and how each factor's relationship with their final grades performance on both courses. As the research stated in the end, more research is also needed in order to understand why and how some variables affect students' performance for these two courses. That is why for our research purposes, we are interested in the relationship and correlations on how parents' education background, occupation, social status, demographic variables may affect students' performances in Mathematics and the Portuguese language classes.

DATA ANALYSIS

In order to do this analysis. We need a dataset that has students' family background information and grades at school, and it will be even better if we can know whether they want to achieve higher education in the future.

We found a dataset from UCI Machine Learning Repository that contains information we need. This dataset approaches student achievement in secondary education of two Portuguese schools. The data attributes include students' grades, demographic, social and school related features, and it was collected by using school reports and questionnaires. There are 2 files included in this dataset, the first one is the math grades and has 395 observations, and the other one is Portuguese grades which has 649 observations, both of the files have 33 attributes.
After some preprocesses on the dataset, we found a major problem, one of the target variables, whether the student wants to get higher education in the future, is extremely imbalanced. This target is a binary variable, students can only answer yes or no. Ideally, in order to build a robust model for classification, there should be a 50% and 50% distribution. But, it turns out more than 90% of the answers are yes and only 10% of the answers are no. This will become a problem if we are going to build a classification model later.

Good thing about this dataset is there are no missing values and no data accuracy problem, so we don't need to remove any value from this dataset.

In order to check if there is any correlation between students' family background and grades at school, we can calculate the correlation between these numerical variables directly. But for another question we want to figure out family background information and whether he/she wants to get higher education in the future. We can't calculate the correlation directly, because they are numerical variables and categorical variables. However, we can build a logistic regression instead to see if we can use the independent variable to classify the target categorical variable. If we can successfully build a logistic regression and F test provide sufficient evidence to conclude that the regression model fits the data better than the model with no independent variables. Then, we can conclude that there is a relationship between them, We calculated the correlation matrix between all the numerical variables, and the result is that there is no evidence indicating that there is a correlation between student family background and their grades at school, neither math grades nor Portuguese grades. When we were building the logistic regression, imbalanced data caused a serious problem, the model even worse than just guessing the result since the AUC-ROC score is less than 0.5. What we did is use weighted logistic regression instead. Basically, weighted logistic regression will give more penalty for wrong prediction of minority class. With this change, the AUC-ROC increased to more than 0.8. Now, we can conclude that we can use the logistic regression to classify the target variable with the family background information.

IMPLICATIONS

` Some implications that can be extended from our project may include the necessary family surveys on parents' specific behavior problems, examples like collecting the reasons behind their parents' high alcoholic consumption, and hold educational forums and social events to help these kinds of issues for parents. Our data can only provide the first glance of what might affect students' level of desire to pursue higher education, but in order to improve the condition and performances, more social behavioral researches are needed in details about the family's social problems and hidden reasons behind their low education attainment, government and family support agencies must take into account of these factors and to allocate resources and programs more specifically.

While it is true that our data and analysis only provides a model or result of influencing factors. Problems still remained for parents and educators to take actions on. Results showed that there exists decent correlation between family backgrounds and students' performance. An important way to be considered as supporting students is building a connection between school and family, teachers and parents. Parents should not throw full educational responsibilities on teachers, which occurs more severely in boarding schools. Depending on disputes of how parents would affect students, we made an analysis on the data from two Portuguese schools. This type of research would be most effective when applied regionally. Educators may implement the same research for public schools that recruit students from the same area to provide detailed guidance on students' educational plans. It would be easier and accessible to do social behavioral research on parents per community. When targeted populations change, the same data collecting and analyzing path mentioned in our project can be followed. Results are both beneficial in Education and Sociology.

CONCLUSIONS & DISCUSSIONS:

In this project, we analyzed the students' achievement in secondary education of two Portuguese schools. We collected the related data, built the logistic regression and used the F test to see the correlations. The preliminary analysis shows that the educational background of the family does not influence the student's grade but is related to the educational level of children, especially the desire to pursue higher education.

However, more data is needed to validate our findings. Because the data we found may be skewed, more than 90% of the students in the data want to pursue higher education. And only two schools in our data, different schools may have different results.

On the other hand, More research is also needed in order to understand why and how some variables affect student performance. For example, we can use social network analysis to find the common factors of the children who are born in families with lower educational backgrounds. By observing the outstanding students who lived in bad environments (parents do not encourage education or cannot afford to study), we can conclude how to get rid of the negative influence of the family so that students can better seek higher education.

REFERENCES

Anonymous. (2021, March 30). Portugal. Eurydice - European Commission. Retrieved December 18, 2021, from https://eacea.ec.europa.eu/national-policies/eurydice/content/portugal_en

Luan J., 2002. Data Mining and Its Applications in Higher Education. New Directions for Institutional Research, 113, 17–36.

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

Search programmes. EERA. (n.d.). Retrieved December 18, 2021, from https://eera-ecer.de/ecer-programmes/conference/23/contribution/44328/