INTELLECTUAL OUTPUT #5

Student Profile for Enhancing Tutoring Engineering

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

Student Profile
for Enhancing
Engineering Tutoring

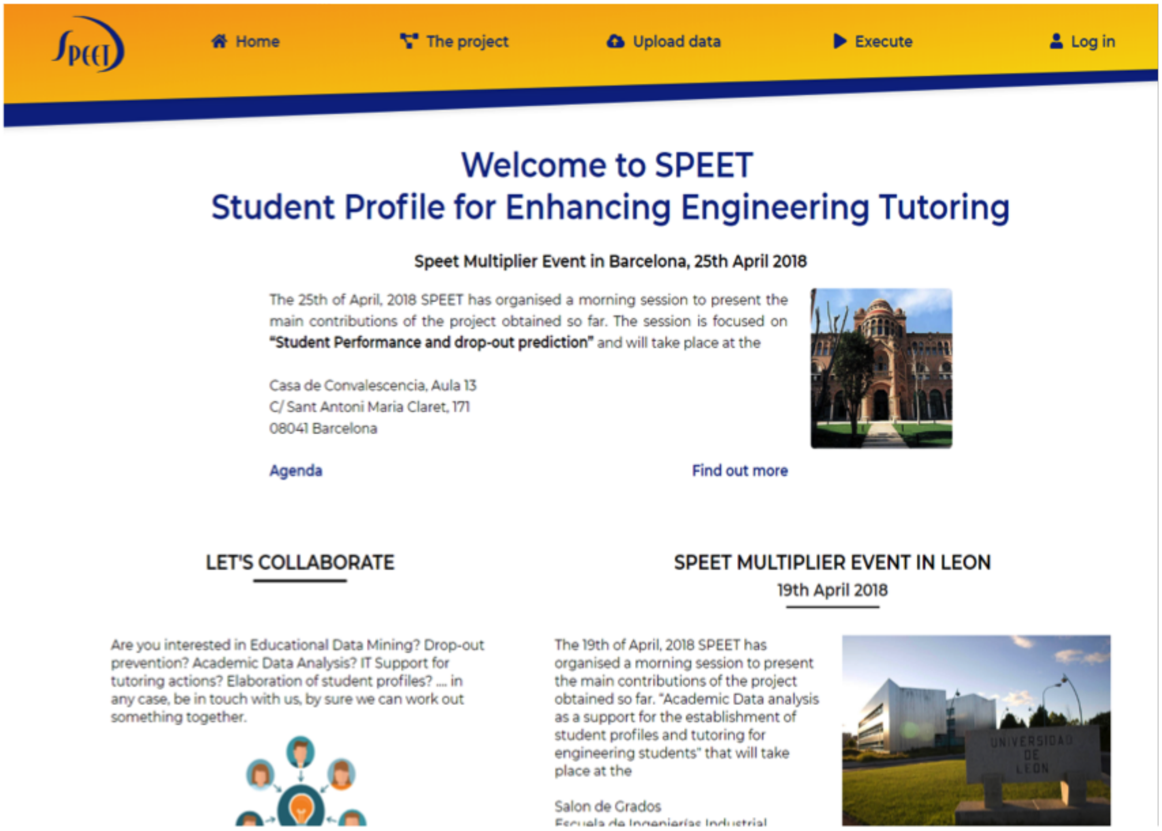
ERASMUS + KA2 / KA203

Data Mining Tool for Academic Data Exploitation

Webtool description and usage

U. Spagnolini, L. Fontana, A. Paganoni, A. Torrebruno,,
M.A. Prada , M. Domínguez, A. Morán, R. Vilanova, J.
Lopez Vicario, M.J. Varanda, P. Alves, M. Podpora
and M. Barbu

March 2019



Welcome to SPEET
Student Profile for Enhancing Engineering Tutoring

Speet Multiplier Event in Barcelona, 25th April 2018

The 25th of April, 2018 SPEET has organised a morning session to present the main contributions of the project obtained so far. The session is focused on **"Student Performance and drop-out prediction"** and will take place at the

Casa de Convalescencia, Aula 13
C/ Sant Antoni Maria Claret, 171
08041 Barcelona

Agenda

Find out more

LET'S COLLABORATE

Are you interested in Educational Data Mining? Drop-out prevention? Academic Data Analysis? IT Support for tutoring actions? Elaboration of student profiles? .... in any case, be in touch with us, by sure we can work out something together.

SPEET MULTIPLIER EVENT IN LEON

19th April 2018

The 19th of April, 2018 SPEET has organised a morning session to present the main contributions of the project obtained so far. "Academic Data analysis as a support for the establishment of student profiles and tutoring for engineering students" that will take place at the

Salon de Grados
Escuela de Ingenierías Industrial

ISBN 978-989-20-8739-9

Data Mining Tool for Academic Data Exploitation

Webtool description and usage

U. Spagnolini, L. Fontana, A. Paganoni, A. Torrebruno,

Scuole di Ingegneria
Politecnico di Milano
Milano, Italy

M.A. Prada , M. Domínguez, A. Morán

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

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 Braganca
Braganca, Portugal

M. Podpora

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

M. Barbu

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

Approved for public release; distribution is unlimited.

ERASMUS + KA2/KA203

Prepared for SPEET Intellectual Output #5
Under 2016-1-ES01-KA203-025452
Monitored by SEPIE

Disclaimer: 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.

Table of Contents

1 Executive Summary

The ultimate goal of SPEET project is the development of an WEB-based tool to disseminate the main intellectual output in form of user-friendly and easily accessible software tool. The WEB-tool is accessible from speet.uab.cat is intended to make accessible by other faculties and schools outside of the SPEET consortium the possibility to make data analysis on students based on the proprietary data after these are organized accordingly.

This Report on Intellectual Output 5 (IO5) is related to the Data Mining Tools for Academic Data Exploitation, and collects the overall architecture of the WEB-tool, the user manual for the use of WEB-tool and the overview of the output to establish clustering, dropout, and their dependency on students' characteristics.

Chapter 2 Covers the details on the WEB-tool architecture. The back-end side is programmed in Python to facilitate the integration of tools developed within the SPEET Projects (IO2 and IO3), and further integrations in the future. The front-end is in charge of interact with users for Students' data upload, analysis and output of the results. Security and privacy are carefully addressed, data transfer is SSL encrypted, files are removed after 48h and all legal requirements are compliant with European GDPR.

Chapter 3 details the Students' profiles format, the reorganization and preparation of the data for consistency across multiple institutions to enable fair comparisons of the analyses, and management of anomalous situations. There are 3 files are in .cvs format for the ensemble of students, in detail: general students' information (sex, age of access, admission score, etc..), their scores on all past exams, and some institution-specific data (degree program, duration, max ECTS credits, etc..).

Chapter 4 is the SPEET WEB-Tool manual. There are four parts of the tool and each one is discussed in detail. In particular: overview of SPEET project, the data uploading interface, the data analysis execution and visualization. There are a set of diagnostics that help the users to pre-process the uploaded datasets. Output are detailed in term of visualization and dropout analysis.

Chapter 5 details the outputs of the WEB-tool either in form of visualization of students' cluster (low/mid/excellent class), their score histogram representations, and classification. Dropout analysis output provides quantitative values that ease the Institution users (e.g., courses and teaching managers) to infer the reliability of the parameters impacting the students' dropout.

The WEB-tool developed within SPEET project represent a product that will remain after the end of the project to let European Universities accessing and comparing their performance metrics. Furthermore, the WEB-tool is a free-of-charge tool that contains the analyses of the SPEET institutions and another engineering institution can autonomously compare with the pre-existing ones to extract comparative analyses, or simply to investigate anomalous dropout situations.

2 Design Considerations

In this chapter, details about the webtool design are presented.

2.1 Webtool Architecture

In Fig. 1, the architecture considered for the SPEET webtool is presented. As observed, this architecture considers different technologies for the Back End and Front End parts of the web-service. At the Back End side, web application is programmed in Python to facilitate the integration of tools developed at the Project (at Intellectual outputs 2 and 3). More specifically, this is done by means of a Flask micro-framework over a Nginx reverse server and a Gunicorn server executing Flask. At the Front End part, HTML5, JS and CSS3 technologies are considered. The request/response functionalites are implemented by means of the HTTP protocol with SSL encryption (HTTPS). HTTP/2 is also implemented for those browsers allowing it (HTTP/2 that is faster and more powerful than the usual HTTP/1.1).

In next subsections, we provide further details about the design considerations.

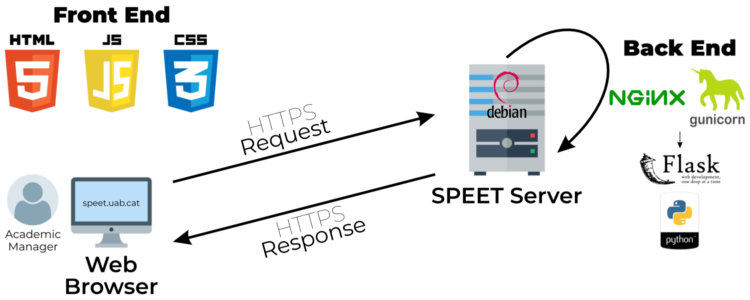


Figure 1. SPEET webtool architecture.

2.1.1 Front End Technologies

The Front End refers to the application that the user interacts it in its web browser. More specifically, the Front End application is in charge of allowing user to upload Student's data for their analysis and, also, showing results obtained by the SPEET data processing engine at the server side (Back End) work. In this project, state-of-the art technologies have been considered for the Front End design:

- HTML5 (`www.w3.org`): This markup language is considered for structuring and presenting the contents of the web-service to the user .
- JS (`www.javascript.com`): This high-level programming language is considered to provide interactivity to the page.
- CSS3 (`www.w3.org`): This style sheet language is considered to provide the style (visual and aural) to the webpage.

2.1.2 Back End Technologies

The Back End refers to the application executed at the server side. As commented in the previous sub-section, the Back End is in charge of implementing the data processing algorithms aimed at analyzing Student's data uploaded by the user. Results of this analysis generate the HTTPS responses to the Front End side to show obtained patterns and results. Since data processing tools were previously generated in Python, the following configuration is considered:

- Nginx (`www.nginx.com`): This web server technology is considered to serve the static files (mainly CSS and JS).
- Gunicorn (`gunicorn.org`): This Python HTTP server executes the main application (Flask) and serves the rest of the files of the SPEET webtool.
- Flask (`flask.pocoo.org`): This micro web framework is in charge of running the Python-written data processing algorithms. It is also in charge of generating the HTML files to present results generated from the data analysis.

2.2 Software Architecture

In the previous Section, we have presented the design of the SPEET webtool. Here, we provide details about the software architecture. More specifically, the application has been programmed using the Model-View-Controller (MVC) paradigm [LR01] represented in Fig.2. As observed in the figure, a router is in charge of receiving all the requests to the webtool to redirect them to a controller. The controller, on the other hand, calls the model when new information from the from the server's database is needed or new results from the data-processing algorithms are required. After executing the model the controller calls the view to show obtained results to the user at web-site (basically, the HTML file).

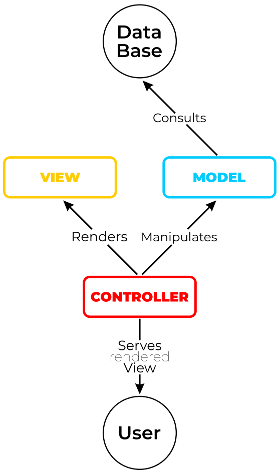


Figure 2. Model-View-Controller scheme.

As a representative example to understand how the MVC model is implemented, we show the procedure carried out when the user (at the Front End) perform a request to the Back End (e.g., `https://speet.uab.cat/profilewebsite`) in Fig. 3. There, files with extensions .py and .html are presented to differentiate Python-based and HTML-based codes, respectively.

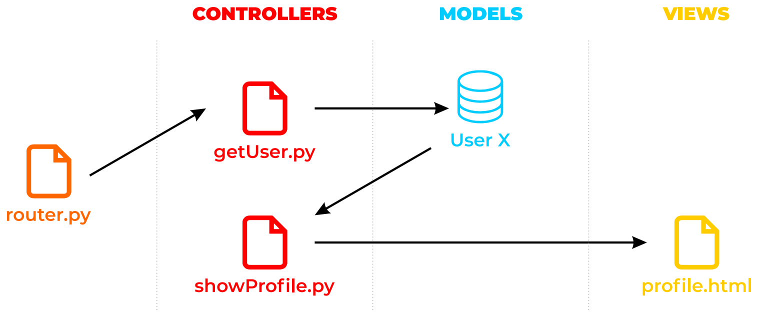


Figure 3. MVC SPEET's files scheme for https://speet.uab.cat/profile request.

