SPEET: SOFTWARE TOOLS 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. This work reflects the outputs of the SPEET project in relation to the data mining tools, specific algorithms developed to deal with the basic problems tackled in the project: Classification, Clustering and Drop-out Prediction.

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 outputs of the SPEET project in relation to the data mining tools, specific algorithms developed to deal with the two basic problems tackled in the project: Classification, Clustering and Drop-out Prediction. First of all, in the next section the SPEET project is presented as well as its main goals. Next, the previously mentioned contributions are detailed:

- 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. This is presented in detail in section 3.
- Drop-out Prediction tool: a dynamic tool based on the drop-out prediction of students based on their early performance such as at the first semester of studies. Details are provided in section 4.

These results are intended for qualified users with knowledge on programming and statistics. Therefore we put at their disposition the building blocks for performing direct data analysis or even generate their own IT tools.

2 SPEET PROJECT

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

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

The objective of this project could be stated in a rather simple way as: determine and categorize the different profiles for engineering students across Europe. The main rationale behind this proposal is the observation that students performance can be classified according to their 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 SET

One of the characteristics of the SPEET project is its transnational nature, since the fact of obtaining (or not) the same student classification and profiles will help identify common characteristics on engineering students coming from different EU institutions. The differences on a country/institution basis will be exposed and leads to deeper analysis. Due to its transnational nature, it is necessary to choose appropriate variables and representation to cover the differences in course organization at a country level. Additionally, the dataset must include students' personal information while complying with privacy regulations of the European Union. As a result, the proposed dataset uses variables obtained from the administrative records of the students, such as demographic data, courses taken and academic performance.

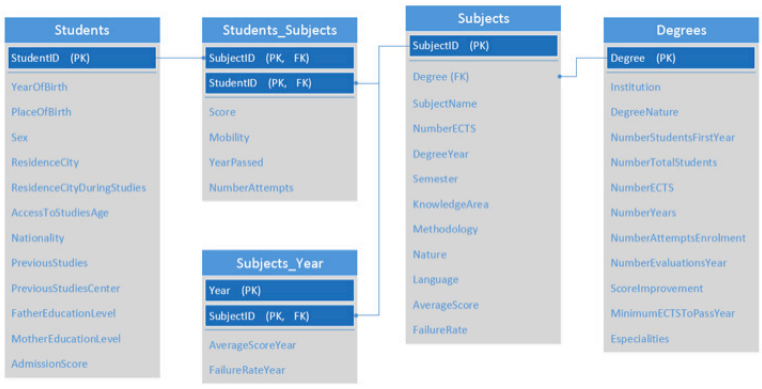

Figure 1: Relational model of the proposed structure of the dataset

Figure 1 shows the initial, minimum core dataset, proposed to perform the analysis. It is also possible to enrich the dataset with other potentially useful additional data sources, e.g., the regional/metropolitan socio-economic indicators provided by organizations such as the Organisation for Economic Cooperation and Development. As a matter of fact the retrieved collections of data includes collateral information regarding the students' origin (year and place of birth, geographical info, previous studies, age, etc), degree information (degree nature, total number of students, number years, etc) as well as student performance on different subjects of the degree (subject score, subject year, subject language, subject nature, etc).

4 STUDENT PERFORMANCE CLUSTERING AND CLASSIFICATION

In this section we present the Clustering and Classification tools. An overview of the architecture and logic of operation is presented in Figure 2. By departing from the datasets presented in the previous section, the Pre-Processing is in charge of adapting data to the Clustering and Classification modules. The Clustering, on the other hand, is aimed at generating three clusters of students based on their performance results. Categorical information is analyzed to obtain profiles of students belonging to different clusters. Finally, Classification is in charge of classifying new students to the clusters generated at the Clustering module. As it will be shown later, this Classification procedure is also useful to obtain insights about the structures of plan studies at the different degrees.

