IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 12, 2020, accepted November 21, 2020, date of publication November 26, 2020,
date of current version December 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3040858

Educational Data Mining for Tutoring Support in Higher Education: A Web-Based Tool Case Study in Engineering Degrees

MIGUEL ÁNGEL PRADA[1], MANUEL DOMÍNGUEZ[1], (Member, IEEE), JOSE LOPEZ VICARIO[2], (Member, IEEE), PAULO ALEXANDRE VARA ALVES[3], MARIAN BARBU[4], (Senior Member, IEEE), MICHAL PODPORA[5], UMBERTO SPAGNOLINI[6], (Senior Member, IEEE), MARIA J. VARANDA PEREIRA[3], AND RAMON VILANOVA[2], (Member, IEEE)

[1]Departamento de Ingeniería Eléctrica y de Sistemas y Automática, Escuela de Ingenierías, Universidad de León, 24007 León, Spain
[2]Departamento de Telecomunicacio i Enginyeria de Sistemes, Escola d'Enginyeria, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain
[3]Research Centre in Digitalization and Intelligent Robotics, Polytechnic Institute of Bragança, 5300-253 Bragança, Portugal
[4]Automatic Control and Electrical Engineering Department, Dunărea de Jos University of Galaţi, 800008 Galaţi, Romania
[5]Faculty of Electrical Engineering, Automatic Control and Informatics, Opole University of Technology, 45-758 Opole, Poland
[6]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milan, Italy

Corresponding author: Miguel ÁNgel Prada (ma.prada@unileon.es)

This work was supported by the Erasmus+ Key Action 2 Strategic Partnerships KA203, funded by the European Commission, under Grant 2016-1-ES01-KA203-025452.

ABSTRACT This paper presents a web-based software tool for tutoring support of engineering students without any need of data scientist background for usage. This tool is focused on the analysis of students' performance, in terms of the observable scores and of the completion of their studies. For that purpose, it uses a data set that only contains features typically gathered by university administrations about the students, degrees and subjects. The web-based tool provides access to results from different analyses. Clustering and visualization in a low-dimensional representation of students' data help an analyst to discover patterns. The coordinated visualization of aggregated students' performance into histograms, which are automatically updated subject to custom filters set interactively by an analyst, can be used to facilitate the validation of hypotheses about a set of students. Classification of students already graduated over three performance levels using exploratory variables and early performance information is used to understand the degree of course-dependency of students' behavior at different degrees. The analysis of the impact of the student's explanatory variables and early performance in the graduation probability can lead to a better understanding of the causes of dropout. Preliminary experiments on data of the engineering students from the 6 institutions associated to this project were used to define the final implementation of the web-based tool. Preliminary results for classification and drop-out were acceptable since accuracies were higher than 90% in some cases. The usefulness of the tool is discussed with respect to the stated goals, showing its potential for the support of early profiling of students. Real data from engineering degrees of EU Higher Education institutions show the potential of the tool for managing high education and validate its applicability on real scenarios.

INDEX TERMS Drop-out prediction, educational data mining, performance prediction, visual analytics.

I. INTRODUCTION

The availability of data is a relevant asset for institutions, because data analysis can be used to help in decision making both in the day-to-day operative as well as strategically. In the educational domain, higher education institutions generate vast amounts of data from different sources. In particular, the universities collect every year data from their students including demographic details (e.g., age, address or socio-economic status) and information about their admission and

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Mei.

academic performance (school, degree, course path, and even examination results). Sometimes this information is augmented with data obtained from questionnaires and field observations or with information about their career after graduation. Knowledge can be extracted from those data to optimize the education management tasks and improve the students' success rate. Indeed, the European Commission states that ''monitoring students creates a foundation for institutional action'' [1].

A. LEARNING ANALYTICS AND EDUCATIONAL DATA MINING

Nowadays, it is quite common that any interaction between students and the computer-based educational information systems leaves a digital footprint that can be seen as complementary data. Learning management systems, apart from providing access to the course contents, might include support for management and evaluation of tasks, student tracking and reporting that allows to assess their learning performance [2] and to predict the risk of dropping out [3]. Intelligent tutoring systems [4] are computer-assisted instruction systems which record all student-educator interaction and consequently customize the teaching process. Data stored by all these systems will have higher granularity, at the course level, because it is related to specific activities or events, such as the results to exercises and quizzes.

In this sense, two fields are receiving increasing attention: Learning Analytics (LA) and Educational Data Mining (EDM). Both fields are multidisciplinary and cover a common ground, but are focused on different targets [5]–[10]. The focus of learning analytics is the collection, analysis and knowledge extraction from learning-related data to better understand and optimize learning results and environments [11]. The expected advantages of learning analytics can include customized learning and course offerings, curriculum adjustments and improvement of faculty performance or research [12]. On the other hand, educational data mining stress the research and development of automated and data-driven methods to discover patterns in large volumes of educational data [13]. Methods in educational data mining can be classified in terms of its aim, i.e., prediction, clustering, relationship mining, distillation of data for human judgment and discovery with models [14]. Nevertheless, there is in any case an overlap with regard to the problems LA and EDM are trying to solve, such as student behavior modeling or drop-out prediction [15].

B. POTENTIAL IMPACT

Although there have been already many studies applying data analysis to learning in higher-education, it is still an emerging field that requires more attention from university administration, instructors and other stakeholders [16]. The prediction of drop-out risk would be useful to identify tutoring needs and define early instructional and counseling actions, which are agreed to be beneficial for students' retention [16]. The number of tutors or counselors is usually small compared to the number of students, so support systems will be needed to help these tutors in their diagnostic activities, alleviating the needed effort to carry out personalized retention actions [17]. However, tutoring staff usually does not have a data scientist background and ignores the potential of data analysis. This is one of the major difficulties that prevents the adoption of those approaches.