It is worth noting that the software application developed in this project consist of 250 files. Indeed, the application has more than thirty controllers that manage static file serving, tool executions, serving views, etc. Due to security issues, neither the code nor the complete software architecture is provided. However, a simplified version of the structure is presented In Fig.4. When the user upload Students's information (in three CSV files as presented in Chapter 3), the following steps are performed:

- Data preparation: uploaded Students's data is pre-processed to be prepared for data algorithms (dataPreparation.py file). More details about this procedure are presented in Section 2.3. Indeed, new dataframes are generated from the uploaded files (SubjectsFrame.csv, GeneralFrame.csv, DropoutFrame.csv and CoordinatedViewsFrame.csv) and ShowExecute.py is called to address the required data processing tool.

- Algorithms execution: as commented, considered algorithms are those developed in Intellectual Output 2 (Students' Clustering, Classification and Drop-out Prediction - [VVB+18]) and Intellectual Output 3 (Coordinated Views - [PMD+18]). So, depending on user's choice, executeAlgorithm.py (Intellectual Output 2 algorithms) or showCoordinatedView.py (Intellectual Output 3) is executed.
- Outputs generation: finally, results generated by algorithms are presented in results.html and coordinated_views.html for Intellectual Output 2 and 3 cases, respectively. Further details about these results can be found in Chapter 5.

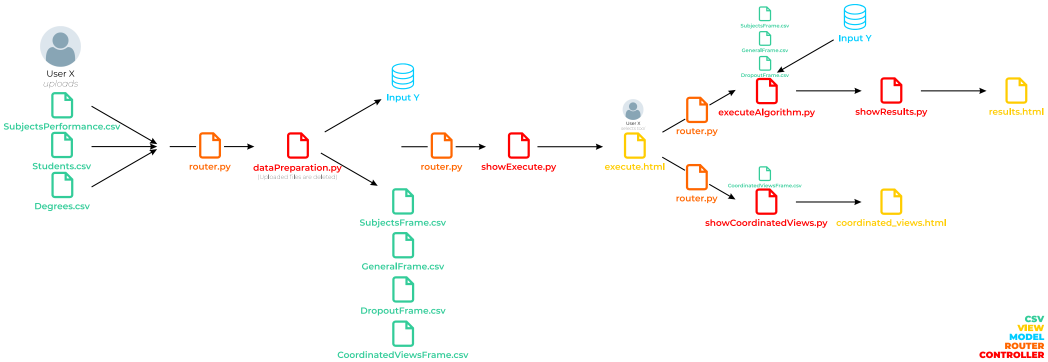


Figure 4. Simplified version of SPEET software structure.

Concerning the new .csv files generated by dataPreparation.py, SubjectsFrame.csv gathers a data frame with a new student's data reallocation adopted by the Clustering and Classification mechanisms of Intellectual Output 2. DropoutFrame.csv is adopted by the Drop-out prediction functionality (additional pre-processing is required). Concerning CoordinatedViewsFrame.csv and GeneralFrame.csv files, the former is the dataframe adopted by the Coordinated Views (Intellectual Output 3) and the latter is a new file available for download in the results page (where the user can consult the integrated data-frame resulting from the data preprocessing of the uploaded information).

2.3 Data Preparation Considerations

As commented in the previous section, uploaded students information should be processed to accommodate them to the data processing algorithms of the

SPEET webtool. Also, preprocessing is also required to detect errors and/or missing data. In order to do this data preprocessing, the Pandas library ([McK11]) of Python has been considered and the following steps have been performed:

- Columns revision: the first step is based on the review of all the columns of the three .csv uploaded by the user. These columns must fulfill the required format (also provided at the SPEET web page). Three cases are addressed:
  - Obligatory columns are missing: an error message is generated and the tool is not executed.
  - Categorical columns (student age, previous studies, etc.) are missing: a warning message is generated but the tool is executed.
  - Unnecessary columns are present: these columns are discarded, a warning message is generated but the tool is executed.
- Data Homogenization: subjects scores are normalized to 0-10 numerical evaluation.
- Missing Value Imputation: this block checks the scores obtained by students at different subjects and assigns reference score values when missing values are detected. These occurrences are assumed to be done due to procedures related to the recognition of subjects from previous studies. For this reason, the value of "PASS" (numerical score equal to 5 for graduated students and 0 to students that did not finished their degrees) are adopted as reference scores. This procedure is performed when there are more than 50%+1 of valid marks for the subject. Conversely, the subject is directly discarded. Columns related to the number of ECTS of missing subjects are also filled. In this case the maximum number of that column (related to other students) is considered as reference value.
- Data Gathering: new .csv formats are generated by gathering the data required for each functionality of the webtool. As presented in the previous section, this new .csv files are: SubjectsFrame.csv, GeneralFrame.csv, DropoutFrame.csv and CoordinatedViewsFrame.csv.

2.4 Security and Privacy Protection

Finally, Security and Privacy design considerations have also been carefully addressed in this project. The most important points are:

- External Libraries: all external libraries and fonts used in the Front End are served from SPEET Server, so no connection to external server are adopted.
- SSL Encryption: the standard HTTPS (HTTP+SSL encryption) recommends a 2048 bits encryption but, in order to improve the security level, a 4096 bits encryption has been considered.
- Uploaded Files: the three CSV files uploaded by the user are removed just after data preparation (this should happen in less than 30 seconds).
- Generated Files: the files generated during data preparation and execution are deleted when user logs out or in the next 48 hours (if there is no change in the files).
- GDPR: all the required legal requirements are completed in order to comply with the European General Data Protection Regulation (GDPR).

3 Data Preparation

Webtool users should preprocess institution data to follow the template that is described in this chapter. Indeed, exploration of project partners data underlined the need to perform those preprocessing steps to reduce inconsistencies and allow application of SPEET webtool. Main sources of differences among institution data are language, variable names, availability of specific variables, scoring system or categorical variables labelling. These inconsistencies are magnified when data is provided by istitutions operating in different nations. Therefore, data preparation is necessary to reorganize available data, to align it to the aim of each visualization and its algorithms, to deal with missing data or to compute additional values whenever necessary.

Three .csv files containing data from the institution are necessary to use the webtool. The files are:

- `SubjectsPerformance.csv`: it includes all the scores obtained by students across their careers. The triplet (StudentID, DegreeID, SubjectID) is expected as the KeyAttribute of the table (Table 1).
- `Students.csv`: it includes student information that is available as soon as the student enrolls in the institution. StudentID is expected as the KeyAttribute of the table (Table 2).
- `Degrees.csv`: it includes information about the degrees that are to be analyzed. Therefore, DegreeID is expected as the KeyAttribute of this table (Table 3).

To be correctly processed by the tool, those files have to comply with the following general rules:

- the encoding format has to be UTF-8;
- the column separator must be ; (semicolon);
- the decimal separator must be . (dot).

In addition, Tables 1-3 describe the variables required in the three files and their labels/ranges. An example of each file (with fictional data) is also provided.

Variable	Description	Type of variable
StudentID	student ID	string
DegreeID	degree ID	string
SubjectID	subject ID	string
SubjectName	subject name	string
SubjectYear	subject year within the degree study plan	integer [1,4]
SubjectNumberECTS	total number of ECTS of the subject	integer
SubjectScore	score obtained by the student in the subject	integer [0,10]
SubjectSemester	subject semester within the study plan	integer {1,2}
SubjectNature	subject nature; possible values are: Mandatory, Elective, Thesis, Internship	string

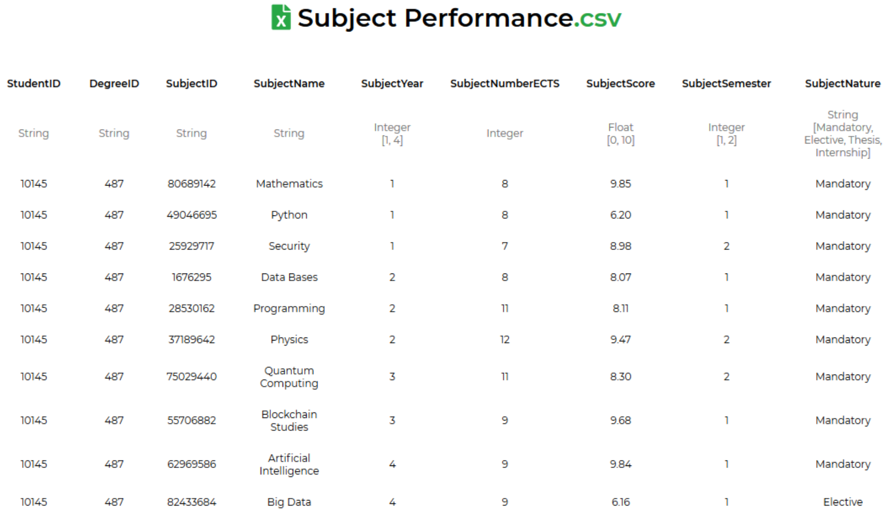
Table 1. Set of variables included in **SubjectsPerformance.csv**

Variable	Description	Type
StudentID	student ID number	string
DegreeID	degree programme attended by student ID	string
Sex*	gender	string [M,W]
AccessToStudiesAge*	age at the beginning of the studies	integer
Nationality*	nationality	factor
PreviousStudies*	type of high school studies before attending university	string
AdmissionScore*	admission test result	float [0,10]
Status	career status at the time of data collection (Active, Graduated, Dropout)	string [A,G,D]

Table 2. Set of variables included in **Students.csv**. *: optional variables

Variable	Description	Type
DegreeID	degree programme ID	string
Istitution	name of the institution organising the programme	string
DegreeNature	degree study programme (e.g. Mechanical Engineering, . . .)	string
DegreeNumberECTS	total number of ECTS of the programme	integer
DegreeYears	duration of the degree programme	integer

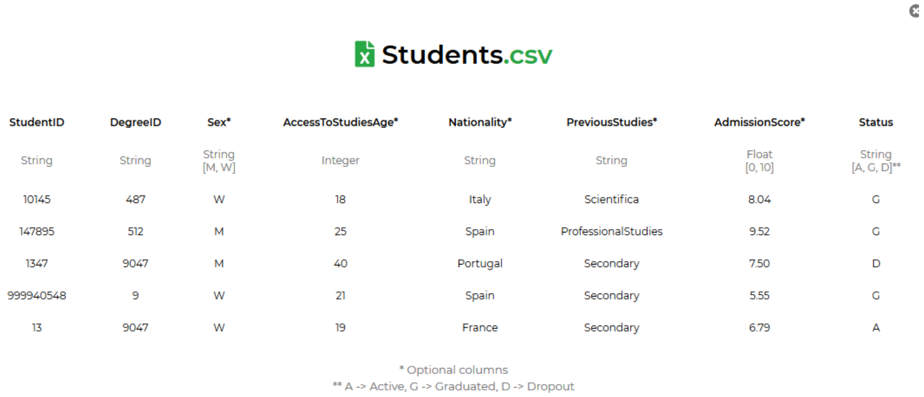
Table 3. Set of variables included in **Degrees.csv**



Subject Performance.csv

StudentID	DegreeID	SubjectID	SubjectName	SubjectYear	SubjectNumberECTS	SubjectScore	SubjectSemester	SubjectNature
String	String	String	String	Integer [1, 4]	Integer	Float [0, 10]	Integer [1, 2]	String [Mandatory, Elective, Thesis, Internship]
10145	487	80689142	Mathematics	1	8	9.85	1	Mandatory
10145	487	49046695	Python	1	8	6.20	1	Mandatory
10145	487	25929717	Security	1	7	8.98	2	Mandatory
10145	487	1676295	Data Bases	2	8	8.07	1	Mandatory
10145	487	28530162	Programming	2	11	8.11	1	Mandatory
10145	487	37189642	Physics	2	12	9.47	2	Mandatory
10145	487	75029440	Quantum Computing	3	11	8.30	2	Mandatory
10145	487	55706882	Blockchain Studies	3	9	9.68	1	Mandatory
10145	487	62969586	Artificial Intelligence	4	9	9.84	1	Mandatory
10145	487	82433684	Big Data	4	9	6.16	1	Elective

Figure 5. Example of **SubjectsPerformance.csv**

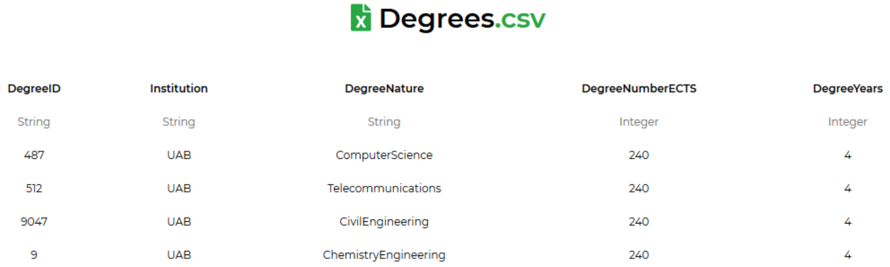


Students.csv

StudentID	DegreeID	Sex*	AccessToStudiesAge*	Nationality*	PreviousStudies*	AdmissionScore*	Status
String	String	String [M, W]	Integer	String	String	Float [0, 10]	String [A, G, D]**
10145	487	W	18	Italy	Scientifica	8.04	G
147895	512	M	25	Spain	ProfessionalStudies	9.52	G
1347	9047	M	40	Portugal	Secondary	7.50	D
999940548	9	W	21	Spain	Secondary	5.55	G
13	9047	W	19	France	Secondary	6.79	A

