Data-driven tool for monitoring of students performance ⋆

R. Vilanova [*] **M. Dominguez** [**] **J. Vicario** [*] **M.A. Prada** [**]
M. Barbu [***] **M. J. Varanda** [****] **P. Alves** [****] **M. Podpora** [†]
U. Spagnolini [‡] **A. Paganoni** [‡]

[*] *Universitat Autonoma Barcelona, Spain*
[**] *Universidad de León, Spain*
[***] *University Dunarea de Jos, Galati, Romania*
[****] *Instituto Politecnico de Bragança, Portugal*
[†] *Opole University of Technology, Poland*
[‡] *Politecnico di Milano; Italy*

Abstract: In today's education, school success is defined as ensuring achievement for every student. To reach this goal, educators need tools to help them identify students who are at risk academically and adjust instructional strategies to better meet these students' needs. Student progress monitoring is a practice that helps teachers use student performance data to continually evaluate the effectiveness of their teaching and make more informed instructional decisions. This paper reflects the main output of the SPEET project as an IT tool that implements specific algorithms developed to deal with the basic problems tackled in the project: Classification, Clustering and Drop-out Prediction.

Keywords: Student performance, monitoring, data-mining education

1. INTRODUCTION

Performance, an outcome of education, is the extent to which a student, teacher or institution has achieved their education goals. Academic achievement is commonly measured by examination or continuous assessment but there is no general agreement on how it is best tested or which aspects are most important. Education, in general sense, is the means through which the aims and character of a group of people living from one generation to the next is achieved. In every educational institution academic performance needs to be controlled quantitatively. The method and procedures to evaluate the student performance always demand tremendous efforts ranging from student's assessment to result processing, which is the best method to control student performance. Examination, in an academic or professional context, is a test which aims at determining the ability of student or prospective practitioners. Examination are usually written test although some may be practical and vary greatly in structure, contents and difficulty depending on the subject, the age group or level of the tested persons and profession. See A. A. Akinrefon (2014).

The purpose of academic feedback, whether by design or accidentally, is complex and far from singular in nature. Feedback can be an encouragement to the recipient, it can help to instill confidence in any marks given and it can help to focus the mind of the assessor, as well as providing the necessary insight to facilitate improvement both for tutor and student; Carless (2006). Much of this work has relied on data collection from both students and academics, with subsequent analysis and conclusions being drawn. See Bailey (2009); Burke (2009).

Many studies into the nature, use and value of feedback have been carried out from both the student and academic perspective. This work reflects the main output of the SPEET [1] project as an IT tool that implements specific algorithms developed to deal with the two basic problems tackled in the project: Classification, Clustering and Drop-out Prediction. First of all, in the next section the SPEET project is presented as well as its main goals. Next, the previously mentioned contributions are detailed:

- Performance analysis algorithms: student performance analysis on the basis of categorical and/or performance data; performance for upcoming semesters on the basis of initial information; for explanatory analysis, etc
- Drop-out prediction algorithms: drop-out prediction on the basis of selected categorical information and first semester grades. A statistical model is elaborated that provides a quantitative evaluation of the student being at risk of drop-out.

⋆ This work has received partial support from the National Programme of R&D aimed at the Challenges of Society, co-funded by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund. The data preprocessing and development of the tools and evaluation testbeds was conducted in the computing infrastructure provided by project DPI2015-69891-C2-1-R, whereas analysis of the best suited data mining has benefited from those provided by project DPI2016-77271-R.

[1] Student Profile for Enhancing Engineering Tutoring

- Visualization tools: for visual inspection of the pre-existing data relationships. Dimensional reduction and histogram techniques are applied to project the data on appropriate dimensions suitable for analysis. The tool provides a complete interactive, on-the-fly visual representation of the data.

