SPEET: WEB BASED IT TOOL FOR ACADEMIC DATA ANALYSIS

R. Vilanova[1], J. Vicario[1], M.A. Prada[2], M. Barbu[3], M. Dominguez[3], M.J. Varanda[4], M. Podpora[5], U. Spagnolini[6], P. Alves[4], A. Paganoni[6]

[1]Universitat Autonoma Barcelona (SPAIN)
[2]Universidad de León (SPAIN)
[3]University Dunarea de Jos, Galati (ROMANIA)
[4]Instituto Politecnico de Bragança (PORTUGAL)
[5]Opole University of Technology (POLAND)
[6]Politecnico di Milano (ITALY)

Abstract

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

Keywords: International projects, International Cooperation, Educational Data Mining.

1 INTRODUCTION

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

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

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

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

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

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

This work reflects the main output of the SPEET project as an IT tool that implements specific algorithms developed to deal with the two basic problems tackled in the project: Classification, Clustering and Drop-out Prediction. First of all, in the next section the SPEET project is presented as well as its main goals. Next, the previously mentioned contributions are detailed:

- Performance analysis algorithms: student performance analysis on the basis of categorical and/or performance data; performance for upcoming semesters on the basis of initial information; for explanatory analysis, etc
- Drop-out prediction algorithms: drop-out prediction on the basis of selected categorical information and first semester grades. A statistical model is elaborated that provides a quantitative evaluation of the student being at risk of drop-out.
- Visualization tools: for visual inspection of the pre-existing data relationships. Dimensional reduction and histogram techniques are applied to project the data on appropriate dimensions suitable for analysis. The tool provides a complete interactive, on-the-fly

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

2 SPEET PROJECT

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

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

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

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

On the basis of the preceding scenarios, this project's goal emerges from the potential synergy among a) the huge amount of academic data actually existing at the academic offices of faculties and

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

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

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

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

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

3 DATA PREPARATION

Webtool users should preprocess institution data to the template that is described in this section. 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.
- *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.
- *Degrees.csv*: it includes information about the degrees that are to be analyzed. Therefore, DegreeID is expected as the KeyAttribute of this table.

Example of the *SubjectsPerformance* csv as it is shown in the webtool can be seen in Figure 1. In addition to the upload of the data according to these formats, pre-processing is also required to detect errors and/or missing data. In order to do this data-preprocessing, the Pandas library (pandas. pydata.org) of Python has been considered and the webtool performs the following steps:

- 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.
- Non necessary 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 degress) 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.

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.

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 1: Example of the StudentsPerformance.csv template

4 WEBTOOL DESCRIPTION

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 Figure 2 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 on uploaded data and specific visualizations, 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.

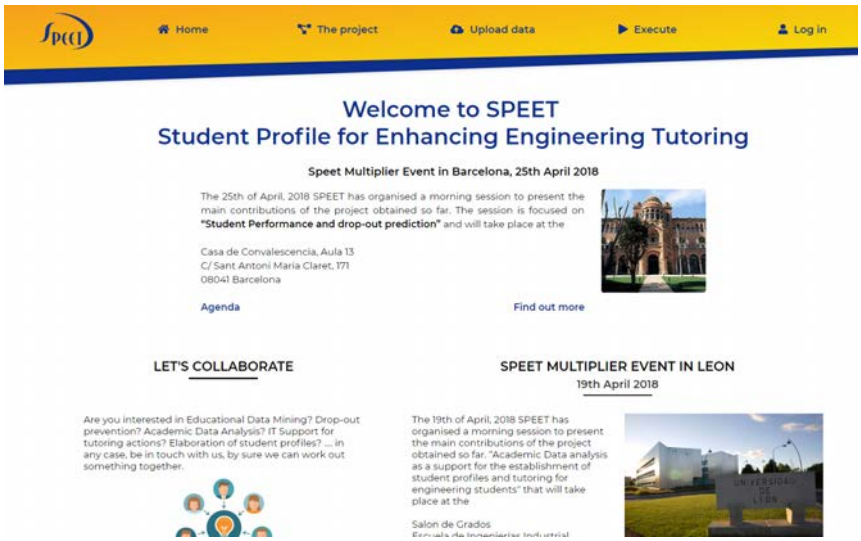

Figure 2: SPEET Webtool Homepage

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, 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, the user must upload a new dataset. The package provides suggestions.