* Optional columns
** A -> Active, G -> Graduated, D -> Dropout

Figure 6. Example of **Students.csv**



Degrees.csv

DegreeID	Institution	DegreeNature	DegreeNumberECTS	DegreeYears
String	String	String	Integer	Integer
487	UAB	ComputerScience	240	4
512	UAB	Telecommunications	240	4
9047	UAB	CivilEngineering	240	4
9	UAB	ChemistryEngineering	240	4

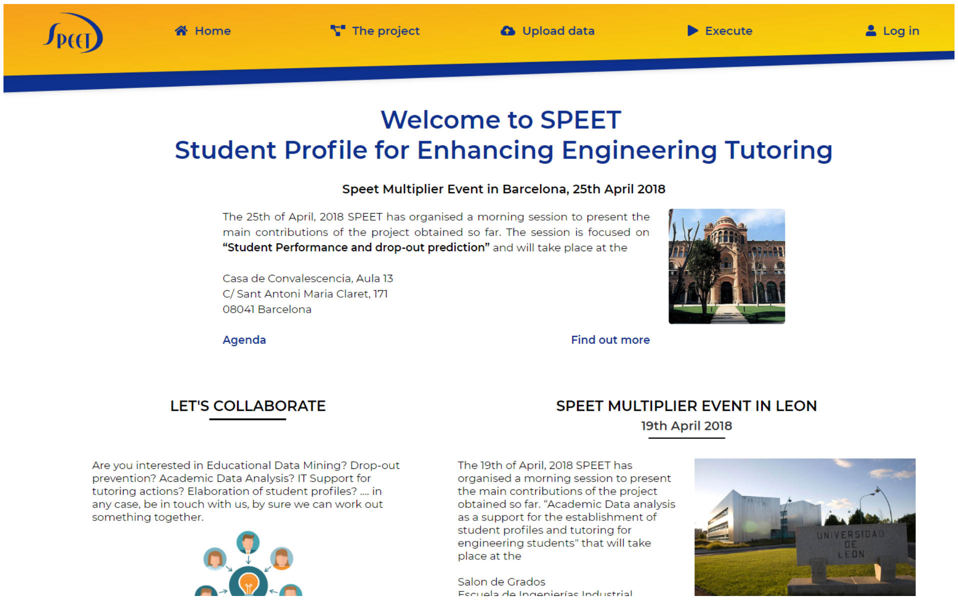
Figure 7. Example of **Degrees.csv**

4 SPEET Webtool manual

4.1 Overview

The SPEET webtool is accessible online at speet.uab.cat website. From the site homepage, a user can navigate the four parts of the tool website (screenshot is in Fig. 8 below):

- The Project: it includes a summary of the SPEET project: goals, members, and an example of the tool output;
- Upload Data: one can upload the specific data and this step is available after completing the access by credentials;
- Execute: run the analysis and specific visualization on uploaded data. This step is available after completing the access by credentials;
- Log In: it allows to register the user (or institution) on the webtool or to log in.



Welcome to SPEET
Student Profile for Enhancing Engineering Tutoring

Speet Multiplier Event in Barcelona, 25th April 2018

The 25th of April, 2018 SPEET has organised a morning session to present the main contributions of the project obtained so far. The session is focused on **"Student Performance and drop-out prediction"** and will take place at the

Casa de Convalescencia, Aula 13
C/ Sant Antoni Maria Claret, 171
08041 Barcelona

Agenda Find out more

LET'S COLLABORATE

Are you interested in Educational Data Mining? Drop-out prevention? Academic Data Analysis? IT Support for tutoring actions? Elaboration of student profiles? .... in any case, be in touch with us, by sure we can work out something together.

SPEET MULTIPLIER EVENT IN LEON
19th April 2018

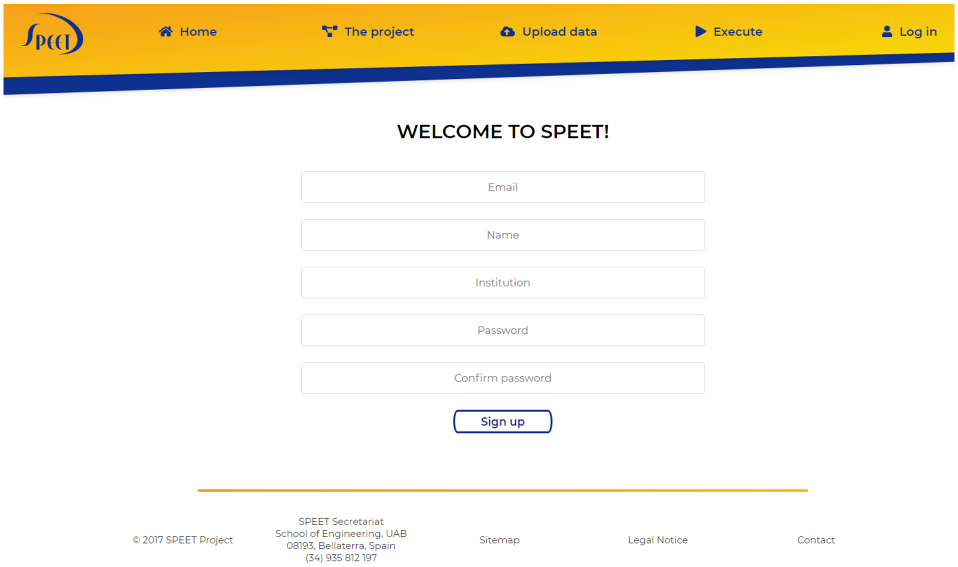
The 19th of April, 2018 SPEET has organised a morning session to present the main contributions of the project obtained so far. "Academic Data analysis as a support for the establishment of student profiles and tutoring for engineering students" that will take place at the

Salon de Grados
Escuela de Ingenierías Industrial

Figure 8. SPEET tool Homepage.

4.2 Creating a new account

A new user can create a new account by completing the registration form, accessible from the Log In dropdown menu. NOTE: it is advised to use Google Chrome browser to guarantee the full compatibility. Other browsers might have limitations as not tested.



WELCOME TO SPEET!

© 2017 SPEET Project

SPEET Secretariat
School of Engineering, UAB
08193, Bellaterra, Spain
(34) 935 812 197

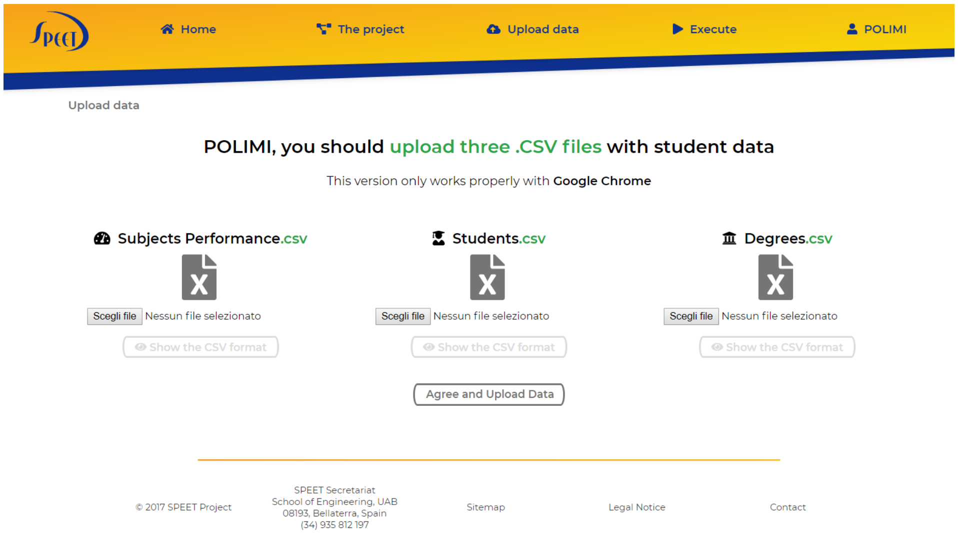
Figure 9. New user Registration page.

4.3 Log In/Upload Data

After logging in the tool website, the user is redirected on the Upload Data page illustrated below by the screenshot. Here one can upload three files in .CSV format containing data from the institution. The files are:

- `SubjectsPerformance.csv`: here, the user is expected to upload a table that includes all the scores obtained by students across their careers. The triplet (StudentID, DegreeID, SubjectID) is expected as the KeyAttribute of the table.
- `Students.csv`: this file includes student personal information that is available as soon as the student enrolls in the institution. StudentID is expected as the KeyAttribute of the table.
- `Degrees.csv`: here, the user must upload a table including information about the degrees that is required to enable the analysis. Therefore, DegreeID is expected as the KeyAttribute of this table.

For each file, an example (with fictional data) is provided by clicking on the corresponding "Show the CSV format" button. Those examples detail which columns are optional (such as personal information about the student - Sex, Age, Nationality). For a more detailed description of the three .CSV files one is referred to Chapter 3 of this document.



POLIMI, you should upload three .CSV files with student data

This version only works properly with Google Chrome

Subjects Performance.csv

Scegli file Nessun file selezionato

Show the CSV format

Students.csv

Scegli file Nessun file selezionato

Show the CSV format

Degrees.csv

Scegli file Nessun file selezionato

Show the CSV format

Agree and Upload Data

© 2017 SPEET Project

SPEET Secretariat
School of Engineering, UAB
08193, Bellaterra, Spain
(34) 935 812 197

Sitemap

Legal Notice

Contact

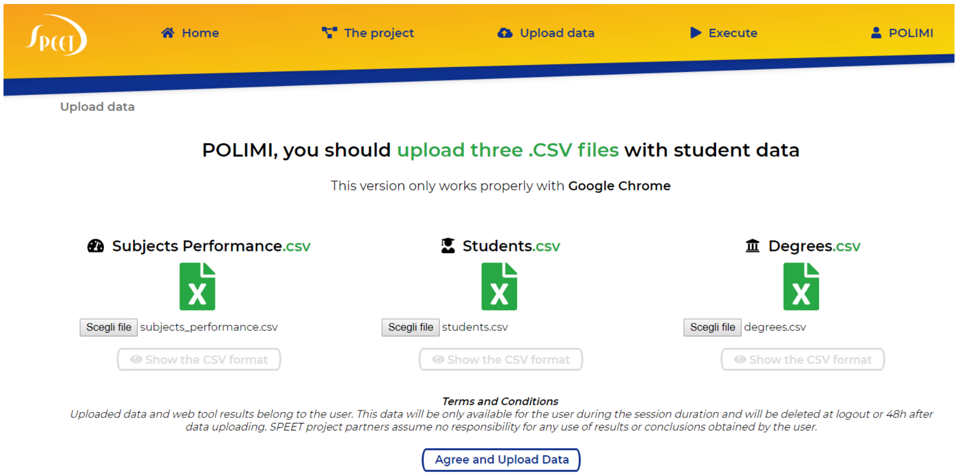
Figure 10. Upload Data page - before the upload.

4.3.1 Terms and Conditions

As soon as the user selects the three files to upload, the site shows Terms and Conditions. These must be accepted to upload data and show results. Specifically:

> Uploaded data and webtool results belong to the user. This data will be only available for the user during the session duration and will be deleted at logout or 48h after data uploading. SPEET project partners assume no responsibility for any use of results or conclusions obtained by the user.

Therefore, all data is fully available during the session in which the user uploads it. It is deleted either at the end of the session or after two days the tool is left unused, in case the user does not explicitly logs out of the website.



POLIMI, you should upload three .CSV files with student data

This version only works properly with **Google Chrome**

Subjects Performance.csv

Students.csv

Degrees.csv

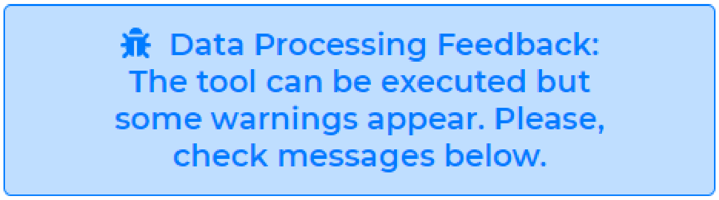
Terms and Conditions

Uploaded data and web tool results belong to the user. This data will be only available for the user during the session duration and will be deleted at logout or 48h after data uploading. SPEET project partners assume no responsibility for any use of results or conclusions obtained by the user.

Figure 11. Upload Data page - Terms and Conditions.