2. THE SPEET PROJECT

SPEET (Student Profile for Enhancing Engineering Tutoring) is an European project funded under the ERASMUS+ programme as an Strategic Partnerships for higher education. The partnership includes higher education institutions from Spain, Portugal, Italy, Poland and Romania:

- Spain: Universitat Autonoma Barcelona (UAB) and Universidad de León (ULEON)
- Romania: University Dunarea de Jos, Galati (GALATI)
- Portugal: Instituto Politecnico de Bragança (IPB)
- Poland: Opole University of Technology (OPOLE)
- Italy: Politecnico de Milano (POLIMI)

The objective of this project could be stated in a rather simple way as: determine and categorize the different profiles for engineering students across Europe. The main rationale behind this proposal is the observation that students performance can be classified according to their behavior while conducting their studies. After years of teaching and sharing thoughts among colleagues from different EU institutions it seems students could obey to some pretty stable classification pattern according to the way they face their studies. Therefore, if it was be possible to know what kind of student is each student according to these patterns, this would be of valuable help for tutoring her/him in the early stages before drop-out..

On the other hand, after years of having been offering engineering curricula and a sufficiently large number of students having been enrolled, it turns out that academic records of all such students are now stored on the academic offices of our Engineering Schools/Faculties. These records include the performance of the student on the different subjects of the degree as well as, usually, collateral information regarding the student's origin (geographical info, previous studies, age, etc). All this information, taken altogether, should be enough to help characterize the student and be able to determine *what categorical class of student are we dealing with.*

On the basis of the preceding scenarios, this project's goal emerges from the potential synergy among a) the huge amount of academic data actually existing at the academic offices of faculties and schools, and b) the maturity of data science in order to provide algorithms and tools to analyse and extract information from what is more commonly referred to as Big Data analytics. A rich picture can be extracted from this data if conveniently processed. Therefore, the main objective of SPEET is to apply data mining algorithms to process this massive set of student profiles in order to extract information about and to identify common features in each of these student profiles. An idea of the student profile we are referring to within the project scope is, for example: students that completed the degree on time, students that are blocked on a certain set of subjects, students that leave degree earlier, etc. Data analytics are very common in many fields such as customer profiling over internet for shopping, and what is investigated in SPEET is somewhat adapter to help tutors to better know their students and improve counselling actions.

A transnational approach will provide rich information as considered data can be analysed on a country basis and also at transnational level. The fact of obtaining the same student classifications and profiles will show engineering students are likely to be statistically the same all across EU. If instead differences arise, this will show that a more detailed analysis country per country should be carried out and main differences can be exposed as well as a deep analysis of the reason that causes such differences (either in positive or negative perspective). A study like the one envisaged on this project, if carried out just on a local country basis would not be able to provide the beneficial EU perspective.

The main use of this student profile analysis is that of being embedded on supporting IT tools for tutoring. Once key labels for the different profiles are determined, there will be the need to determine the profile each student complies with as it starts. The first results along with collateral data should allow the IT tool to identify the student's profile (or potential profiles when in doubts) and help the tutor to know how to provide the student with the appropriate addressing in order to increase performance and satisfaction with the studies. An immediate step further is that of extending the analysis to other disciplines than engineering (social sciences, medicine, etc) and compare (if any difference) the student profiles that arise. The comparison can be done country and discipline wise [2] .

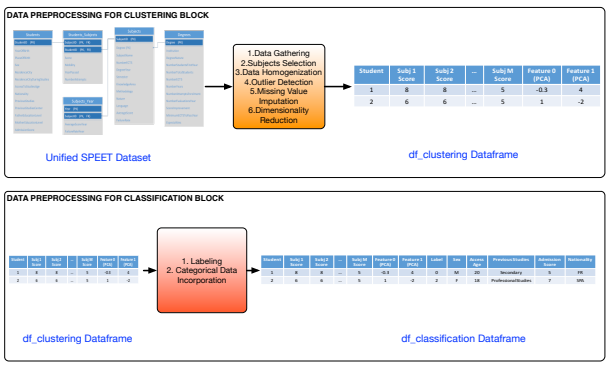


Fig. 1. Preprocessing steps to obtain dataframes used by the Clustering Block (*df_clustering* dataframe) and the Classification Block (*df_classification* dataframe).