Furthermore, it also needs to be recognized that tracking, collection and evaluation of data is challenging. For that reason, previous works are usually constrained to, at most, data from one institution. However, the joint analysis of students' behavior at different institutions could lead to interesting insight about their common aspects and their differences that might be rooted in the institutional characteristics. Even more if those institutions are heterogeneous enough, with different sizes, demographic circumstances or countries of origin. There are currently few reports in the learning analytics literature of deployment at scale [18].

For the previous reasons, the aim of the work reflected in this paper is the development of an web-based software tool,[1] to be used for support of the predictive modeling activities of tutoring staff without a data scientist background. This work has been developed in the context of a joint educational project, *Student Profile for Enhancing Tutoring Engineering*, with the participation of 6 European institutions of higher education. The proposed web tool (SPEET tool) is focused on the analysis of students' performance in Engineering Bachelor degree programs, because the problem of dropout is common in this stage and disciplines. Performance, for that purpose, would be defined in terms of observable scores and completion of studies. It is also necessary that the data on which the tool are based are easily acquired and processed [19], so that any faculty or school could collect and organize their own data in a format that matches with the one proposed here, gaining meaningful benefits from the resulting tool analysis with a remarkable benefit arising from inter-institutions comparison.

Finally, the proposed approach needs also to have a transnational nature, since obtaining similar student profiles among different EU institutions might help to identify common characteristics of European engineering students and the differences on a country/institution basis could also be exposed and lead to deeper analysis. However, this transnational context imposes some constraints on the targets that are studied. For that reason, the work focuses in a global and transnational degree-wide view of performance, rather than focusing on a course-wise analysis. Higher granularity is impractical due to multiple reasons: courses from different institutions would hardly be comparable unless they were specifically designed for that purpose, the usage and particular implementation of course tracking software would differ among institutions and findings would not be easily generalizable. Moreover, the need for a simple and easily available data set brings further constraints. For these reasons,

[1] The SPEET tool 1.0 is available at http://speet.uab.cat

the proposed common data set and representation, accounting for the national and institutional differences in degree organization, uses only variables obtained from the administrative records of the students, such as demographic data, courses taken, or academic performance [20].

C. OUTLINE OF THE PAPER

The paper is structured as follows. Section II examines the previous work on analysis of students' performance and dropout. Section III presents the data analysis and visualization approaches that are proposed to exploit data for tutoring support. The data set, the implementation of the tool and the experiments are described in depth in Section IV. Results are presented and discussed in Section V. Finally, Section VII contains the conclusions, limitations and future work.

II. BACKGROUND

One of the most interesting uses of data analysis in the educational field is the exploration of data to discover patterns and derive knowledge. For this purpose, it is useful to involve the human analyst in the process. Therefore, interactive visual analytics, which blends information visualization and advanced computational methods to provide a semi-automated analytical process driven by interaction, is an interesting option [21]. The ability of visual analytics to augment data analysis with human perceptual and cognitive abilities is valuable as a tool to manage educational data [15], because these techniques allow people to discover trends, gaps or groups. Most applications of visual analytics in education have been constrained to the analysis of data obtained from the interaction of students with learning management systems and other learning support platforms. For instance, in [22], interactive visualizations were used for the analysis of the correlations between activity patterns in MOOCs (massive open online courses) and dropout.

Nevertheless, most previous works face educational data analysis from a predictive perspective, aiming to forecast future academic outcomes and to obtain a better understanding of the factors that play a part in academic success. The factors related to students' performance are still the subject of debate among educators, academics, and policy makers. Some authors [23] found that academic achievement is related to the student's ability and adaptation (also described in relation to motivation and perseverance). The challenge is to acquire quantitative data for those factors, because questionnaires could be used for that matter but students' responses might not reflect faithfully their latent abilities or attitudes. Other studies examining this problem also point out that environmental factors such as previous schooling, parents' education or family income have a significant effect on the students behavior. The institutional factors can also influence academic success, specifically the degree of adaptation and support that the institution provides, its structure, as well as the clarity on the communication of expectations and requirements, such as the admissions criteria [24].

In this sense, the joint analysis of data from multiple sources of the university, such as academic records, the activity on a LMS, the prior academic history or demographic variables, has been used to predict the likelihood of being unsuccessful and the retention rate [25]. In any case, a non-trivial stage of data preprocessing is necessary, where aspects such as the hierarchical structure, context, granularity and time range of data must be considered [13].

The goal behind the prediction of students' performance is generally explanatory, i.e., to obtain a better understanding that guides educational actions that would hopefully result in enhanced outcomes. For that reason, sometimes performance prediction is rather posed as a classification problem, either binary [26] or with multiple classes, ranging from low to high performance [27]–[29]. That is also the case in the approach presented in [30], which is also aimed at finding courses that are good predictors of students' performance and their progression. In this application, it is necessary to find a trade-off between classification performance and interpretability [30]. Widely-used classification techniques have been used for this purpose, including decision trees [30], Bayesian networks [25], [31], k-nearest neighbors [30], naïve Bayes [26], [28], [30] and random forests [30].

The prediction of dropout, which pertains to the fact or risk of not completing the degree due to academic failure, voluntary withdrawal or transfer to other institution, is useful not only to help faculty in understanding its causes but also to provide an early alert that might lead to corrective interventions. Student retention is an important aim, because dropout has undesirable consequences both for individuals and society. For that reason, dropout has been extensively studied in the literature, trying to analyze its predicting variables [32].