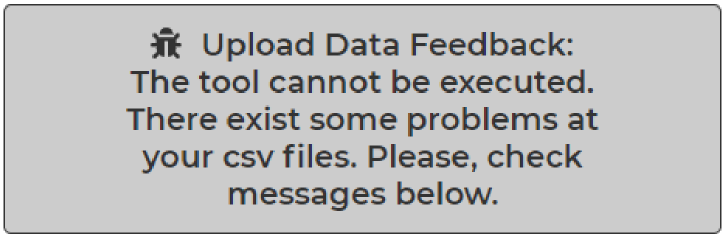
4.4 Execution

After uploading data, the user is redirected on the Execute page. Here, the page provides a Data Processing Feedback about the data preprocessing that is performed before the execution. The preprocessing may fail, may have warnings or may be fully completed. If some warnings appear (12), the tool can still be executed. However, the analysis might be improved after the user checks the suggestions provided by the page. In case the preprocessing fails (13), the user must upload a new dataset. The package provides suggestions on how to fix major errors.



Data Processing Feedback: The tool can be executed but some warnings appear. Please, check messages below.

Figure 12. Example of warning message.



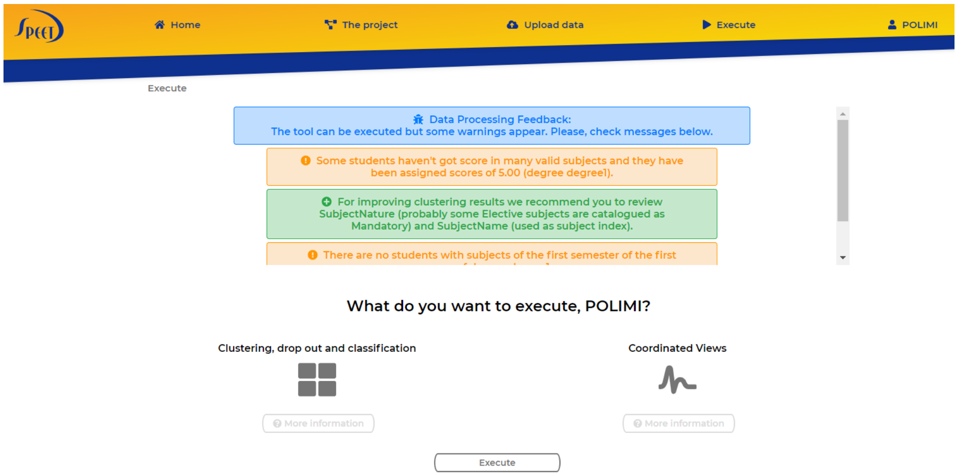
Upload Data Feedback: The tool cannot be executed. There exist some problems at your csv files. Please, check messages below.

Figure 13. Example of error message.

Once the preprocessing is completed, the user can choose to execute one of the two tools available by clicking on the corresponding icon (Figure 14):

- Clustering, drop out and classification
- Coordinated views

More information about each tool is available under the corresponding icon. In case previous results have already been executed during the same session, the user can also explore them on the Execute page or get back to the Upload Data page to upload a new dataset.



Data Processing Feedback:
The tool can be executed but some warnings appear. Please, check messages below.

Some students haven't got score in many valid subjects and they have been assigned scores of 5.00 (degree degree1).

For improving clustering results we recommend you to review SubjectNature (probably some Elective subjects are catalogued as Mandatory) and SubjectName (used as subject index).

There are no students with subjects of the first semester of the first

What do you want to execute, POLIMI?

Figure 14. Execution of the tools is now possible. On top, there are some warning messages and the corresponding suggestions to fix them.

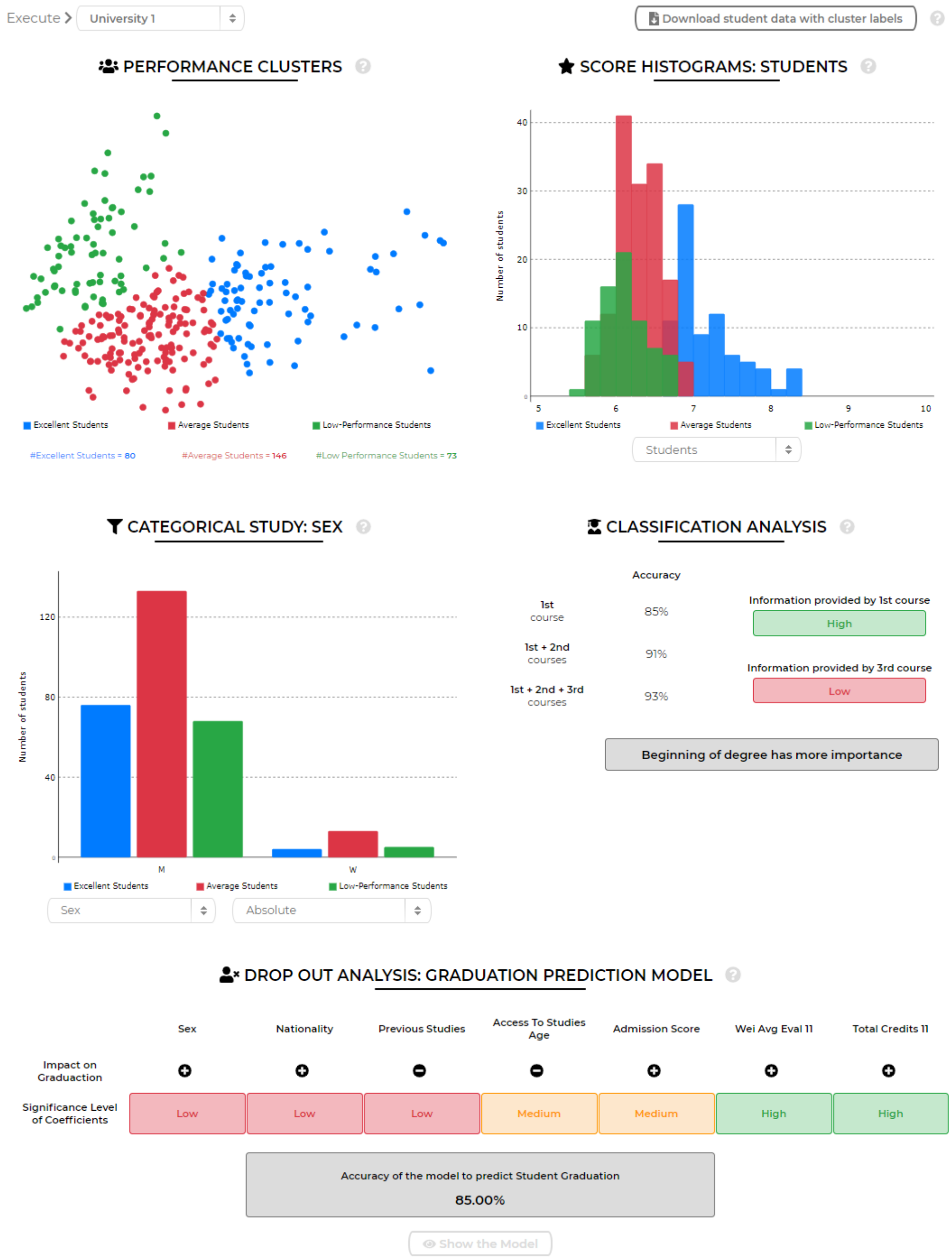
4.4.1 Clustering, Dropout and Classification

Multiple analyses are carried out for homogeneous dataset referred to one course of one degree. Therefore, in this degree-by-degree analysis the user must choose a degree from the dropdown selection before any execution. After one degree is executed, the user can select a different degree from the corresponding dropdown selection. This tool provides five different visualizations, as in Figure 15:

- Performance Clusters
- Scores Histograms (users can interact with this visual by choosing the reference variable - Students or Subject)
- Categorical Study (users can interact with this visual by choosing both the categorical dimension and the normalization setting)
- Classification Analysis

- Dropout Analysis: Graduation Prediction Model

A recap of this tool graphical outputs is included in Section 5.1. The full presentation of the methodologies of these analysis is presented in [VVB+18].



	Accuracy
1st course	85%
1st + 2nd courses	91%
1st + 2nd + 3rd courses	93%

Information provided by 1st course: High

Information provided by 3rd course: Low

Beginning of degree has more importance

	Sex	Nationality	Previous Studies	Access To Studies Age	Admission Score	Wei Avg Eval 11	Total Credits 11
Impact on Graduaction	+	+	−	−	+	+	+
Significance Level of Coefficients	Low	Low	Low	Medium	Medium	High	High

Accuracy of the model to predict Student Graduation

85.00%

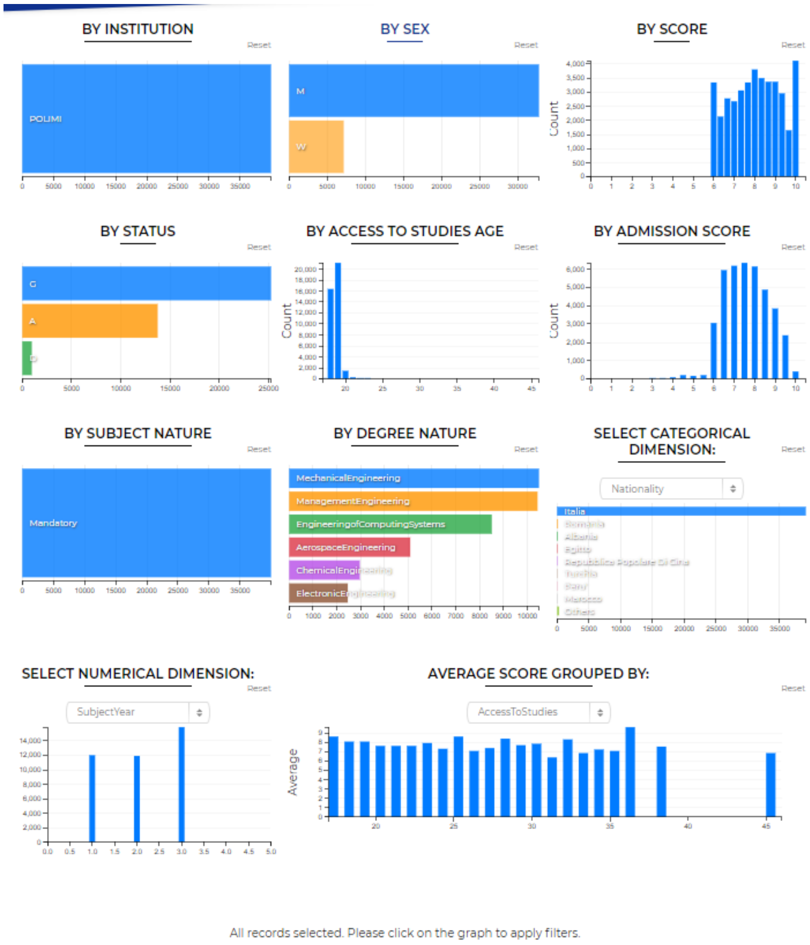
Figure 15. Example of execution output of the Clustering, Drop out and Classification Tool

4.4.2 Coordinated Views

This tool analyzes the whole data uploaded in the session. The statistical unit is a single student-subject interaction (exam score). The distribution of exam scores is visualized across different variables using coordinated histograms and barplots. In addition, the distribution of the average score across a single variable (to be selected from a dropdown) is shown in the bottom right panel (Figure 16).

The user can interact with Coordinated Views in different ways:

- hovering over a barplot column to check the corresponding relative frequency;
- applying a filter by a categorical variable, by clicking on a barplot column (multiple selection is possible, holding CTRL key);

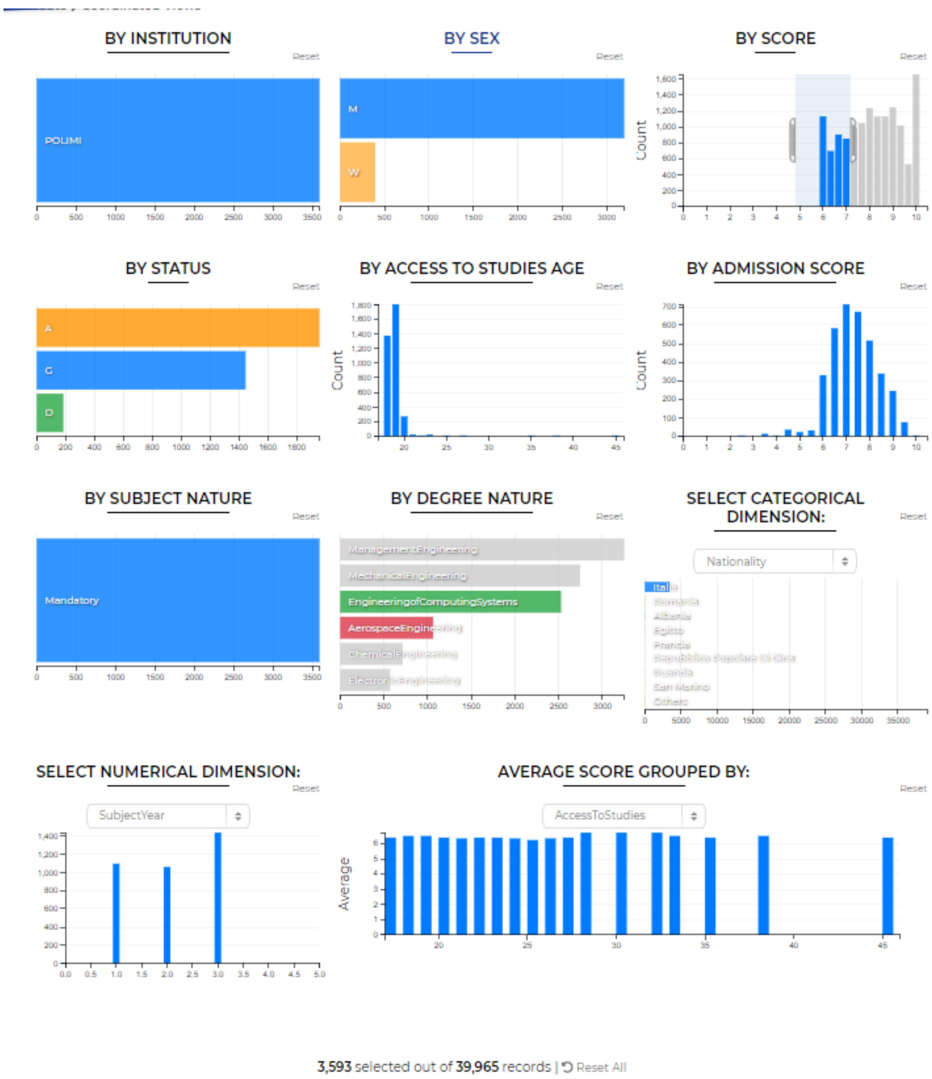


All records selected. Please click on the graph to apply filters.

