Co-funded by the
Erasmus+ Programme
of the European Union

INTELLECTUAL OUTPUT #4

Student Profile for Enhancing Tutoring Engineering

Student Profile
for Enhancing
Engineering Tutoring

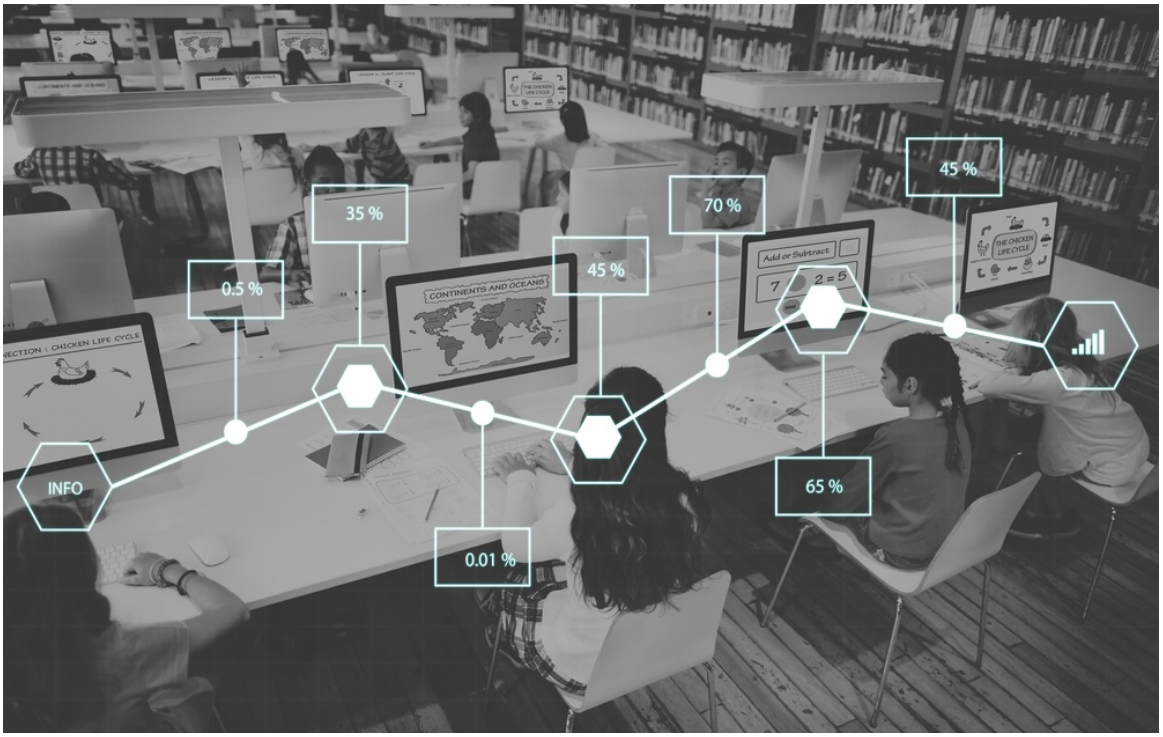
ERASMUS + KA2 / KA203

Data Mining Tool for Academic Data Exploitation

Publication Report on Engineering Students Profiles

M. Barbu (Coordinator), R. Vilanova, J. Lopez Vicario, M.J. Varanda, P. Alves, M. Podpora, A. Kawala-Janik, M.A. Prada, M. Dominguez and U. Spagnolini, L. Fontana

February 2019



ISBN 978-989-20-9286-7

Data Mining Tool for Academic Data Exploitation

Publication Report on Engineering Students Profiles

M. Barbu (Coordinator)

Automatic Control and Electrical Engineering Department
"Dunarea de Jos" University of Galati
Domneasca 47, 800008
Galati, Romania

R. Vilanova, J. Lopez Vicario

Dept. de Telecomunicacio i Enginyeria de Sistemes
Escola d'Enginyeria, UAB
Carrer de es Sitges 08193 Bellaterra
Barcelona, Spain

M.J. Varanda, P. Alves

Escola Superior de Tecnologia e Gestao
Instituto Politecnico de Bragança
Bragança, Portugal

M. Podpora, A. Kawala-Janik

Faculty of Electrical Engineering, Automatic Control and Informatics
Opole University of Technology
Opole, Poland

M.A. Prada, M. Dominguez

Dept. de Ingeniería Eléctrica y de Sistemas y Automática
Escuela de Ingenierías Industrial e Informática
Universidad de León
León, Spain

U. Spagnolini, L. Fontana

Scuole di Ingegneria
Politecnico di Milano
Milano, Italy

Final Version
Approved for public release; distribution is unlimited.

ERASMUS + KA2/KA203

Prepared for SPEET Intellectual Output #4

Under 2016-1-ES01-KA203-025452

Monitored by SEPIEE

Table of Contents

1 Executive Summary

This report summarizes the findings of the project. It rely on the initial document generated as Intellectual Output #1 [BVV+17] and the results obtained by application of the IT tools developed in Intellectual Output #2 [aRVBP+18] and Intellectual Output #3 [PDM+18] to the academic data provided by the partner institutions.

Data has always been a significant asset for institutions, and has been used to inform their day-to-day operational decisions as well as longer-term business and strategic decisions. From a more purely educational point of view, the available academic data can be collected, linked together and analyzed to provide insights into student behaviours and identify patterns to potentially predict future outcomes [BVV+17].

In case of academic institutions the main objectives of applying analytic techniques to evaluate the data sources can be categorized as follows [BVV+17]:

- Improve Student Results: The overall goal of big data within the educational system should be to improve student results. During his or her student life however, every student generates a unique data trail. This data trail can be analyzed in real-time to deliver an optimal learning environment for the student as well to gain a better understanding in the individual behaviour of the students;
- Create Mass-customized Programs: All this data will help to create a customized program for each individual student. Providing mass customization in education is a challenge, but thanks to algorithms it becomes possible to track and assess each individual student;
- Improve the Learning Experience in Real-time: Each student learns differently and the way a student learns affects the final grade of course. When the course materials are available online, it can be monitored how a student learns. This information can be used to provide a customized program to the student or provide real-time feedback to become more efficient in learning and thus improve their results;
- Reduce Dropouts, Increase Results: Using predictive analytics on all the data that is collected can give educational institute insights in future stu-

dent outcomes. These predictions can be used to change a particular program if bad results are predicted or even run scenario analysis on a program before it is started.

An important issue regarding the use of student data is Privacy and Data Protection [BVV+17]. EU data protection law has undergone a long-awaited, rigorous and comprehensive revision. After long discussions in the various committees, on 16 April 2016, the EU Parliament formally approved the General Data Protection Regulation (GDPR or Regulation) and it became effective in May 2018 in all EU member states. The GDPR was adopted by the European Commission "to strengthen online privacy rights and boost Europe's digital economy", recognizing that "technological progress and globalization have profoundly changed the way our data is collected, accessed and used" (European Commission, 2012).

In what follows in Section 2 and 3 we will the IT tools developed in Intellectual Output #2 and Intellectual Output #3, respectively. Section 4 contains the results obtained by applying these software tools to the academic data from each of the partners. Finally, some conclusions are draw regarding the engineering students profiles in the different countries of partner organizations and the ability of the developed software tools to capture the particularities existent in each country/institution.

2 Overview on the IT Tool for student data processing

In this Chapter, we present an overview of the data processing tools resulting from Intellectual Output # 2, which have been considered for the identification for students' profiles. As presented in [aRVBP+18], 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.

2.1 Clustering and Classification tool

This tool is in charge of generating three clusters of students based on their performance results (Clustering) and, also, to derive a classification mechanism able to classify new students to the clusters generated (Classification). This tool also provides an analysis of students belonging to different clusters in terms of histogram-based representation of categorical information (Clustering Explanation), which is used to obtain student's patters.

2.1.1 Data format

As presented in the Intellectual Output # 1 document [BVV+17], a unified dataset format has been considered for the project. 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).

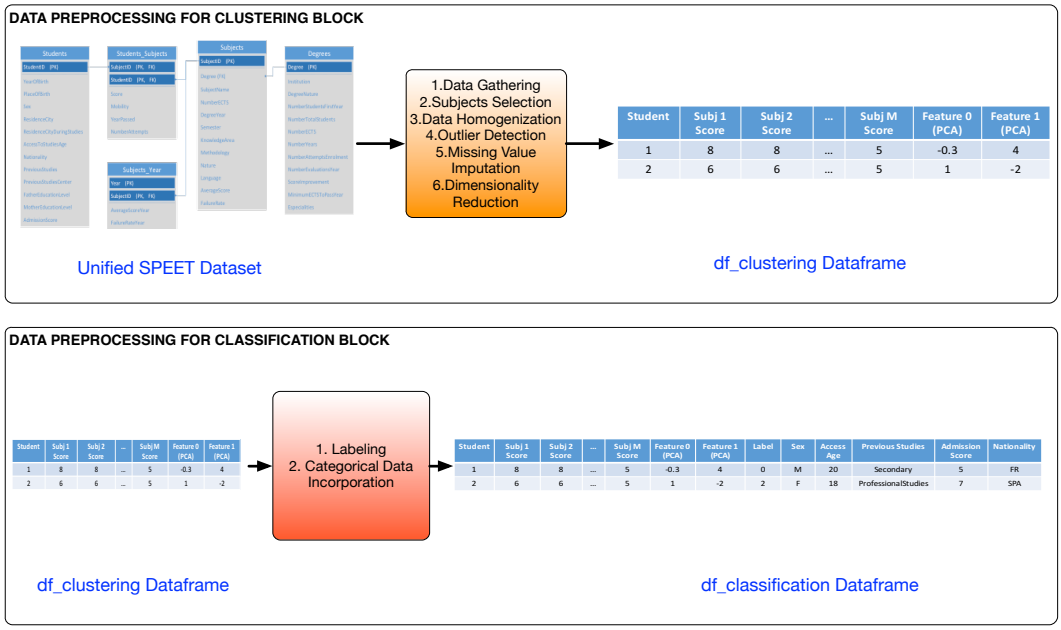


Figure 1. Preprocessing steps to obtain dataframes used by the Clustering Block ($df\_clustering$ dataframe) and the Classification Block ($df\_classification$ dataframe).

Data frame $df\_classification$ is also adopted to perform the histogram-based Clustering Explanation.

2.1.2 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 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.

2.1.3 Classification

Finally, 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

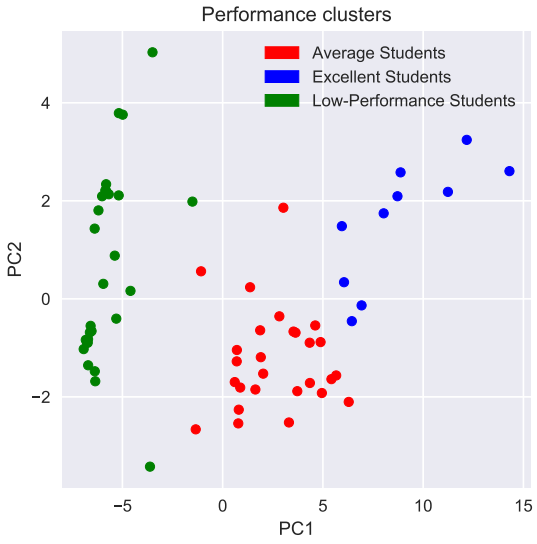


Figure 2. Performance clusters of students.

is not adopted to obtain students' patterns. Its purpose here is to extract degrees' patterns. This can be done by analyzing 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.

2.2 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.

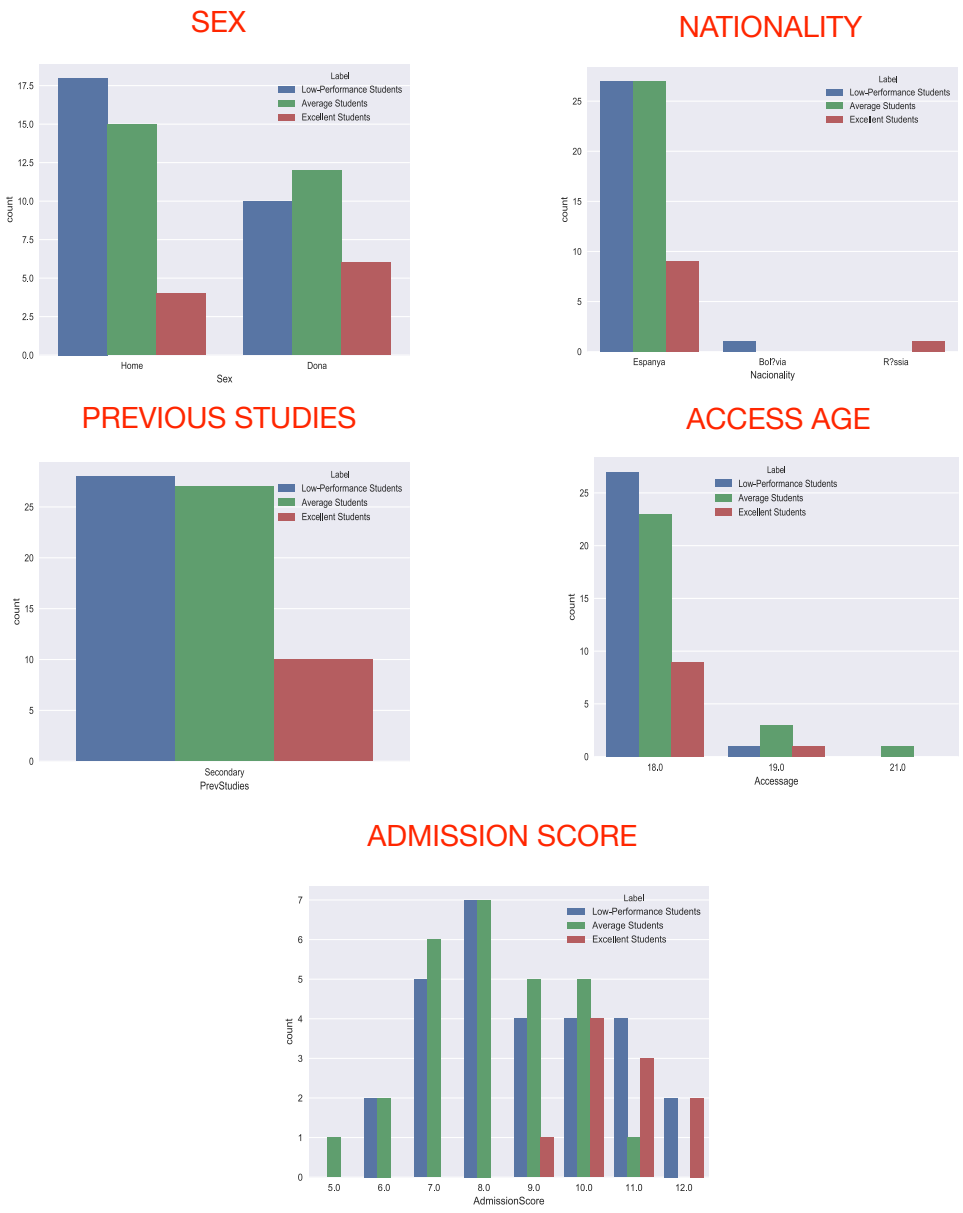


Figure 3. Clustering Explanation based on Histogram analysis of Categorical variables.

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

Figure 4. Degree Analysis based on Classification Accuracy results.

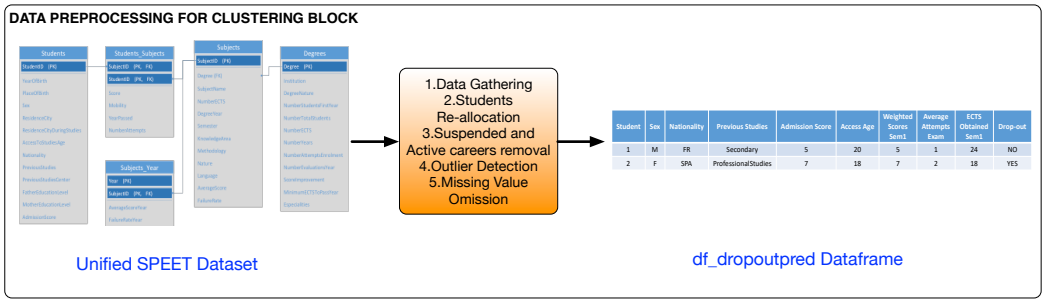


Figure 5. Preprocessing steps to obtain dataframe used by the Drop-out Prediction tool (*df _dropoutpred* dataframe).

2.2.1 Data format

Details about the data format at Drop-out Prediction tool are also presented in the Intellectual Output # 1 document [BVV+17]. 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. 5). 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 (Weigh Scores Sem1, where weighting is based on the number of credits per subject).

2.2.2 Graduation Probability

In Fig. 6), 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 the Intellectual Output # 1 document [BVV+17].

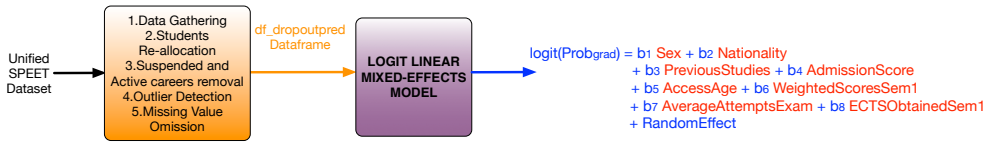


Figure 6. Block diagram of the drop-out prediction tool.

2.2.3 Pattern identification and obtained results

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.

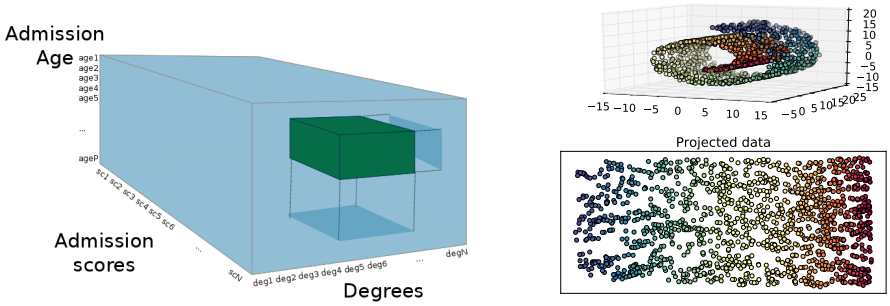
3 Overview on the IT Tool for Graphical Data Analysis and Visualization

In this Chapter, we present an overview of the data visualization tools resulting from Intellectual Output #3, which have been conceived for the support of the exploratory analysis conducted by tutoring staff. As presented in [PDM+18], 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.

Two types of visualization tools have been implemented in this project:

- 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.
- Data projection tool: This tool provides a 2D scatterplot of the high-dimensional students' data, obtained by means of dimensionality reduction. It also takes advantage of the graphical properties of the points to convey additional information and its parameters can be interactively adjusted. This tool has been applied to two cases: one where data have been organized by year and another where data are grouped by degree.

As presented in the Intellectual Output # 1 report [BVV+17], a unified data set format has been considered to be used in the tools. Some pre-processing tasks are performed to accommodate this data set to the visualization 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. 7.



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

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

Figure 7. Data interpretation for both visualization tools

3.1 Coordinated view tool

3.1.1 Method

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 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.

