A data mining approach to study gender differences in scientific degrees courses

Renza Campagni, Donatella Merlini and M. Cecilia Verri

Dipartimento di Statistica, Informatica, Applicazioni
Università di Firenze, Italy
[renza.campagni,donatella.merlini,mariacecilia.verri]@unifi.it

Abstract. In this paper we present an analysis of the productivity of students attending scientific degree courses using data mining techniques and focusing the study on gender. Particular attention is given to the degree course in Computer Science in which the gender gap is extremely high in order to see if there are different behaviors compared to other courses in the same area. This study proves in an analytic way the existence of three categories of students with similar characteristics in terms of test results and productivity, transversal to gender.

Keywords: Educational Data Mining, Students Productivity, Gender Gap

1 Introduction

Educational data mining (EDM) is a recent research area that explores and analyzes, by using machine learning and data mining algorithms, both large repositories of data usually stored in the schools and universities databases for administrative purposes and large amounts of information about teaching-learning interaction generated in e-learning or web-based educational context. EDM seeks to use all this information to better understand the performance of the student learning process and can be used by the university or school management to improve the entire educational process. Over the last years, several data mining models have been designed and implemented to analyze the performance of students and we refer to [1,5,6] for recent surveys about the state of the art of EDM and to [4] for a recent study related to gender gaps. In particular, the existing literature about the use of data mining in education is concerned with techniques such as clustering, classification, association rules mining and sequential pattern analysis.

Such techniques have been recently applied to data concerning students in the Computer Science laurea degree of the University of Florence. For example, [2] proposes a data mining methodology, based on clustering and sequential pattern analysis, to study the student behavior by comparing student careers with the ideal career of a virtuous student who takes every examination just after the end of the corresponding course. In [3] a cluster analysis is used to classify students according to the results of the self-assessment test and the first year performance. This study highlights three groups of students strongly affected by the results of the first year: high achieving students who start high and maintain their performance over the time, medium-high achieving

students throughout the entire course of study and, low achieving students unable to improve their performance who often abandon their studies.

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 the scientific degree courses sharing the same self-assessment test required to students before enrolling in the University of Florence. The work focuses on the study of gender differences, both from a numerical point of view (number of enrollments in degree programs in the scientific area), and from a productivity point of view during the first year. In this regard, the number of credits acquired during the first year and the average grade are analyzed, divided by gender, in the various degree courses and in the different academic years under consideration. Despite the great variability of the studies that these students have undertaken, the analysis seems to identify also in this case 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.

2 Data preprocessing

In this section, we describe how university students' data are organized, referring to the scientific degree courses of the University of Florence, Italy, under the Italian Ministerial Decree n. 270/2004. These academic degrees are structured over three years and every academic year is organized with several courses, each course has assigned some credits (CFU) for an amount of 60 credits in each year. Data under analysis concern university students enrolled from academic year 2010-2011 (afterwards cohort 2010) up to 2016-2017 (afterwards cohort 2016). Each student, before enrolling in the degree course, has to take an entrance test to self-evaluate his background in mathematics. This test consists of some multiple choice questions on mathematics topics usually studied in high school.

Initially, we worked with two different data sets: the first contains information about students and their school career before entering university, together with information on the entrance test; the second contains information about grades and credits in the exams taken by students. We performed an important preprocessing phase to fix errors and to reorganize data, before applying the various analysis techniques. During this preprocessing phase we joined and aggregated the previous data sets to obtain the productivity of the student in a year in terms of total credits and the average grade. Since the self-assessment test has changed over the years, the corresponding results have been standardized to make the scores comparable.

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 (for example, December 2011 for students of cohort 2010) and without validated exams, that is, exams taken in previous laurea degrees. For some students, this corresponds to passing only an exam without grade such as English. At the end of the preprocessing phase we obtained data organized as in Table 1, where *Id* is the student identifier, *degree* is the laurea degree, *cohort* is the student cohort, *gender* is the student gender, *school* is the high school that the student attended and, *test_n* is the standardized value of the grade obtained in the

entrance test, *cfu_g* is the number of credits corresponding to exams with a grade, and *grade* is the average grade, varying in the range 18..30.

