CHARACTERIZATION OF ENGINEERING STUDENT PROFILES AT EUROPEAN INSTITUTIONS BY USING SPEET IT-TOOL

R. Vilanova[1], J. Vicario[1],M. A. Prada[2] ,M. Barbu[3],M. Dominguez[3],
M. J. Varanda[4],M. Podpora[5],U. Spagnolini[6], P. Alves[4], A. Paganoni[6]
[1]Universitat Autonoma Barcelona (SPAIN)
[2]Universidad de León (SPAIN)
[3]University Dunarea de Jos, Galati (ROMANIA)
[4]Instituto Politecnico de Bragança (PORTUGAL)
[5]Opole University of Technology (POLAND)
[6]Politecnico di Milano (ITALY)

Abstract

The international ERASMUS+ project SPEET (Student Profile for Enhancing Engineering Tutoring) aims at opening a new perspective to university tutoring systems. Before looking for its nature, it's recommended to have a look on the current use of data in education and on the concept of academic analytics basically defined as the process of evaluating and analysing data received from university systems for reporting and decision making reasons. The provided tools are freely available to anyone that has academic data to explore. The paper will present the architecture that is behind the presented IT tool, input data needed to operate and main functionalities as well as examples of use to show how academic data can be interpreted.

Keywords: International projects, International Cooperation, Educational Data Mining

1 INTRODUCTION

For the last 20 years, statistical analysis in education is a growing area that aims to offer high quality education that produces well-educated, skilled, mannered students according to needs and requirements of the dynamically growing market. The use of statistical analysis in education has grown in recent years for four primary reasons: a substantial increase in data quantity, improved data formats, advances in computing and increased development of tools available for analytics.

Higher education institutions are not an exception and the use of analytics in education has grown in recent years for four primary reasons [1]. The available academic data can be collected, linked together and analysed to provide insights into student behaviours and identify patterns to potentially predict future outcomes. In this paper available data will be described as well as its potential use for the benefit of academic managers. The use of academic data for supporting tutoring action is where we will put the focus on.

In recent years, the sophistication and ease of use of tools for data analytics make it possible for an increasing range of researchers to apply data mining methodology without needing extensive experience in computer programming. Many of these tools are adapted from the statistical data analysis for massive datafield. Higher education institutions have always operated in an information-rich landscape, generating and collecting vast amounts of data each day. A coarse classification of the types of data that higher education institutions deal with every day is: Student record data, Staff data, Admissions and applications data, Financial data, Alumni data, Course data, Facilities data, etc.

Although the SPEET project goal is very clear (i.e. determine and categorize different profiles for engineering students across Europe), the approach to achieve student profiles in such a situation raises several questions and problems arising from the difficulty of the challenge assumed by the project partners, namely

- the official data reported by universities are quantitative/numerical. The social context of the student is not investigated because of the fact that it is related with the education level of the environment he lives with, health habits and financial support.
- the phenomenon of dropout from university studies has multiple causes which can be grouped at least into two major categories of factors: internal factors related to the student's personality and her/his level of bio-psycho-social development and external factors related to the socioeconomic, cultural and educational environment in which the student lives.

However, the official data reported by universities about students are enough to 1) identify different patterns of students in terms of their performance and 2) detect students with educational risk of dropout. This information is precious to raise the attention of educators, teachers and management levels of the university to initiate some tutorial actions, counselling and failure avoidance. Tutoring and counselling will later complete the student profile by obtaining qualitative data about the student with dropout risk. Namely, for example, information generated by tools such as questionnaire, interview, checklist, structured essay, etc. The data collected duly analysed and classified will enable a personalization of the profile and identification of other causes of socio-emotional and attitude-behavioural nature not found in official data statistically reported by universities

This work reflects the outputs of the SPEET project in relation to the data mining tools, 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.

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.

2 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 behaviour 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.

In this paper, the first steps conducted within the SPEET project are presented. It describes the conceptualization of a practical tool for the application of EDM/LA (Educational Data Mining / Learning Analytics) techniques [1],[2],[3] to currently available academic data. The paper is also intended to contextualize the use of Big Data within the academic sector, with special emphasis on the role that student profiles and student clustering do have in supporting all tutoring actions. Finally, the proposal of the key elements that conform a software application that is intended to give support to this academic data analysis is presented. Three different key elements are presented: data, algorithms and application architecture.