3.1.2 Implementation

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

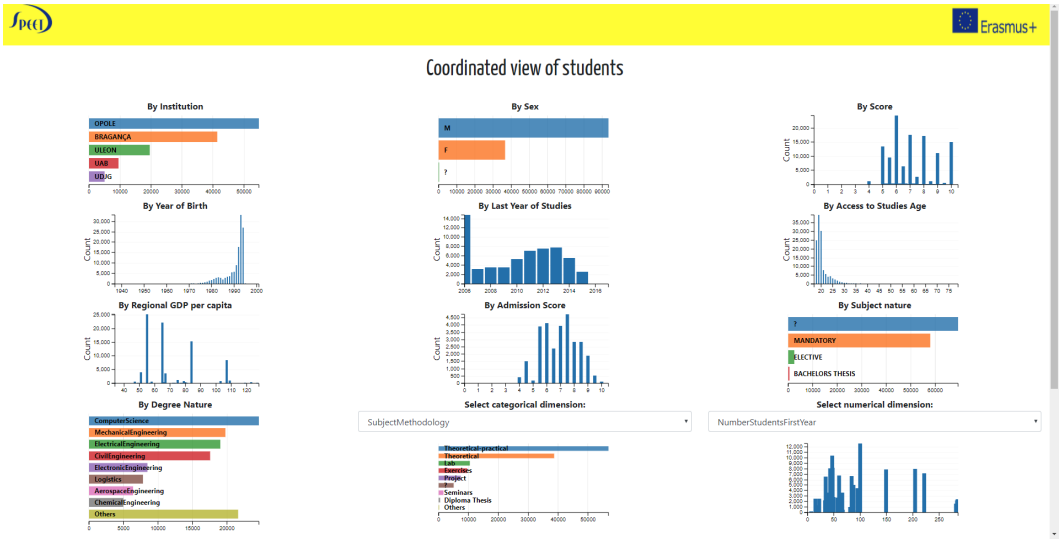


Figure 8. "Coordinated view" tool.

another explanatory variable and a choropleth map are included. In Fig. 8, a screenshot of this tool is provided.

3.2 Data projection tool

3.2.1 Method

A geometrical interpretation of data, where the values of each variable are understood as the coordinates of a high-dimensional space, is the starting point of many machine learning methods. However, the visualization of student data is not directly possible because each student will be represented by a point with a dimensionality much higher than 3. For that reason, it is necessary to use a transformation known as dimensionality reduction, which aims at representing high-dimensional data in low-dimensional spaces while preserving most of its structure.

Dimensionality reduction is performed by several different approaches. Among them, manifold learning algorithms are a class of techniques that perform non-linear projections of data onto a low-dimensional space by preserving distances or divergences. A manifold learning algorithm that is known to provide good visualization results in real data is the t-SNE (t-Distributed Stochastic Neighbor Embedding) [MH08], which aims to find the data projection that minimizes the mismatch between the probabilities computed from the pairwise high-dimensional and low-dimensional distances. This technique will be used in the data projection tools.

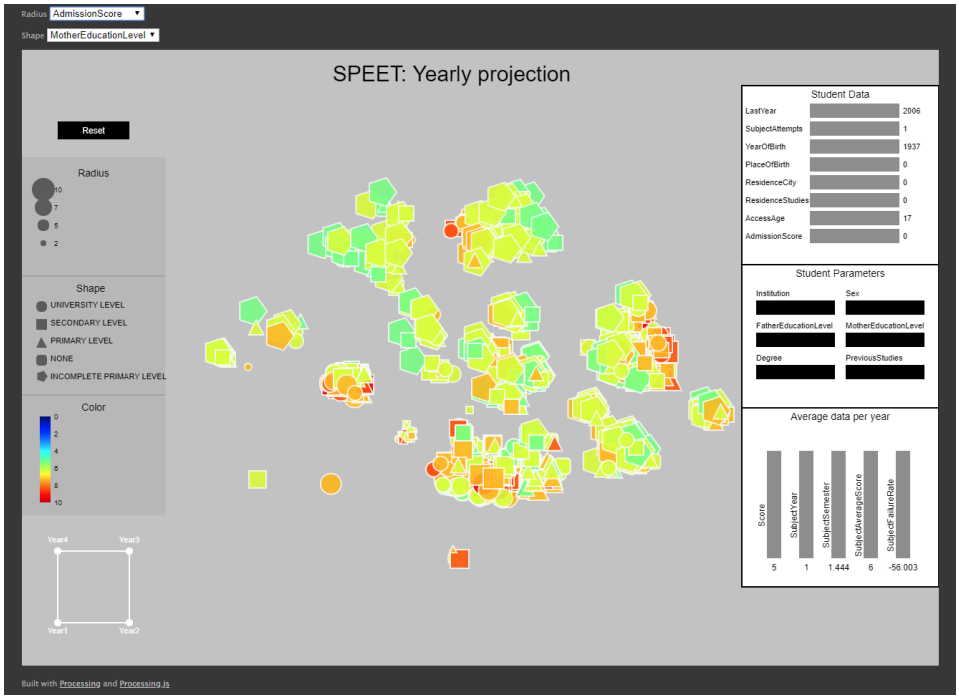


Figure 9. Screenshot of the "yearly projection" tool.

The low-dimensional projection obtained through this approach can be visualized as a two-dimensional scatterplot where the relative distances between points are interpretable, assuming that closeness in the representation can be assimilated to high similarity in the original space. The analysis of these scatterplots, especially when the visualization takes advantage of interactivity and additional visual information, might be useful to better understand the data structure.

Two applications of this approach have been considered for the problem at hand:

- The projection of a common set of students, represented by their descriptive variables and the average score for each academic year, in order to understand common characteristics in institutions.
- The projection of several data sets of students (one for each degree), represented by their descriptive variables and the scores of all the subjects, allowing potentially missing data.

3.2.2 Implementation

Each one of the visualizations described in the previous section has been implemented as an interactive web application. In both cases, the applica-

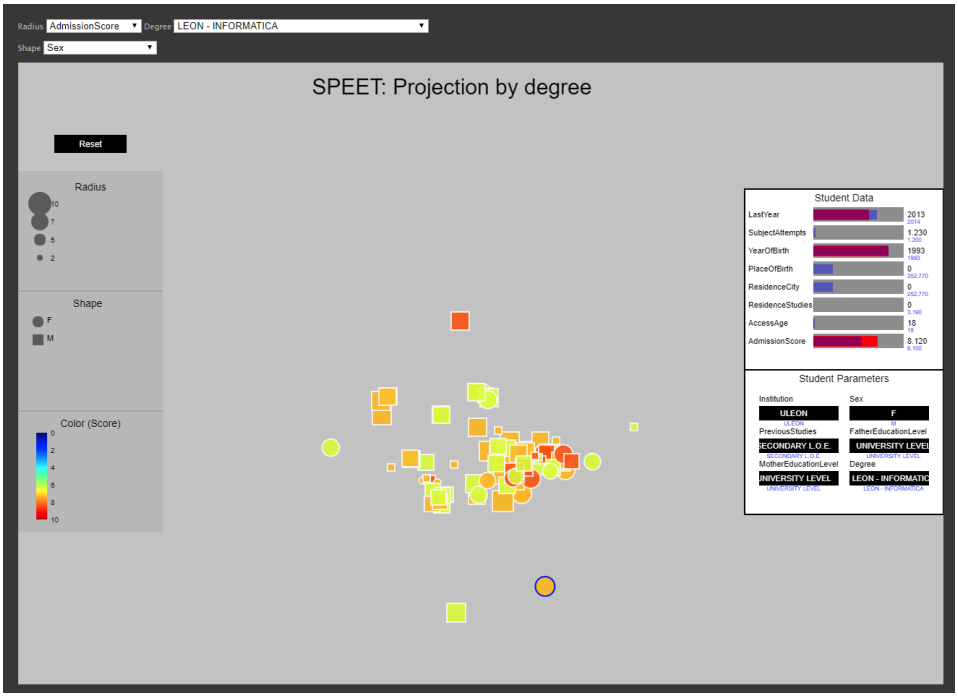


Figure 10. Screenshot of the "projection by degree" tool.

tion presents a complete dashboard that displays, in the central panel, a 2D scatterplot.

The visual channels that can be associated to a point of the scatterplot (i.e., radius, shape, and color) are also used to show values from the original variables. Indeed, the variables that are linked to radius and shape can be customized by means of two dropdown menus on top of the visualization. Additionally, when a user hovers a certain point, the value of important explanatory and average performance variables is shown in a table in the right side. A certain point can also be fixed as a reference for comparison with the other ones by simply clicking on it.

In the first case, data has been organized by year. An additional square is displayed in the bottom left side to allow users to select the weight of each year in the projection. In Fig. 9, a screenshot of the tool is provided.

In the second case, a different visualization is provided for each degree. The projected data is essentially constituted by the scores of every course for each student. For this visualization, an additional menu to select the degree is provided. Fig. 10 shows an example of the tool in use.

4 Applying the Profiling IT Tools for the Data Collected from the Partner Organizations

Each of the partners applied the IT Tools implemented in the project with their own set of data. In what follows the obtained results are presented.

4.1 Universitat Autònoma de Barcelona Degrees Analysis

4.1.1 Clustering and Classification

Four degrees have been considered for Universitat Autònoma de Barcelona (UAB) case:

- UAB 951 - Chemical Engineering (65 students)
- UAB 956 - Telecommunications Systems Engineering (25 students)
- UAB 957 - Telecommunications Electronics Engineering (28 students)
- UAB 958 - Computer Engineering (197 students)

Although UAB degrees do not have a high number of students, the tool help to identify some patterns. The performed analysis is presented next, where degrees' codes will be considered for the sake of brevity (i.e., UAB 951, UAB 956, etc.).

Clusters Analysis

Three very clear clusters in terms of student's performance behavior have been observed for the cases UAB 951 and UAB 956. As presented at Fig. 11 and Fig. 12, where both performance clusters and Average Scores obtained by students are presented.

Concerning the UAB 957 and UAB 958 cases, it is observed that clusters are not very clear (see Figures 13 and 14). Low-performance and Average students are not so well separated as in the previous cases. Indeed, it seems that two clusters could be better option. By observing Average Score of students at both cases, Low-performance and Average students present some overlap. The

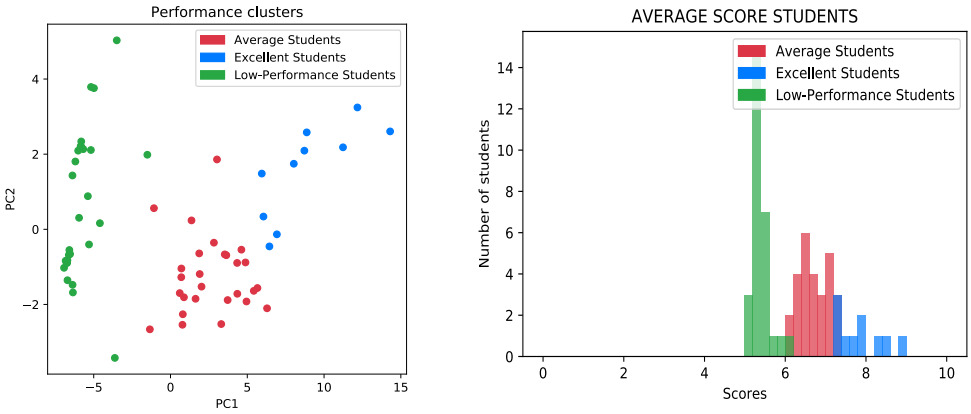


Figure 11. Performance clusters and Average Score of students (UAB 951).

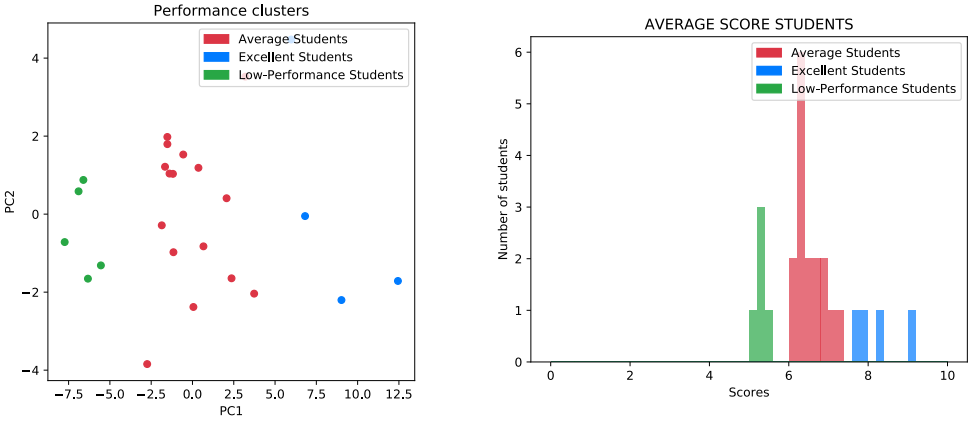


Figure 12. Performance clusters and Average Score of students (UAB 956).

possible explanation is that Low-performance students can have a similar or better performance than Average students in a set of subjects and vice versa.

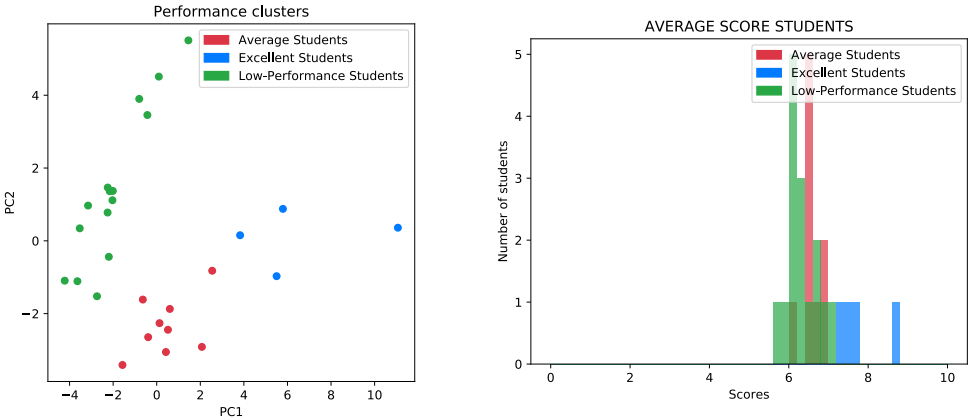


Figure 13. Performance clusters and Average Score of students (UAB 957).

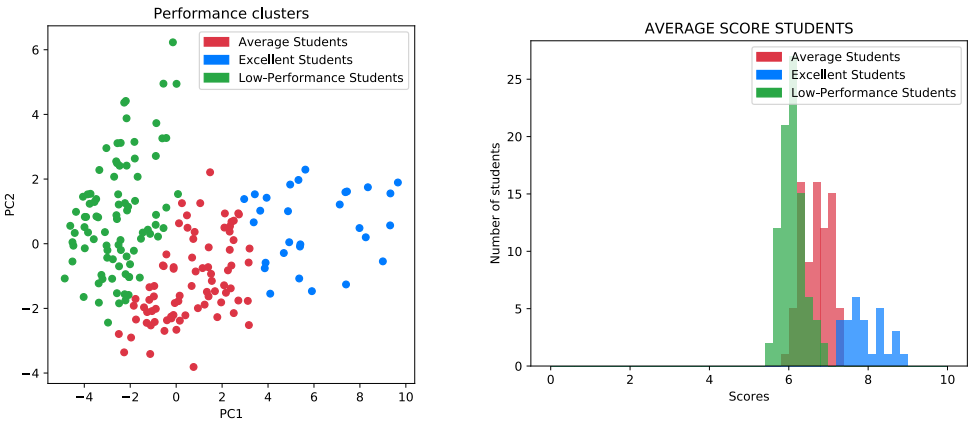


Figure 14. Performance clusters and Average Score of students (UAB 958).

Student-wise characterization

Obtained clusters have been analyzed in terms of Categorical Variables and the following trends have been observed:

- UAB 951 (see Fig. 15): In this case, very homogeneous students' patterns are found, but some conclusions can be extracted:
  - Sex: excellent students tend to be Women (Dona in Catalan).
  - Access Age: excellent students tend to be younger.
  - Admission Score: excellent students tend to have higher admission scores.
- UAB 956 (see Fig. 16): Quite homogeneous students at nationality and previous studies. Some trends:
  - Sex: very few women.
  - Access Age: most part of students tend to have 18 years.
  - Admission Score: excellent students tend to have higher scores but not clear pattern.
  - Previous studies: excellent students tend to come from secondary. All Low-performance students come from there.
- UAB 957 (see Fig. 17): It is also observed a quite homogeneous students at nationality and previous studies. Some trends:
  - Sex: very few women.
  - Access Age: excellent students tend to be younger.

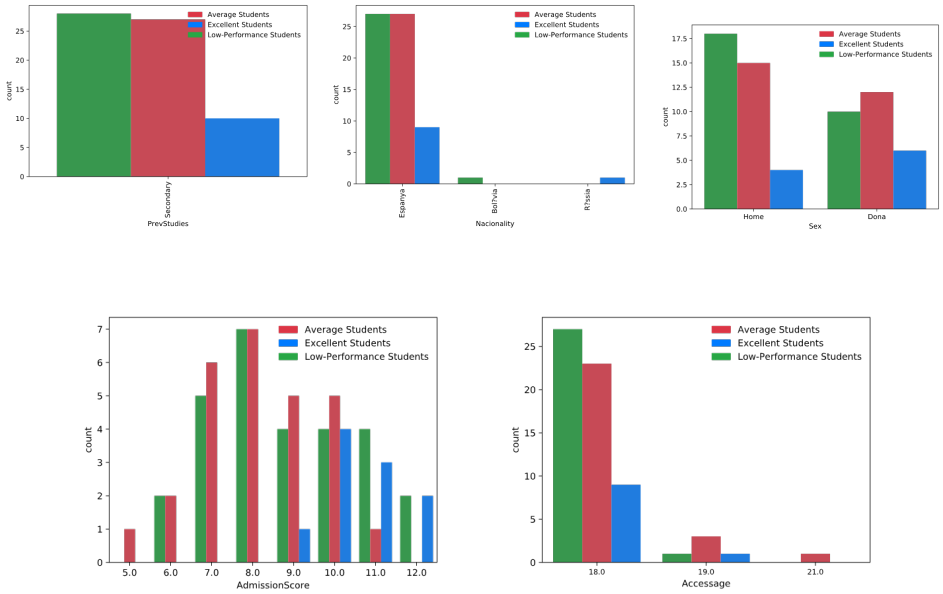


Figure 15. Categorical Variable Analysis (UAB 951).

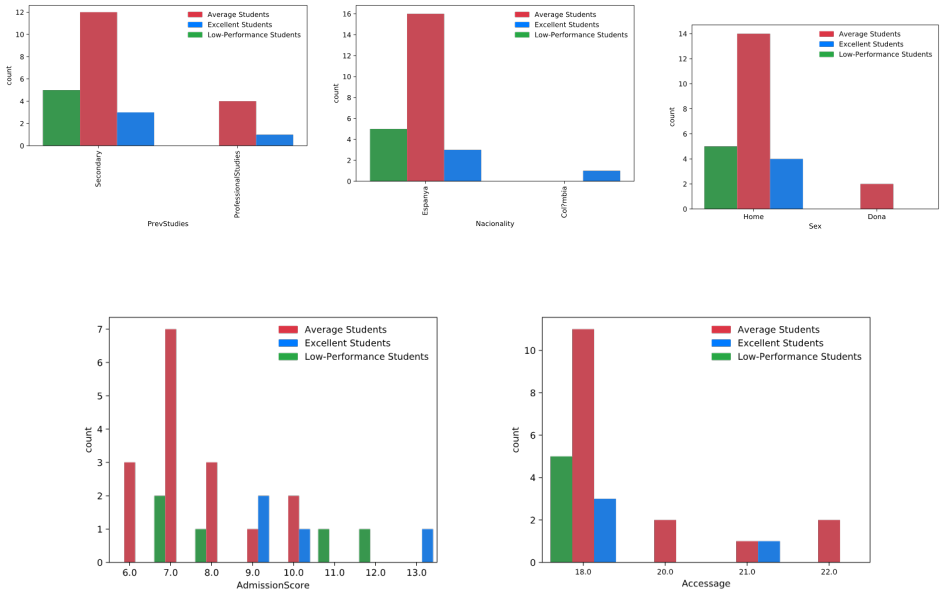


Figure 16. Categorical Variable Analysis (UAB 956).