Several factors are assumed to have an impact on the drop-out rate. Among the external ones, one is the socio-economic environment, which includes variables such as family income, fees, availability of financial support, need for a supporting job, parents' previous education, cultural differences or social disadvantages [33]. Apart from that, low performance in previous studies, poor results at the first year or simultaneous enrollment in multiple programs can be relevant factors of dropout. On the other hand, there are additional internal factors, related to the student's personality and development, including at least the students' general attitude towards studying, their confidence and beliefs about themselves as learners, the anxiety with certain subjects, the perception of value, the interest in a subject, and the enjoyment [20]. Loss of motivation is usually linked to situations where a student cannot master fundamental concepts and skills, due to alienation or disengagement from learning.

Data stored by universities about students only reflect these categories in part, because they are essentially quantitative. The available data for drop-out analysis are generally demographic, such as gender, age at enrollment or parents' level of education, and scores of previous university courses or pre-university exams. When possible, attendance and information about the development of the course,

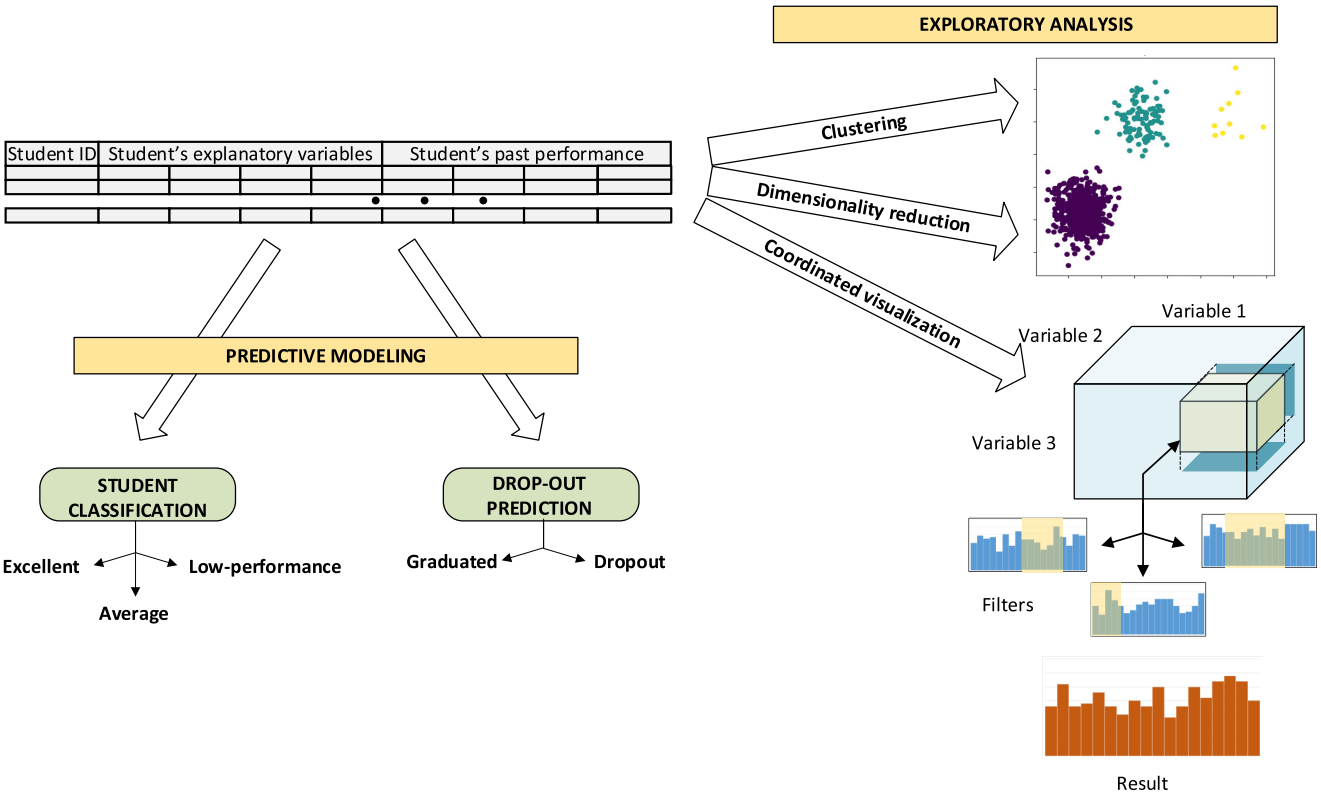

FIGURE 1. Summary of the proposed approach.

obtained from learning management systems, are also useful. Tutoring and counseling should later complete the student profile by obtaining further information about the students at risk from questionnaires, interviews and checklists, providing a better understanding of each personal case and the ability to identify other causes of socio-emotional and behavioral nature.

Drop-out prediction has been addressed using different techniques, such as for example random forests [17]. The reported accuracy of drop-out prediction depends strongly both on the definition of dropout and on the level of studies. The ultimate aim, nevertheless, is that this analysis leads to the development of early warning systems.

It is also important to note that the aggregation and processing of educational data raises ethical and legal concerns. Since the data set includes demographic variables, it is important to safeguard the student's privacy. As a result, data are anonymized so that the individuals are unidentifiable. Researchers also need to comply with personal data protection regulations such as the General Data Protection Regulation (GDPR) in the European Union [34], if they are applicable to the collection, aggregation and analysis of the data under study.

III. PROPOSED METHODS

As stated in the introduction, the ultimate aim of this work is the development of a web-based software tool for the support of predictive modeling activities of tutoring staff in a transnational context focused on engineering students, where basic data from university administration is assumed to be available. That is, the analyses focus on two kinds of data: performance variables, in terms of the scores of students in each subject; and explanatory variables, including features such as student demographic data (age of access, gender, nationality), educational background (previous studies) or access conditions (access score).