Table 1. A sample of postprocessing data.

Id	*degree*	*cohort*	*gender*	*school*	*test_n*	*cfu_g*	*grade*
100	01	2010	F	LS	0.89	57	28
200	01	2010	M	IT	0.71	48	28
…	….	…	…	…	…	…	…
800	05	2016	M	LC	0.85	60	26
900	06	2016	M	IT	0.70	48	27

3 Data understanding

The degree programs of the scientific area present a very varied distribution of students according to gender: as can be seen in Fig.1[1], they range from the degree course in Computer Science which has a female percentage of less than 14%, to courses with a peer gender distribution as Mathematics and Natural Sciences, up to courses, such as Optics, with a female percentage of almost 70%.

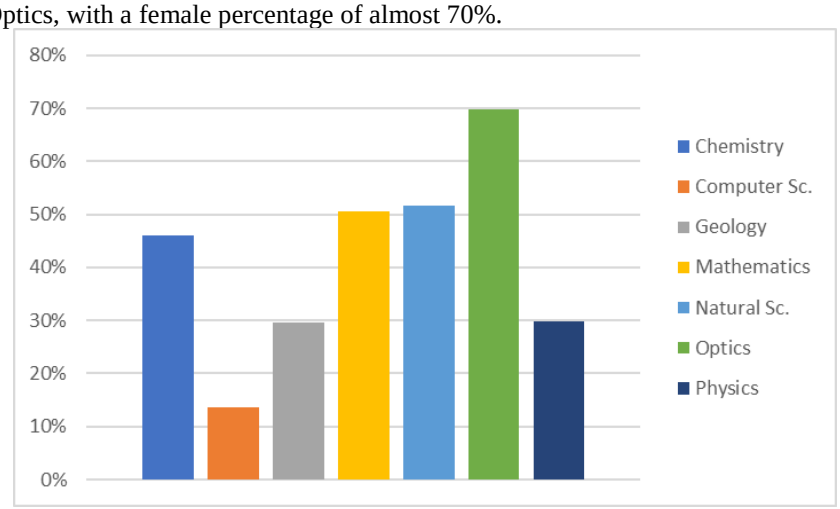

Fig. 1. Female percentage of students belonging to the scientific degree courses of the University of Florence in the years [2010,2016].

The female percentage has small fluctuations over the years, but the gender distribution is quite characteristic of each degree course, as shown in Fig. 2. We wish to point out

[1] For interpretation of the references to colors in the figures, the reader is referred to the electronic version of this paper.

that the line for the Computer Science degree shows the lower values over the entire interval. The low percentage of women enrolled in the Computer Science course is a phenomenon that has spread both in Europe and in the United States since the mid-1980s: many hypotheses have been made about the reasons for this phenomenon and, in recent years, several initiatives from Universities and private companies have been taken to try to reverse the trend [7,8].

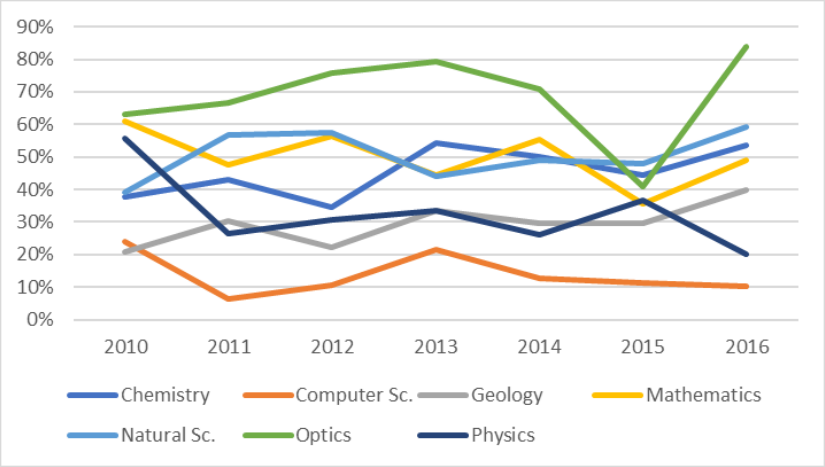

Fig. 2. Female percentage of students belonging to the scientific degree courses of the University of Florence from the academic year 2010-2011 up to 2016-2017.