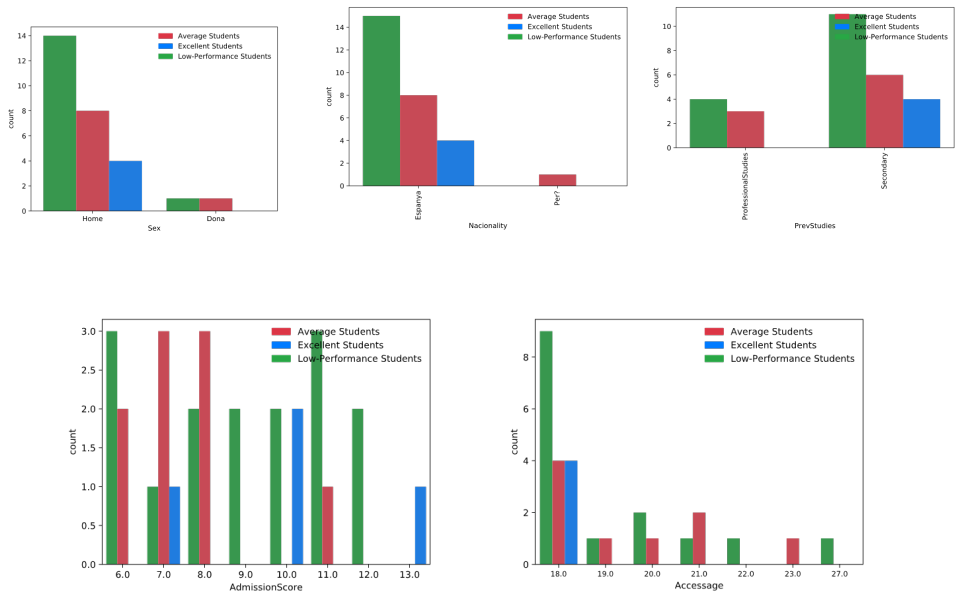


Figure 17. Categorical Variable Analysis (UAB 957).

- Admission Score: maximum score is for an excellent student but not clear pattern.
- Previous studies: excellent students come from secondary, but no clear pattern for the rest of students (come from both Secondary and Professional Studies).

• UAB 958 (see Fig. 18): in this case students are heterogeneous when analizing access age. Trends:

- Sex: very few women.
- Previous studies: excellent students tend to come from secondary.
- Access Age: excellent students tend to be younger but some older students perform quite well. Access age tend to be younger but the number of older students is not negligible.
- Admission Score: excellent students tend to have higher admission scores but the trend is not clear (Average and Low-performance obtain a high range of Admission scores).

Institution-wise characterization

At this point, if one analyse the trends observed in terms of Clustering behaviour and distribution of Categorical variables, two patterns are observed:

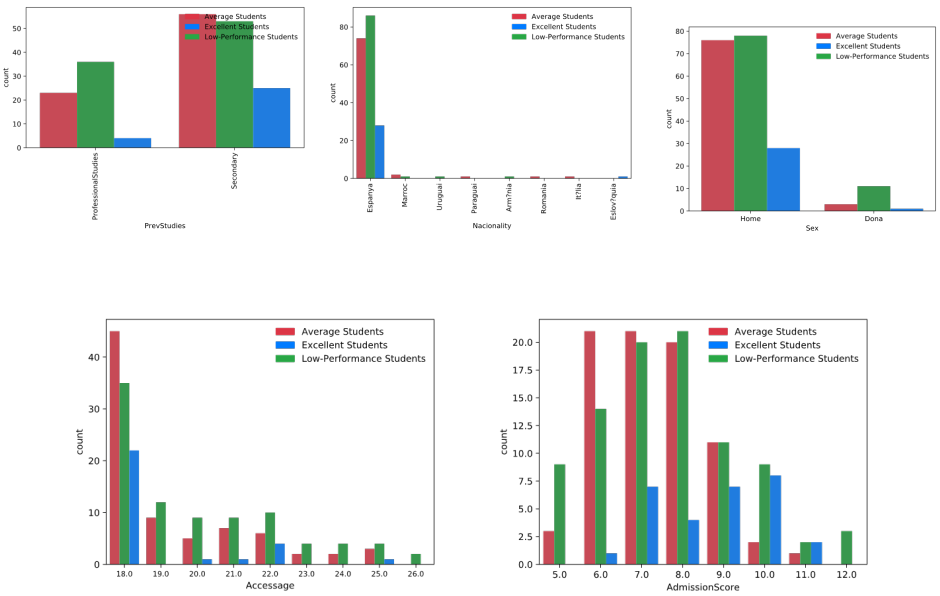


Figure 18. Categorical Variable Analysis (UAB 958).

- UAB 951 and UAB 956: Good Clustering behaviors are observed and student's patterns in terms of categorical variables are quite homogeneous.
- UAB 957 and UAB 958: Worse Clustering behaviors are present and heterogeneity is observed in terms previous studies (UAB 958) and access ages (UAB 957 and UAB 958). It was also commented that Low-performance and Average students show a similar performance. Some students are better in a set of subjects than the other cluster, and vice versa. Heterogeneity could be one of the reasons.

Besides, after analyzing the four UAB degrees, one can observe common trends in terms of Access Age, Admission Score and Previous studies:

- Access Age: excellent students tend to be younger.
- Admission Score: excellent students tend to obtain higher scores.
- Previous studies: excellent students come from secondary, but no clear pattern for the rest of students (come from both Secondary and Professional Studies).

Classification Analysis

As commented at IO2 report [aRVBP+18] (Section 3.3 in this document), classification analysis can also be adopted to analyze the course-dependency

behavior of students at the different degrees. In the case of UAB degrees, it is observed that all the cases present a behavior where the first course is very important. In other words, the classification accuracy obtained when analyzing only the subjects at the first are considerably high when compared with accuracies obtained by adding the rest of the courses information. Classification obtained with performance attained at the first course is kept along the studies. When analyzing the studies programs in detail, it is observed how most of mathematical and physics subjects fall into the first course.

4.1.2 Coordinated Views and Dimensionality Reduction

IO3 provides an efficient tool to graphically compare students' performance with a wide set of explanatory variables. When analyzing UAB degrees the following trends have been observed:

- UAB 951:
  - Higher admission scores show higher subjects performance.
  - Subject year: 1st year shows the worst results and 4st shows the best results.
  - Subject Failure Rate: higher failure rate at the scores significantly penalizes the performance of students.
  - Subject Number of Attempts tend to be 1 in this case.
- UAB 956:
  - Higher admission scores show higher subjects performance.
  - Subject year: 3st year shows the worst results and 4st shows the best results.
  - Subject Failure Rate: higher failure rate at the scores significantly penalizes the performance of students.
  - Subject Number of Attempts: it is clearly observed how students performing higher attempts has worsen scores.
  - Students coming from Professional Studies have lower admission scores.
- UAB 957:
  - Higher admission scores show higher subjects performance.
  - Subject year: 1st and 2nd years show the worst results and 4st shows the best results.

  - Subject Failure Rate: higher failure rate at the scores significantly penalizes the performance of students.
  - Subject Number of Attempts: it is clearly observed how students performing higher attempts has worsen scores.
  - Students coming from Professional Studies have lower admission scores (but these are more distributed than in UAB 956 case).

- UAB 958:
  - Higher admission scores show higher subjects performance.
  - Subject year: 1st year shows the worst results and 4st shows the best results.
  - Subject Failure Rate: higher failure rate at the scores significantly penalizes the performance of students (this is the clearest case).
  - Subject Number of Attempts: it is clearly observed how students performing higher attempts has worsen scores, but only 1 or 2 attempts are observed in this case.
  - Students coming from Professional Studies have similar admission scores than students coming from Secondary.

In the analysis observed above, one can see how 4st course subjects show the best results at all the cases. It is worth noting that the structure of this course is the same at UAB degrees: most of elective courses and degree's final project. Besides, some common patterns are:

- Higher admission scores show higher subjects performance (as observed at the Clustering/Classification IO2 tool).
- Subject Failure Rate: higher failure rate at the scores significantly penalizes the performance of students.
- Subject Number of Attempts: it is clearly observed how students performing higher attempts has worsen scores.

Finally, one can see how at UAB 958 Professional Studies have similar admission scores than Secondary students. When aligning this to the Clustering behavior previously observed, one can see how two kinds of students with different backgrounds and admission ages (i.e., Secondary vs. Professional Studies) access to the Computer Engineering degrees with similar admission

scores. This could explain why Low-performance and Average students show a similar performance and reinforce the idea that heterogeneity could be one of the reasons.

Concerning the dimensionality reduction tool, UAB degrees have a number of students that is not high enough to see clear patterns.

4.2 Politecnico de Milano Degrees Analysis

4.2.1 Clustering and Classification

In this section Politecnico de Milano (POLIMI) reports the analysis performed through the tool described in IO2 of the following six POLIMI degree engineering track:

- POLIMI 7 - Aerospace Engineering (2847 students)
- POLIMI 13 - Chemical Engineering (1623 students)
- POLIMI 37 - Electronic Engineering (1184 students)
- POLIMI 44 - Computer Science Engineering (5213 students)
- POLIMI 49 - Mechanical Engineering (5168 students)
- POLIMI 62 - Automation Engineering (1412 students)

This analysis covers all careers that started between Academic Year (A.Y.) 2010/2011 and A.Y. 2015/2016. On average, POLIMI BSc degrees have a high number of students: this allows the tool to identify some significant patterns. From now on, each degree program is referenced through its code indicated above (i.e., POLIMI 7, POLIMI 13 etc.).

Clusters Analysis

After performing the dimensionality reduction step, POLIMI clustered the observations into groups using a K-means algorithm (with K=3). In all six POLIMI degrees, the application of this algorithm is nearly equivalent to splitting the observations into three groups according to the value of the 1st Principal Component. Therefore, observations are grouped into three clusters, with

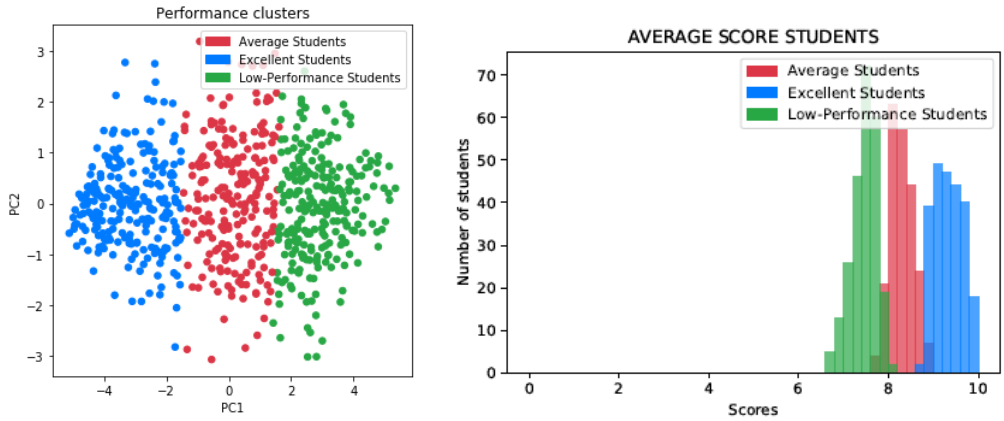


Figure 19. Performance clusters and Average Score of students (POLIMI 7).

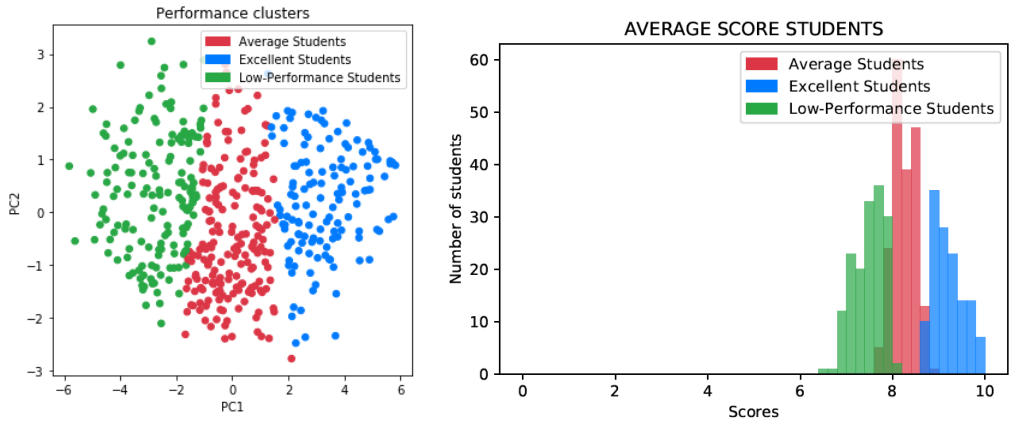


Figure 20. Performance clusters and Average Score of students (POLIMI 13).

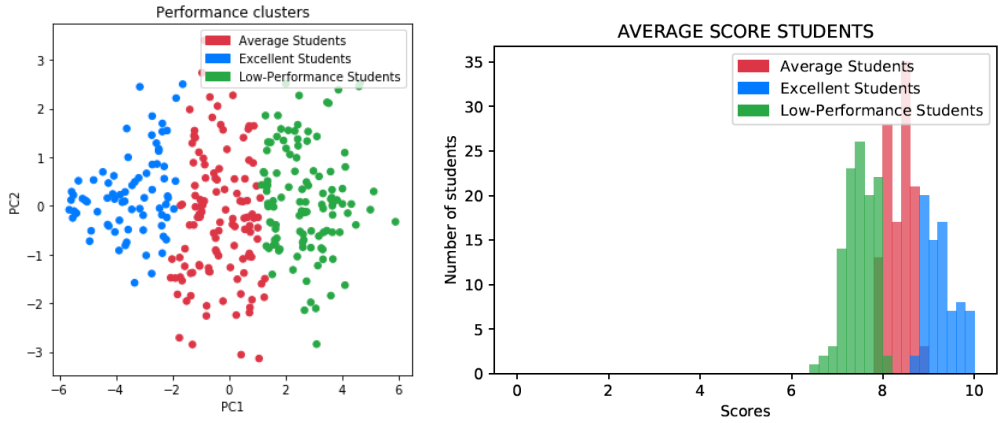


Figure 21. Performance clusters and Average Score of students (POLIMI 37).

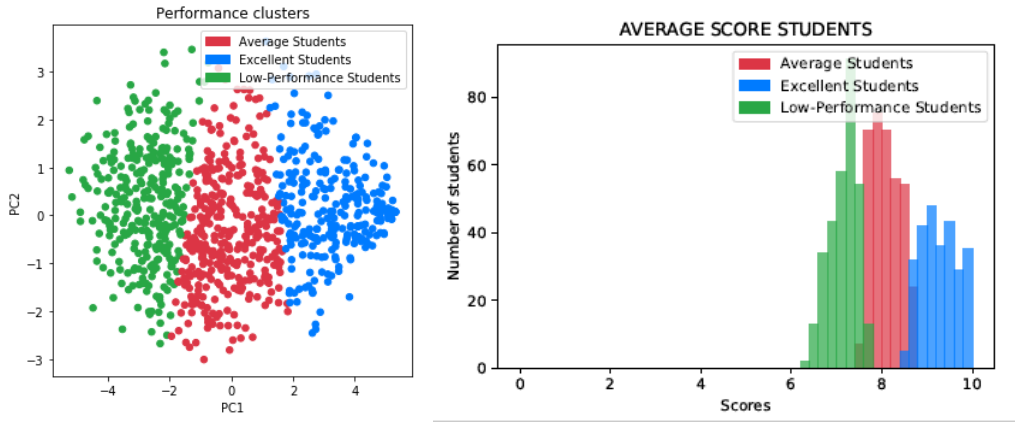


Figure 22. Performance clusters and Average Score of students (POLIMI 44).

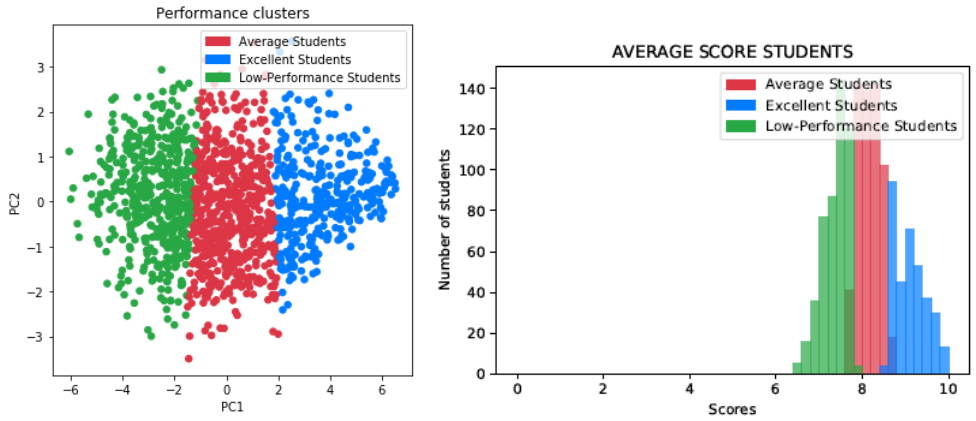


Figure 23. Performance clusters and Average Score of students (POLIMI 49).

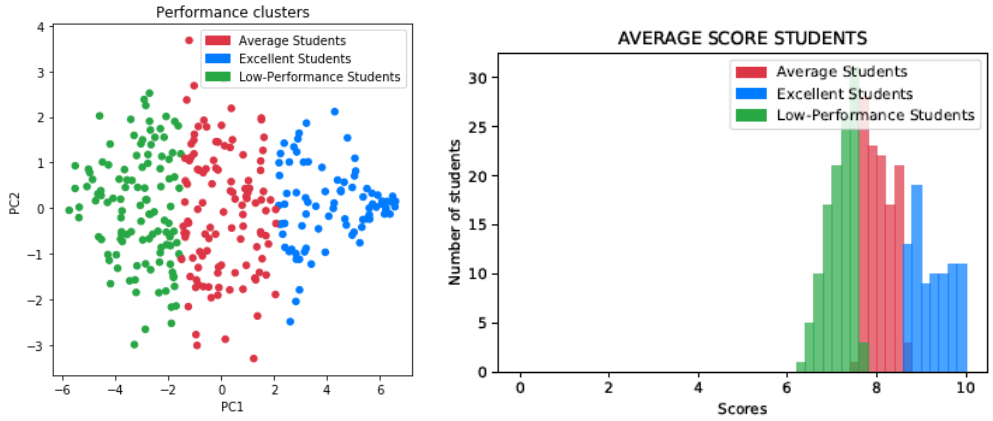


Figure 24. Performance clusters and Average Score of students (POLIMI 62).

very few overlaps. In addition, the three different groups show a clear separation according to the average score obtained by the student, as reported in the Figures 19 to 24.

