Gender gap in scientific studies: a data mining approach

Renza Campagni, Donatella Merlini, M. Cecilia Verri
Dipartimento di Statistica, Informatica, Applicazioni
University of Florence, Italy
[renza.campagni,donatella.merlini,mariacecilia.verri]@unifi.it

KEYWORDS

Educational data mining, computer science, gender gap, clustering

1 INTRODUCTION

Educational data mining (EDM) is a research area that explores and analyzes large repositories of data usually stored in the schools and universities databases for administrative purposes and large amounts of information generated in e-learning or web-based educational context. The aim is to full understand the performance of the student learning process and improve the entire educational process. Several data mining models have been designed and implemented to analyze the performance of students, mainly concerned with techniques such as clustering, classification, association rules mining and sequential pattern analysis: [1, 5, 6] are recent surveys illustrating the state of the art of EDM and [4] is a recent study related to gender gap.

In the present work we use clustering techniques to study seven cohorts of students, from the academic year 2010-2011 up to 2016-2017, belonging to scientific degree courses sharing the same self assessment test required to students before enrolling in the University of Florence. In our analysis we concentrated on active pure students, that is, students who have taken at least an exam within December of the second year and without a prior university career. The data set contains the following attributes for 2283 students: the student identifier, the laurea degree, the student cohort, the student gender, the high school attended by student, the standardized value of the grade obtained in the entrance test, the number of credits corresponding to exams with a grade and, finally, the average grade. By using an approach similar to that presented in papers [2, 3], the present work focuses on the study of gender differences, in terms of the number of enrollments in the scientific degree programs and of the students productivity during the first year. The goal is to use data mining techniques to give analytical evidence that the result in scientific studies does not depend on gender and then use these results in tutoring activities to encourage girls enrollment.

2 DATA UNDERSTANDING AND MINING

The degree programs of the scientific area present a varied distribution of students according to gender: Computer Science has a female percentage of less than 14% against a 40% on the totality of the students of all scientific degrees. Both in Europe and in the United States the number of women enrolled in Computer Science courses is extremely low and the problem is the subject of many studies. During the analysis of students productivity, we found that on average male students seem to be able to achieve a greater number of credits during the first year but the difference with female productivity is minimal and does not involve the grades; this is true for all students and in particular for those of Computer Science. The correlation between the result of the entrance test and the students productivity shows a slightly higher value for girls enrolled in Computer Science. We performed a cluster analysis of students by using the k-means implementation of the software WEKA, by trying several values of k and different attributes. In our analysis we measured cluster validity by computing the Pearson's correlation between the proximity and incidence matrices. We obtained the best results in terms of correlation with k=3 and using as clustering attributes the number of credits corresponding to exams with a grade, the average grade and the grade of the self-assessment test. Despite the great variability of the studies undertaken by students, the analysis seems to identify three groups of students affected by the results of the test that repeat fairly similar in all the degree courses under examination regardless of gender, with values of correlation ranging from -0.66 to -0.8, where -1 means a perfect negative linear relationship: this applies in particular to the Computer Science students.

Clustering identifies, independently of gender, a first group of students with high results both in the test and in the exams taken in the first year of study, a second group with poor results in the test but good results in the exams and, finally, a group with poor results everywhere. Data Mining techniques applied to the study of the productivity of students of scientific degrees at the University of Florence show that there is not substantial difference in results between males and females. Even in Computer Science course, where the presence of women is low, the few girls have the same results as males, if not better: we can think that the few girls have a higher level of motivation.

The analysis conducted on this data set analytically shows that the success of studies in the scientific area does not depend on gender. Other techniques and characteristics of the students could be considered to highlight that the low percentages of girls attracted by scientific studies are mainly due to cultural factors.

ACM womENcourage 2019, September 16–18, 2019, Rome, Italy
2019.

REFERENCES

[1] R.S.J.D. Baker. 2014. Educational data mining: an advance for intelligent systems in education. *IEEE Intelligent Systems* 29, 3 (2014), 78–82.
[2] R. Campagni, D. Merlini, R. Sprugnoli, and M.C. Verri. 2015. Data Mining models for student careers. *Expert Systems with Applications* 42, 13 (2015), 5508–5521.
[3] R. Campagni, D. Merlini, and M.C. Verri. 2018. The influence of first year behaviour in the progressions of university students. In *Communications in Computer and Information Science*. Springer International Publishing, 343–362.
[4] S. Chopra, H. Gautreau, A. Khan, M. Mirsafian, and L. Golab. 2018. Gender Differences in Undergraduate Engineering Applicants: a Text Mining Approach. In *Proceedings of EDM'2018*. 44–54.
[5] A. Pena-Ayala. 2014. Educational data mining: a survey and a data mining-based analysis. *Expert Systems with Applications* 41 (2014), 1431–1462.
[6] C. Romero, J. R. Romero, and S. Ventura. 2014. Educational data mining: an advance for intelligent systems in education. *Educational Data Mining Studies in Computational Intelligence* 524 (2014), 29–64.