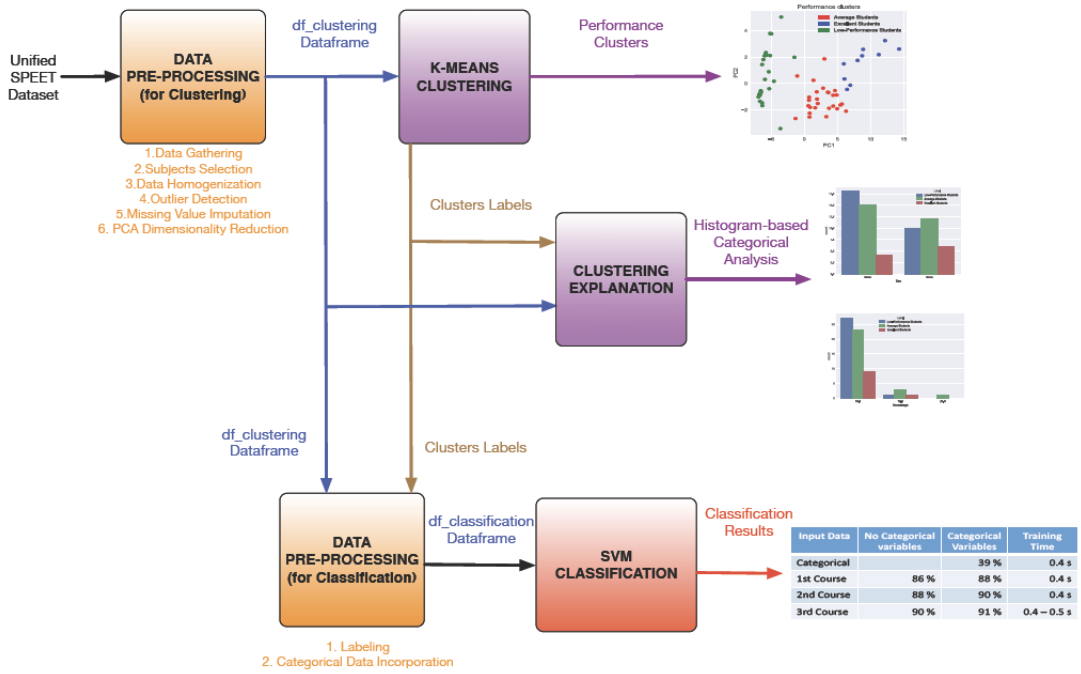

Figure 2: Architecture of the Clustering and Classification tool.

4.1 Categorical *vs* Numerical/Performance data

From the dataset presented in Fig. 1., the algorithms developed in this Tool focuses on the use of two kinds of data:

- Performance data: this data refers to the scores obtained by students at the different subjects. The nature of this data is numerical.
- Categorical data: this data refers to collateral information related to students. This includes features such as student demographic data (access age, gender, nationality), educational background (previous studies) or access conditions (access score).

4.2 Data pre-processing

The idea behind this procedure is to organize students in different groups (clusters) based on their performance results. To do so, a classical k-means clustering approach will be adopted, based on

gathering in a cluster those elements with the highest similarity. The goal is to obtain a small number of clusters, typically three or four clusters.

4.3 Clustering and Classification of students' profiles

In this section, we present specific details about the Machine Learning Algorithms considered for Clustering and Classification and their implementation.

4.3.1 Clustering

This module is in charge of grouping the different students based on their performance behavior. It is also in charge of providing explanations about the resulting clusters (i.e., identifying students' profiles belonging to the different clusters). Further details about these two functionalities are provided below:

- **K-means based Clustering**: As shown in Fig. 2, we adopt the k-means algorithm [4] as Clustering algorithm. It is worth recalling that inputs to this block are based on the PCA components of subject scores for each student to focus the clustering on a 2 dimensional problem. It is worth noting that we performed tests with clustering directly applied to the full dimensional problem and similar results were obtained. The clear advantage of working with PCA-based compressed data is that clustering computation time is significantly reduced.
- **Histogram based Clustering Explanation**: The second functionality of the Clustering tool 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 age and nationality. For each Categorical Variable, three histograms are generated to show the students' distributions associated to the different clusters. This methodology is inspired by the customer segmentation procedures applied in marketing applications [5]

4.3.2 Classification

In this module the objective is to develop a classification mechanism able to classify new students in terms of the Performance Clusters obtained at the previous module. In this project we consider Support Vector Machine (SVM) [6]. It is a supervised algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, each data item is represented as a point in *n*-dimensional space (where *n* is number of features you have) with the value of each feature being the value of a particular coordinate. Then, a classification is identified by finding the hyper-plane that differentiate classes in the best way. This project considers the SVM implementation of library scikit-learn version 0.18.2 in Python 2.7.13. (C-Support Vector Classification) and different configurations were tested:

4.3.3 Case Study Applications

In this Section, we present a summary of results we obtained with both Clustering and Classification algorithms:

Clustering Evaluation: Concerning the Clustering part, we first show two representative cases. These are degrees from two institutions of the SPEET consortium (see Fig. 3). These two cases are representative as provide a bad and a good example in terms of Clustering behaviour, respectively. This is reflected at the Average Score of Students histograms. Clearly, the Degree A does not present as clear performance groups as the Degree B does. Indeed, one can readily observe that the Degree A case is a scenario where students are better grouped by taking into account two Clusters (i.e., "Low-Performance Students" and "Average Students" should belong to the same cluster). This is a common pattern observed with the degrees considered at this project: when the Clustering behaviour is bad, it means that two clusters is a better option. However, to improve the robustness of the proposed tool, we focus on the three cluster configuration as baseline.

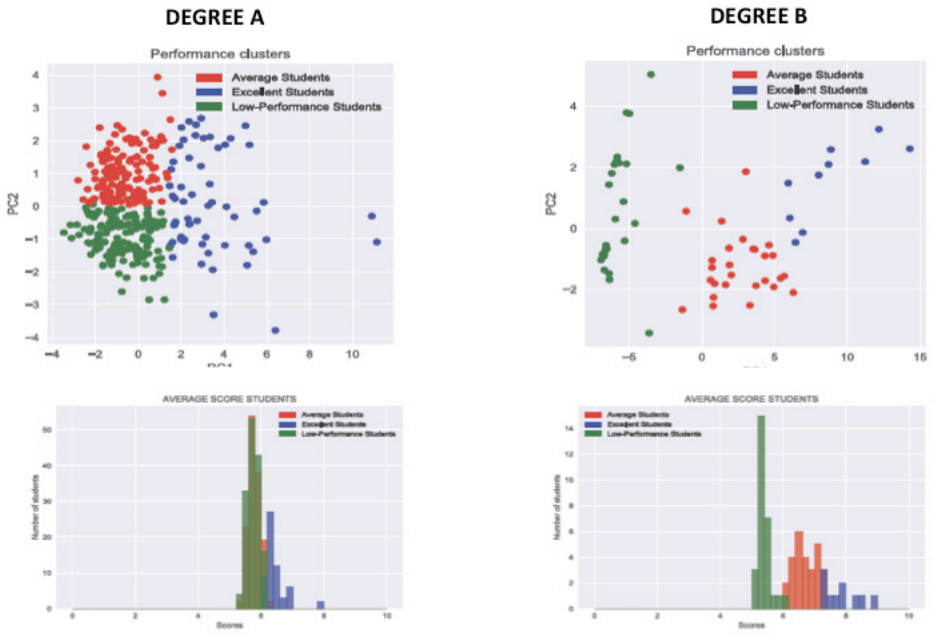

Figure 3: Performance Clusters and histogram for two cases in terms of Clustering Behavior: Left. Bad Clustering behavior, Right. Good Clustering behavior.

Classification Evaluation: in Figs. 4, we present classification results for different degrees in one university. As previously commented, the SVM-based classifier is adopted. Once the results are analyzed, one can verify that classification performance presents satisfactory results depending on the kind of degree/institution considered. But it is shown that the adoption of a SVM-based classifier provides a good trade-off in terms of accuracy vs. training time. On the other hand, one can also observe that categorical variables are not enough to classify students. Here it is worth pointing out that categorical variables are useful to understand profiles belonging to different clusters (by means of the Histogram-based Clustering explanation provided by the Clustering tool) but, however, this is somewhat different to try to accurately classify students beforehand. This is because performance at their studies are significantly affected by a complex set of factors not fully determined by the categorical variables.