Student-wise characterization

Obtained clusters have been analyzed in terms of four Categorical Variables:

- Access To Studies Age: age at the time of enrollment to POLIMI.
- Previous Studies: type of High School studies. POLIMI students generally attend Liceo Scientifico, Istituto Tecnico or Liceo Classico before the enrollment.
- Sex: gender of the student.
- Admission Score: before enrolling in POLIMI, each student must attempt a compulsory admission with score up to 100. Generally, if a student obtains a score greater than 60.00, he can enroll in the engineering programme. However, if the maximum number of students for each of the programme has not been saturated, students with an admission score lower than 60 might be admitted to those engineering track. This score has been standardized in [0, 10] before the analysis.

POLIMI 7 (see Fig. 25):

- Access Age: no clear pattern can be observed, since almost all students enrolls at 18/19 years old.
- Previous Studies: the majority attended Liceo Scientifico; students from Technical studies has a slightly lower performance.
- Sex: few women, the proportion of excellent women is lower, but not significantly.
- Admission Score: very clear pattern, the higher the result, the higher the performance. At a score of 8.0, there is a balance across the three groups. Students at 10.0 usually belong to excellent cluster.

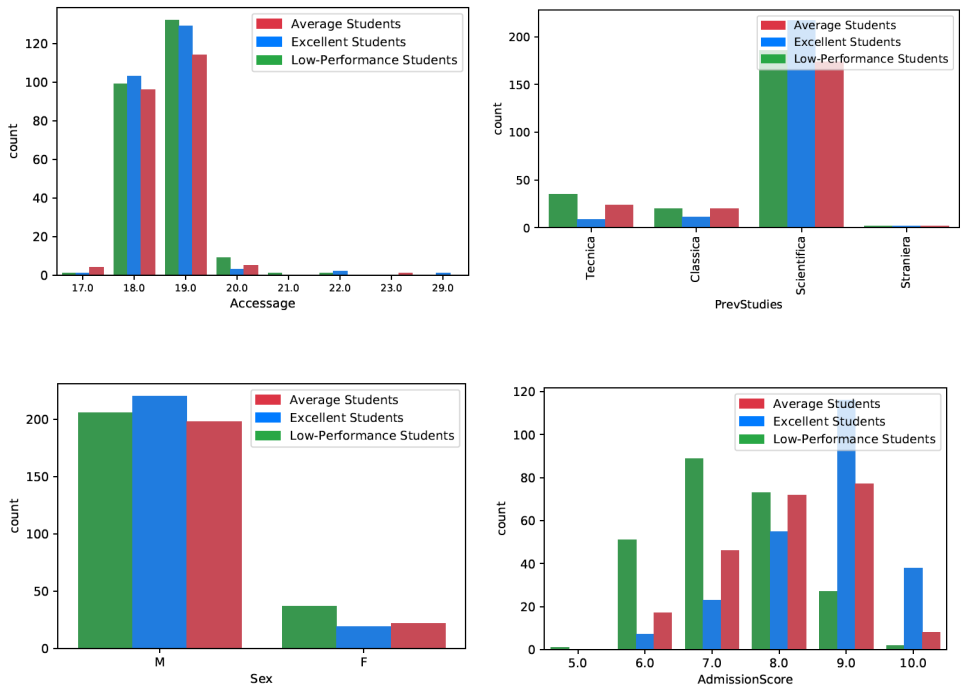


Figure 25. Categorical Variable Analysis (POLIMI 7).

POLIMI 13 (see Fig. 26):

- Access Age: no clear pattern can be observed since almost all students enrolls at 18/19 years old.
- Previous Studies: the majority attended Liceo Scientifico, students from Technical studies also have excellent performance.
- Sex: the proportion of female is higher than in other Engineering programmes; the performance is the same across both genders.
- Admission Score: very clear pattern, the higher the result, the higher the performance. At a score of 8.0, there is a balance across the three groups. Students at 10.0 usually belong to excellent cluster.

POLIMI 37 (see Fig. 27):

- Access Age: no clear pattern can be observed since almost all students enrolls at 18/19 years old.
- Previous Studies: there is a higher percentage of students from Technical studies than in other engineering, but these students have lower performance.

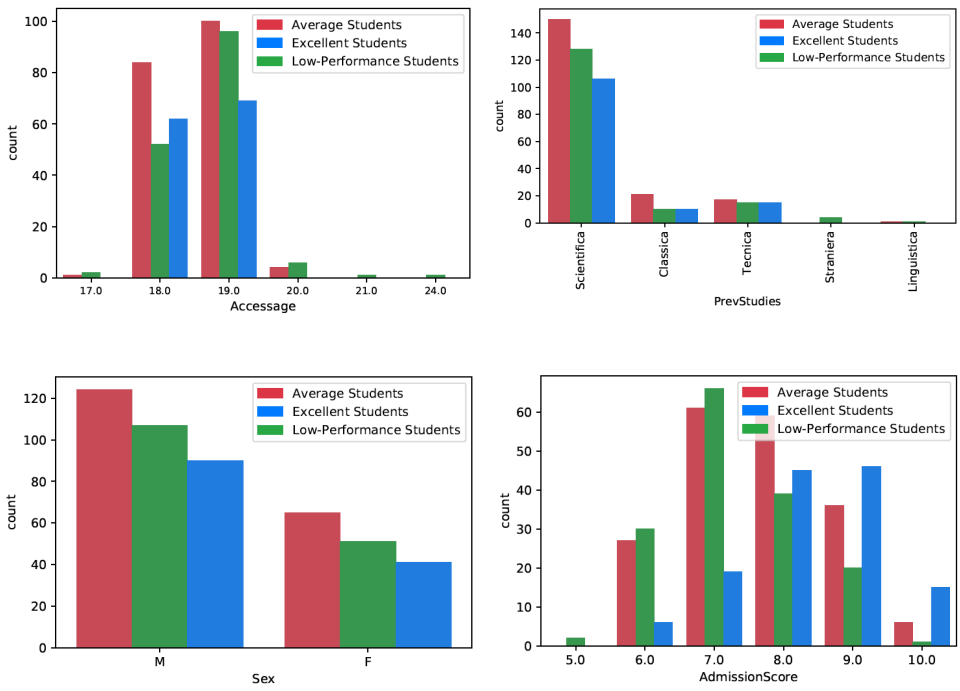


Figure 26. Categorical Variable Analysis (POLIMI 13).

- Sex: few women attend this programme, but the performance seems the same across both genders.
- Admission Score: clear pattern, the higher the result, the higher the performance. At 8.0 the proportion of "Low-performance students" is significant, while students at 10.0 usually belong to excellent cluster.

POLIMI 44 (see Fig. 28):

- Access Age: no clear pattern can be observed since almost all students enrolls at 18/19 years old.
- Previous Studies: there is a higher proportion of Technical studies than in other Engineering. No clear pattern according to previous studies can be observed.
- Sex: very few women attend this programme, but the performance seems the same across both genders.
- Admission Score: At a score of 8.0, there is a balance across the three groups. Students at 10.0 usually belong to excellent cluster.

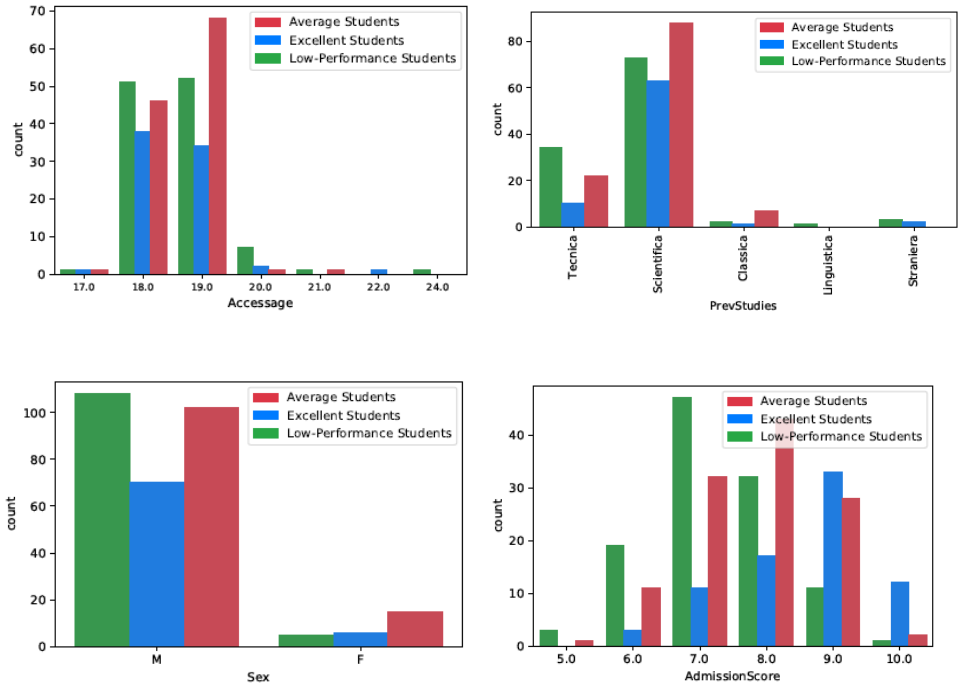


Figure 27. Categorical Variable Analysis (POLIMI 37).

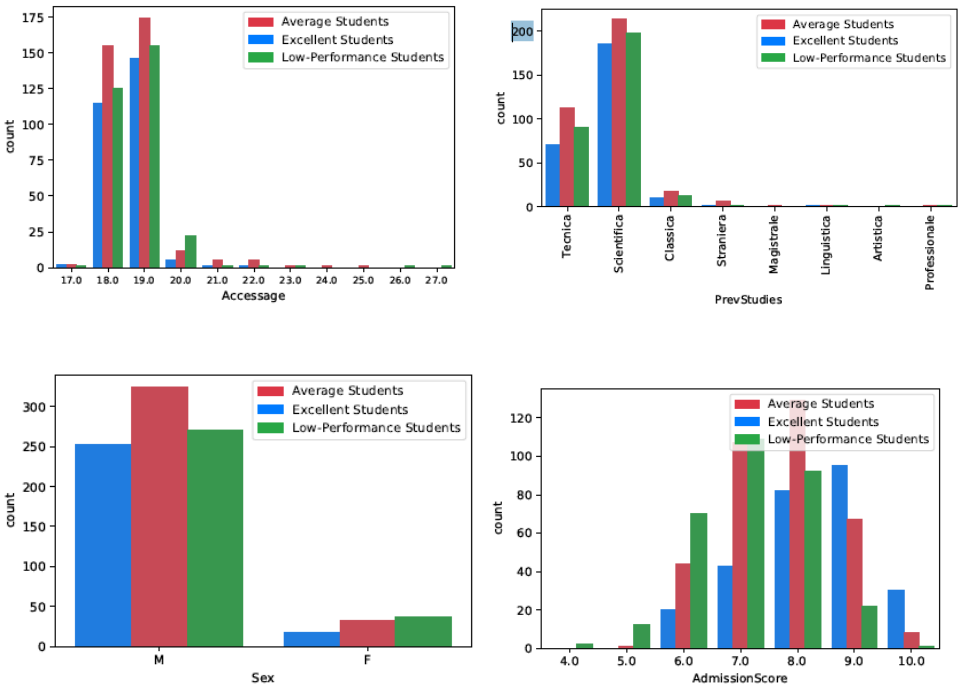


Figure 28. Categorical Variable Analysis (POLIMI 44).

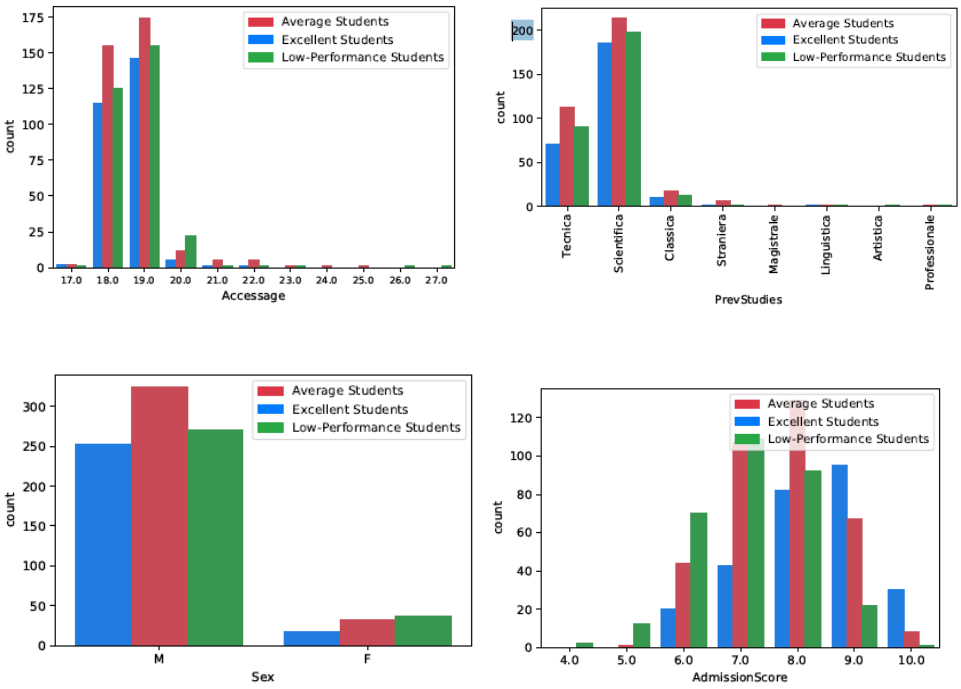


Figure 29. Categorical Variable Analysis (POLIMI 44).

POLIMI 49 (see Fig. 29):

- Access Age: no clear pattern can be observed since almost all students enrolls at 18/19 years old.
- Previous Studies: the majority attended Liceo Scientifico. No clear pattern according to previous studies can be observed.
- Sex: very few women attend this programme, but the performance seems the same across both genders.
- Admission Score: like in other programmes, at a score of 8.0 there is a balance across the three groups. Students at 10.0 usually belong to excellent cluster.

POLIMI 62 (see Fig. 30):

- Access Age: no clear pattern can be observed since almost all students enrolls at 18/19 years old.
- Previous Studies: the majority attended Liceo Scientifico, but the percentage of students from Istituto Tecnico is significant. No clear pattern according to previous studies can be observed.

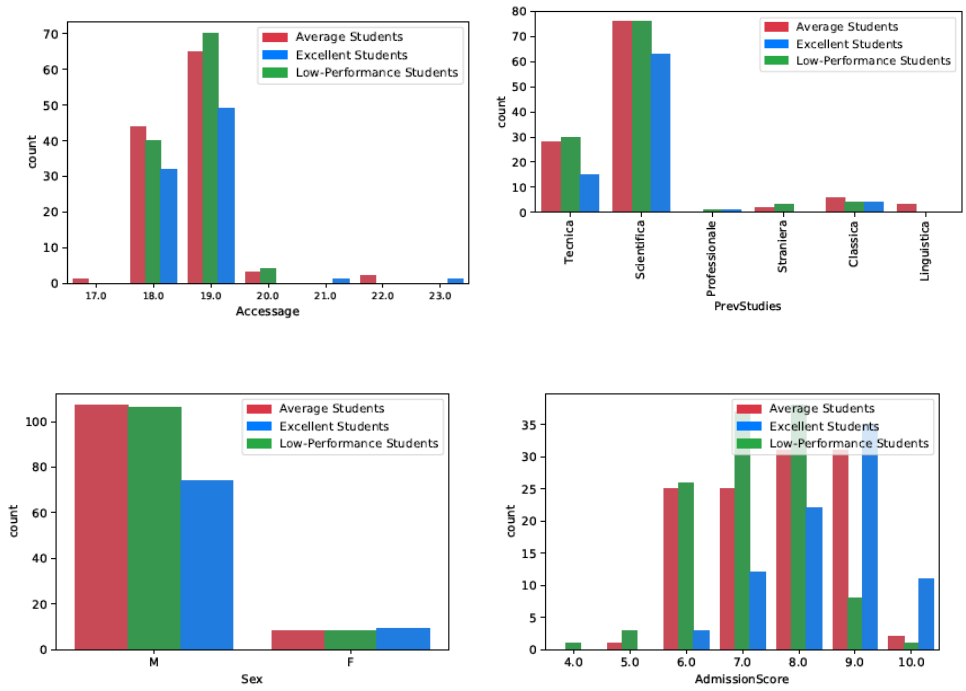


Figure 30. Categorical Variable Analysis (POLIMI 44).

- Sex: very few women attend this programme, but the proportion of "Excellent" women is high with relation to the other groups.
- Admission Score: the usual pattern is observed. Unbalance at 8.0 where the proportion of "Low" is significant, students at 10.0 usually belong to excellent cluster.

Institution-wise characterization

As noted before, in all six POLIMI degrees, observations are split into three clusters with very few overlaps. In addition, those clusters shows a clear separation according to the average score obtained by the student. Silhouette coefficients are either medium or good, according to our baseline.

After analyzing the 6 POLIMI degrees, we can explore the common trends of the four categorical variables:

- Age at the time of admission: in all degrees, the majority of students is either 18 or 19. Therefore, no clear pattern could be observed due to the low heterogeneity in ages.
- Previous Studies: in all degrees, the majority of students graduated at Liceo Scientifico. The proportion of students from Technical Studies is

variable across degrees. Specifically, it is higher in Computing Systems (POLIMI 44) and Mechanical (POLIMI 49) as expected, since those are main topics in Technical Institutes. Generally, no clear pattern according to previous studies could be observed. The proportion of "Low performance" students is slightly higher on average in Technical Studies, and especially in Electronic Engineering (POLIMI 37).

- Sex: the proportion of Male students is always higher, especially in POLIMI 44 and POLIMI 49. The performance seems the same across both genders.
- Admission score: a very clear pattern can be observed across all degrees: the higher the result, the higher the performance. In some degrees, there is a balance at 8.0, while in Electronic and Automation, the proportion of "Low" is still significant. Very few students have low performance and high Admission Score, and vice versa.

Classification Analysis

As commented at IO2 report (Section 3.3), we compare the performance of the classification algorithm in two different situations. First, we consider only information available after the first year of studies, both exam scores and categorical variables. Second, we consider the information available at the end of the second year of studies. In the case of POLIMI 13, POLIMI 37 and POLIMI 62, the classification accuracy obtained when including the score at the 1st year is considerably lower when compared with accuracies obtained by adding the scores of the 2nd year. In POLIMI 7, POLIMI 44 and POLIMI 49, the accuracy gap is lower, but still significant. If we take into account the career at the end of the 1st semester, students are almost always correctly classified into the three clusters, as reported in the following table.

ACCURACY	categorical + 1st	categorical + 1st + 2nd
POLIMI 7	0.857	0.969
POLIMI 13	0.751	0.962
POLIMI 37	0.726	0.937
POLIMI 44	0.794	0.979
POLIMI 49	0.820	0.979
POLIMI 62	0.726	0.956

4.3 Instituto Politecnico de Bragança Degrees Analysis

Five degrees have been considered for the case of Instituto Politecnico de Bragança (IPB):

- IPB 9123 - Mechanical Engineering (266 students)
- IPB 9089 Civil Engineering (346 students)
- IPB 9112 Electrical Engineering (193 students)
- IPB 9119 Computer Engineering (236 students)
- IPB 9126 Chemical Engineering (165 students)

4.3.1 Clustering and Classification

Clusters Analysis

For each degree considered at IPB is possible to identify three clusters based on average score. Depending on the degree these clusters are more visible or more overlapped. When the columns are overlapped is difficult to identify the average score by subject or by student.

IPB 9123 (see Fig. 31):

In this degree the three identified clusters are clearly separated, there is no overlapping. The average score students can be much better distinguished than the average score subjects. It is possible to have in each subject several kind of students getting good scores.

IPB 9089 (see Fig. 32):

In this degree there is a strong overlapping between low and average students. There is no distinction between these two groups of students. The student performance in each subject doesn't depend on the scores he/she got in the other subjects neither on the student profile.