In order to stay up-to-date about the project, the website http://www.speet-project.eu can be accessed.

3 DATA SET

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). A unified dataset format has been considered for the project as described in [4]. 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.

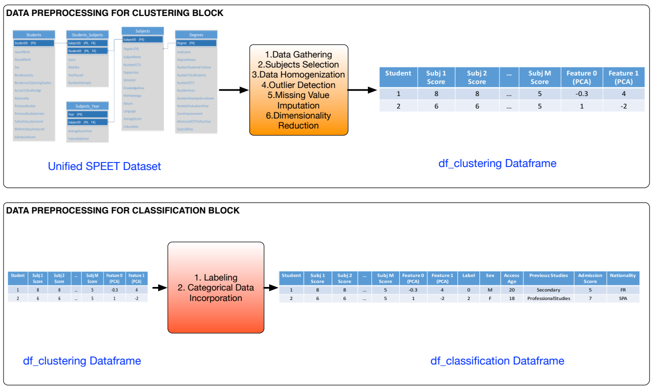


Figure 1: Preprocessing steps to obtain dataframes used by the Clustering Block (df_clustering dataframe) and the Classification Block (df_classification dataframe).

4 OVERVIEW OF THE STUDENT ANALISYS TOOL

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 [6], 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 tackled in 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. Here, the axes are the principal components with respect to where the clusters are projected. The decision of just considering three clusters was made for simplicity reasons. Main motivation was to identify students that may need of some extra tutoring action. Too much extra granularity may not help too much into this respect.

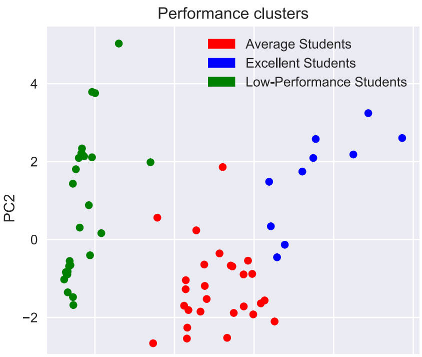


Figure 2: Performance clusters of students.

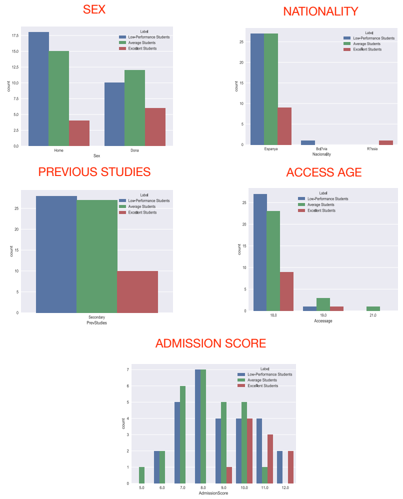


Figure 3: Clustering Explanation based on Histogram analysis of Categorical variables.

Once the Clusters are generated, Clustering Explanation is performed by analyzing 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 analyzing 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.

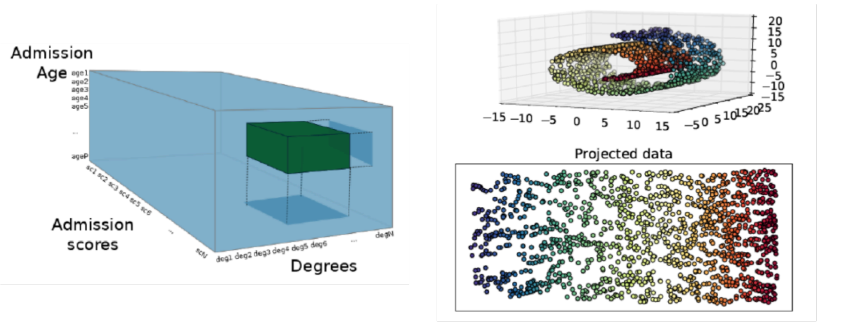
In Fig. 4, we provide an example. The first row is related to the accuracy obtained classifying new students when only the performance at the first course is considered, the second row refers to the case where first plus second course performance is considered and so on. In the example provided, it is 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 %

Figure 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 [5], 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.