3. DATA FORMAT

First of all academic data is conveniently divided into categorical and performance data of the student as it progresses on the semesters of the degree the student is enrolled on. The main idea is to be able to predict student information as soon as possible by joining the categorical data (static) and the semesters performance (dynamic).

[2] The project can be accessed at the website http://www.speet-project.eu

A unified dataset format has been considered for the project as described in Barbu et al. (2017). From this dataset, some pre-processing tasks are performed to accommodate data to the Clustering and Classification tools. This is represented in Fig. 1, where data frames *df_clustering* and *df_classification* are the inputs to Clustering and Classification blocks, respectively. As observed, Clustering is only based on performance data (scores of students at the different subjects), whereas classification data frame includes categorical variables (Sex, Access Age, Previous Studies, Admission Score and Nationality) along with the Clustering Label (0 - *Average Students*, 1 - *Excellent Students* and 2 - *Low Performance Students*). Data frame *df_classification* is also adopted to perform the histogram-based Clustering Explanation.

4. OVERVIEW ON THE IT TOOL FOR STUDENT DATA PROCESSING

In this section, we present an overview of the data processing tools which have been considered for the identification for students' profiles. As presented in Vicario et al. (2018), two data mining tools have been implemented in this project:

- **Classification and Clustering tool**: this is a stationary-based tool consisting in the grouping of students at clusters based on their performance during their studies.
- **Drop-out Prediction tool**: a dynamic tool based on the drop-out prediction of students based on their performance at the first semester of studies.

In this section we concentrate on the classification and clustering tool whereas the drop-out prediction is tackle din Section 6.

4.1 Clustering and Clustering Explanation

As commented, the Clustering mechanism is in charge of organizing students in three Clusters based on their performance: *Average Students*, *Excellent Students* and *Low Performance Students*. In Fig. 2, one example is provided where the three clusters can be clearly observed:

Once the Clusters are generated, Clustering Explanation is performed by analysing each of the categorical variables for each group of students. In Fig. 3, one can observe an example where it is observed how Excellent Students tend to be women, younger and with a high admission score. Then students patterns are obtained by means of analysing what categorical variables influence each of the clusters.

4.2 Classification

The Classification block is in charge of classifying new students to the clusters generated at the Clustering block. Concerning the pattern identification, however, this Classification procedure is useful to obtain insights about the structures of plan studies at the different degrees. So, here the tool is not adopted to obtain students' patterns. Its purpose here is to extract degrees' patterns. This can be done by analysing the amount of classification accuracy provided by each of the curses at the degree.

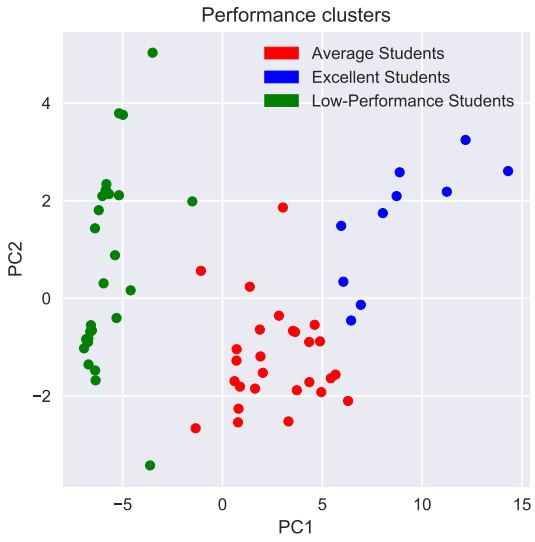


Fig. 2. Performance clusters of students.

Fig. 3. Clustering Explanation based on Histogram analysis of Categorical variables.

In Fig. 4, we provide an example. The first row is related to the accuracy obtained classifying new students when only the performance at the first course is considered, the second row refers to the case where first plus second course performance is considered and so on. In the example provided, it is observed how the first course provides a high level of accuracy w.r.t the other cases. The meaning of this is that the first course influences the way students are grouped in terms of performance. Those students obtaining good results just at the beginning of the degree will also obtain good results at the rest of courses. Therefore, the first year is very important at this degree.