IPB 9112 (see Fig. 33):

In this degree there are less students and the distribution of the clusters are bigger. Low performance and average can be better distinguished but there is a small overlap.

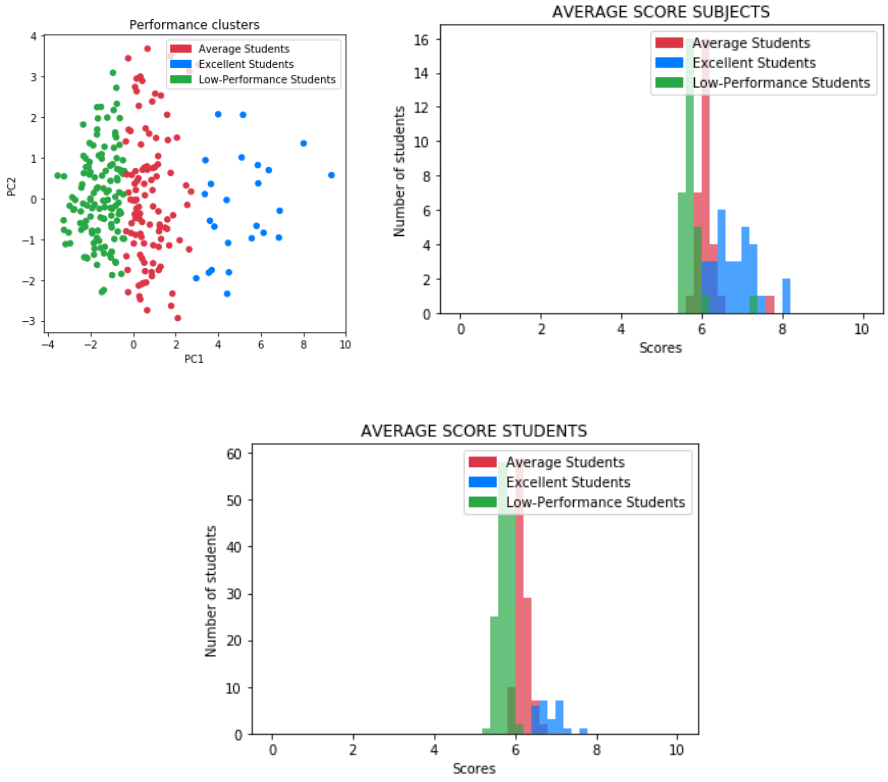


Figure 31. Performance clusters and Average Score of students (IPB 9123).

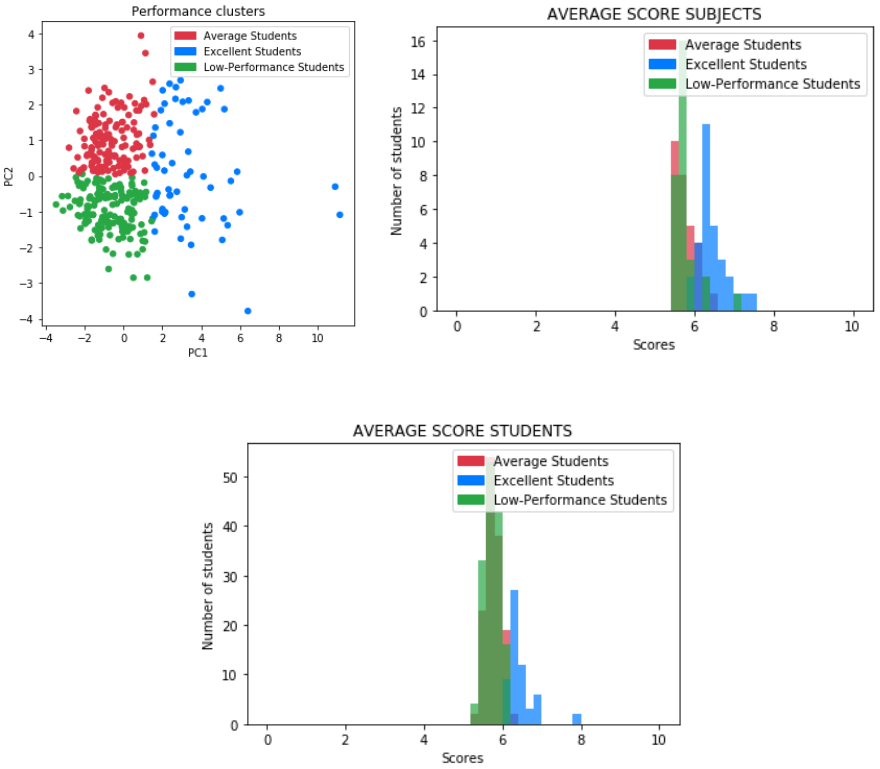


Figure 32. Performance clusters and Average Score of students (IPB 9089).

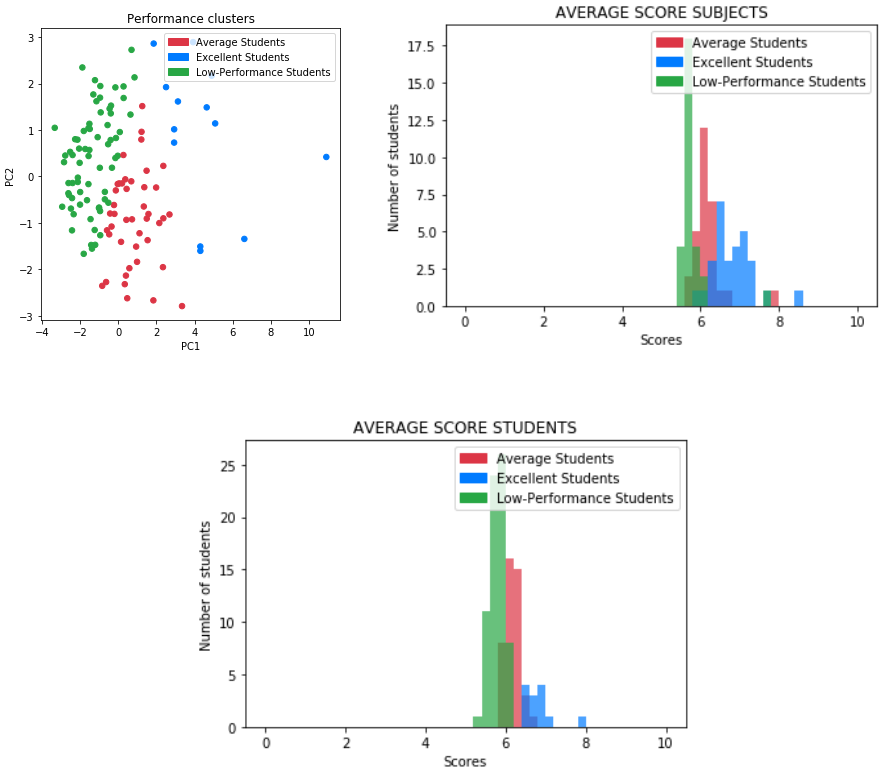


Figure 33. Performance clusters and Average Score of students (IPB 9112).

IPB 9119 (see Fig. 34):

In this degree the three identified clusters are separated however there is a very small overlapping. The average score students can be much better distinguished than the average score subjects.

IPB 9126 (see Fig. 35):

In this degree better students can be found in general but there is a overlap between average and excellent students. And this overlapping can also be seen in the average score students and average score subjects.

Student-wise characterization

We clearly observe and compare the student profile in each degree based on their age, previous studies and nationality. This allows us to take some interesting conclusions.

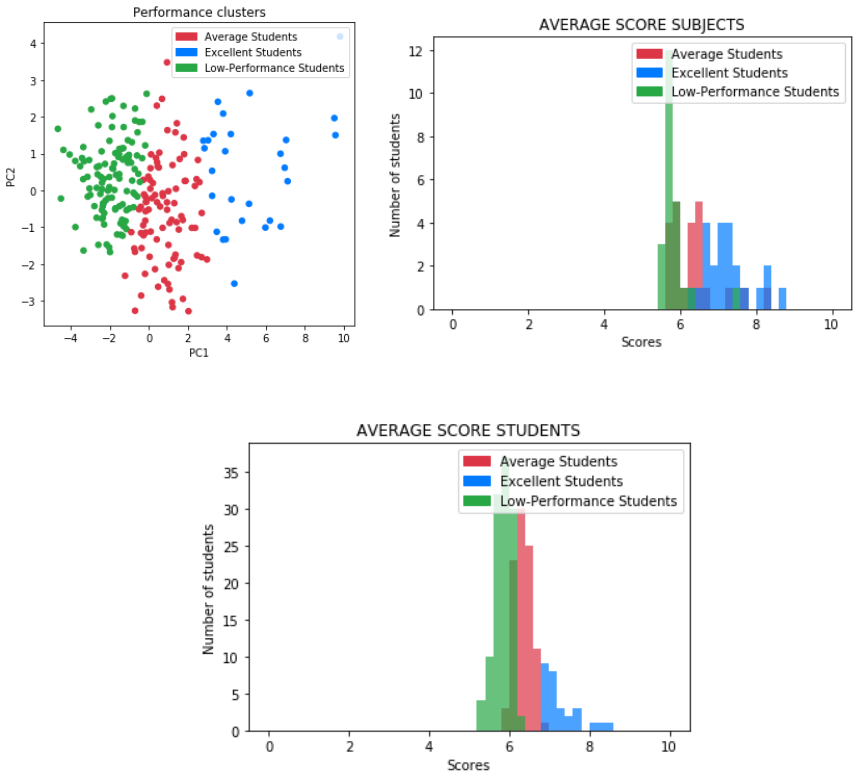


Figure 34. Performance clusters and Average Score of students (IPB 9119).

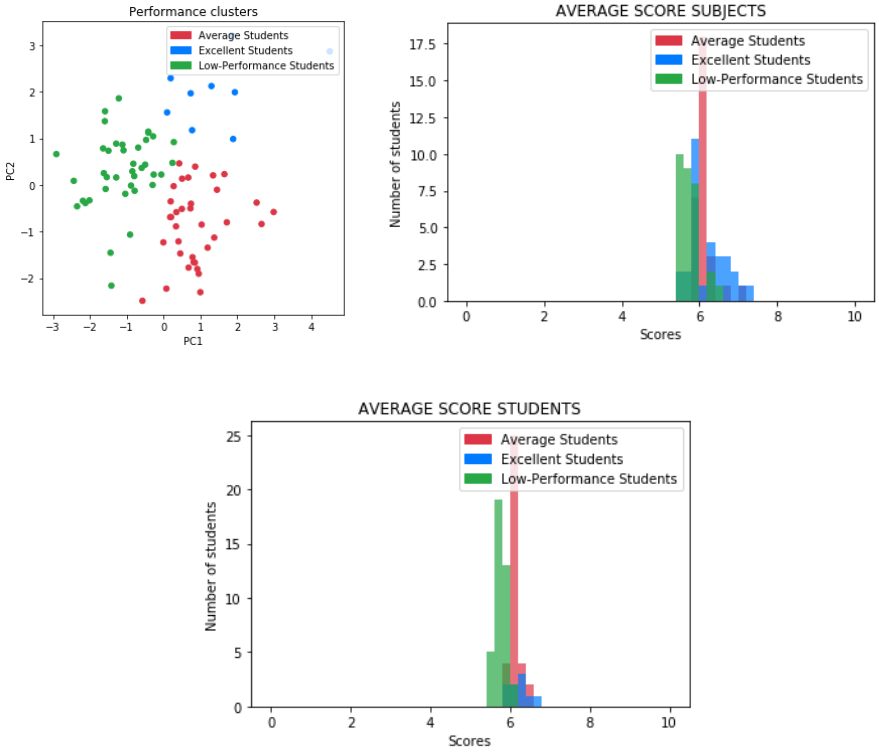


Figure 35. Performance clusters and Average Score of students (IPB 9126).

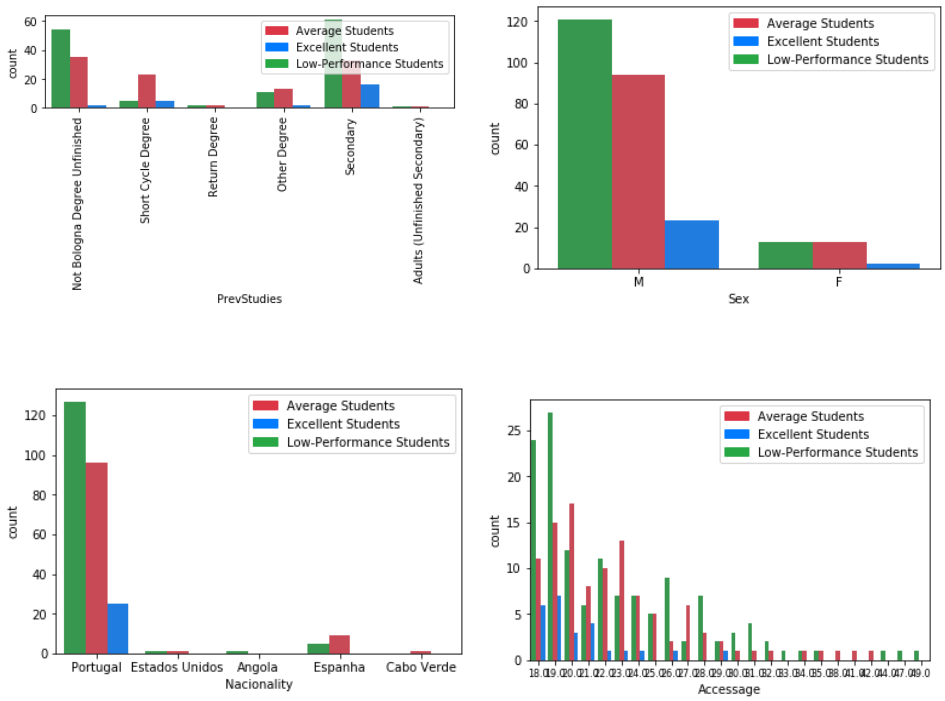


Figure 36. Categorical Variable Analysis (IPB 9123).

IPB 9123 (see Fig. 36):

Students are heterogeneous when analyzing access age. The trends are:

- Sex: Very few women. There are no significant differences in terms of performance between male and female gender.
- Previous studies: The students that came from short cycles are medium students. The best students came from secondary school. Most of the students came from Secondary and the others mainly from Not Bologna Degree Unfinished.
- Access age: Most of the students have an access age between 18 and 20 years old.

IPB 9089 (see Fig. 37):

Students are heterogeneous when analyzing access age. The trends are:

- Sex: more males than females. There are no significant differences in terms of performance between male and female gender.

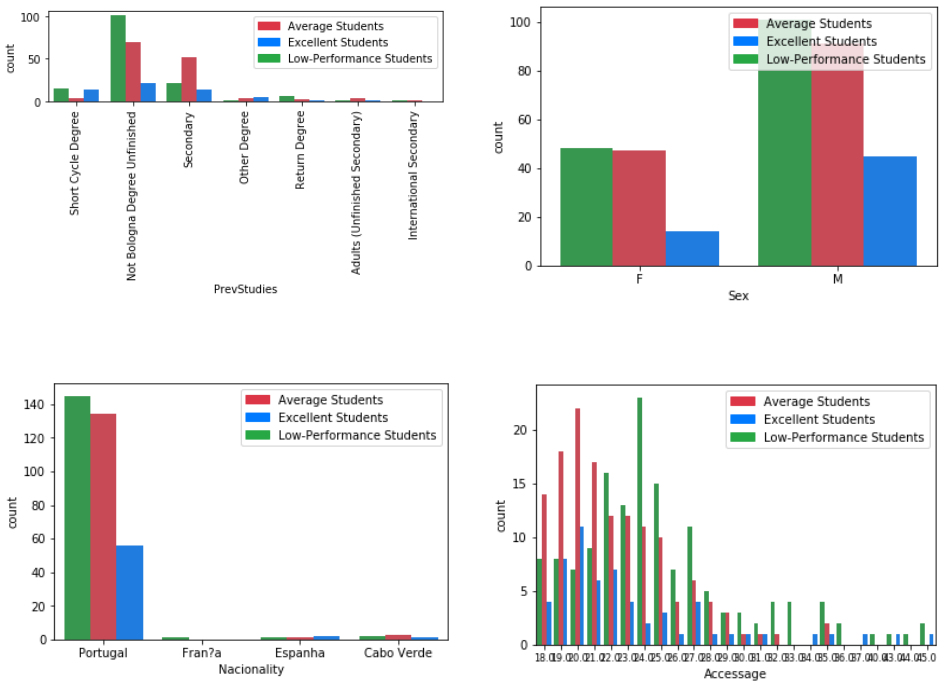


Figure 37. Categorical Variable Analysis (IPB 9089).

- Previous studies: The students that came from short cycles are low performance and excellent students. Most of the students came from Not Bologna Degree Unfinished and the others came mainly from Secondary.
- Access age: Most of the students have an access age between 18 and 25 years old. Average students tend to be the younger ones. Older ones tend to be low-performance students.

IPB 9112 (see Fig. 38):

This degree has changed its name in 2012 and this analyzed data is related with the years between 2006 and 2011. Students are heterogeneous when analyzing access age. The trends are:

- Sex: more males than females. There are no significant differences in terms of performance between male and female gender.
- Previous studies: The students that came from short cycles are mainly low performance students. Most of the students came from Not Bologna Degree Unfinished and the others came mainly from Secondary.
- Access age: Is very heterogeneous. Older students tend to be low-performance students.

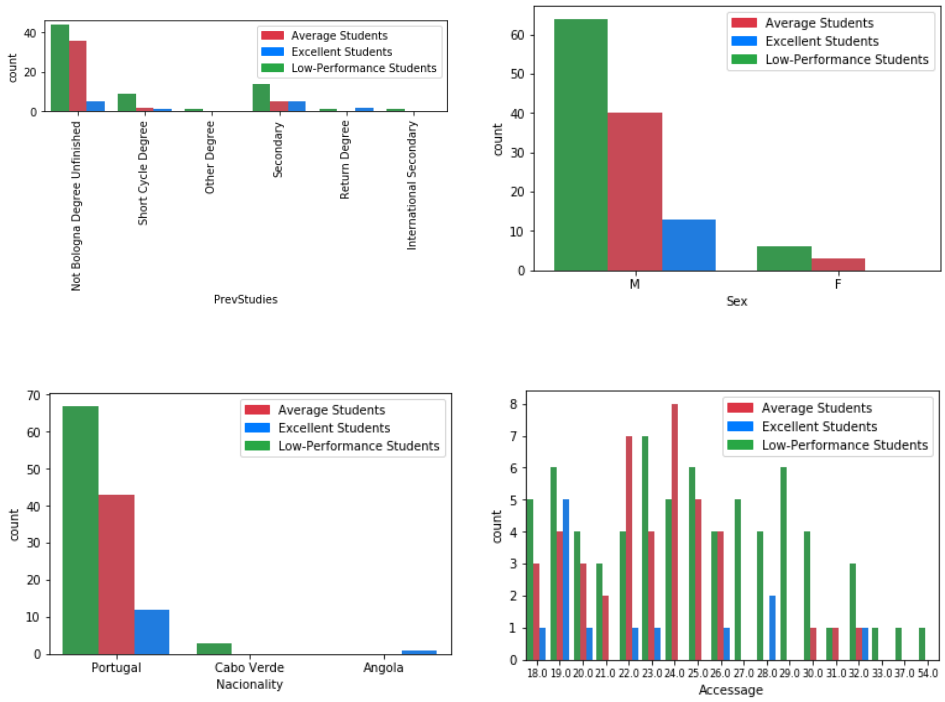


Figure 38. Categorical Variable Analysis (IPB 9112).