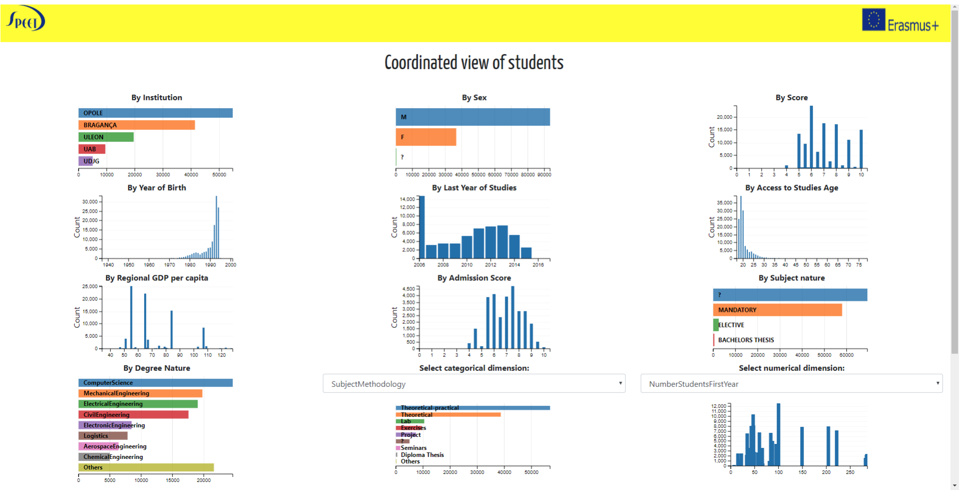
(a) Data cube perspective for the coordinated view tool (b) High dimensional data and its projection to enable visualization

Figure 5: Data interpretation for both visualization tools.

As presented in [4], 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 multidimensional 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 analyse the distribution of any variable, subject to certain filters on the others. But when the histograms or bar charts of the variables are visualized 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.

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



Coordinated view of students

Select categorical dimension: SubjectMethodology

Select numerical dimension: NumberStudentsFirstYear

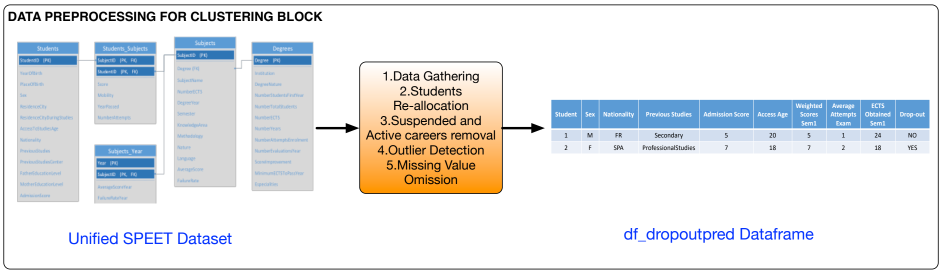
Figure 6: "Coordinated view" tool.

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 [4]. Departing from the SPEET's unified dataset format, some additional pre-processing actions are performed here. Besides the categorical variables 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_Obtained_Sem1), 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 e_ects 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 bi, 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 [4].



Student	Sex	Nationality	Previous Studies	Admission Score	Access Age	Weighted Scores Sem1	Average Attempts Exam	ECTS Obtained Sem1	Drop-out
1	M	FR	Secondary	5	20	5	1	24	NO
2	F	SPA	ProfessionalStudies	7	18	7	2	18	YES

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

Besides the information in terms of graduation probability provided by the tool, the weights bi 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 analysing the different weights of the model one can identify the effects of both categorical and performance variables and, by doing so, identify students' profiles.

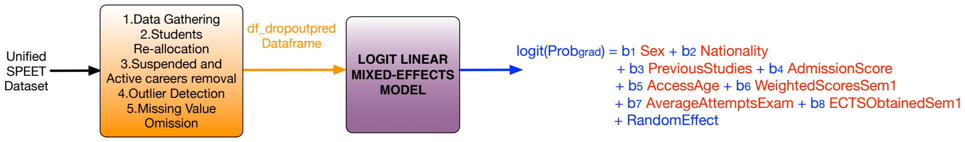


Figure 8: Block diagram of the drop-out prediction tool.