Input Data	No Categorical variables	Categorical Variables	Training Time
Categorical		34 %	1 s
1st Course	77 %	76 %	1.1 – 1.4 s
2nd Course	88 %	86 %	1.1 – 1.3 s
3rd Course	93 %	94 %	1.2 – 1.4 s

Input Data	No Categorical variables	Categorical Variables	Training Time
Categorical		51 %	1.5 s
1st Course	68 %	68 %	2.1 – 2.8 s
2nd Course	73 %	74 %	2.8 – 3.3 s
3rd Course	93 %	92 %	1.6 – 1.9 s

Input Data	No Categorical variables	Categorical Variables	Training Time
Categorical		46 %	0.6 s
1st Course	69 %	71 %	0.6 - 0.7 s
2nd Course	78 %	78 %	0.6 – 0.8 s
3rd Course	93 %	92 %	0.6 – 0.7 s

Input Data	No Categorical variables	Categorical Variables	Training Time
Categorical		51 %	0.8 s
1st Course	73 %	75 %	0.9 – 1 s
2nd Course	85 %	85 %	0.8 – 0.9 s
3rd Course	93 %	91 %	0.8 – 0.9 s

Input Data	No Categorical variables	Categorical Variables	Training Time
Categorical		42 %	0.5 s
1st Course	64 %	63 %	0.6 s
2nd Course	85 %	87 %	0.5 – 0.6 s
3rd Course	86 %	91 %	0.5 – 0.6 s

Input Data	No Categorical variables	Categorical Variables	Training Time
Categorical		51 %	0.5 s
1st Course	76 %	70 %	0.5 – 0.6 s
2nd Course	83 %	84 %	0.5 s
3rd Course	89 %	88 %	0.5 s

Figure 4: Classification Results obtained with the degrees of one of the institutions

5 STUDENT DROP-OUT PREDICTION

The other issue the SPEET project is tackling is that of student drop-out prediction. It has been decided to continue the analysis conducted with the classification and clustering tools by analysing the distinction between students completing their study programme graduating and those who instead decide to abandon studies. The student profiles we are referring to within the SPEET project scope are:

a) *dropout*: students that leave degree studies

b) *graduate*: students that get the degree sooner or later

distinction which will be defined by a variable called ”status”. The choice to analyse such factor are clear as student drop-out is becoming a generalized manner. As an example, from the cumulative data in Italy, almost one student out of two withdraw from his engineering degree before the end of the studies. According to National Council of Engineers statistics related to the students that choose to study engineering and the numbers underline how the rate of abandonment is elevated, even if the graduates' number in the sector continues to increase in the years.

In this case, the data set for elaboration of the drop-out prediction will not be the same as per the classification and clustering. The reason for this distinction is that the specific goal is to predict the academic success of a student as soon as possible. In order to predict a final student's class (graduate/dropout) through the early student' performance information, we decide to focus the attention to the information available at the end of the first semester of the first year classified in term of different degree

In order to obtain a single model that takes into account the ”grouped” nature of the data, a generalized linear mixed-effects model (GLME) is implemented with the purpose of describing the relationship between the success probability (getting the degree) and the covariates using exactly the data as ”grouped” according to one classification factor (the Engineering School). In the next subsections we present the regression model that accounts for fixed and random effects in the prediction with the corresponding data variables associated to them.

5.1 Logistic Mixed-Effects Model

In logistic regression, a categorical dependent variable *y* having two unique values is regressed on a set of *k* independent variables x_1, x_2, … , x_k. The mean of the response variable *p*, in terms of explanatory variables x_1, x_2, … , x_k could be modelled relating *p* and x_1, x_2, … , x_k. The logistic by the *logit* models the natural log odds as a linear function of the explanatory variables

$$l = logit(p) = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

A simple logistic model can be applied to examine the relationship between the success probability (getting the degree) and a set of attributes for each student such as sex, year of birth etc., but the analysis has to be conducted independently for each one of the Engineering Schools within the same university, in order to not to lose the grouped nature of our database. A prototype of the suggested model is the following:

$$p_j = P(status_j = graduate) = P(status_j = 1)$$

$$logit(p_j) = ln\left(\frac{p_j}{1-p_j}\right) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj}$$

where p_j is the graduating probability for student *j*, x_1, x_2, … , x_k are the corresponding explanatory variables and β_1, β_2, … β_k are the coefficients to be estimated from the training set..