In the sequel we will study, with a typical data mining process, the behavior of incoming students in the scientific degree courses of the University of Florence and their productivity at the end of the first year of study, according to the gender and the degree course.

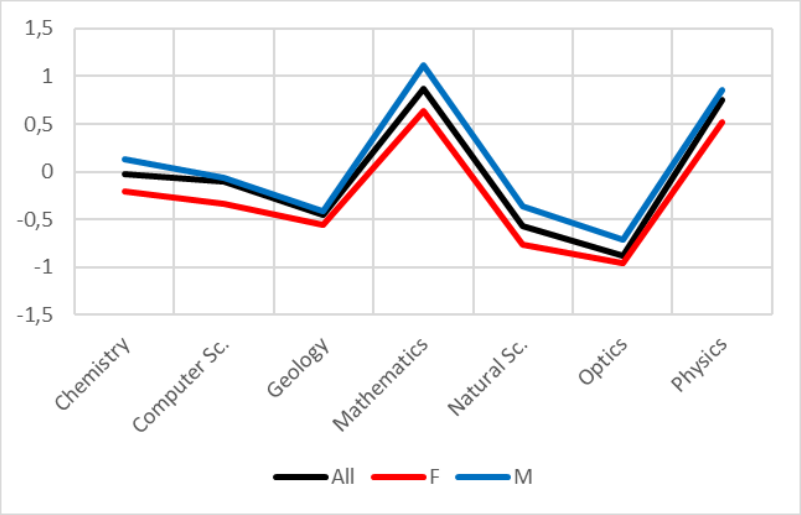

Fig. 3. Average results of normalized test values by degrees.

The degree programs examined share the same self-assessment test, as already illustrated in Section 2; Fig. 3 shows the normalized average results of the test by gender and degree course. Gender differences are minimal while variations among courses are much more evident. The Computer Science degree fits almost perfectly the average value.

3.1 Productivity

In this section we want to analyze student productivity during the first year of the course of study and determine if there are correlations between the result of the self-assessment test and the results achieved in the same period. As mentioned in the Section 2, we consider only active students, that is, those who have passed at least one exam by December of the year following enrollment. To compare productivity we have calculated the average number of credits acquired during the first year of study (Fig. 4) and the average grade obtained in the corresponding exams (Fig. 5)

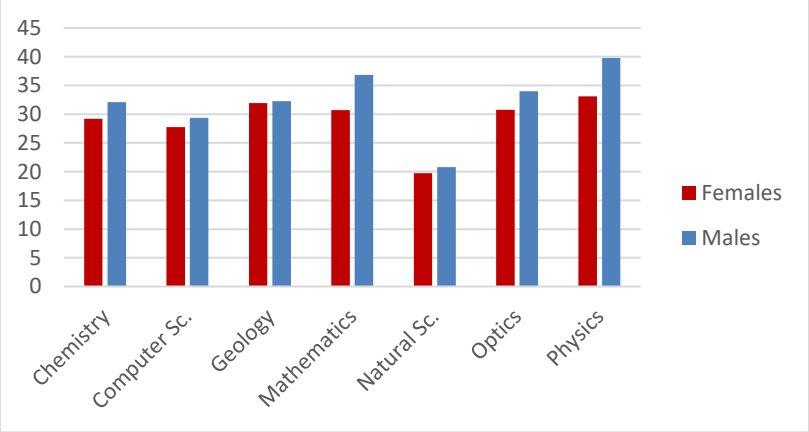

Fig. 4. Average number of credits acquired during the first year of study.