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 two of them have been able to collect this information and process some results. For this reason, a full drop-out analysis have not been addressed but, in order to provide some insights, the main patterns observed at both institutions 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 STUDIES

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 GENERAL REMARKS

In this section we are trying to draw some conclusions regarding the engineering students profiles in the different countries of partner organizations. For this reason each of the partners answered to a set a questions, the resulted conclusions being presented below.

- *Could we separate students at different groups (clusters) based on their performance behaviour?* In all studied cases, it has been reported that for each degree is possible to identify three clusters based on the average score. Usually these clusters are clearly separated. In some cases the Low-performance and Average clusters can present some overlapping. A possible explanation is that Low-performance students can have similar performance than Average students in a set of subjects. This is shown in most cases also from the score analysis at clusters where Average Score Students presents a clear separation with few overlaps, compared with Average Score Subjects where some overlap clusters trends can be observed.
- *Could we observe clear students' profiles at these groups based on categorical variables such as age, admission score, sex, previous studies?* Following we will present the conclusions regarding each one of these categorical variable.
    - *Age*: we have two cases. For the degrees were almost all students are 18/19, no clear pattern can be observed. If the number of older students allows some patterns to be observed, Excellent students tend to be younger.
    - *Admission score*: we have a clear pattern: the higher the admission score, the higher the obtained performance.
    - *Sex*: usually the number of women enrolled in engineering degrees is low. Nevertheless the proportion of Excellent students tend to be higher for women for most of the partners,
    - *Previous Studies*: we have a very clear pattern: the best students come from secondary school.
- *The quality of cluster separation (clearly or badly separated clusters) can be explained by means of the way categorical variables (age, admission score, sex, previous studies) are distributed (homogeneous vs. heterogeneous students' profiles)?* In most cases we have observed that homogeneous students' patterns offer good Clustering behavior. In some cases it was observed that Low- performance and Average show similar performance where some students are better in a set of subjects than the other cluster, and vice versa. An- other observation is that if the separation of clusters is not clear, it would be more appropriate to consider only two groups (low and high-performance).
- *Could we see if one or several courses determine the behavior of students at one degree?* Based on the obtained results we can conclude that in most cases there are subjects in a specific year that have a strong influence in student performance.
- *Could we formulate any hypothesis about the relationship between explanatory variables and performance through histogram filtering?* It is possible to compare the student score with other categorical variables draw conclusions for each degree (e.g. better students are younger, with a low access age and from secondary school).
- *Does any score distribution grouped by an explanatory variable show an evident trend?* Yes, it is possible to relate the score with the Admission Score.

9 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. So, finally we can conclude that the tools developed in this project can offer some significant information in detecting different profiles and the relationship between these profiles and categorical variables such as age, admission score, sex, previous studies.

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

REFERENCES

[1] G. Siemens and R.S. Baker. Learning analytics and educational data mining: towards communication and collaboration. In Proceedings of the 2nd international conference on learning analytics and knowledge, pages 252–254, 2012.

[2] O. Scheuer and B.M. McLaren. Encyclopedia of the Sciences of Learning, chapter Educational data mining, pages 1075–1079. Springer, 2012.

[3] C. Romero and S. Ventura. Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3:12–27, 2013.

[4] Barbu, M., Vilanova, R., Vicario, J.L., Varanda, M., Alves, P., Podpora, M., Prada, M., Moran, 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.

[5] Prada, M., Domínguez, M., Moran, 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.

[6] Vicario, J.L., Vilanova, R., Bazzarelli, M., Paganoni, A., Spagnolini, U., Torrebruno, A., Prada, M., Moran, 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.