Considered courses	Classification Accuracy
1st	86 %
1st + 2nd	88 %
1st + 2nd + 3rd	90 %

Fig. 4. Degree Analysis based on Classification Accuracy results.

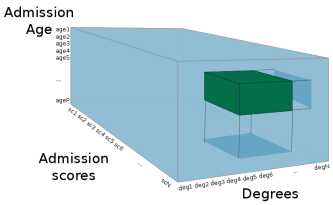
5. OVERVIEW ON THE IT TOOL FOR GRAPHICAL DATA ANALYSIS AND VISUALIZATION

In this section, we present an overview of the data visualization tools, which have been conceived for the support of the exploratory analysis conducted by tutoring staff. As presented in Prada et al. (2018), a visual analytics approach is used in those tools, in order to involve human analysts in the task of knowledge discovery through the blend of information visualization, advanced computational methods and interaction. Thus, these tools take advantage of the ability of humans to understand and interact with complex visual presentations to facilitate their process of hypotheses generation and confirmation. The visualization tool implemented in this project is a Coordinated view tool. This interactive tool provides a set of coordinated histograms where a user can filter by one or more variables, causing the other charts to update accordingly. The coordinated histograms enable the exploration of the distributions of the variables and of the links between them.

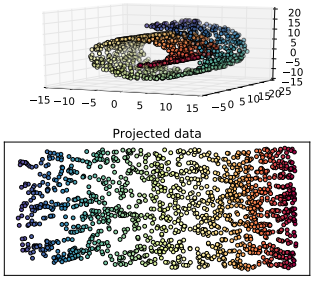
As presented in Barbu et al. (2017), a unified data set format has been considered to be used in the tools. First, it is necessary to eliminate the inconsistencies found in the variable values. Later, we need to create a multi-dimensional array, where each variable can be interpreted as a dimension. This data structure is suitable for the different views of data that are used in the visualization tools, which are represented in Fig. 5.

If each explanatory or performance variable is considered as a dimension, the multi-dimensional array that contains the students' data can be interpreted as a data (hyper-)cube. This is a well-known approach, similar to that of online analytical processing (OLAP) in the business intelligence field, which enables operations such as slicing or dicing (range selections in one or more dimensions). Following this idea, it seems interesting to visually analyze the distribution of any variable, subject to certain filters on the others. But when the histograms or bar charts of the variables are visualised jointly and in a coordinated way, it is not only possible to obtain a global view of the data set but also to explore the correlations between variables. Furthermore, interactive and real-time filtering can be used to facilitate the rapid validation or rejection of hypotheses about a set of students.

Implementation The coordinated view approach has been implemented as a web application that displays an interactive dashboard. The tool shows a set of coordinated



(a) Data cube perspective for the coordinated view tool.



(b) High-dimensional data and its projection to enable visualization

Fig. 5. Data interpretation for both visualization tools

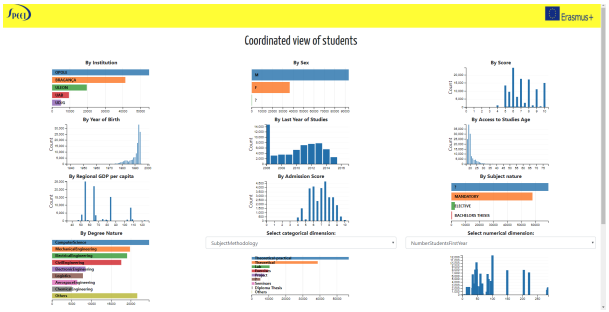


Fig. 6. "Coordinated view" tool.

histograms where a user can filter by one or more variables, causing that the rest of the charts to update accordingly. The charts are fixed or customisable and show the count of student-subject records binned by interval/category. The filters are applied by means of a range selection for the numeric variables and by means of a one-click selection for the categorical ones. Additionally, a histogram of the score grouped by another explanatory variable and a *choropleth* map are included. In Fig. 6, a screenshot of this tool is provided.