IPB 9119 (see Fig. 39):

Students are heterogeneous when analyzing access age. The trends are:

- Sex: more males than females. There are no significant differences in terms of performance between male and female gender.
- Previous studies: The students that came from short cycles are average. Most of the students came from Not Bologna Degree Unfinished and the others came mainly from Secondary. The students that came from Secondary are better than the students that came from Not Bologna Degree Unfinished.
- Access age: Most of the students have an access age between 18 and 26 years old. Average and excellent students tend to be the younger ones.

IPB 9126 (see Fig. 40):

Students are heterogeneous when analyzing access age. The trends are:

- Sex: more females than males. Female students in terms of performance are better than male.

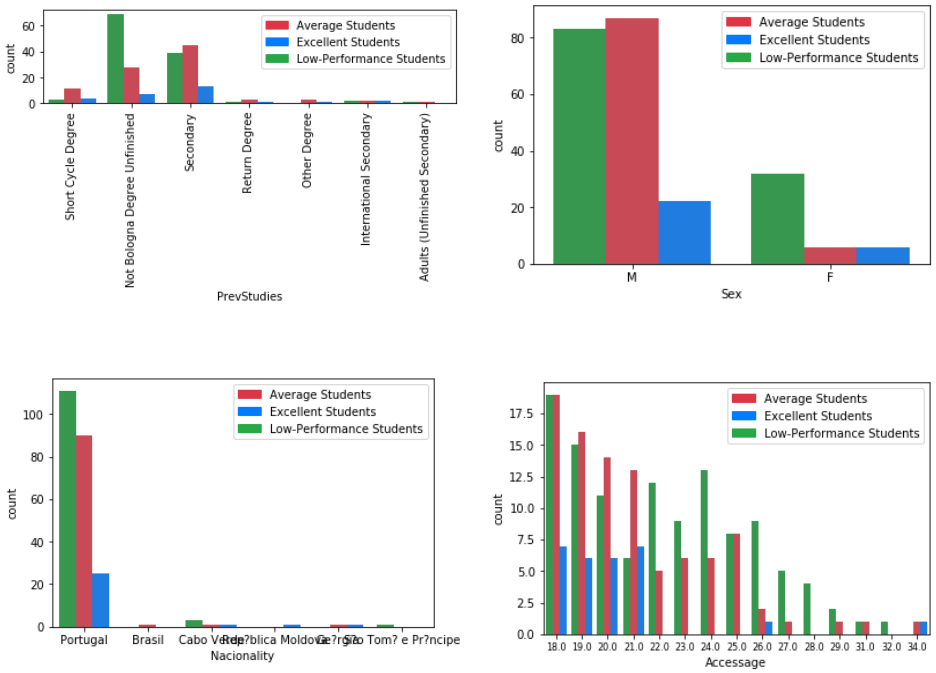


Figure 39. Categorical Variable Analysis (IPB 9119).

- Previous studies: Most of the students come from Not Bologna Degree Unfinished.
- Access age: Most of the students have an access age between 22 and 27 years old. Average and excellent students tend to be younger within this range.

With the IO2 tool for IPB degrees it is only possible to get 3 clusters with medium silhouette value in 3 degrees (Computer Engineering, Electrical and Computer Engineering and Chemical Engineering). For all the others we got only 2 clusters (but with a good silhouette value). Computer Engineering is the only degree where we can observed homogeneous students' pattern. In IPB, we do not have information about admission score, so we cannot compare using this parameter. Concerning access age, younger students are better student in all degrees. The boys are better students than girls (except in Chemical engineering). The most important categorical variable is, at first, the previous studies and then the access age.

Several courses determine the behavior of students at one degree:

- IPB 9123: 1st course Medium, 3rd course Medium
- IPB 9089: 1st course Low, 3rd course High

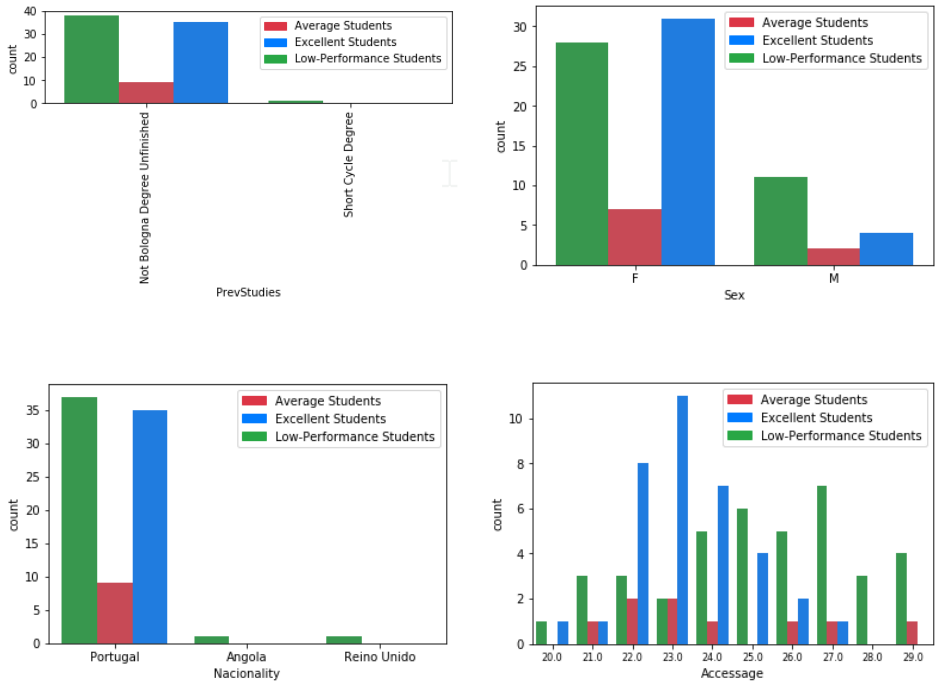


Figure 40. Categorical Variable Analysis (IPB 9126).

- IPB 9112: 1st course Low, 3rd course Low
- IPB 9119: 1st course Low, 3rd course Medium
- IPB 9126: 1st course Medium, 3rd course Low

We conclude that there are some subjects on the 3rd year of studies in Civil Engineering and Computer Engineering that have a strong influence in student performance. In Mechanical Engineering and Electrical Engineering the subjects are better distributed over the years. In Chemical Engineering the last years are not so critical, perhaps because there are lots of laboratory subjects.

In the analysis with IO3 tool it is possible to compare the student score with other categorical variables and we notice that the better students are younger, with a low access age and they are graduated. It is possible to make this kind of comparison to all degrees. The sex does not have a significant influence. It is possible to relate the score with year of birth and the access age. When we want to compare two different degrees, we can visualize them both simultaneously using different windows but we do not have a way to make a direct comparison. Groups can be distinguished in the dimensionality reduction by year, but there is a high overlap in data representation because

it includes all institutions, which presents similar data at some extend. It is possible to verify the differences between the institutions by the difference of shapes and colors. In the dimensionality reduction by degree is it possible to verify, for instance, that the subject attempts are strongly related to the variables Previous Studies and Year of Birth.

4.4 Universidad de León Degrees Analysis

Four degrees have been considered for the case of Universidad de León (ULEON):

- ULEON 707 - Electronics Engineering (88 students)
- ULEON 708 - Mechanical Engineering (131 students)
- ULEON 709 - Computer Science (107 students)
- ULEON 710 - Aerospace Engineering (166 students)

Clusters analysis

For all the degrees at ULEON, it is possible to identify clusters based on their performance. Nevertheless, average students are, in general, highly overlapped with the other clusters, especially with the low-performance cluster, causing a low silhouette value. It can be clearly seen in the Figures 41 to 44, especially in the results obtained for ULEON 710.

On the other hand, overlapping is stronger in the subjects' histograms than in the students' histograms. In fact, from the observation of the principal components, we can suggest a clearer separation in only two clusters (high-performance and non-high-performance students) in the four degrees analyzed. Furthermore, we are not able to explain cluster separation with respect to the homogeneity of the profiles, but at this institution, all the degrees are quite homogeneous.

Student-wise characterization

It is possible to find patterns with regard to access age, sex or previous studies, but not regarding nationality due to the low number of foreign students. Furthermore, the patterns are quite similar for all the degrees, although there are interesting differences for Computer Science (ULEON 709) and Aerospace Engineering (ULEON 710).

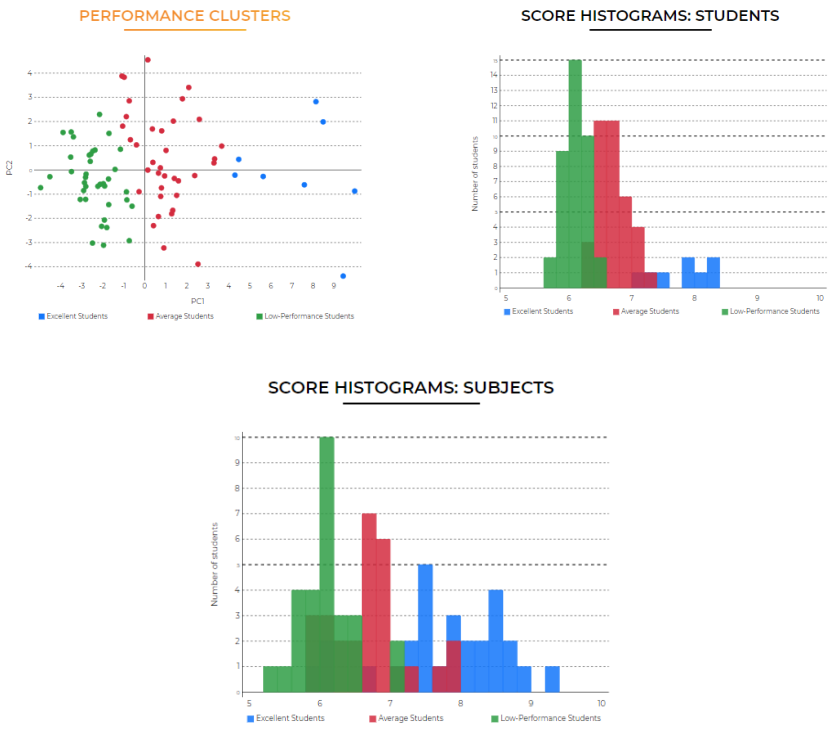


Figure 41. Performance clusters and score histograms (ULEON 707, Electronics Eng.)

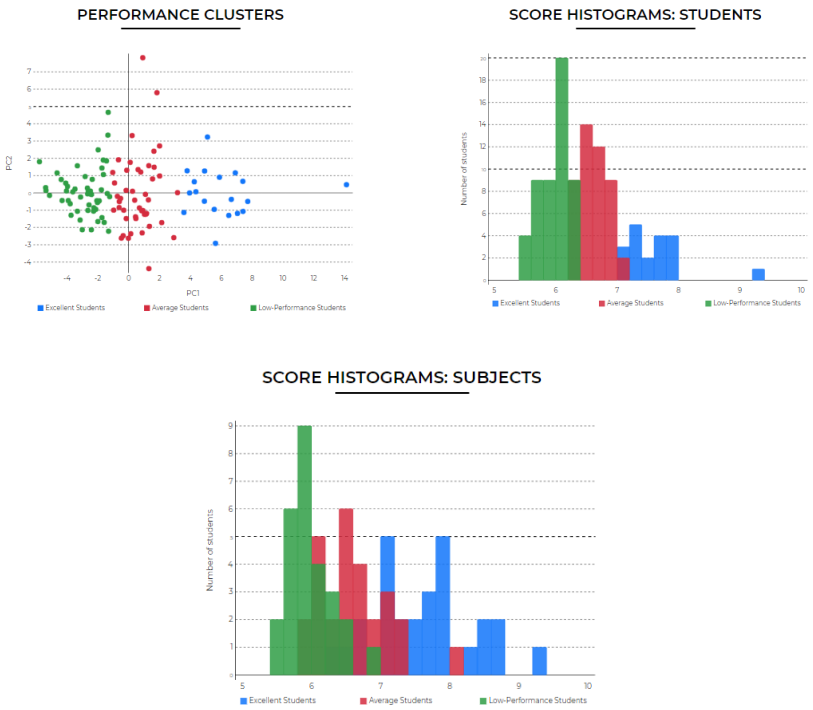


Figure 42. Performance clusters and score histograms (ULEON 708, Mechanical Eng.)

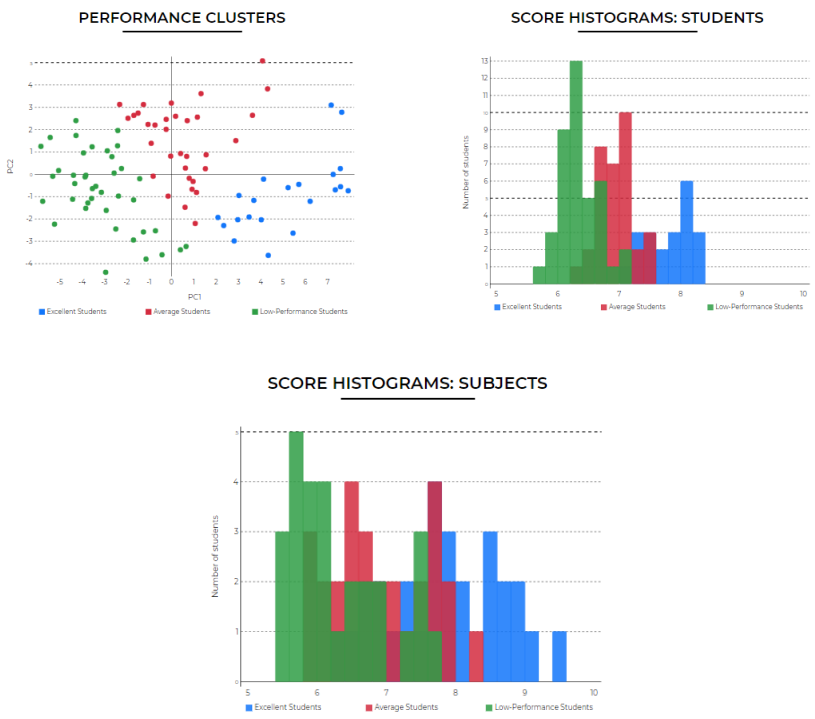


Figure 43. Performance clusters and score histograms (ULEON 709, Computer Science)

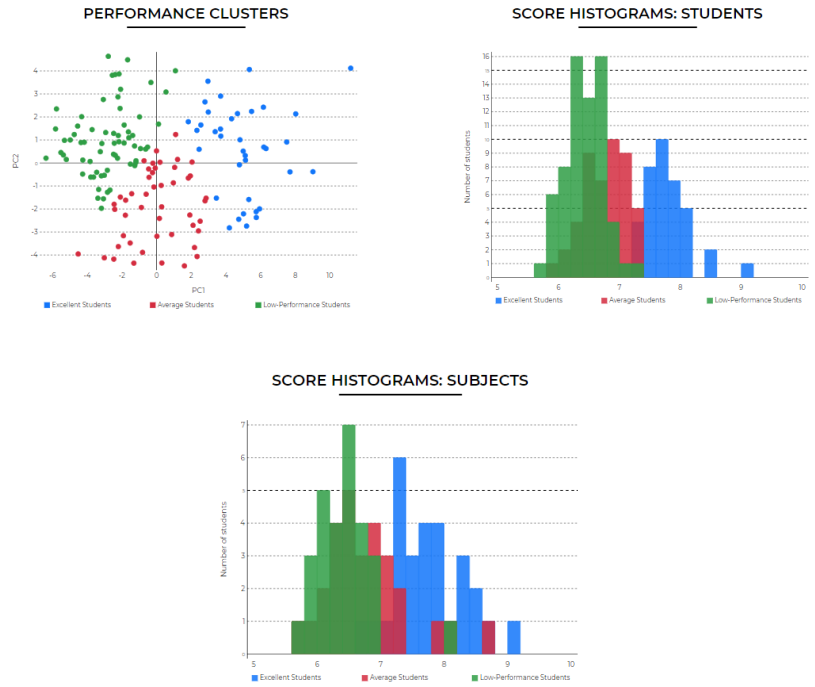


Figure 44. Performance clusters and score histograms (ULEON 710, Aerospace Eng.)

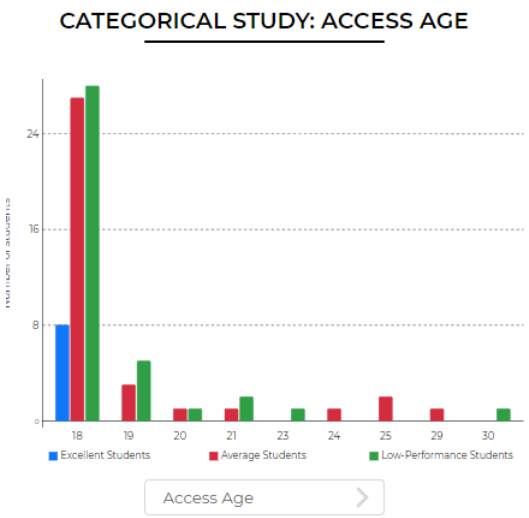


Figure 45. Influence of the access age in Electronics Engineering (ULEON 707)

The general trend is that the proportion of excellent students tend to be higher for women, the youngest students, those with higher admission scores and those coming from secondary. Nevertheless, in Computer Science (ULEON 709), the profiles are more heterogeneous and some tendencies disappear or are mitigated. For instance, there is also a high proportion of excellent students coming from professional studies and the range of access ages and admission scores related with excellent performance is much wider. On the other hand, in Aerospace Engineering (ULEON 710) sex does not seem to be so relevant.

Classification analysis

If we focus on the most determinant categorical variables at the institution level, it is clear that admission scores and access age are determinant for all of them, since the youngest students with highest admission scores tend to be the ones with a higher performance. Furthermore, sex and previous studies are also determinant in some cases (women with previous secondary studies seem to become excellent students more often). The data set from this institution does not allow to draw conclusions with regard to nationality, due to the extremely small percentage of international students. Nevertheless, the preliminary results showed that the effect of categorical variables in the classification was negligible.

On the other hand, with regard to the influence of the first courses in the classification accuracy, we can observe different behaviors for each degree. E.g., for Mechanical Engineering, the accuracy is already high when considering only the first course, whereas the second course improve slightly the results (see Fig. 46).

	Accuracy
1st course	76%
1st + 2nd courses	80%
1st + 2nd + 3rd courses	90%

Figure 46. Classification accuracy for Mechanical Engineering (ULEON 708)

	Accuracy
1st course	58%
1st + 2nd courses	80%
1st + 2nd + 3rd courses	93%

Figure 47. Classification accuracy for Computer Science (ULEON 709)

On the contrary, the addition of the second course in Computer Science strongly improves the accuracy, which was quite low for only the first course (see Fig. 47).

The third course has a stronger influence on the classification accuracy for the Electronics Engineering degree, which achieves the higher result (see Fig. 48).

Finally, for the classification in Aerospace Engineering, the first course is important, whereas the second and third courses improve slightly the accuracy (see Fig. 49).

Coordinated views

First, we need to discuss whether it is possible to formulate hypotheses about the relationship between explanatory variables and performance through histogram filtering. We found that it is, taking as examples single explanatory variables that intuitively are supposed to have an influence on the scores.