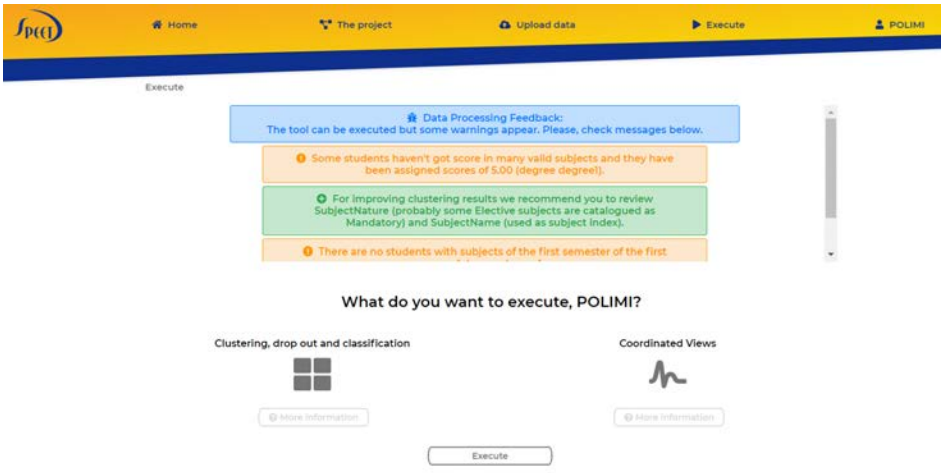

Figure 3: Execution of the tools is now possible. On top, there are some warning messages and the corresponding suggestions to fix them.

Once the preprocessing is completed, the user can choose to execute one of the two tools available by clicking on the corresponding icon (Figure 3):

- Clustering, drop out and classification
- Coordinated views

More information about the 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.

4.1 Clustering, drop out 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.

Figure 4: Example of execution output of the Clustering, Drop out and Classification Tool

This tool provides five different visualizations, as in Figure 4.

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

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. 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);
- 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 5. The number of filtered unit 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

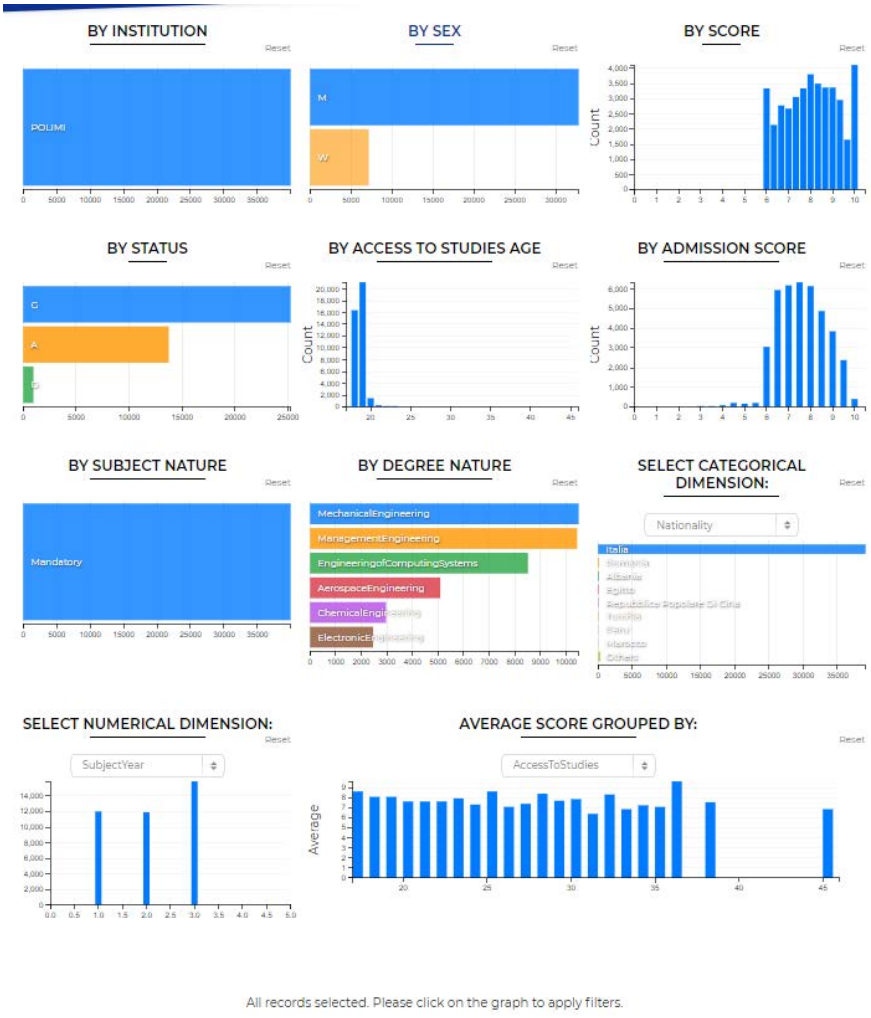

Figure 5. Example of execution output of the Coordinated Views Tool

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 visualization shown in this chapter are based on fictional data.

5.1 Clustering and classification

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. In Fig. 6, we present an example of the Performance Clusters plot generated by the web tool. This figure shows how the students have been organized in three clusters: Blue cluster (Excellent), Red cluster (Average) and Green (Low-Performance).