6. DROP-OUT PREDICTION TOOL

This tool is in charge of generating a model able to estimate the probability of graduation of students based on categorical and performance variables. Besides providing this probability, which could help to predict potential drop-outs, the parameters obtained with the generated model also help to understand which students' profiles are more sensitive to early drop-out.

This part of the tool has some specific needs for the data needed to perform the prediction. Details about the data format at Drop-out Prediction tool are also presented in Barbu et al. (2017). Departing from the SPEET's unified dataset format, some additional pre-processing actions are performed here. Besides the categorical variables

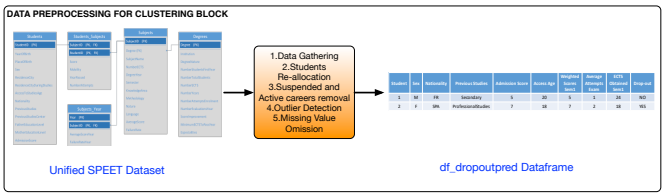


Fig. 7. Preprocessing steps to obtain dataframe used by the Drop-out Prediction tool (*df_dropoutpred* dataframe).

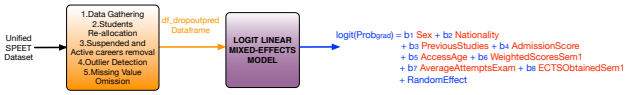


Fig. 8. Block diagram of the drop-out prediction tool.

also addressed at the Clustering block (i.e., Sex, Access Age, Previous Studies, Admission Score and Nationality), student's performance information is considered here but following a different approach. Only information concerning the first semester of the first course is considered (see *df_dropoutpred* dataframe format in Fig. 7). More specifically, three variables are adopted: the number of credits passed at the first semester (ECTS_Obtaine_dSem1), the average number of exam attempts per subject (Average Attempts Exam) and the weighted average score obtained by the student at this semester (Weighted_Scores_Sem1), where weighting is based on the number of credits per subject.

In Fig. 8), we present the block diagram of the drop-out prediction tool. As observed, the tool generates a graduation probability model by considering the variables collected at the *df_dropoutpred* dataframe. This model is based on the Logit-linear mixed effects approach, where variables are linearly combined to generate the logit of the graduation probability. Besides, a random term is also included to address differences between students belonging to different degrees studies. The model obtains the optimal weights b_i, indicating each of them the contribution to its associated variable to graduation probability (e.g., a positive weight for "Admission Score" means that this variable contributes to increase the probability of graduation). Further technical details can be found in Barbu et al. (2017).

Besides the information in terms of graduation probability provided by the tool, the weights b_i generated by the model can be used to search for patterns of drop-out students. As commented above, the weights indicate the contribution to graduation probability of the associated variables. By keeping the same example of the Admission Score variable, to have a positive weight means that students with low scores will potentially present an early drop-out. In summary, by analyzing the different weights of the model one can identify the effects of both categorical and performance variables and, by doing so, identify students' profiles.

It is worth noting that this tool requires information about the status of the students (Graduated, Drop-out or In Progress). This information is not directly available at all the institutions of this project. Indeed, only UAB and POLIMI have been able to collect this information and process some results. For this reason, drop-out analysis have not been addressed at Chapter 4 but, in order to provide some insights, the main patterns observed at both POLIMI and UAB are summarized below:

- Access Age (Negative Impact): Graduated Students tend to be younger.
- Admission Score (Positive Impact): Graduated Students tend to have higher scores.
- Weigh Scores Sem1 (Positive Impact) and ECTS Obtained Sem1 (Positive Impact): the average performance on Semester 1 has a big impact on Graduation/Drop-out.
- The rest of variables do not show a remarkable impact on the model.