	Accuracy
1st course	75%
1st + 2nd courses	77%
1st + 2nd + 3rd courses	95%

Figure 48. Classification accuracy for Electronics Engineering (ULEON 707)

	Accuracy
1st course	80%
1st + 2nd courses	85%
1st + 2nd + 3rd courses	89%

Figure 49. Classification accuracy for Aerospace Engineering (ULEON 710)

For instance, taking the Computer Science Degree of the ULEON, we can observe that the variables Admission Score, Subject Nature, Subject Methodology, Mobility, Knowledge Area and Subject Year have a clear influence on the distribution of scores.

Indeed, a glimpse of the distributions shows that scores tend to be higher for: elective courses, courses with a theoretical-practical methodology, courses taken during mobility programs, specific courses (i.e., those belonging to the areas of computer architecture, computer languages or systems engineering) courses of the last years (see Fig. 50).

Similar relationships between performance and AdmissionScore SubjectNature, Mobility or SubjectYear can also be found for the other degrees.

Another important finding would be to discover trends in the score distributions when it is grouped by an explanatory variable. Both the grouping by Admission Score and Subject Year show a direct correlation with the scores in the Degrees of the ULEON (see Fig. 51). The influence of Admission Score is especially clear for the Mechanical Engineering Degree.

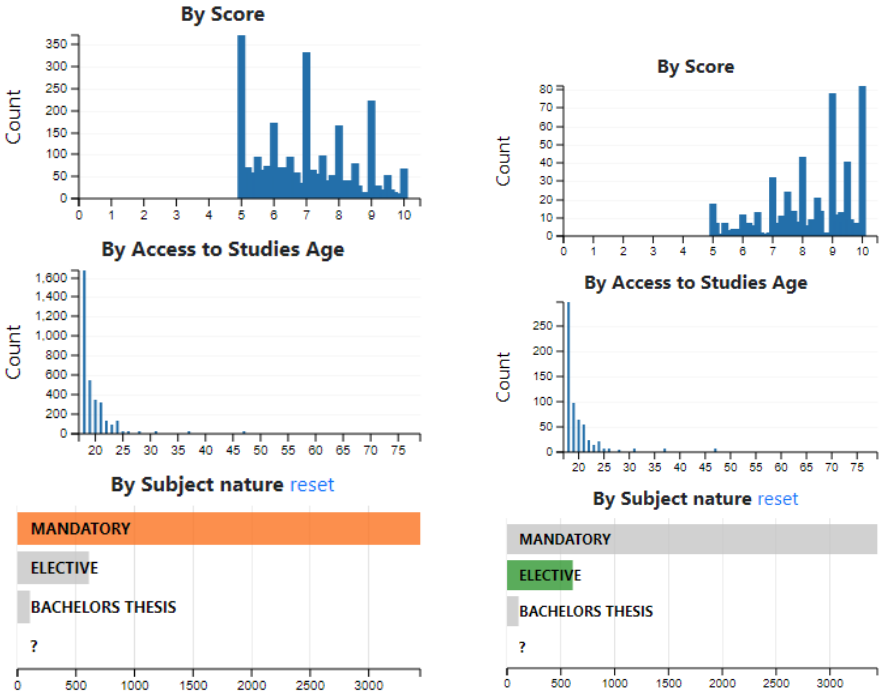


Figure 50. Influence of different variables on the distribution of scores.

Figure 51. Score distributions grouped by an explanatory variable.

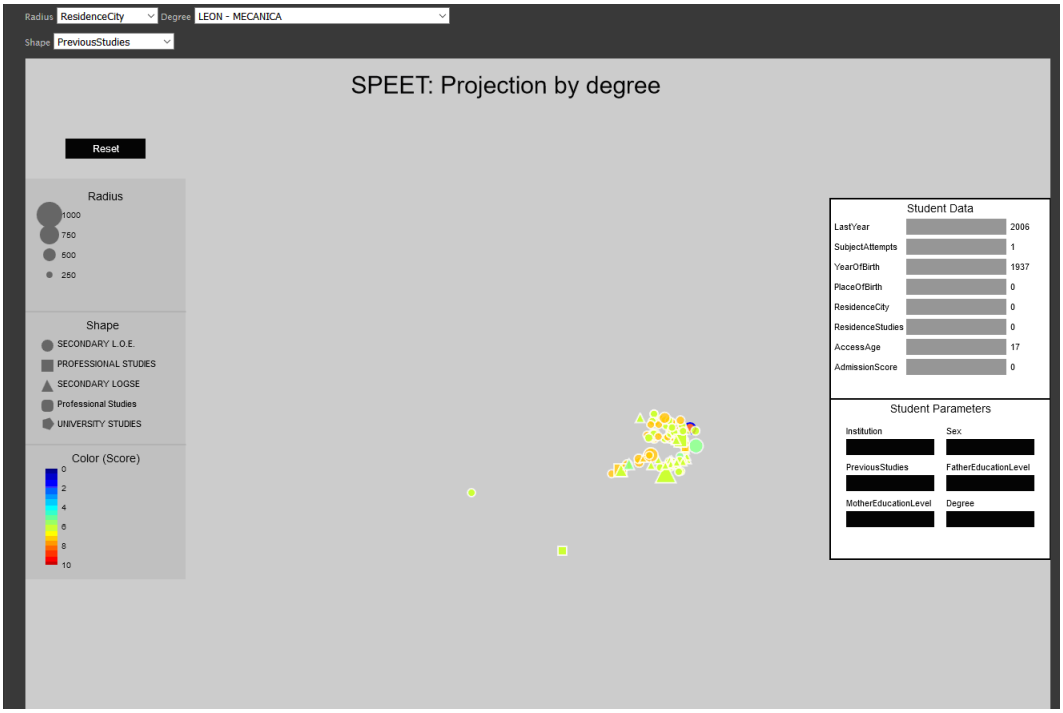


Figure 52. Projection of the data from the Mechanical Engineering Degree

Dimensionality Reduction

Although groups can be clearly distinguished in the dimensionality reduction by year, which includes data from different institutions and degrees, it is not possible for the dimensionality reduction by degree, probably due to the low number of samples.

It is nevertheless possible to distinguish outliers, as can be seen in the figure for the case of Mechanical Engineering at the ULEON.

Although the small data set hinders the ability to distinguish groups, we still might be able to draw some conclusions using variables not considered for the projection. For instance, again, for the case of Mechanical Engineering at the ULEON, the center of the group displays students whose previous studies are secondary L.O.E. (i.e., the latest secondary regulation in Spain), whereas students with other previous studies are shown at the boundaries of that group.

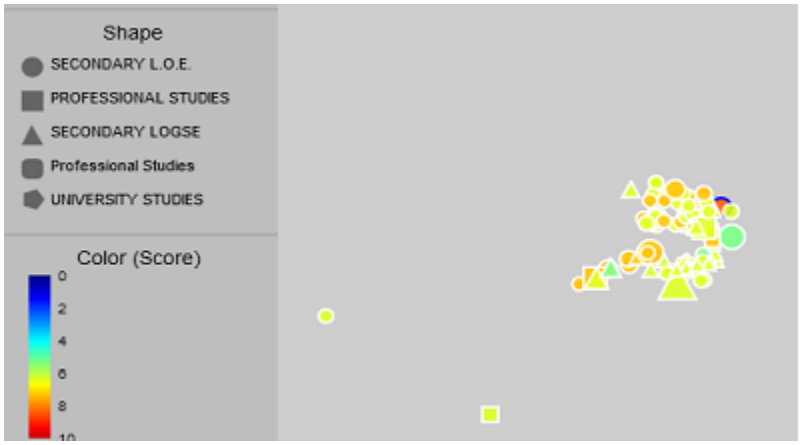


Figure 53. Zoom on the Projection of the data from the Mechanical Engineering Degree

4.5 "Dunarea de Jos" University of Galati

Two degrees have been considered for "Dunarea de Jos" University of Galati (UGAL) case:

- UGAL CS - Computer Science (66 students)
- UGAL AIA - Automatic Control and Applied Informatics (21 students)

Although UGAL degrees do not have a high number of students, the tool help to identify some patterns.

Clusters Analysis

For the two degrees from UGAL, three clusters can be observed based on the student's performance (see Figures 54 and 55). In the case of the CS degree there is a small overlap between the low-performance and average-performance clusters. this can be observed also from the Average Score Students histograms. In the same time from the Average Score Subjects we can observe a strong overlapping of the histograms, which indicates heterogeneity in the subjects evaluation in both degrees.

Student-wise characterization

Obtained clusters have been analyzed in terms of four Categorical Variables and some interesting trends are obtained. These Categorical Variables are: Sex, Nationality, Access Age and Admission Score.

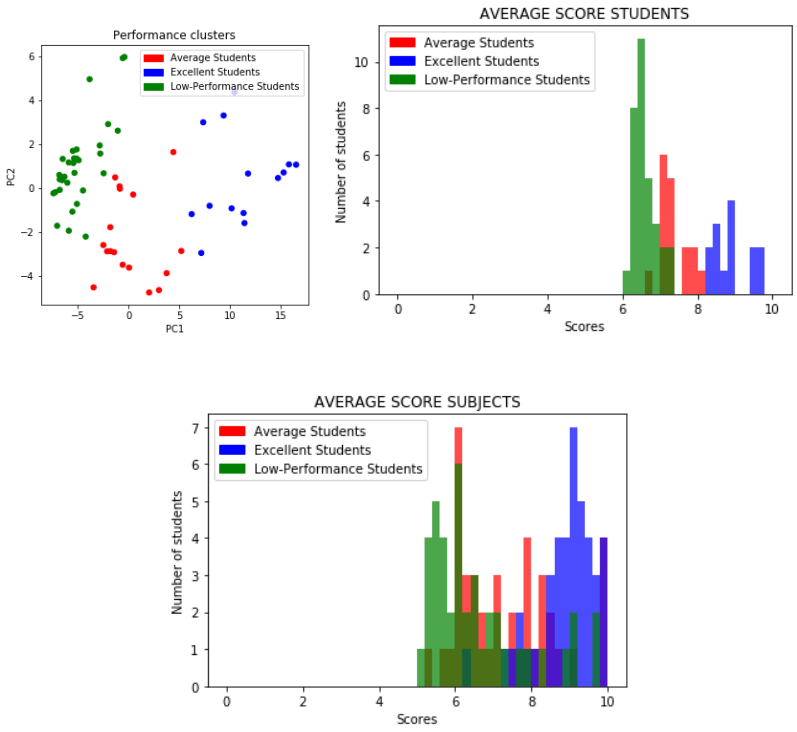


Figure 54. Performance clusters and Average Score of students (UGAL CS).

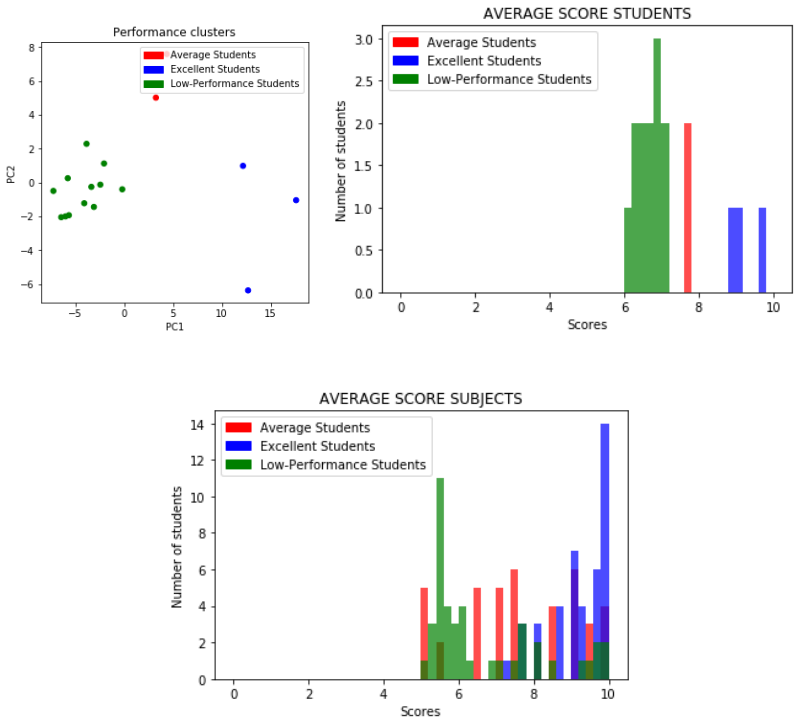


Figure 55. Performance clusters and Average Score of students (UGAL AIA).

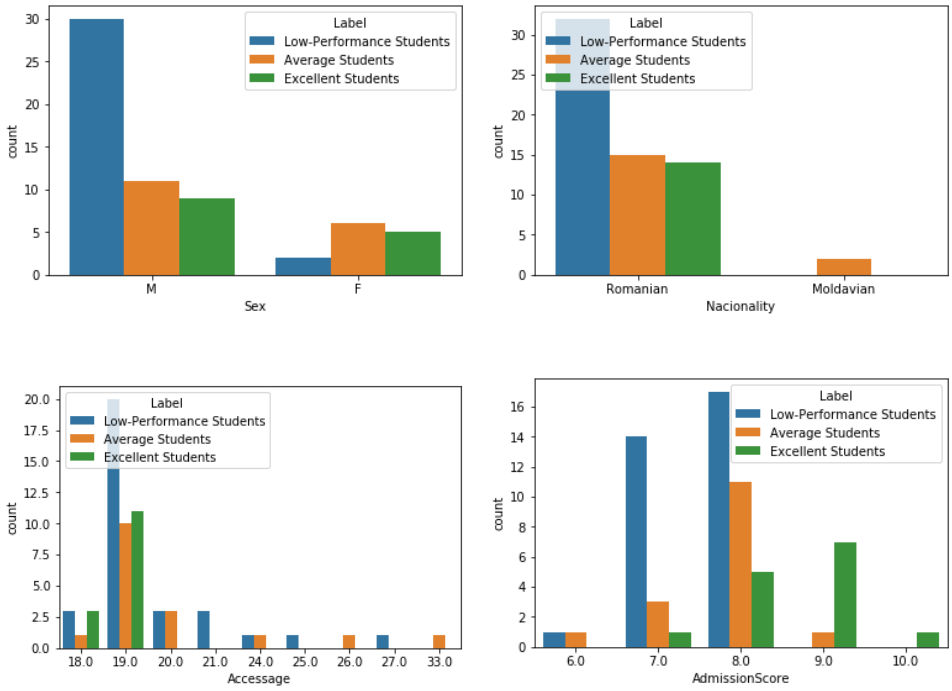


Figure 56. Categorical Variable Analysis (UGAL CS).

UGAL CS (see Fig. 56):

- Sex: few women but the proportion of low-performance women is significantly lower.
- Nationality: the limited number of foreign students prevents us from drawing any conclusion.
- Access Age: almost all students enrolls at 18/19 years old. A clear pattern is that older students are not included in the Excellent performance cluster.
- Admission Score: very clear pattern, the higher the admission score, the higher the performance. Although we can find students included in the Excellent performance cluster, even they have an admission score of 7.0.

UGAL AIA (see Fig. 57):

- Sex: few women but the proportion of low-performance women is lower.
- Nationality: only Romanian students.
- Access Age: almost all students enrolls at 18/19 years old. A pattern is that older students are usually included in the Low performance cluster.

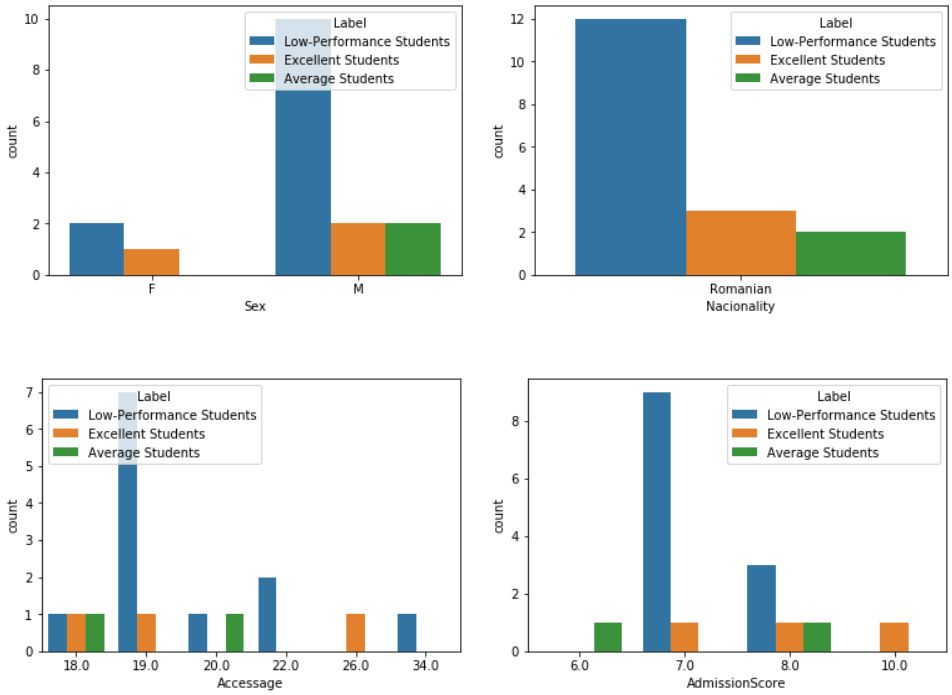


Figure 57. Categorical Variable Analysis (UGAL AIA).

- Admission Score: very clear pattern, the higher the admission score, the higher the performance. Although we can find students included in the Excellent performance cluster, even they have an admission score of 7.0.

Institution-wise characterization

After analyzing the two degrees from UGAL, we can observe common trends in terms of Access Age, Admission Score and Sex:

- Sex: women tend to be better students.
- Access Age: excellent students tend to be younger.
- Admission Score: students with high admission score tend to have a higher performance.

4.6 Opole University of Technology

This section presents the analysis performed for the following engineering degrees at Opole University of Technology (UTOPOLE):

- EI-SI - Computer Engineering (74 records)
- EA-SI - Control Engineering and Robotics (51 records)
- BB-SI - Civil Engineering (62 records)
- BA-SI - Architecture (30 records)
- IL-SI - Logistics (62 records)
- IT-SI - Food Technology and Human Nutrition (44 records)
- MM-SI - Mechanical Engineering (62 records)

The analysis and the figures presented below cover only such cases, in which the students had complete records and the records were in accordance with current study plan of their degree. This means that if there has been any change to the degree study plan (and there have been many changes), the historical data of students studying in accordance to the previous study plan would not have been loaded. The reason for this is to avoid any ambiguities in data, subjects, tables and results.

Clusters analysis

Figures 58 to 64 present the output of the Performance Clusters section of the tool developed as a part of the SPEET project, applied to the anonymized student data provided by the Opole University of Technology. Only the engineering degrees have been taken into consideration, excluding the newest degrees (while having insufficient number of records) and excluding the historical study plans (while having changes in subject relations which may give the appearance of influencing the result or indeed influence the result).

In some of the figures above the groups proposed by the algorithm of the tool are easy to be seen and clearly separated, e.g. in Fig. 61. On the other hand there are figures in which the resulting groups are not clearly visible, e.g. in Fig. 63. The most intriguing figure is the Fig. 64, in which the Average Students seems to be a linear continuation of a part of the Excellent Students group.

Figure 58. Performance Clusters view calculated for EI-SI (Computer Engineering).