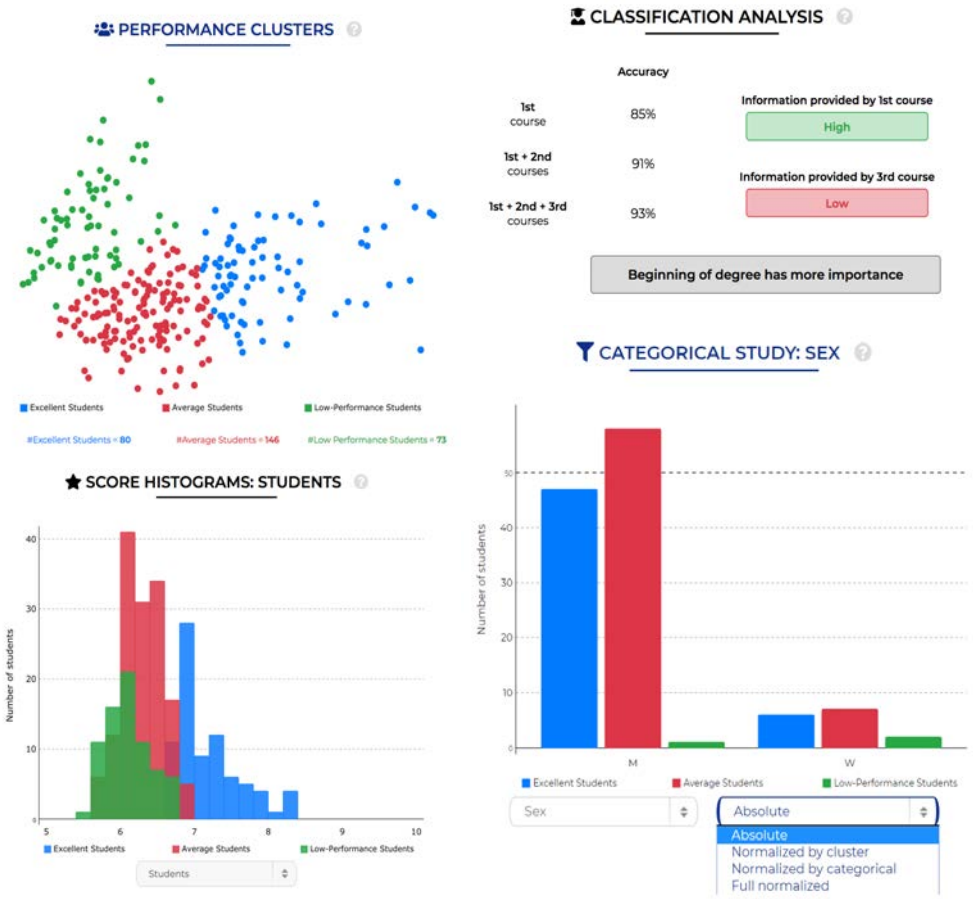

Figure 6 : Outputs example for the categorical, performance and classification studies

The categorical study 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.

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).

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 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.

5.2 Drop-out analysis. Graduate 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 includes those available at the time of the enrollment and those recorded after the first semester of the first year of study.

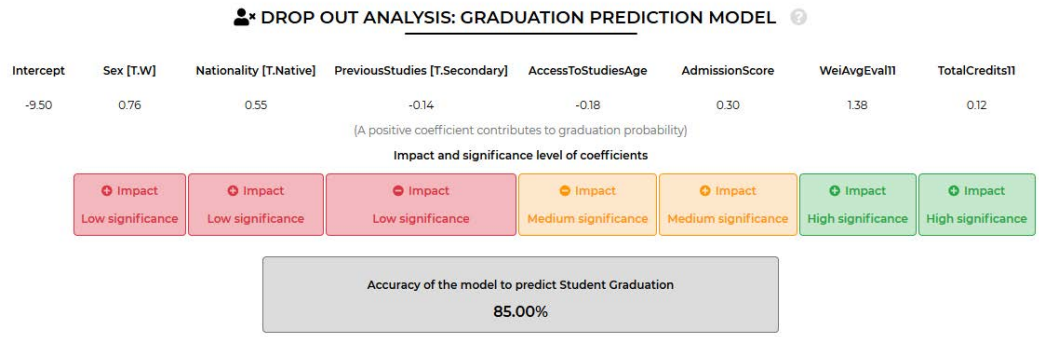

Figure 7: Example of output from the Dropout Analysis: Graduation Prediction visual

An example of this visual is shown in Figure 7. 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. The example shows [W] - woman - for variable Sex, so its other level [M] - man - is used as reference. In the middle, 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.

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).

6 CONCLUSIONS

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

ACKNOWLEDGEMENTS

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

REFERENCES

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

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

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

[4] J. MacQueen. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967.

[5] P. Agell and J.A. Segarra. Escuchando la voz del mercado: Decisiones de segmentacion y posicionamiento (original document in Spanish). EUNSA: Manuales IESE, 2001

[6] N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines. Cambridge University Press, 2000.