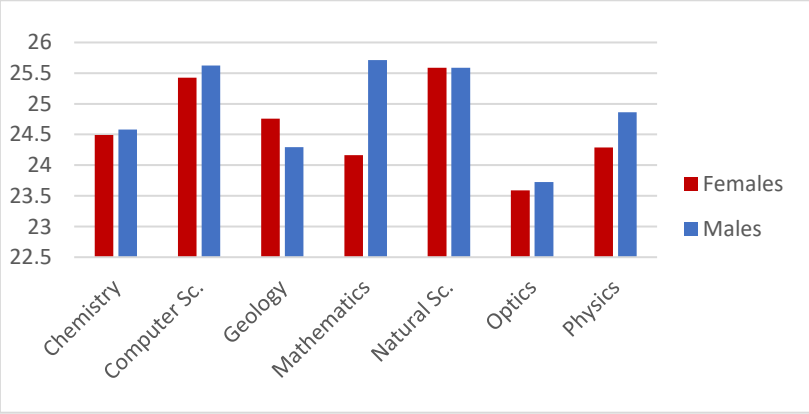

Fig. 5. Average grade in exams taken during the first year of study.

As can be seen from the figures, male students are able to achieve on average a greater number of credits: in general the difference is minimal and only in two cases (Mathematics and Physics) this difference reaches about 6 CFU i.e. the weight of an exam. This small difference in productivity does not always correspond to higher marks.

In the following Fig. 6 and 7 we can see the correlation between the result of the entrance test and the productivity of the students: we note that there is not a great difference between males and females while this gap is more marked among the various degree courses.

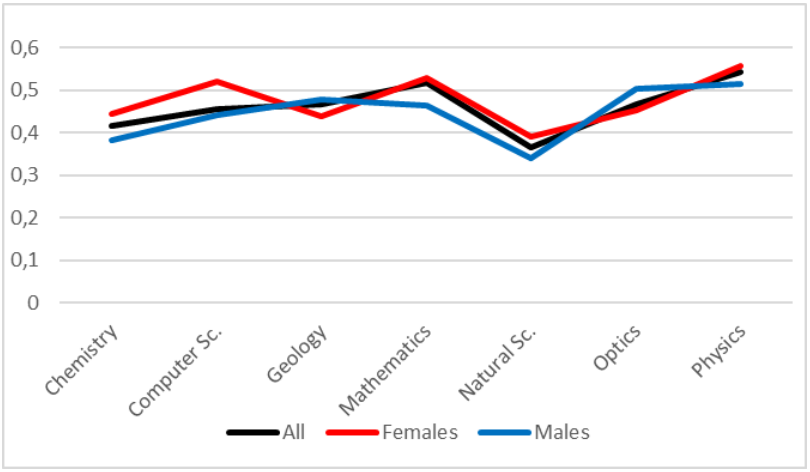

Fig. 6. Correlation between test result and credits acquired during the first year.

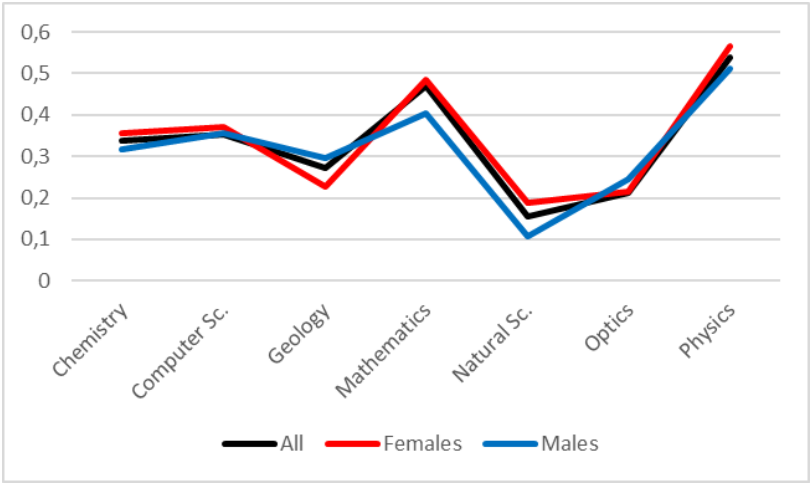

Fig. 7. Correlation between test result and average grade in exams taken during the first year.