To achieve the stated purpose, we propose two approaches. A first one of exploratory analysis, that lets a tutor obtain insight from data through unsupervised machine learning and interactive visualization techniques. The second one is centered on the classification of students, either with respect to their performance or risk of dropout (see Figure 1).

A. STRUCTURE OF THE DATA SET

Due to the national and institutional differences in degree organization, a common set and representation structure needs to be chosen. As a result, the data set proposed here uses variables obtained only from the administrative records of the students. Note that data obtained from learning management systems would not be universally available for all the courses and would be highly dependent on the university, platform, educational approach and context. The data gathered by university administrations include demographic information,

courses and grades. Specifically, the following variables are considered in our approach:

- Student explanatory information: gender, age, age of access to studies, nationality, type of previous education, admission score, student status, parents' education level and residence.
- Degree information: institution, nature and length.
- Student's performance and subject information: name, type and length of the subject, score, number of attempts, semester, year, knowledge area, language, nature, average score, failure rate and mobility status.

Depending on the aim of the analysis, the set of students to be considered might only consider current students and also students that have graduated or dropped out.

Similar data sets can be found in the literature. For instance, features such as student demographic data (age, gender, country of residence, citizenship), educational background (secondary school, highest education level), or current and past information about study units (study area, enrollment mode, delivery method, average grade) are considered in [35] for student profile modeling. The publicly available Open University Dataset [36] also includes this information, along with student interaction data tracked by a virtual learning environment. In [37], only the final grades of the courses are used to predict dropout and performance. Although the application process for student enrollment differs in different countries, the admission scores are known to be useful for academic performance prediction. It needs to be noted that other authors have included some variables that are more difficult to be obtained consistently in different institutions, such as scholarship, marital status, special needs, type of dedication, debt situation [26], number of siblings, household income, working status [28], parents' occupation [31] or English proficiency [27].

B. EXPLORATORY ANALYSIS

To understand the data structure, it is interesting to perform *clustering*, i.e., grouping the students associated variables for their similarity. The use of clustering to find similar groups of students has already been used in previous works [30]. The most widely used clustering technique is the K-means algorithm [38], [39], which is used in this work. This approach can be applied to find distinguishable student profiles if each student is considered a data item, where their n variables are seen as the coordinates of a point in an n-dimensional space from a geometrical perspective [40]. The students are clustered with respect to their performance, using the subjects scores as features. Students can fall into several clusters that would correspond to groups including from excellent to low-performing students.

This unsupervised labeling of students is useful for their profiling, but results would be more explanatory if the end-user could visually analyze them. However, when data dimensionality is higher than 3, visualization is not directly possible. For that reason, it is necessary to use a transformation aimed at faithfully representing high-dimensional data onto low-dimensional spaces, called *dimensionality reduction*. Let $\mathbf{E}$ be the $n \times D$ matrix that consists of n vectors, each one describing a student. Therefore, a student l is described by D variables, i.e., $\mathbf{e_l} = [e_{1l}, e_{2l}, \ldots, e_{Dl}]$. The goal of dimensionality reduction is to project $\mathbf{E}$ onto a $n \times d$ matrix $\mathbf{F}$ where $d < D$. Indeed, $d \leq 3$ when the purpose is visualization. The dimensionality process aims at maximizing the preservation of the properties of matrix $\mathbf{E}$. For that purpose, certain metrics and assumptions about the geometrical structure of data need to be considered, leading to different approaches [41].

A well-known alternative is the PCA (Principal Components Analysis) transformation [42] that aims at converting the high dimensional data $\mathbf{E}$ into a low-dimensional set $\mathbf{F}$ of uncorrelated components that preserve most of the variance. It achieves that purpose by computing a linear mapping W through the eigenvalue decomposition of the correlation matrix $\mathbf{E^T E}$. Manifold learning algorithms are another well-known class of techniques to perform nonlinear projections of data onto a low-dimensional space by preserving distances or divergences. t-SNE (t-Distributed Stochastic Neighbor Embedding) [43] is a manifold learning algorithm that is known to provide good visualization results with real data. Its purpose is to find the data projection that minimizes the Kullback-Leibler divergence between joint probabilities P and Q computed respectively from the high-dimensional and low-dimensional pairwise Euclidean distances. A Gaussian distribution models the high-dimensional space, but a a Student t-distribution with one degree of freedom is used in the low-dimensional space to accurately model the local data structure.

The projection of data onto a 2D space (i.e., $d = 2$) by means of dimensionality reduction is useful to facilitate their visual interpretation, as itself or jointly with clustering. The lower-dimensional projection obtained through dimensionality reduction is shown as a 2D scatterplot where the relative distances between points are interpretable, because proximity in the low-dimensional space corresponds to high similarity in the original space. Since every student is represented by a point, visual properties, such as color or shape, can be used to convey additional information for visual inspection related to labels or values of relevant variables. The analysis of the scatterplots, especially if the visualization takes advantage of interactivity and additional visual cues, might be useful to better understand the data structure. This approach could be applied to project the whole set of students, represented by their descriptive variables and the average score for each academic year, or a set of students per degree, including their descriptive variables and the scores of all the subjects and allowing missing data [44], [45].

An alternative approach is considered to provide an interactive visualization for exploratory analysis. Using a single student-subject interaction (subject score) as the statistical unit, we propose a *joint and coordinated visualization of the histograms or bar charts* corresponding to each variable that can be automatically updated subject to custom filters

set interactively on the other ones [45], [46]. Users can also interact with histograms that are aggregated with respect to a categorical dimension. This approach would enable operations such as slicing or dicing (range selections in one or more dimensions) similar to the ones used by On-Line Analytical Processing (OLAP) in the business intelligence field. For that reason, this approach is useful for the visual analysis of the distributions of variables, providing a global view of data and letting a user explore the correlations between variables. Thus, the interactive and real-time filtering can be used to facilitate the rapid validation or rejection of hypotheses about a set of students.