7. CASE STUDY

Each of the partner institution of the SPEET project applied the IT Tools implemented in the project with their own set of data. Therefore collecting real data of students from their organisation information services. In what follows the obtained results for one of the institutions are presented in order to exemplify the performance monitoring that the tool provides in a real case. The analysis performed by using the described tools on the data was applied to a series of engineering degrees from the partner universities. Because of space constraints, just results of three of the degrees are showed here.:

- Aerospace Engineering (2847 students)
- Chemical Engineering (1623 students)
- Computer Science Engineering (5213 students)

This analysis covers all careers that started between Academic Year (A.Y.) 2010/2011 and A.Y. 2015/2016. On average, the accessed degrees have a high number of students: this allows the tool to identify some significant patterns. Figures (9,10, 11) show the outputs generated by the IT tool regarding the performance clusters and average score of students for the previous degrees as well as the explanatory terms for such clusters.

8. CONCLUSIONS

This paper has presented the developments achieved within the SPEET project in the elaboration of software tools for the analysis of academic data. Specific algorithms developed to deal with the basic problems tackled in the project: classification, clustering and drop-out Prediction have been presented. These results are intended for qualified users with knowledge on programming and statistics. Therefore we put at their disposition the building blocks for performing direct data analysis or even generate their own IT tools.

ACKNOWLEDGEMENTS

Co-funded by the Erasmus+ Programme of the European Union. The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

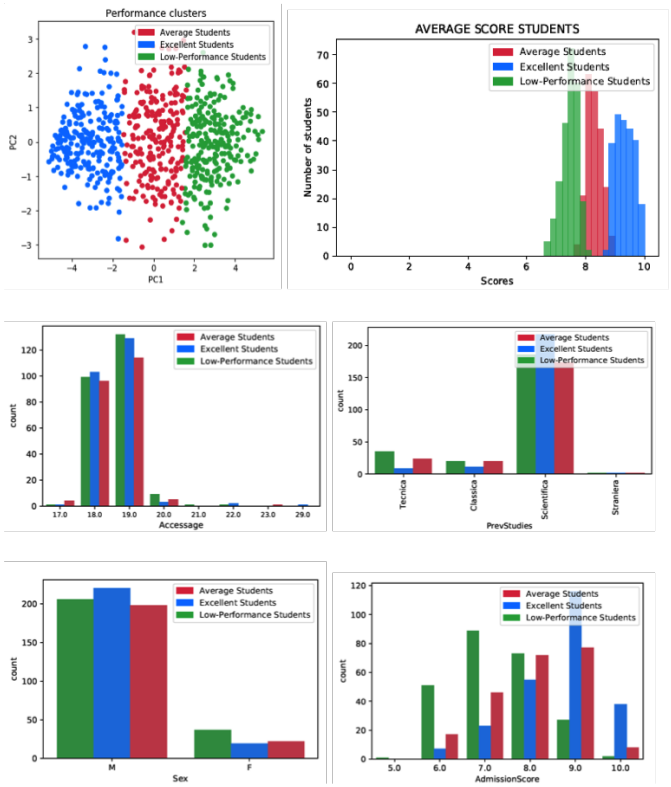


Fig. 9. Performance clusters and Average Score of students for Aerospace Engineering.

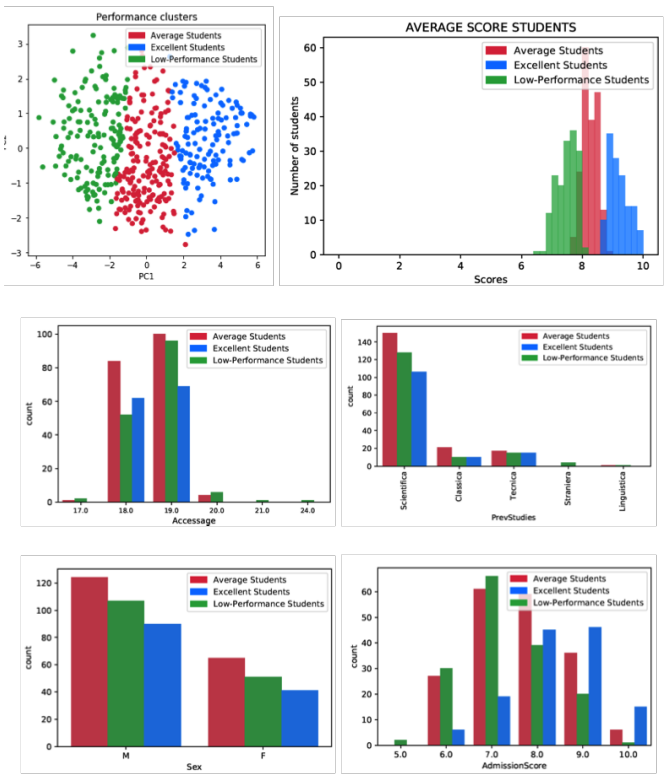


Fig. 10. Performance clusters and Average Score of students Chemical Engineering.