4 Clustering students

In this section we perform a cluster analysis of students by using the k-means implementation of the software WEKA. We tried the k-means algorithm with several values of k and with $k = 3$ we obtained the clusters for the students of the seven cohorts from 2010 up to 2016, illustrated in Fig. 8. As cluster attributes we used the number of credits corresponding to exams with a grade, *cfu_g*, the average grade, *grade*, and the grade of the self-assessment test, *test_n*. Other choices of attributes are possible but in the various tests carried out these are the ones that have given the best results. In our analysis we measure cluster validity with correlation, by using the concept of proximity and incidence matrices: in the proximity matrix $P = (P_{i,j})$ each element $P_{i,j}$ represents the Euclidean distance between elements i and j in the data set; in the incidence matrix $I = (I_{i,j})$, each element $I_{i,j}$ is 1 or 0 if the elements i and j belong to the same cluster or not. We then compute the Pearson's correlation between the linear representation by rows of matrices P and I and we expect to find a negative value, where -1 means a perfect negative linear relationship.

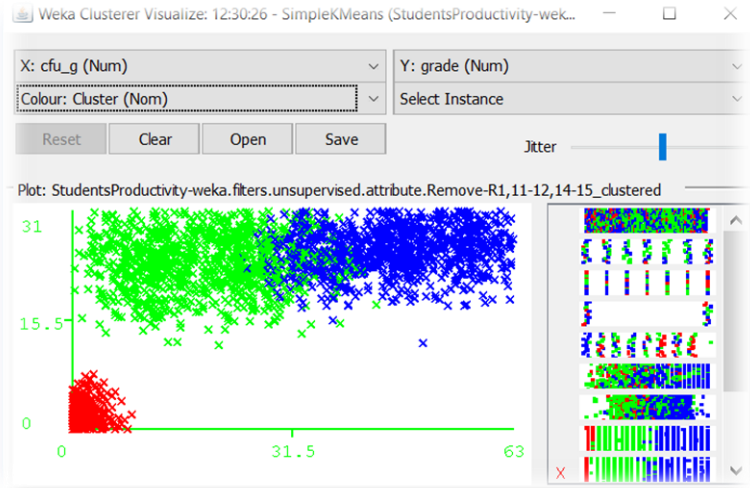

Fig. 8. Clusters for all the 2283 students of cohorts 2010-2016 with respect to *cfu_g*, *grade*, *test_n* and their projection with respect to *cfu_g* and *grade*. In red the students that during the first year had success only with some exam without a grade, such as English, and therefore have no credits and no grade; in green medium and in blue high achieving students.

The centroids of the clusters are illustrated in Table 2 and, in particular, cluster **Low** corresponds to students that during the first year had success only with some exam without grade, such as English, and therefore have no credits and no grade in this clustering, **Medium** identifies medium achieving students and, finally, **High** identifies high achieving students. The clusters are characterized by colours red, green and blue in Fig. 8, respectively. The *Pearson*'s correlation between the linear representation of the proximity and incidence matrices is -0.70, a good value of correlation.

Table 2. Centroids of clustering in Fig. 8: the correlation between the incidence and proximity matrices of all students is -0.70.

Attribute	**Full Data** (2283)	**Low** (285)	**Medium** (1015)	**High** (983)
test_n	0	-0.4906	-0.4092	0.5648
cfu_g	28.2527	0	18.8926	46.1089
grade	21.7525	0	23.9192	25.822

We went on in a similar way to examine the male and female students separately; in Fig. 9 and 10 the three clusters identified by k-means are shown while in Tables 3 and 4 the corresponding centroids are illustrated. The *Pearson*'s correlation between the linear representation of the proximity and incidence matrices for the male students clustering is -0.64 while for the female students clustering it reaches the value -0.80. Fig. 8, 9 and 10 present three different groups of students with very similar characteristics.