C. STUDENT CLASSIFICATION AND DROP-OUT PREDICTION

As in most previous works, performance prediction is posed as a *classification* problem where students are matched to classes ranging from low to high performance. In this case, performance is defined in terms of the scores obtained in each subject. For that purpose, experimentation should focus on analyzing the accuracy obtained in the classification of students that have already graduated, using only partial information [47]. Apart from explanatory variables, three different setups are considered herein, where the amount of available information varies:

1) Only the scores of the subjects corresponding to the first course are considered.
2) The data set includes the scores of the subjects at the first and second courses.
3) The scores of the three first courses are included.

The reason is that it is interesting to analyze the discriminability of the scores from the initial courses, to understand the degree of course-dependency of students' behavior at different degrees [40]. The results obtained in this analysis might provide insight about the study program of the degree.

Concerning the classification algorithm, the *Support Vector Machine* (SVM) [48] is considered. SVM is a well-known supervised algorithm that aims to find the hyper-plane that better separates classes, when each data item is interpreted as point in n-dimensional space (where n is number of features). The adoption of multi-layer perceptrons [49] was also addressed at the beginning of this work. However, SVM was finally selected as the reference classifier because it was approximately 10 times faster and provided similar results. It is worth noting that the goal of this work was to develop an online and real-time tool, so the reduction of complexity was a main concern.

We also propose to predict, from the analysis of the explanatory variables and the performance in early stages of the degree, the probability that a student will graduate or, conversely, drop out. The analysis again includes early performance in the degree besides the explanatory variables, under the assumption that it might be significant to predict students' success. In this case, early performance variables are two variables constrained to the first semester of the degree: the number of credits achieved and the weighted average score obtained by the student at this semester (with the weight depending on that number of credits). The success prediction must therefore be addressed again as a classification problem, defining models of the relationship between a set of input variables and the graduation status, which is considered a binary variable (graduated = 1, dropout = 0). For that reason, only students labeled as graduated or dropout are used for training. However, it must be noted that the aim is not only to predict early dropout but also to understand which students' profiles are more sensitive to that situation.

Two statistical models are used to compute, for each input variable, its impact on graduation probability (positive or negative) and its level of significance (low, medium or high) [47]. The proposed methods can lead to the identification of students' profiles, through the analysis of the model weights that explain the effects of both explanatory and early performance variables on the graduation probability. This way, we can find patterns of the influence of variables in student dropout.

A first simple choice is the use of logistic regression [50], which models the dependency of a binary response variable y on a set of k explanatory independent variables $x_1, x_2, \ldots, x_k$. Being p the probability that $y = 1$ (i.e., the probability of successful completion of the degree), its natural log odds is modeled as a linear function of the explanatory variables, where β_k are the coefficients estimated from training:

$$\begin{aligned} logit(p) &= ln\left(\frac{p}{1-p}\right) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \end{aligned} \quad (1)$$

This model can be used to examine, independently for each degree within the same institution, the relationship between the probability of successful completion of the degree and a set of attributes. However, the logistic model assumes that the predictors are independent. This assumption does not hold when repeated measures or clusters are found in data and, in this context, it is interesting to jointly analyze students of different degrees. Alternatively, in that case, a logistic mixed-effects model can be used to analyze grouped data [40], [51]. Generalized mixed-effects models [52] take into account the grouped nature of data, describing relationships between a response variable and some covariates in data that are grouped according to one or more classification factors. They associate common random effects to observations that share the same level of a classification factor in order to represent the covariance structure induced by the grouping of the data.

As a particular case of this approach, the logistic mixed-effects model is an extension of ordinary logistic regression that describes the relationship between the graduation probability and the covariates of variables grouped according to the degree as classification factor. Let i be the groups that specify the different degrees and j the students. A random term $\mathbf{b}_i$ is included to address differences between students belonging to different degrees, so the model for a single observation y_{ij} for $j = 1, \ldots, n_i$ in group i,

TABLE 1. Structure of the data set.

Students		
Variable	Description	Type
StudentID	student ID number	string
DegreeID	degree programme attended by student ID	string
Gender	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	string [(A)ctive,(G)raduated,(D)ropout]
Degrees		
Variable	Description	Type
DegreeID	degree programme ID	string
Institution	name of the institution organizing the programme	string
DegreeNature	degree study programme (e.g. Mechanical Engineering, . .)	string
DegreeLength	total number of ECTS (European Credit Transfer and Accumulation System) of the programme	integer
DegreeYears	duration of the degree programme	integer
Subjects Performance		
Variable	Description	Type
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]
SubjectLength	total number of ECTS credits 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	[Mandatory, Elective, Thesis, Internship] string

$i = 1, \ldots, M$ can be written as

$$logit(p_{ij}) = log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \sum_{k=0}^{p-1} x_{i,jk}\beta_k + \sum_{h=0}^{q-1} z_{i,jh}b_{ih} \quad (2)$$

being

$$p_{ij} = P(y_{ij} = 1|\mathbf{b}_i, \boldsymbol{\beta}) = \frac{exp(\sum_{k=0}^{p-1} x_{i,jk}\beta_k + \sum_{h=0}^{q-1} z_{i,jh}b_{ih})}{1 + exp(\sum_{k=0}^{p-1} x_{i,jk}\beta_k + \sum_{h=0}^{q-1} z_{i,jh}b_{ih})} \quad (3)$$

In this model, $x_{i,jk}$ is an element of matrix $\mathbf{X}_i$ (with size $n_i \times p$) that contains the values of the explanatory variables for fixed effects model parameters of ith group, and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots \beta_{p-1}]^T$ is the p-dimensional vector of fixed effects coefficients. In turn, $z_{i,jh}$ is an element of the matrix $\mathbf{Z}_i$ (with size $n_i \times q$) of random effects model parameters for the ith group and $\mathbf{b}_i = [b_{i0}, b_{i1} \ldots b_{iq-1}]^T$ is the q-dimensional vector of random effects coefficients, which has a multivariate normal distribution $N(0, \sigma^2)$.