Figure 16. Example of execution output of the Coordinated Views Tool

- applying a filter by a numerical variable, by selecting a range on a histogram (double-click and drag horizontally over the columns within the desired range).

After applying a filter, the visual automatically updates across all its panels as illustrated in Figure 17. The number of filtered units is reported at the bottom of each visual. A single filter can be reset from its corresponding visual, while all filters can be reset in one shot at the bottom of the visual. A recap of this tool graphical outputs is included in Section 5.2. The full presentation of the methodologies of these analysis is presented in [PMD+18].



3,593 selected out of 39,965 records | Reset All

Figure 17. Example of a filtered the Coordinated Views Tool. A categorical filter is applied on DegreeNature and a continuous filter is applied on Score.

5 Detailed tool outputs

This chapter describes in detail the different visuals that are produced by the SPEET webtool for visual analytics applied to the academic data. An example of output is provided for each visualization, along with the insight it produces. All the visualizations shown in this chapter are based on fictional data.

5.1 Clustering, Dropout and Classification

5.1.1 Performance Clusters

This visual is in charge of representing the three groups of students generated by the Clustering Block. Groups are generated based on the performance results obtained by students in terms of subjects scores. Groups generated are: Excellent Students, Average Students and Low-Performance students. Representation is based on a 2D dimensional reduction to facilitates the visual interpretation. Clustering algorithm adopted is based on the k-means approach. Further details can be found at the Intellectual Output 2 report (available at the SPEET website).

In Fig. 18, we present an example of the Performance Clusters plot generated by the webtool. This figure shows how the students have been organized in three clusters: Blue cluster (Excellent), Red cluster (Average) and Green (Low-Performance).

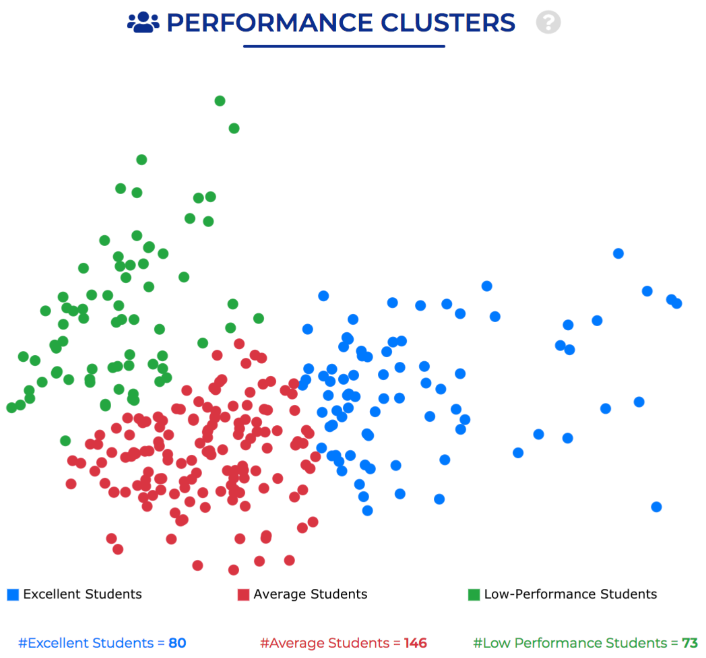


Figure 18. Example of output from the Performance Clusters visual.

5.1.2 Score Histograms

This visual shows a Histogram-based representation of the subjects scores by taking into account two possible choices: Subjects-based and Students-based. The Subjects-based approach considers a given cluster and averages all the scores for each subject.The Students-based option takes all the students belonging to a cluster and computes the average score of all the subjects of each student.

In Fig. 19 and 20 we provide examples of both possible Score Histograms: by Students and by Subjects. In the Students case, one can observe how Excellent Students (blue) tend to have the higher average scores. In this specific case (the toy example provided at the website), Low-performance and Average students present some overlap. The possible explanation is that Low-performance students can have a similar or better performance than Average students in a set of subjects and viceversa. This effect can be clearly observed when the Subjects is activated (see left part of the figure).

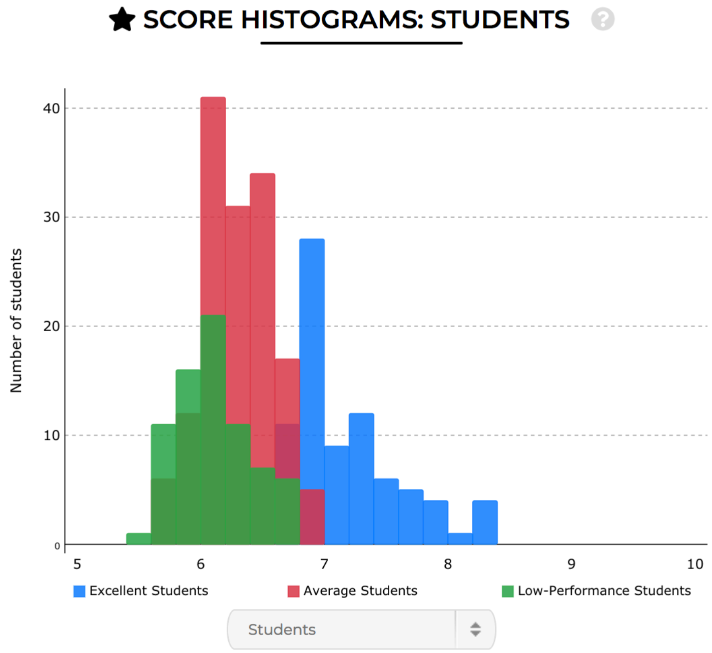


Figure 19. Example of output from the Score Histograms visual (Students-based option).

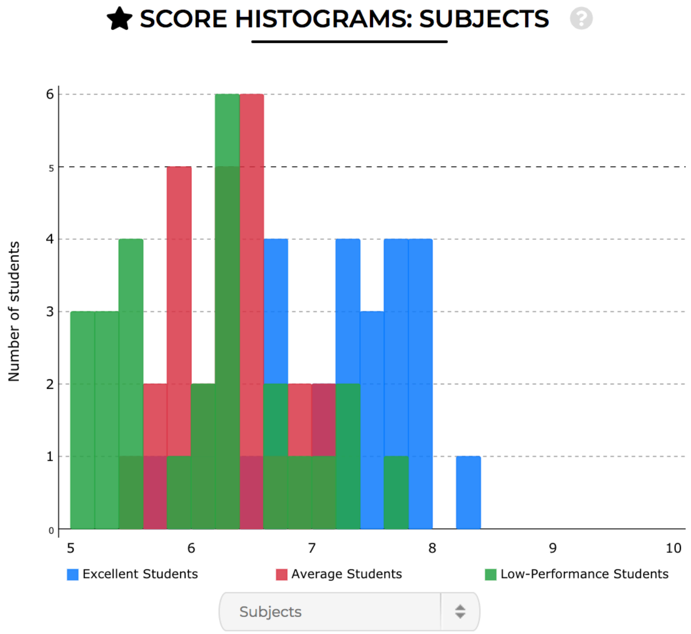


Figure 20. Example of output from the Score Histograms visual (Subjects-based option).

5.1.3 Categorical Study

This visual is based on the generation of Histograms to analyze the patterns of students at different clusters. More specifically, these patterns are analyzed by considering a set of categorical variables: Sex, Previous Studies, Admission Score, Access to Studies Age and Nationality. Different modes are available for the Histograms representation in this case:

- Absolute: Each column represent the absolute value of students belonging to each cluster and categorical variable.
- Normalized by cluster: Each column represents the percentage of students of a given cluster belonging to a categorical variable. By adding all the columns belonging to the same cluster (columns with the same color), the 100% value is obtained.
- Normalized by categorical: Each column represents the percentage of students of a given categorical variable belonging to a cluster. By adding all the columns belonging to the same categorical variable (e.g., in Sex categorical, all the columns with Sex equal to Female), the 100% value is obtained.
- Full normalized: Each column represents the percentage of students belonging to a given categorical variable/cluster pair. By adding all the columns of the representation, the 100% value is obtained.

In Fig. 21, a visualization example is presented where the categorical Sex is selected. One can also notice how the four representations options can be selected, being the Absolute one selected in this case. In Fig. 22, we provide an example by selecting PreviousStudies as categorical option and visualization based on Normalized by cluster option. As observed, a high percentage of Excellent Students (blue) comes from Secondary in this case.

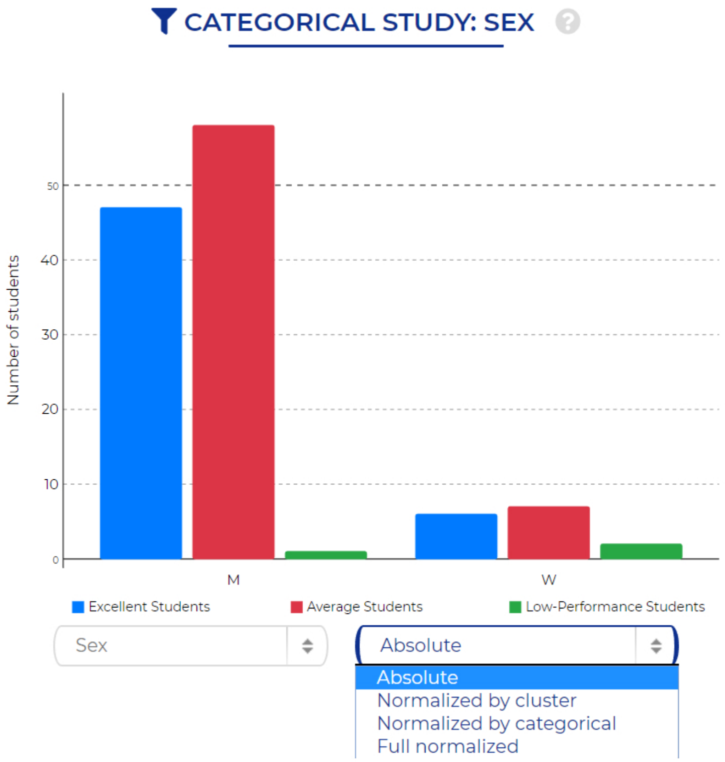


Figure 21. Example of output from the Categorical Study visual.

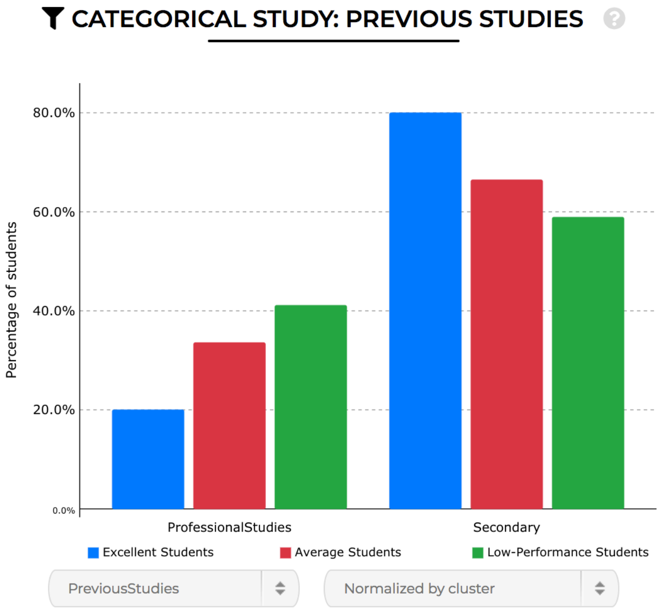


Figure 22. Example of output from the Categorical Study visual based on Normalized by Cluster representation.