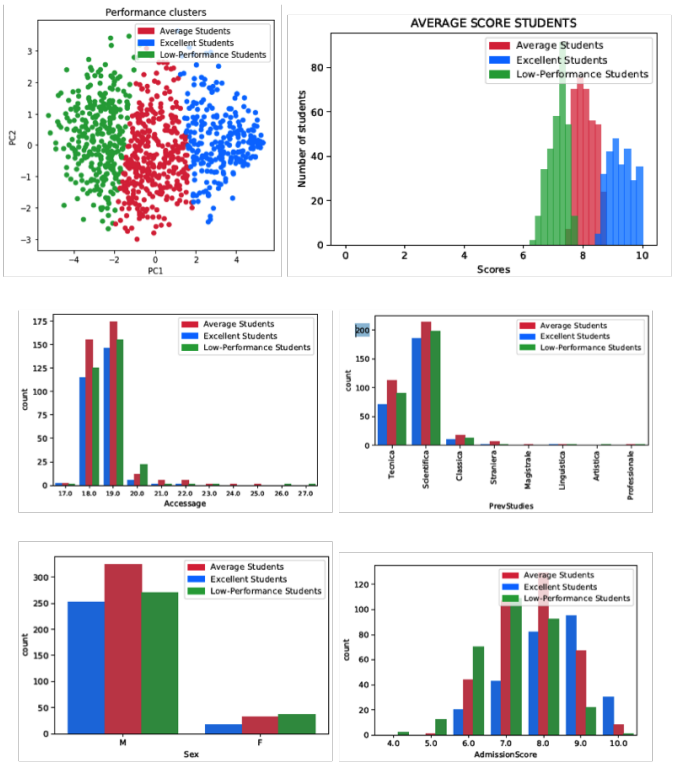


Fig. 11. Performance clusters and Average Score of students Computer Science Engineering.

REFERENCES

A. A. Akinrefon, O.S.B. (2014). Use os shewart control chart technique in monitoring student performance. *Bulgarian Journal of Science and Education Policy (BJSEP)*, 8, 311–324.

Bailey, R. (2009). Undergraduate students? perceptions of the role and utility of written assessment feedback. *Journal of Learning Development in Higher Education*, 1.

Barbu, M., Vilanova, R., Vicario, J.L., Varanda, M., Alves, P., Podpora, M., Prada, M., Morán, A., Torrebruno, A., Marin, S., and Tocu, R. (2017). Data mining tool for academic data exploitation. literature review and first architecture proposal. Technical report, ERASMUS + KA2 / KA203 SPEET Project.

Burke, D. (2009). Strategies for using feedback students bring to higher education. *Assessment & Evaluation in Higher Education*, 34(1), 41–50.

Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education*, 2, 219–233.

Prada, M., Domínguez, M., Morán, A., Vilanova, R., Vicario, J.L., Varanda, M., Alves, P., Podpora, M., Barbu, M., Torrebruno, A., Spagnolini, U., and Paganoni, A. (2018). Data mining tool for academic data exploitation. graphical data analysis and visualization. Technical Report IO3, ERASMUS + KA2 / KA203 SPEET Project.

Vicario, J.L., Vilanova, R., Bazzarelli, M., Paganoni, A., Spagnolini, U., Torrebruno, A., Prada, M., Morán, A., Domínguez, M., Varanda, M., Alves, P., Podpora, M., and Barbu, M. (2018). Io2 - data mining tool for academic data exploitation. Technical report, ERASMUS + KA2 / KA203 SPEET Project.