IV. EXPERIMENTS

A. EXPERIMENTAL CONDITIONS

A relational structure with three tables was designed as input for data: one for the Students (including the information that is available as soon as the student enrolls in the institution), one for the Degrees (including information about the degree nature and length) and one for the Subjects (with student and degree IDs as foreign keys and information about the performance). Only a subset of the variables considered in Section III-A is included in these tables, to ensure a wider availability for any user. These variables are described in Table 1. For experimentation, data are acquired from the university administrations. Data from the six universities participating in the project, from five countries of the European Union, were collected between 2011 and 2017. As a result, data from approximately 21,000 graduated and non-graduated students were available. The anonymity of data subjects has been ensured, because all the variables that might identify a student were removed by the data custodian before they were shared with the analysts.

During the definition of the experiments, the relevance of preprocessing to reduce inconsistencies became clear, especially among different institutions. Main sources of discrepancies were found in the labeling of categorical variables, the availability of specific variables or the scoring system. It is necessary to deal with missing values, outliers and corrupted values, generally caused by user-entry errors or different data dictionary definitions. The students' courses scores are normalized to a 0-10 range. Missing value imputation is performed for subjects where more than half of the scores are present, using the average for students who have graduated and zero for dropouts. If a fewer percentage of scores is available, the subject is discarded to avoid the use of a high number of imputed values. Before drop-out analysis, it is also necessary to perform further preprocessing to account for students who have studied more than one degree, omit the degrees with a small sample of available data and remove

students that are still active or known to have only temporarily interrupted their studies.

Furthermore, it is necessary to perform further processing to accommodate data to the needs of the proposed algorithms. In all the cases, the three tables are joined into a single one. For the coordinated histograms, each row represents a student-subject record. In the clustering and classification experiments, there are as many data items as students. The drop-out analysis requires aggregated values related to the performance in the first semester of the degree, which need to be computed from the original data set. The drop-out analysis also requires information about the student's status (Graduated, Dropout or In Progress), which was not directly available at all the institutions of this project [53].

Experiments were organized in two phases. First, each one of the methods was preliminarily assessed on data from the participating institutions, to validate their applicability and interest for the problem at hand [44], [47]. Later, experiments were performed on the SPEET tool that included the data analysis techniques as well as all the requirements with regard to user interface and automated processing needed to achieve the stated purposes, i.e., to be used for support of the predictive modeling activities of third-party tutoring staff without a data scientist background [53].

Preliminary experiments led us to several decisions. First, an elbow analysis on the within-groups sum of squares was used to choose the number of clusters, which was set to three. This number of clusters is manageable enough to obtain students' patterns and also provides flexibility to adapt tutoring actions to segments of students, which could generally be seen as ''excellent'', ''average'' and ''low-performance students'' if we take into account the average of the scores obtained by all students at each cluster. For that reason, the number of classes in the performance classification experiments has also been set to three. Second, to provide a quicker execution of the clustering algorithm, a 2D projection of the original data was considered, i.e., the components obtained from PCA transformation. Preliminary results consistently showed similar results to those obtained with the full dimensional data.

B. TOOL

Two interactive visualizations have been implemented in this project [44]. The first is a coordinated view, which provides a set of interactive coordinated histograms where a user can establish filters by one or more variables, accordingly triggering the update of the other charts. The second is a data projection view, which provides a 2D scatterplot of the high-dimensional students' data, obtained by means of dimensionality reduction.

The aim of the coordinated view is to enable the exploration of the distributions of variables and of the links between them. This view has been implemented as an interactive dashboard, showing a set of coordinated histograms and barplots where a user can filter by one or more variables, causing that the rest of the charts update automatically. The charts, some fixed and the others customizable, show the count of student-subject records binned by interval/category. In order to define a filter by a variable or a set of variables, a user can click on one or more barplot columns, when variables are categorical, or select a range on a histogram through brushing, when it is numerical. After applying a filter, the visualization across all the panels updates automatically. The number of filtered units is also reported at the bottom of the visualization. On the other hand, a tooltip showing the corresponding relative frequency is shown when hovering over a barplot column.

With respect to data projection, an initial implementation (see Fig. 2) enabled to interactively adjust the parameters and to exploit the visual channels or graphical properties of each point of the 2D scatterplot (i.e., radius, shape, and color) for conveying additional information (i.e., values from the original variables) [44]. This tool can be applied to two cases: one where data are organized by year and another one where data are grouped by degree. An additional square is displayed in the bottom left side to allow users to select the weight of each year in the projection of the first case. In the second case, the degree can be selected by an additional menu.

Nevertheless, it must be noted that, after the feedback obtained from preliminary field tests, it was decided not to include this functionality in SPEET. Nevertheless, a simpler non-customizable 2D projection, using the PCA algorithm, has been kept to facilitate the visual interpretation of clustering.

Another auxiliary visualization is provided to support the interpretation of clustering. In this case, data are visualized as scores histograms, with the possibility of choosing the reference variable between students or subjects. The subject-based approach considers a given cluster and averages all the scores for each subject. The student-based option takes all the students belonging to a cluster and computes the average score of all the subjects of each student.