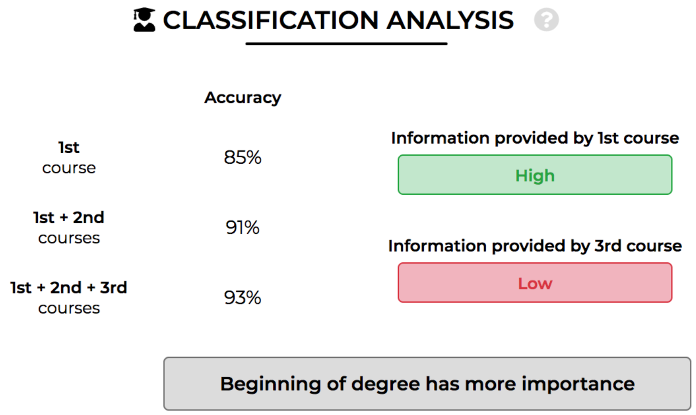
5.1.4 Classification Analysis

This visual shows the Classification results obtained when a new student is classified into existing groups (i.e., Excellent, Average and Low-performance students). Classification algorithm is based on the SVM approach. Further details can be found at the Intellectual Output 2 report (available at the SPEET website). The classifier is trained with 80% of available data and tested with 20% of data. Classification accuracy results are based on this 20% of test data.

Classification is performed by taking into account the performance in terms of subjects score. Three classifiers are implemented based on the amount of available information: 1) only the first course subjects scores, 2) first+second courses scores and 3) first+second+third courses scores. By taking into account, the accuracy differences between these two options, the tool also shows the amount of information provided by the 1st course (high when a significant level of accuracy can be obtained with 1st course results) and the amount of information provided by the 3rd course (high when accuracy is significantly increased when the 3rd course results are included).

In Fig. 23, we provide an example of classification. As observed, 85%, 91% and 93% accuracies are obtained for the 1st, 1st+2nd and 1st+2nd+3rd, respectively. This shows that students can be classified into performance groups from the first course with a high accuracy.

Besides, classification analysis can also be adopted to analyze the course-dependency behavior of students at the different degrees. In this case, it is observed that the first course is very important. In other words, the classification accuracy obtained when analyzing only the subjects at the first course 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.



CLASSIFICATION ANALYSIS

	Accuracy
1st course	85%
1st + 2nd courses	91%
1st + 2nd + 3rd courses	93%

Information provided by 1st course

High

Information provided by 3rd course

Low

Beginning of degree has more importance

Figure 23. Example of output from the Classification Analysis visual.

5.1.5 Dropout Analysis: Graduation Prediction Model

This visual analyzes the differences between two different student profiles: dropout students (D) and graduate students (G). Active students are not considered here. Therefore, the career status is turned into a binary variable (graduate = 1, dropout = 0). The tool explores the relationship between a set of input variables and the career status (binary) through a Logistic Mixed Model. In order to build a model that is useful for prediction, input variables include those available at the time of the enrollment and those recorded after the first semester of the first year of study.

An example of this visual is shown in Figure 24. On top, the list of input variables is reported. If an input variable is categoric, the visual reports its levels apart from one that is used as reference level.

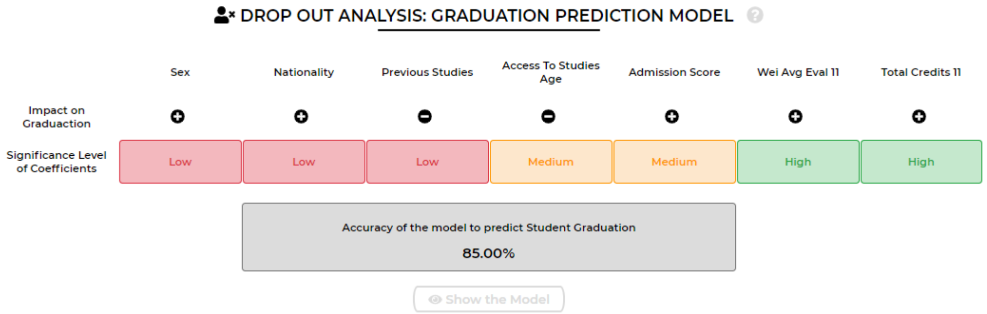
In the middle of the panel, the tool shows, for each input variable, the impact on graduation probability (positive or negative) and the level of significance (low, medium or high) of the variable. In the example:

- `Sex[T.W]` has positive impact: we could say female students perform better than male ones. Hovewer, the significance level is low: this difference is small and the student's gender is not useful to predict graduation.
- `AccessToStudiesAge` has a negative impact: the higher the age at the time of the admission, the lower the probability of graduation. The significance level is medium: this variable improves the prediction of graduation.
- `WeiAvgEval.1.1` has a positive impact: the higher the weighted average in the first semester, the higher the probability of graduation. The signif-

icance level is high: this variable strongly impacts graduation probability and it is really important to predict it.

Similar considerations can be made for the other variables.

The model is build using only a portion of the student data (80% of the students). The remaining data is used to assess the accuracy of the model in predicting student graduation. This value is reported at the bottom of the visual. In the example, the career status is correctly predicted for 85.0% of the students (definitely a good result).



DROP OUT ANALYSIS: GRADUATION PREDICTION MODEL

	Sex	Nationality	Previous Studies	Access To Studies Age	Admission Score	Wei Avg Eval 11	Total Credits 11
Impact on Graduaction	⊕	⊕	⊖	⊖	⊕	⊕	⊕
Significance Level of Coefficients	Low	Low	Low	Medium	Medium	High	High

Accuracy of the model to predict Student Graduation

85.00%

Figure 24. Example of output from the Dropout Analysis: Graduation Prediction visual

5.2 Coordinated Views

This tool analyzes all student-subject interaction (exam score) from data uploaded in the current session. The distribution of exam scores is visualized across different variables using coordinated histograms and barplots. In addition, the user can apply a custom set of filters to the visual by interacting with it.

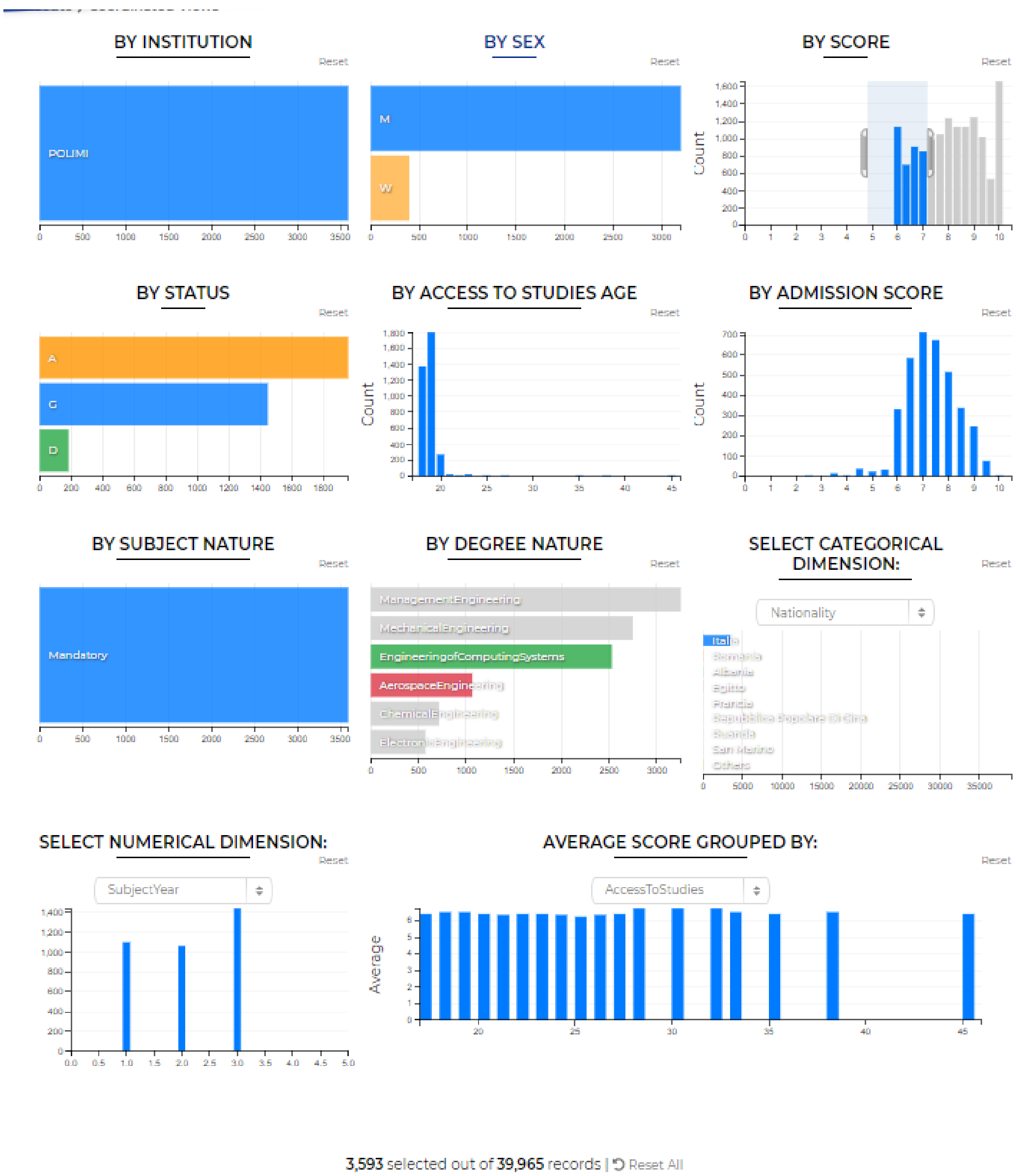
Available numerical variables are:

- Score
- Access to Studies Age
- Admission Score
- Subject Year (from dropdown)
- Subject Semester (from dropdown)
- Subject ECTS (from dropdown)
- Degree ECTS (from dropdown)
- Degree Year (from dropdown)

Available categorical variables are:

- Institution
- Gender
- Career Status
- Subject Category
- Degree Nature
- Previous Studies (from dropdown)
- Nationality (from dropdown)

Filtering in the categorical charts is accomplished by clicking in the horizontal bars. For the numerical ones, filtering is performed by brushing, i.e., dragging the mouse to select a range of vertical bins. For all the cases, the resulting filter is computed during these actions, so that the results driven by the user actions can be immediately perceived. An example of this visual and its filters is reported in Figure 25.



3,593 selected out of 39,965 records | Reset All

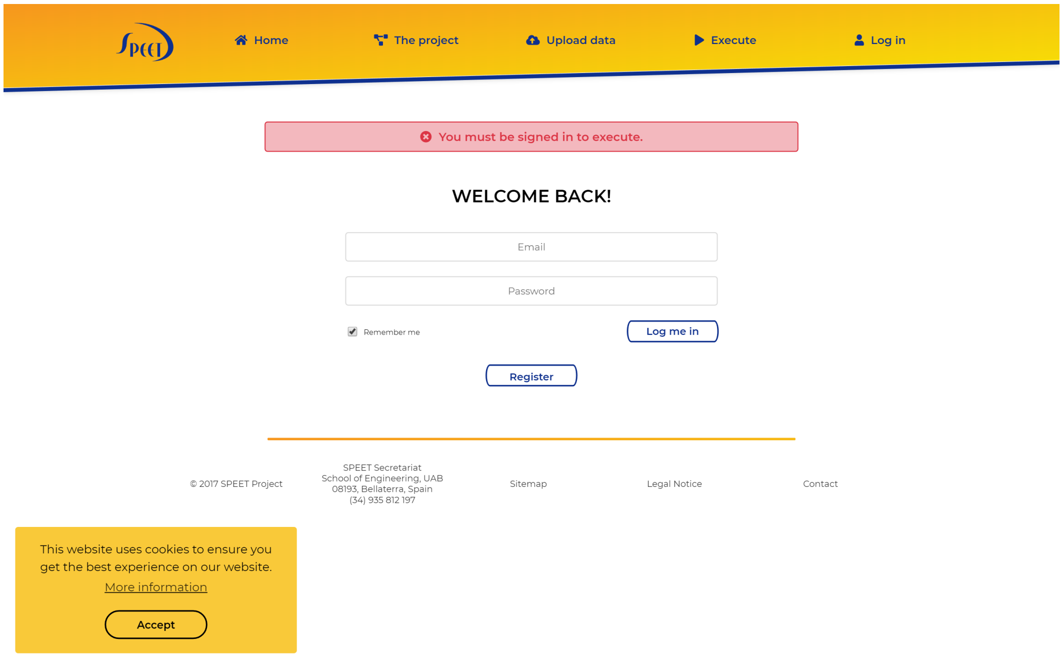
Figure 25. Example of a filtered the Coordinated Views Tool. A categorical filter is applied on DegreeNature and a continuous filter is applied on Score.