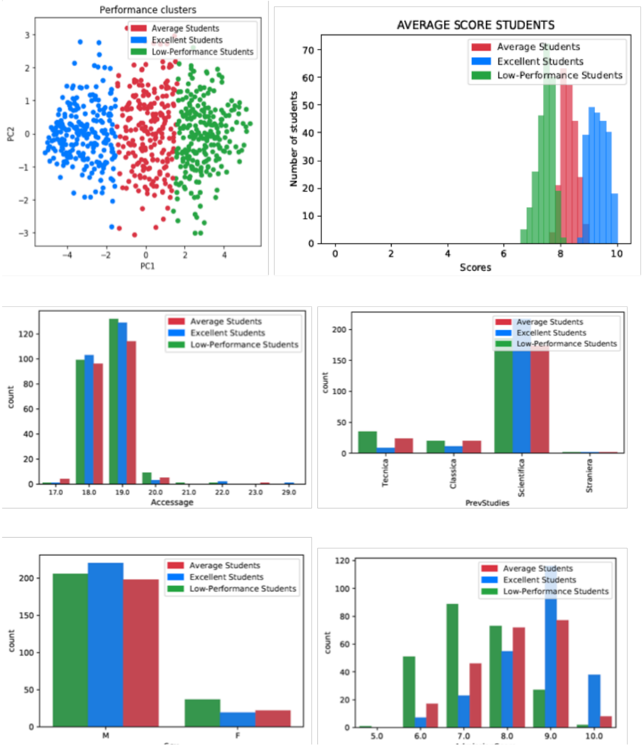


Figure 9: Performance clusters and Average Score of students for Aerospace Engineering.

ç

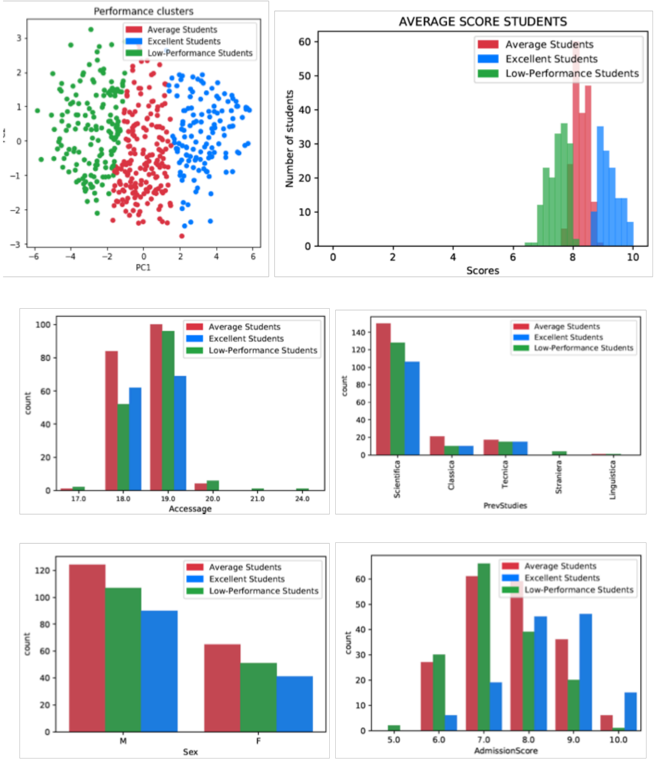


Figure 10: Performance clusters and Average Score of students Chemical Engineering..

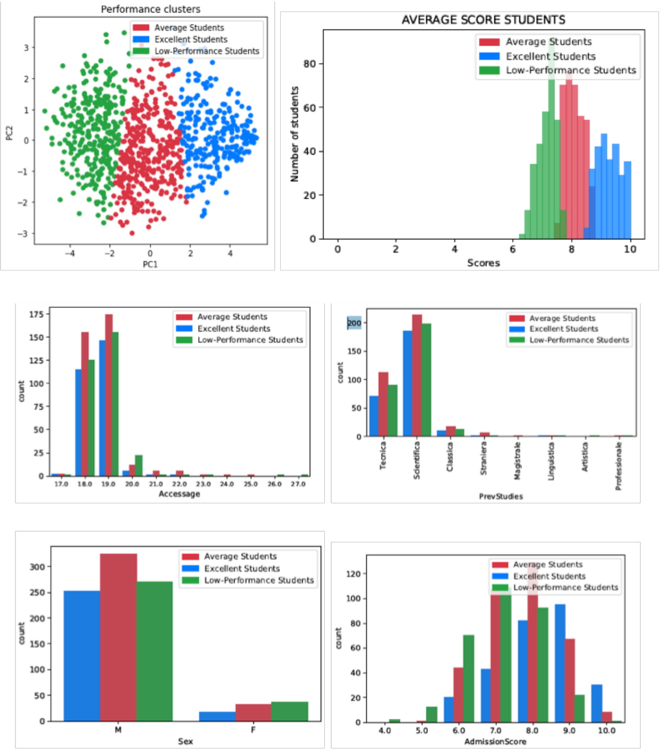


Figure 11: Performance clusters and Average Score of students Computer Science Engineering.