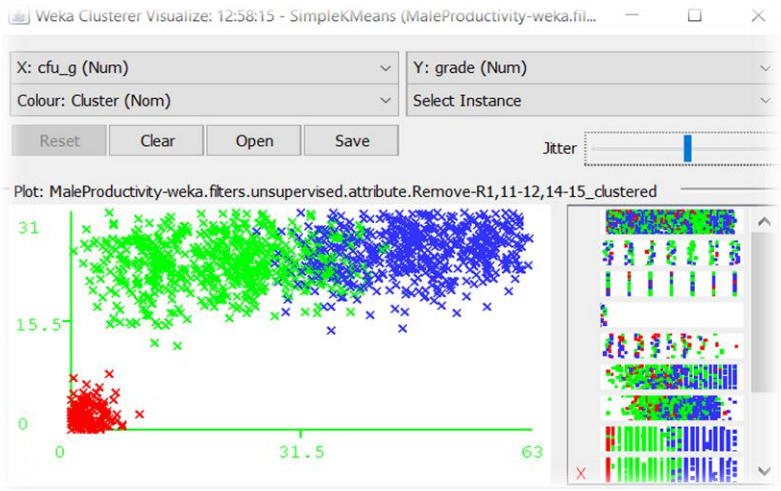

Fig. 9. Clusters for the 1379 male students of cohorts 2010-2016 with respect to *cfu_g*, *grade*, *test_n* and their projection with respect to *cfu_g* and *grade*. In red the male students that during the first year had success only with some exam without a grade, such as English, and therefore have no credits and no grade; in green medium and in blue high achieving male students.

Table 3. Centroids of clustering in Fig. 9: the correlation between the incidence and proximity matrices for male students is -0.64.

Attribute	**Full Data** (1379)	**Low** (150)	**Medium** (650)	**High** (579)
test_n	0.152	-0.4112	-0.223	0.7188
cfu_g	29.8078	0	20.82	47.62
grade	22.3166	0	24.0169	26.1883

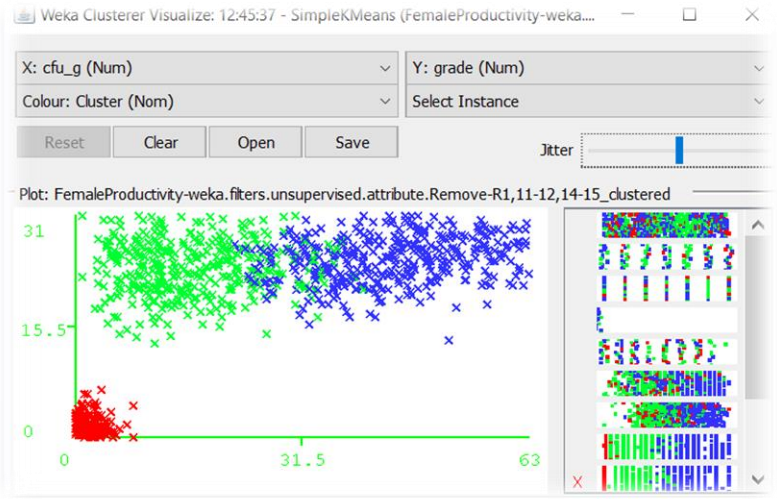

Fig. 10. Clusters for the 904 female students of cohorts 2010-2016 with respect to *cfu_g*, *grade*, *test_n* and their projection with respect to *cfu_g* and *grade*. In red the female students that during the first year had success only with some exam without a grade, such as English, and therefore have no credits and no grade; in green medium and in blue high achieving female students.

Table 4. Centroids of clustering in Fig. 10: the correlation between the incidence and proximity matrices of female students is -0.81.

Attribute	**Full Data** (904)	**Low** (135)	**Medium** (378)	**High** (391)
test_n	-0.2318	-0.5788	-0.7298	0.3694
cfu_g	25.8805	0	16.5952	43.7928
grade	20.8927	0	23.7751	25.3197

Finally, we used k-means algorithm for clustering students of each laurea degree. In Fig. 11 we give the results for Computer Science students and Table 5 reports the centroids of the clusters; the correlation between the incidence and proximity matrices in this case is -0.66. We do not include the analogous figures for the other laurea degrees, however in almost all cases k-means found three groups of students with very similar results in terms of credits with grade, average grade and grade of the self-assessment test. These groups do not change significantly if we study students by degree and gender.