5.3 Messaging System

The SPEET webtool also implements a messaging system to help user during all the web experience. There are three types of messages: errors, warnings and solutions. Messages are represented by different color formats (errors - red, warnings - orange, and solutions - green) and are generated by the server after user's request. There are also help tips messages, which appear as a pop up when user press help buttons (marked with ? or "More information" button).

5.3.1 Message Areas

Messages can appear in the website locations presented at the next figures:



You must be signed in to execute.

WELCOME BACK!

© 2017 SPEET Project

SPEET Secretariat
School of Engineering, UAB
08193, Bellaterra, Spain
(34) 935 812 197

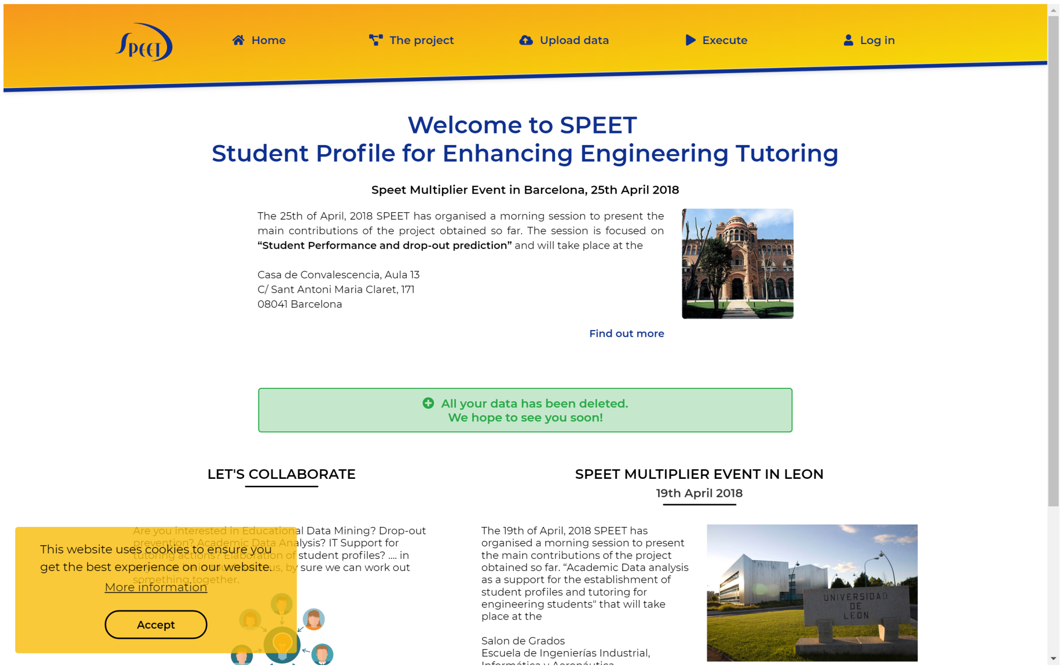
Sitemap

Legal Notice

Contact

This website uses cookies to ensure you get the best experience on our website.

Figure 26. Message related to access to areas restricted to login users or login process fails. Registration page messages are very similar.



Welcome to SPEET
Student Profile for Enhancing Engineering Tutoring

Speet Multiplier Event in Barcelona, 25th April 2018

The 25th of April, 2018 SPEET has organised a morning session to present the main contributions of the project obtained so far. The session is focused on **"Student Performance and drop-out prediction"** and will take place at the

Casa de Convalescencia, Aula 13
C/ Sant Antoni Maria Claret, 171
08041 Barcelona

All your data has been deleted.
We hope to see you soon!

LET'S COLLABORATE

Are you interested in Educational Data Mining? Drop-out prevention? Academic Data Analysis? IT Support for tutoring actions? Elaboration of student profiles? .... in [illegible] us, by sure we can work out something together.

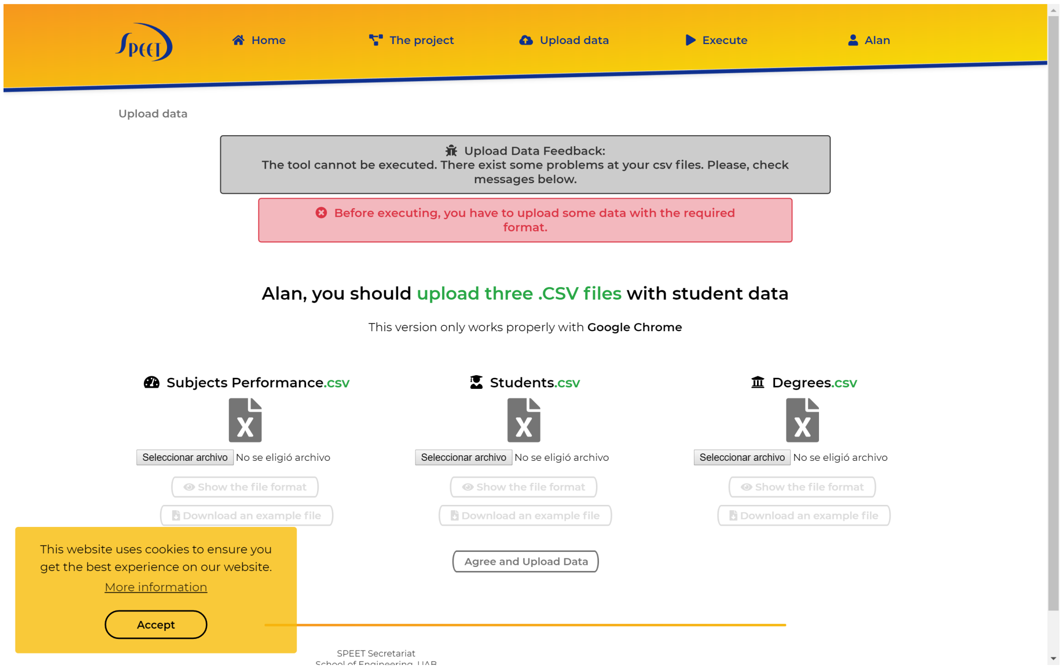
SPEET MULTIPLIER EVENT IN LEON
19th April 2018

The 19th of April, 2018 SPEET has organised a morning session to present the main contributions of the project obtained so far. "Academic Data analysis as a support for the establishment of student profiles and tutoring for engineering students" that will take place at the

Salon de Grados
Escuela de Ingenierías Industrial,

This website uses cookies to ensure you get the best experience on our website.

Figure 27. Message for website log out.



Upload data

Upload Data Feedback:
The tool cannot be executed. There exist some problems at your csv files. Please, check messages below.

Before executing, you have to upload some data with the required format.

Alan, you should upload three .CSV files with student data

This version only works properly with **Google Chrome**

Subjects Performance.csv

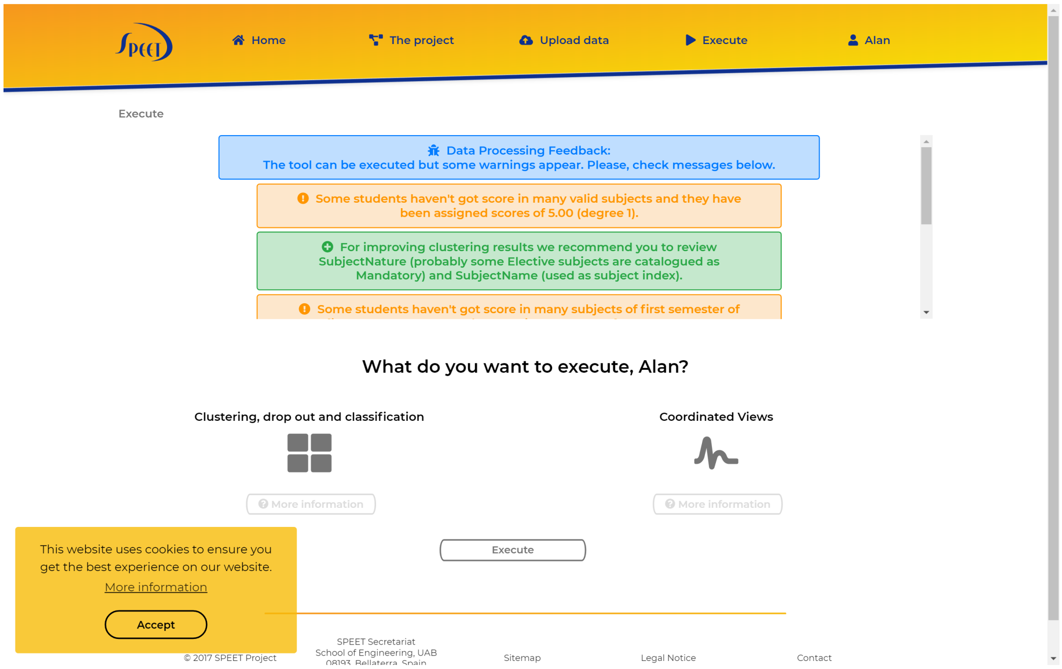
Students.csv

Degrees.csv

This website uses cookies to ensure you get the best experience on our website.

SPEET Secretariat

Figure 28. Messages related to data uploading errors.



Data Processing Feedback:
The tool can be executed but some warnings appear. Please, check messages below.

Some students haven't got score in many valid subjects and they have been assigned scores of 5.00 (degree 1).

For improving clustering results we recommend you to review SubjectNature (probably some Elective subjects are catalogued as Mandatory) and SubjectName (used as subject index).

Some students haven't got score in many subjects of first semester of

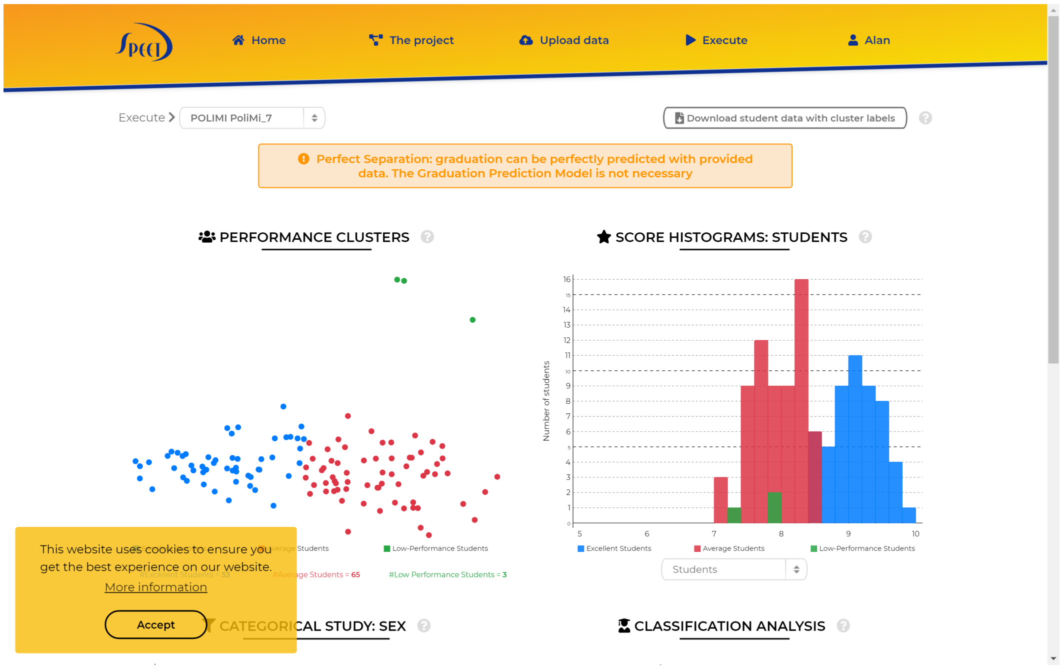
What do you want to execute, Alan?

Clustering, drop out and classification

Coordinated Views

This website uses cookies to ensure you get the best experience on our website.

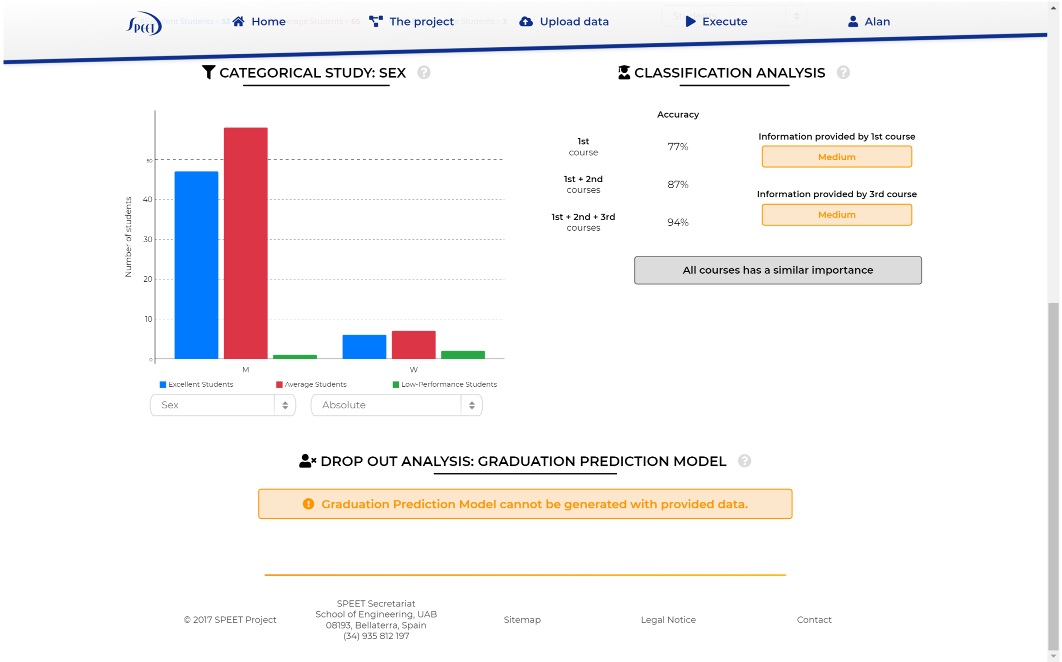
Figure 29. Messages related to tool execution errors.