Preliminary experimentation with regard to classification considered the two approaches commented in Section III (SVM and MLP). Since the SVM approach provided similar results and a reduced complexity, this was the classifier considered to be included in the SPEET tool. For model assessment, data were split into training (80%) and test (20%) sets. Different kernels (linear, third-order polynomial and radial basis function) and different values of the penalty parameter (between 0.5 and 5) were tested. After testing the classifier with degrees at different universities of the consortium, it was found that a linear kernel with a penalty of 1 provided the best results.

Preliminary experiments were performed with the approaches proposed for the graduation prediction model. For model assessment, data have been split again into training (80%) and test (20%) sets. The initial model used a binary logistic mixed-effects model with all the covariates but, even without explicit autocorrelation, the random effects were complex for a fairly small data set. For that reason, in order to simplify the model, a single random effect on the

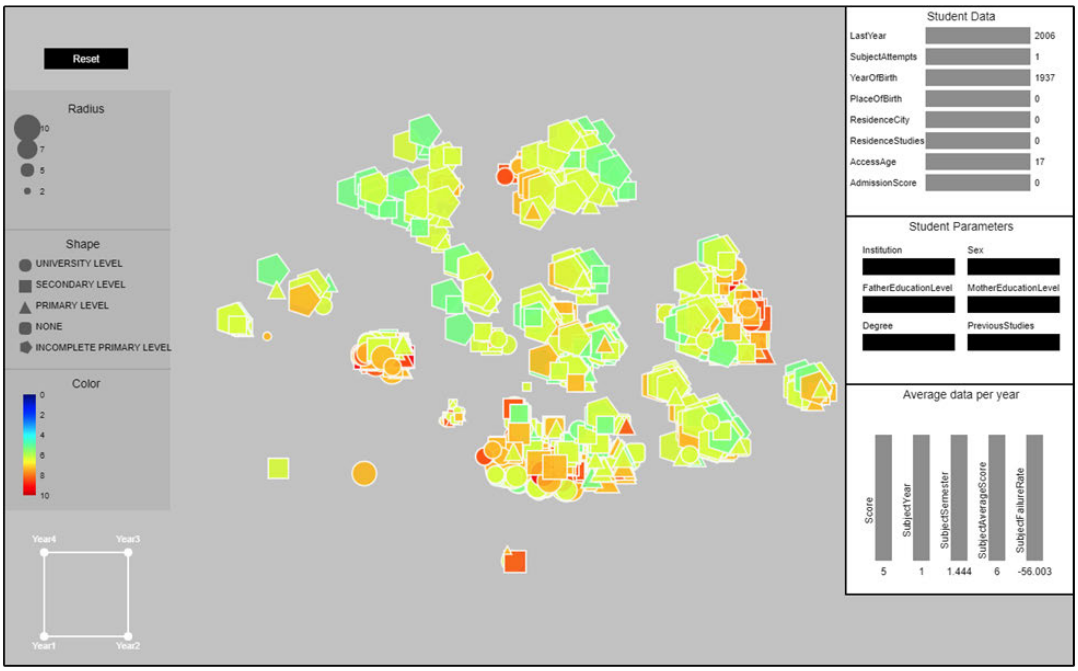

FIGURE 2. Screenshot of preliminary data projection implementation [53].

intercept is considered in model (2,3). The model is further simplified, removing the least significant covariates (a significant p-value is usually taken as ≤ 0.05) and performing a likelihood-ratio test relative to the initial model step-by-step. This is the model that has been implemented in the SPEET tool. The implementation uses a Laplace approximation as the numerical method, which is a compromise between accuracy and computational speed.

In the *SPEET tool*, the three tables described in section IV-A are uploaded as comma-separated values files by the user. Those files contain respectively the information about students, subjects and degrees. The uploaded information is processed to accommodate data to the requirements of data analysis algorithms but also to detect errors and/or missing data, since it is necessary to guarantee that the obligatory columns are not missing. If categorical columns are missing or unnecessary columns are present, warning messages are issued but the tool can still be executed.

C. SOFTWARE DEVELOPMENT

The SPEET tool displays different sections: information about the project, an interface for data uploading, data analysis and data visualization. The visualizations described in the previous subsection have been implemented as an interactive web application, which presents complete dashboards that display information and can be interactively adjusted to the needs of the analyst. The SPEET tool also implements a messaging system to help users and report errors, warnings and suggested solutions, which are represented by different colors.

The application has been developed using a Model-View-Controller (MVC) paradigm. In Figure 3, a simplified schema of the software structure is presented. There, one can observe the flow of the process between data upload and results presentation.

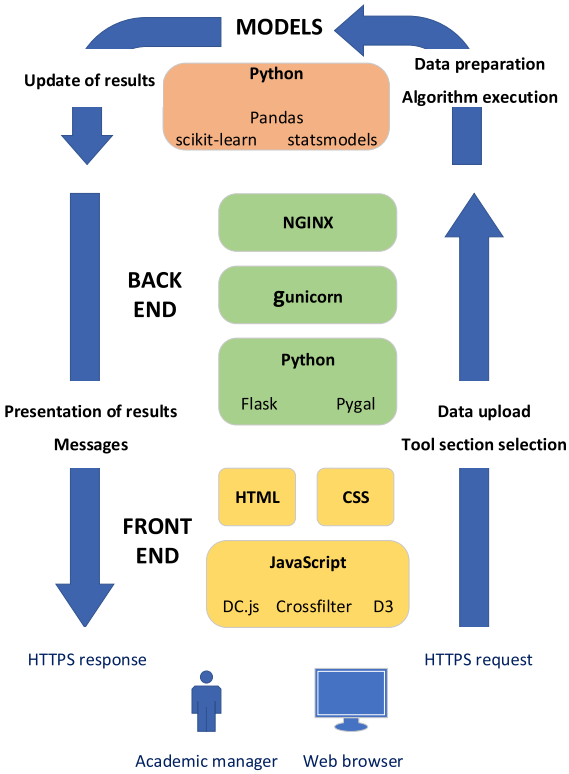

FIGURE 3. SPEET webtool architecture.