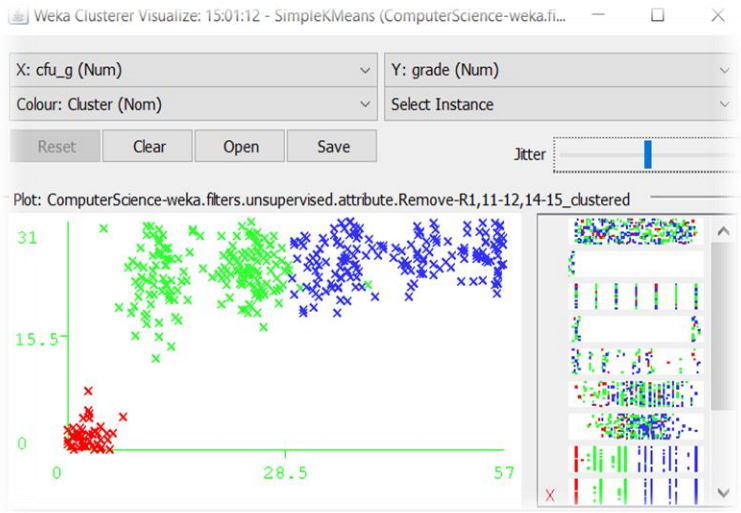

Fig. 11. Clusters for the 395 Computer Science students of cohorts 2010-2016 with respect to *cfu_g*, *grade*, *test_n* and their projection with respect to *cfu_g* and *grade*. In red the students that during the first year had success only with some exam without a grade, such as English, and therefore have no credits and no grade; in green medium and in blue high achieving students.

Table 5. Centroids of clustering in Fig. 11: the correlation between the incidence and proximity matrices of Computer Science students is -0.66.

Attribute	**Full Data** (395)	**Low** (44)	**Medium** (179)	**High** (172)
test_n	-0.097	-0.7088	-0.3206	0.2923
cfu_g	27.5848	0	18.8212	43.7616
grade	22.7468	0	24.581	26.657

5 Conclusions

Despite the great variability of the studies undertaken by the examined students, there is not a substantial difference in the productivity of males and females and, just in the degree courses where the female percentage is low, as Computer Science, it seems that the difference is less pronounced: this could be explained by a higher level of motivation of the (few) girls. Clustering identifies, independently of gender, a group of students with high results both in the test and in the exams taken in the first year of study. A low result in the test, on the other hand, does not necessarily predict poor productivity, but can be an indicator to be used for identifying students to whom guidance and tutoring activities can be directed to try to reduce the dropout rate. The use of data mining techniques can help to take decisions in this direction. In particular, we think that further features of the students could be examined and that other types of algorithms could be used, such as classification and sequential patterns mining, to investigate the reasons for the choice of the studies based on gender.

References

1. Baker, R.S.J.D.: Educational data mining: an advance for intelligent systems in education. IEEE Intelligent Systems, 29(3):78-82 (2014).
2. Campagni, R., Merlini, D., Sprugnoli, R., Verri, M.C.: Data Mining models for student careers. Expert Systems with Applications, vol. 42 (13), 5508-5521 (2015).
3. Campagni, R., Merlini, D., Verri, M.C.: The influence of first year behaviour in the progressions of university students. Communications in Computer and Information Science, 343-362 Springer International Publishing (2018).
4. Chopra, S., Gautreau, H., Khan, A., Mirsafian M., Golab, L.: Gender Differences in Undergraduate Engineering Applicants: a Text Mining Approach. Proceedings of EDM'2018, 44-54 (2018).
5. Peña-Ayala, A.: Educational data mining: a survey and a data mining-based analysis. Expert Systems with Applications, 41:1432-1462 (2014).
6. Romero, C., Romero, J. R., Ventura, S.: A survey on pre-processing educational data. Educational Data Mining Studies in Computational Intelligence, 524, 29-64 (2014).
7. Thompson, C.: The Secret History of Women in Coding, The New York Times Magazine (13 Feb. 2019).
8. Fondazione IBM Italia - Progetto Nerd, last accessed 2019/04/09.