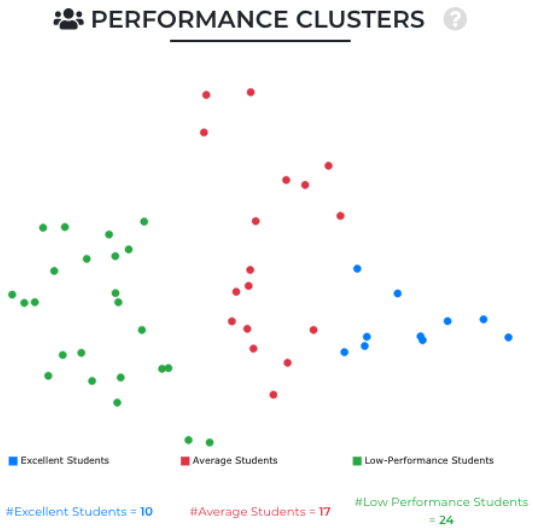


Figure 59. Performance Clusters view calculated for EA-SI (Control Engineering and Robotics).

Student-wise characterization

The clusters have been analyzed considering the following factors (Categorical Variables):

- Age at the beginning of 1st semester (Access To Studies Age)
- Student's gender (Sex)
- Student's nationality (Nationality)

It was not possible to analyze one of the most interesting/promising parameters, i.e. Admission Score, because the candidates records are processed

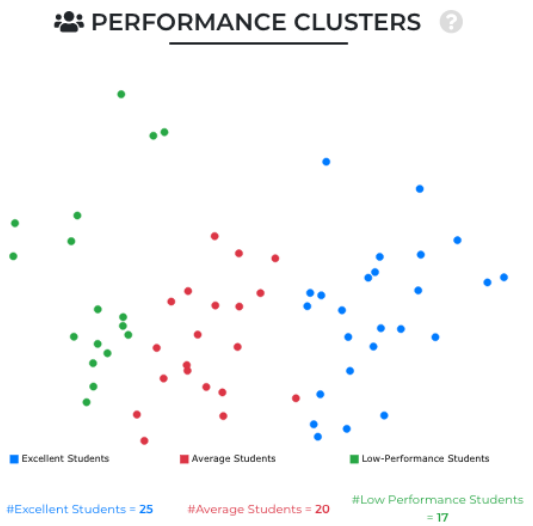


Figure 60. Performance Clusters view calculated for BB-SI (Civil Engineering).

Figure 61. Performance Clusters view calculated for BA-SI (Architecture).

in another data center /system, and when/if they become students, their data is migrated to the university's system. Including the Admission Score (to the set of data being migrated) requires changing the Data Processing Agreement, which makes it not a technical detail but a legal issue.

Figures 65 to 73 present the results of the Categorical Study performed by using the developed tool. Every figure is followed by a brief summary.

Fig. 65 shows that the Computer Engineering degree has a great majority of male students, and that there are some excellent students among them. The excellent students are the majority of every age group, except the youngest students.

Figure 62. Performance Clusters view calculated for IL-SI (Logistics).

Figure 63. Performance Clusters view calculated for IT-SI (Food Technology and Human Nutrition).

The Control Engineering and Robotics (Fig. 66) degree has a very little number of female students, and all of them are excellent (or average) students.

In the Civil Engineering degree (according to Fig. 67) the majority of excellent students is female students, the best students seem to be the youngest students, and the male student group seem to be homogeneous in terms of performance.

Although in some aspects The Architecture degree (Fig. 68) might be considered similar to Civil Engineering, the results presented in the figures show strong differences: BA-SI has only a few male students (while the BB-SI has a similar number of male and female students), the youngest students

Figure 64. Performance Clusters view calculated for MM-SI (Mechanical Engineering).

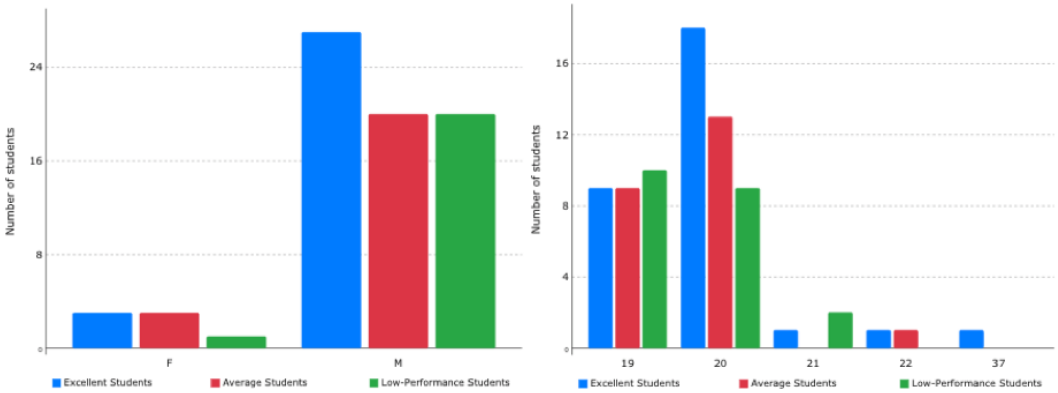


Figure 65. Categorical Study view: Sex (left) and AccessToStudiesAge (right) for the EI-SI - Computer Engineering.

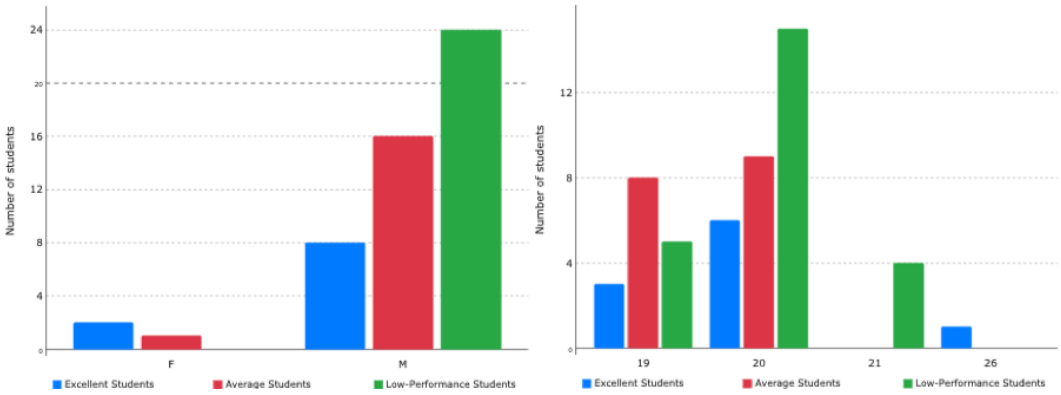


Figure 66. Categorical Study view: Sex (left) and AccessToStudiesAge (right) for the EA-SI - Control Engineering and Robotics.

of BA-SI are mostly average (while the youngest BB-SI students are mostly excellent students).

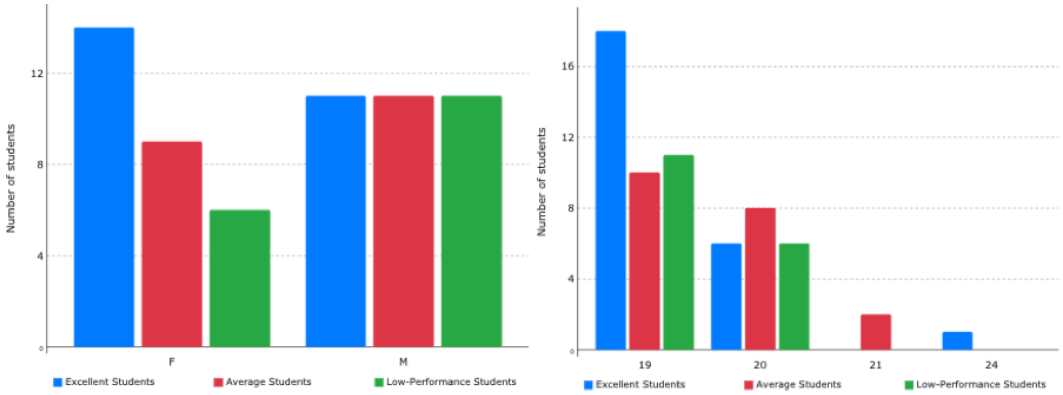


Figure 67. Categorical Study view: Sex (left) and AccessToStudiesAge (right) for the BB-SI - Civil Engineering.

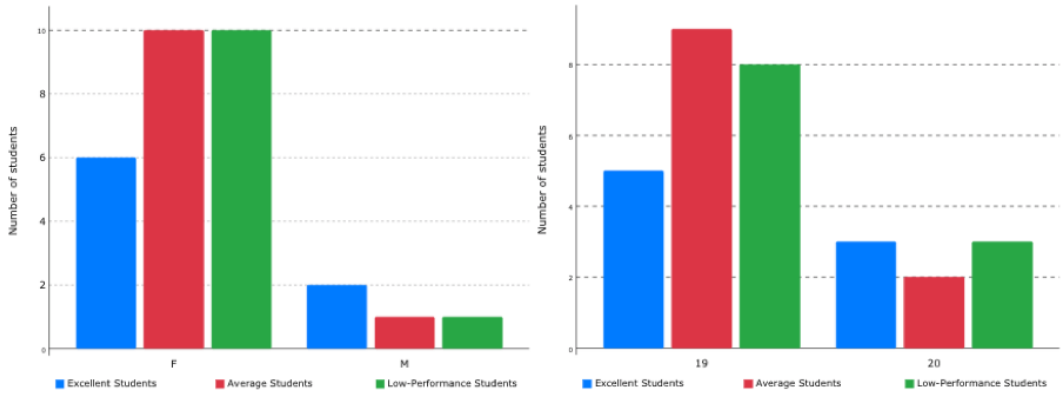


Figure 68. Categorical Study view: Sex (left) and AccessToStudiesAge (right) for the BA-SI - Architecture.

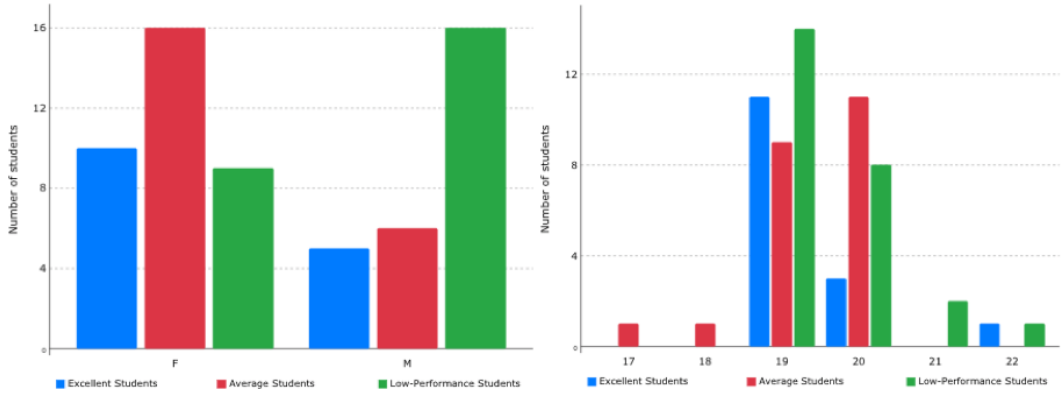


Figure 69. Categorical Study view: Sex (left) and AccessToStudiesAge (right) for the IL-SI - Logistics.

The Fig. 69 shows that the majority of the IL-SI female students belong to the Average Students group, and that there are more excellent than low-performance students among female students of the IL-SI. The graph for the male students shows the opposite.

The Fig. 70 shows that the majority of students of IT-SI belong to the Average Students group. Although the number of excellent and low-performance

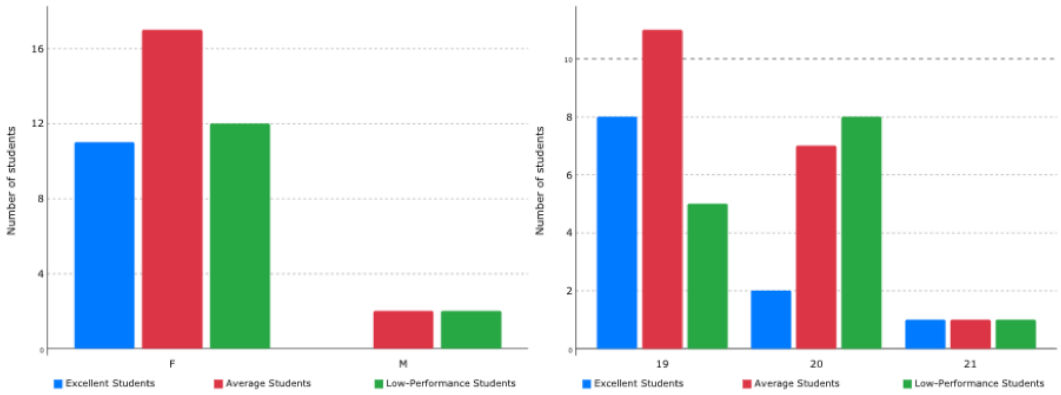


Figure 70. Categorical Study view: Sex (left) and AccessToStudiesAge (right) for the IT-SI - Food Technology and Human Nutrition.

female students seem to compensate, the AccessToStudiesAge graph shows that the issue is much more complex, while the older group has four times more low-performance students that Excellent ones, and the younger group is contrariwise: the number of excellent students exceed the number of low-performance ones.

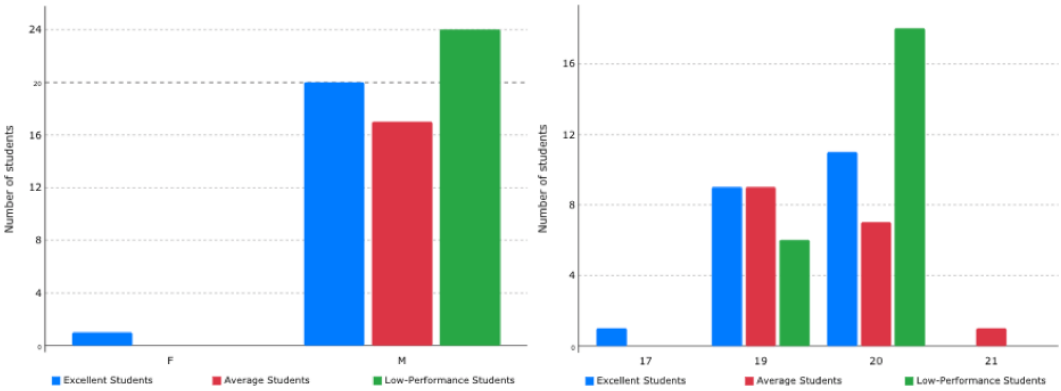


Figure 71. Categorical Study view: Sex (left) and AccessToStudiesAge (right) for the MM-SI - Mechanical Engineering.

The analysis of the MM-SI degree (Fig. 71) shows that older students have a bigger number of low-performance students.

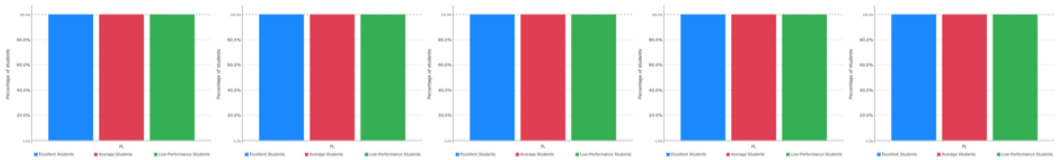


Figure 72. The Nationality view for EI-SI, EA-SI, BB-SI, BA-SI and IT-SI (100% of students are Polish).

The homogeneity of the groups that can be seen in Fig. 72 is a result of a few factors. First of all, Poland did not have any significant political or economical reason to be a target country for young people from neighboring countries in last 5 decades (it has changed recently), especially in small and

medium-sized cities. Young people were rather emigrating to Germany, United Kingdom or Ireland than immigrating. Of course, it is easy to find foreign students in universities, but their data is usually processed separately, by using special groups or labels, mainly due to the individual teaching program. This situation has change drastically in last few years, mainly due to the political situation in Ukraine. All Polish universities, not excluding Opole University of Technology, observe increasing immigration from Ukraine, which involves young people coming to study in Poland. Some of these students speak very good Polish, have Polish ancestors and therefore Polish nationality, which makes it a little complicated to analyze this factor using the tool. But recently even more students come to Poland, so that their presence in the education process as well as education system and databases is no different than Polish students. For this reason the contents of the Fig. 73 few years ago would be considered surprising or exotic, but nowadays it is not.

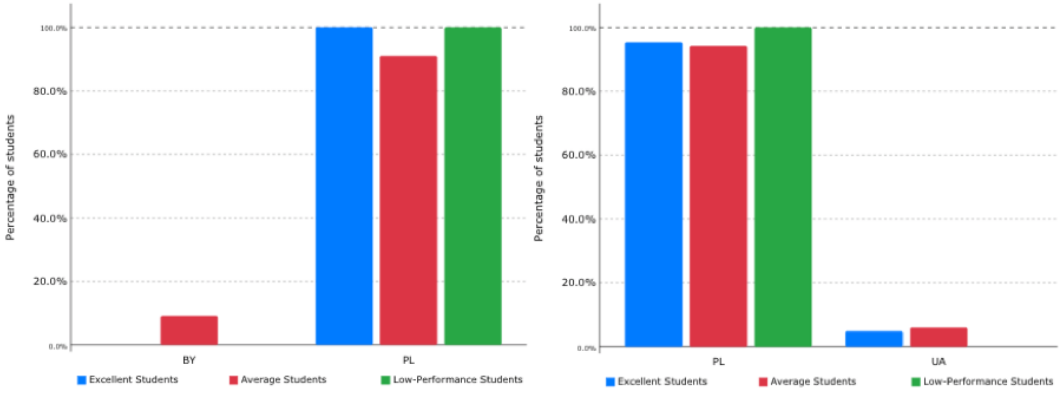


Figure 73. The Nationality view for IL-SI and MM-SI.

Classification Analysis

The table below presents the Classification Analysis using only information available after the second year of studies, both exam scores and categorical variables. The last column uses data obtained from three years of studies. The Classification Analysis shows that students are mostly always correctly classified into right clusters.

	Accuracy →	Categorical; 1st +2nd courses	Categorical; 1st +2nd +3rd courses
EI-SI	Computer Engineering	0,94	0,93
EA-SI	Control Engineering and Robotics	0,91	0,96
BB-SI	Civil Engineering	0,84	0,93
BA-SI	Architecture	0,92	0,94
IL-SI	Logistics	0,84	0,85
IT-SI	Food Technology and Human Nutrition	0,90	0,95
MM-SI	Mechanical Engineering	0,91	0,98

5 Conclusions

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 bellow.

- Could we separate students at different groups (clusters) based on their performance behavior?

  All the partners has 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 (Excepts POLIMI where the performance is the same across both genders).
  - 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 behaviour. 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. Another 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.

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

References

[aRVBP+18] J. L. Vicario amd R. Vilanova, M. Bazzarelli, A. Paganoni, U. Spagnolini, A. Torrebruno, M.A. Prada, A. Morán, M. Domínguez, M.J. Varanda, P. Alves, M. Podpora, and M. Barbu. Io2 - data mining tool for academic data exploitation. Technical report, ERASMUS + KA2 / KA203 SPEET Project, 2018.

[BVV+17] M. Barbu, R. Vilanova, J. Lopez Vicario, M.J. Varanda, P. Alves, M. Podpora, M.A. Prada, A. Morán, A. Torrebruno, S. Marin, and R. Tocu. Data mining tool for academic data exploitation. literature review and first architecture proposal. Technical report, ERASMUS + KA2 / KA203 SPEET Project, 2017.

[MH08] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008.

[PDM+18] M.A. Prada, M. Domínguez, A. Morán, R. Vilanova, J. Lopez Vicario, M.J. Varanda, P. Alves, M. Podpora, M. Barbu, A. Torrebruno, U. Spagnolini, and A. Paganoni. Data mining tool for academic data exploitation. graphical data analysis and visualization. Technical Report IO3, ERASMUS + KA2 / KA203 SPEET Project, 2018.