Perfect Separation: graduation can be perfectly predicted with provided data. The Graduation Prediction Model is not necessary

This website uses cookies to ensure you get the best experience on our website.

Figure 30. Messages related to tool execution warnings.



CATEGORICAL STUDY: SEX

CLASSIFICATION ANALYSIS

	Accuracy
1st course	77%
1st + 2nd courses	87%
1st + 2nd + 3rd courses	94%

Information provided by 1st course: Medium

Information provided by 3rd course: Medium

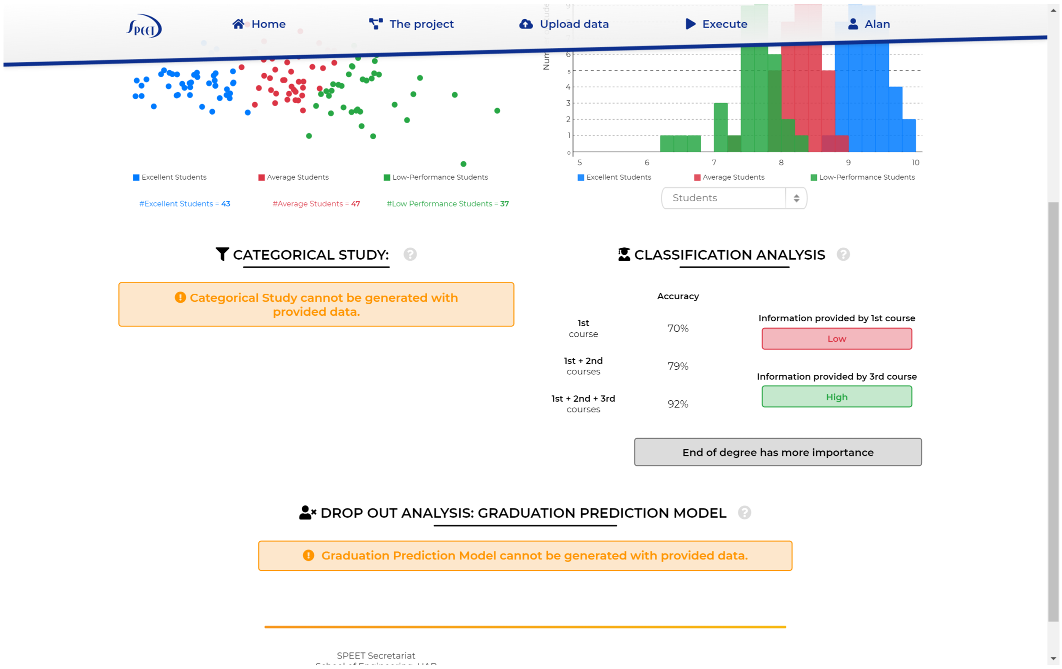
All courses has a similar importance

DROP OUT ANALYSIS: GRADUATION PREDICTION MODEL

Graduation Prediction Model cannot be generated with provided data.

SPEET Secretariat
School of Engineering, UAB
08193, Bellaterra, Spain
(34) 935 812 197

Figure 31. Message related to Drop-out prediction algorithm errors: due to a low number of categorical variables or perfect separation issue.



CATEGORICAL STUDY:

Categorical Study cannot be generated with provided data.

CLASSIFICATION ANALYSIS

	Accuracy
1st course	70%
1st + 2nd courses	79%
1st + 2nd + 3rd courses	92%

Information provided by 1st course: Low

Information provided by 3rd course: High

End of degree has more importance

DROP OUT ANALYSIS: GRADUATION PREDICTION MODEL

Graduation Prediction Model cannot be generated with provided data.

SPEET Secretariat

Figure 32. Message related to non-categorical variables error.

5.3.2 Messages list

In Table 4, we provide a list with all possible messages and recommended solutions.

Table 4. Message outputs, meanings and solutions

Message	Meaning	Possible solution
"You must be signed in to access this page."	You have no session established or it's timed out and it is required to access this page.	Log in to your account or register if you have not done it yet.
"You are not allowed to see this!"	You are trying to access some page or file that you are not allowed to.	Don't try that.
"Input the required data in the given form."	You are trying to input something by changing the HTML (or there is a bug).	Do not try that or, if it's a bug, contact to the administrator.
"Something failed during execution, try it again or contact with administrators."	It may be a bug in the execution algorithm.	Contact the administrator.
"You have exceeded maximum file size (100 MB)."	The limit of the three CSV files is 100MB (each file)	Try splitting the file by degree (just the overweighted).
"The uploaded files have to be CSV."	The uploaded file is not a CSV file.	Save the file in the required .csv format.
"You have to upload three CSV files with the required data."	You are trying to do some action without uploading the three valid CSV files required.	Upload the three CSV files with a valid format.
"Before executing, you have to upload some data with the required format."	You are trying to execute without uploading the three valid CSV files required.	Upload the three CSV files with a valid format.
"User, you are already logged in."	You are trying to sign in but you are already logged in.	Don't try to sign in.
"All your data has been deleted. We hope to see you soon!"	You are now logged out and all the data generated during the session has been removed.	It's all ok. It's important to always click on log out because if you don't the generated data will remain in the server 48 hours.
"You are already logged out."	You are trying to log out but you are already logged out.	Don't try to log out.
"Invalid email or password"	The credentials entered to log in are not correct.	Try another email and/or password.

Message	Meaning	Possible solution
"This tool is not available yet."	You are trying to execute a tool that is not available at this moment	Maybe is maintenance or a tool is being upgraded. You can ask the administrator for more information.
"Web tool cannot be executed, please review messages above and try it again or contact with administrators."	Maybe there is a bug or something happened with your session.	Log out and begin again if this message appears again contact with the administrator.
"Something failed while reading your X CSV file."	The application cannot read X CSV file.	Maybe it is because the encoding format of the file is not UTF-8. You can change the encoding, for example, with Sublime Text (open the file and in File select: Save with Encoding: UTF-8).
"Something is wrong in X CSV file. It seems like Y column is missing."	The Y column is missing in the X file.	Review columns names and that CSV separator is ; and not , or other.
"There are no students of Y degree in X CSV."	The application detects no students of Y degree in X CSV.	Check if you entered a degree by mistake in degrees.csv.
"An error happened during the conversion of C column values to type T (Y degree and X CSV)."	Is not a common error but maybe your CSV has got some cell in column C that cannot be converted to the required type T.	Check if all the cells of column C are in the required format specified in upload data examples.
"An error happened during indexing StudentID in X CSV."	The application cannot index the X CSV (by StudentID).	Check StudentID column (search something different, strange). If this does not result, contact with the administrator.
"There are no students of Y degree with marks in Mandatory subjects."	Maybe you have not got any Mandatory subject.	Review SubjectNature column of subjects performance CSV.
"There are no students of Y degree, that do subjects of more than 2 ECTS."	Maybe there is an error in SubjectName or in SubjectNumberECTS	Review SubjectNumberECTS column of subjects performance CSV.
"There are no students of Y degree in both Students and SubjectsPerformance CSVs."	Maybe the SubjectID anonymize that you applied (if you have not done it, do it) generated different StudentID in each file.	Review StudentID columns of students and subjects performance CSVs.

Message	Meaning	Possible solution
"Some students have not got score in many valid subjects and they have been assigned scores of 5.00 (Y degree)."	Check the chapter 2	For improving clustering results we recommend you to review SubjectNature (probably some Elective subjects are catalogued as Mandatory) and SubjectName (used as subject index).
"Columns C of X CSV are empty."	C columns have not got data.	Check C columns of X CSV.
"It has been impossible to save the data frame of Y degree."	Something happens while the application is trying to save the generated data frame.	If you are executing and uploading data for the same degree don't do that. If this is not your case, contact with the administrator.
"It has been impossible to save the valid subject's information of Y degree."	Something happens while the application is trying to save the generated subjects file.	If you are executing and uploading data for the same degree do not do that. If this is not your case, contact with the administrator.
"At least one column of X CSV is empty (Y degree)."	One or more columns of Y degree are empty in X CSV.	Check X CSV file.
"There are no graduate (G) or Drop Out (D) students in the Status column of Y degree in X CSV."	For executing Drop Out D and G students are required, for Clustering and Classification just G students.	Review the Status column.
"It has been impossible to save the valid information of Drop Out (degree X)."	Something happens while the application is trying to save the generated Drop Out data.	If you are executing and uploading data for the same degree do not do that. If this is not your case, contact with the administrator.
"There are no students with subjects of the first year of Y degree."	The Drop Out require students of the first year.	Review the SubjectYear column of subjects performance CSV but this is just a warning (your execution will not include Drop Out).
"There are no students with subjects of the first semester of the first year of Y degree."	The Drop Out require students of the first semester of the year.	Review SubjectSemester column of X CSV but this is just a warning (your execution will not include Drop Out).
"Some students have not got X in many subjects of the first semester of the first year and they have been assigned Xs of Y value (degree Z).	Some value, for example, the score is missing.	For improving Drop Out results it is recommended to review C column of W CSV.

Message	Meaning	Possible solution
"The scores should be normalized to [0.00, 10.00] (Y degree, X CSV."	Scores should be between 0 and 10.	Normalize the scores.
"There less than X valid Y of degree W."	A number of X values (for example 20) is required of Y (for example students).	Review your CSV files but maybe you cannot execute that W degree (sometimes this happen if there is old and new data together, with different student plan or something similar).
"The degree cannot be well processed."	The application cannot process any degree.	Review all the CSV files.
"The Drop Out Analysis of Y degree will not be generated because Status column (X CSV) is a single value. Drop Out tool requires both graduated (G) and Drop Out (D) students for working well."	For executing Drop Out D and G students are required, for Clustering and Classification just G students.	Review the Status column.
"The execution failed in X, try it again or contact with administrators."	Something failed during the execution process. It is not usual, maybe it is a bug.	Try again or contact with administrator.
"Something is wrong in X CSV file. It seems like C column is not necessary."	The application advises when you upload an unnecessary column and removes it.	Review columns names and that CSV separator is ; but it is just a warning.
"Something is wrong in your DegreeID or Institution name."	Maybe you put a character in the degree or institution fields that generates a conflict. It is not usual.	Review them and try not to use - or _.
"Something failed during the preparation of the required data for Coordinated Views."	There is an error in your data for coordinated views. That is not usual.	If there are more messages review them. If not, contact with the administrator.
"At least three categorical columns required to execute Drop Out in students CSV file."	Drop Out requires at least three categorical valid columns in students CSV.	Add some valid categorical column. This is just a warning (you execution will not include Drop Out).
"Your previous executions of Y degree have been deleted because the categorical variables have changed."	If you upload new data of a degree that was uploaded first application shows you that warning.	This is just a warning, you should execute again the affected degree.
"Perfect Separation: graduation can be perfectly predicted with provided data. The Graduation Prediction Model is not necessary"	If your data is so clear you do not need to execute the drop out algorithm.	This is just a warning. You can review the categorical columns but this is not usual if you introduce real data.

Message	Meaning	Possible solution
"Congratulations, you are now a registered user!"	Welcome to the SPEET web tool.	Now you can upload some CSV files and then execute the Clustering, Classification and Drop Out algorithm for a degree or have a general view with the Coordinated Views tool.

6 Summary

References

[LR01] Avraham Leff and James T Rayfield. Web-application development using the model/view/controller design pattern. In Proceedings of the Fifth IEEE International Enterprise Distributed Object Computing Conference, EDOC'01., pages 118–127. IEEE, 2001.

[McK11] Wes McKinney. pandas: a foundational python library for data analysis and statistics. Python for High Performance and Scientific Computing, pages 1–9, 2011.

[PMD+18] M.A. Prada, A. Morán, M. Domínguez, J. L. Vicario, R. Vilanova, A. Paganoni, U. Spagnolini, A. Torrebruno, M.J. Varanda, P. Alves, M. Podpora, and M. Barbu. Io3 - data mining tool for academic data exploitation. Technical report, ERASMUS + KA2 / KA203 SPEET Project, 2018.

[VVB+18] J. L. Vicario, 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.