Mixed-effects models are primarily used to describe relationships between a response variable and some covariates in data that are grouped according to one or more classification factors. Examples of such grouped data include, repeated measures data, multilevel data, and block designs. By associating common random effects to observations sharing the same level of a classification factor, mixed-effects models flexibly represent the covariance structure induced by the grouping of the data. A logistic mixed-effects model is exploited therein to examine the relationship between the success probability (getting the degree) and a set of attributes for each student such as sex, year of birth etc

according to the grouping factor DegreeNature (= Engineering School within the same university). A prototype of the suggested model is the following:

$$logit(p_{ij}) = \sum_{k=0}^{p-1} x_{ijk}\beta_k + \sum_{h=0}^{q-1} z_{ijh}b_{ih}, \qquad i = 1,...,M \qquad j = 1,...,n_i$$

$$\boldsymbol{b}_i = [b_{i0} \quad b_{i1} \quad ... \quad b_{iq-1}]^T \sim N(0,\Psi)$$

where p_{ij} is the graduating probability for student j in group i, x_{ijk} and z_{ijh} represent, respectively, the (j,k) element of matrix X_i (of size $ni \times p$) and the (j,h) element of Z_i (of size $n_i \times q$), that is, the values of explanatory variables for fixed and random effects model parameters. While, $\beta = [\beta_0, \beta_1, \ldots \beta_{p-1}]^T$ is the p-dimensional vector of fixed effects, and $\beta_i = [\beta_{i0}, \beta_{i1}, \ldots \beta_{ip-1}]^T$ is the q-dimensional vector of random effects.

On the basis of this setting, we specify a final model whose estimates for the fixed effects and variance component with the corresponding p-value, are the following:

Table 1. *P-values and coefficients estimates for the final Logit Mixed-Effects Model.*

	Variable	Estimate	p-value
Fixed effects	(Intercept)	-2.322716	2.47e-05
	Sex(male)	-0.292086	0.000447
	Nationality (foreign)	-0.423296	0.020026
	AccessToStudiesAge	-0.054090	0.029240
	AccessToStudiesYear(2010)	-0.022170	0.801236
	AccessToStudiesYear(2011)	-0.344638	8.32e-05
	AccessToStudiesYear(2012)	-0.844056	< 2e-16
	WeightedAverageEvaluations 11	0.060766	< 2e-16
	AverageNumbAttemptsPerExam 11	0.028752	0.562176
	NumbSubjectsPassed 11	1.709591	< 2e-16
	Change(yes)	-0.373339	0.011962
Random effects	Grouping factor	Variable	Variance
	DegreeNature	(Intercept)	1.062

5.1.1 Prediction

The logistic model is the basis for the individual student behavior prediction. The predicted values are the success probabilities (p) and are therefore restricted to (0,1). Our decision boundary will be 0.5. If p > 0.5 then y = 1 := graduate, otherwise y = 0 := dropout. this choice is arbitrary, for some applications different thresholds could be a better option.

A logistic model was fitted to the available data by using R. As an example of the effect of small differences in the student categorical data can determine variations in code example shows the application of the prediction regression models to a couple of students with a single difference in the data. As an example of how a minimum difference can change the estimate of the percentage of success, we consider two student profiles that can be explicitly read directly from the R-code. For the first student the estimated success percentage is 25.75%, while for the second student 31.71%. Note that the only difference between the profiles is the sex of the student.

5.1.2 Model Accuracy

When developing models for prediction, the most critical metric regards how well the model does in predicting the target variable on out of sample observations. This is typically done by estimating accuracy using data that was not used to train the model such as a test set, as we have done with simple logistic models. The process involves using the model estimates to predict values on the training set. Afterwards, we will compared the predicted target variable versus the observed values for each observation. We use 12859 student data for model development and keep 3200 sampling to check model accuracy. Based on the proposed model, we compute predicted graduating probability, then looking at the difference between observed and predicted, for those 3200 cases, we find

Table 2. *P-values and coefficients estimates for the final Logit Mixed-Effects Model.*

Observed	Predicted	
	1	0
1	2221	68
0	224	687

Thanks to the misclassification error we can obtain an estimate of the model accuracy = 1/(68+224)=(2221+68+224+687) = 90.87% accuracy on the test set is a very good result. Moreover, we can consider sensitivity=2221/(2221+68)=0.9702927 and specificity =687/(224+687)=0.7541164. High sensitivity and specificity indicate a good fit of the model.

6 CODE AVAILABILITY

As previously commented, this tool has been developed in Python. All the code can be found at a *bickbucket* repository that can be accessed from the project website (http://www.speet-project.eu)

7 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.