At the back-end side, data analysis was implemented in Python to facilitate the integration of methods presented in Section III, as well as future integrations. A Flask micro-framework is used as controller, in charge of running the Python-written data processing algorithms. For its deployment, a Gunicorn server and a Nginx reverse server have been used. With regard to the model, the Python Pandas library [54] was used to preprocess and structure data for all the experiments whereas scikit-learn [55] was used for the clustering, dimensionality reduction and classification algorithms. Initial implementations of the logistic and the mixed-effect model used, respectively, the stats and lme4 R packages. However, for its final implementation in SPEET, the generalized linear models (binomial family) available in the statsmodels Python library were used.

At the front-end part, HTML5, JavaScript and CSS3 were used to provide the interactivity and style required for visualization and reporting. Indeed, the front-end was designed to provide the desired interaction so that users can easily upload students' data sets, analyze them and obtain an output with the results. The interactive visualization that provides the coordinated view is implemented using several JavaScript libraries such as DC.js[2] and Crossfilter,[3] both of them based on D3.js [56]. Although preliminary visualizations obtained from nonlinear dimensionality reduction were implemented with Processing.js, the graphical results results of clustering and classification in SPEET are produced with the pygal Python library. Additionally, the results obtained for drop-out prediction were also summarized graphically using that tool.

Apart from usability, security and privacy were also addressed. In that sense, data transfer is SSL encrypted, files are removed after 48h and the requirements of the European General Data Protection Regulation are followed. Source code can be found at a public repository.[4]

V. RESULTS

A. PRELIMINARY RESULTS

In this section, we show a summary of the results we obtained by applying the different algorithms to the data of the project consortium. These preliminary results, explained in detail in the project technical reports [44], [47], [53], are compared with the ones reported in previous works.

1) EXPLORATORY ANALYSIS

a: VISUALIZATION

The application of nonlinear dimensionality reduction to data from the partners produced a 2-dimensional visualization where students are grouped according to the explanatory variables. However, this visualization was only interesting for the common set of students. The projections per degree generally contained a too small number of students, so visualizations were uninformative [44]. Furthermore, the feedback obtained in preliminary field tests from the members of the project and the exchange of opinions with the external audience of educational staff that participated in a set of dissemination events proved that users without a data science background found it difficult to interpret a 2-dimensional visualization where information is conveyed through relative pairwise distances rather than absolute positions. Since the main target of the project are tutors with no prior training, this visualization was not included in the first version of SPEET.

The usefulness of coordinated views was qualitatively assessed through the evaluation of hypotheses about the relationship between explanatory variables and performance. For that purpose, the teaching staff involved in the project formulated hypotheses about variables that they intuitively understood to have great influence on performance [53]. The use of filters on those explanatory variables showed, for instance, that students with high admission scores always provided better score distributions, that scores were higher for elective than for mandatory subjects and that other explanatory variables such as PreviousStudies or SubjectYear had a strong influence on performance but their relationship was highly dependent on the degree.

b: CLUSTERING

After applying the proposed clustering approach to all the degrees of the partners' consortium, it was possible to highlight three clusters of students with regard to their performance (i.e., low, mid, high-performing students) for all the institutions, which are generally well separated [53]. However, sometimes there is some overlapping between the low-performance and average clusters that is due to indistinguishable performance of both groups of students for a certain set of subjects. This situation can be observed through the analysis of the score histograms, where average-score students present a clear separation with few overlaps but average-score subjects show certain overlapping.

To complement the clustering analysis, the distribution of categorical variables (Gender, Previous Studies, Admission Score, Access to Studies Age and Nationality) at the different clusters was also analyzed. Some conclusions obtained from the students analyzed at the Universities of the project were the following:

1) Age: two cases were observed. For the cases where almost all students are 18/19 years old when they access the degrees, no clear pattern can be observed. If there is a non-negligible number of older students, it is observed that high-performance students tend to be young.
2) Admission score: a clear pattern is observed. The higher the admission score, the higher the obtained performance.
3) Gender: the number of women enrolled in the analyzed engineering degrees was low. Nevertheless, the proportion of high-performance students tend to be higher among women for most of the partners.

[2] https://dc-js.github.io/dc.js/
[3] https://github.com/crossfilter/crossfilter
[4] https://bitbucket.org/SPEET_PROJECT/speet_code

4) Previous Studies: a rather clear pattern of higher performance for students coming from secondary school was observed.

2) CLASSIFICATION

Concerning the classification accuracy, that is, the ability of the SVM algorithm to classify a new student into the obtained clusters, it was verified that classification performance presented satisfactory results but depended both on the degree type and the institution [53]. When the scores of the first three courses along with the categorical variables were considered for classification, accuracies of at least 85% were obtained (being higher than 90% in most of the cases). In the cases where only the first course or first+second courses were addressed, quite different accuracies were obtained depending on the type of degree and university. Although there were degrees reflecting a high accuracy level with categorical features and first course scores (up to 90%), in other cases accuracies rose 20% by adding the scores of the second course. An interpretation of these results is that the degree plan and/or nature has a great importance of student classification and, therefore, this analysis might be useful to extract insights and patterns through the comparison of institutions and degrees.

It is difficult to compare these results with the ones obtained in previous literature, due to the different experimental conditions. However, it can be noted that previous examples of academic performance classification have provided different results, with a reported effectiveness that ranges between 66.67% and 96.47%, depending on the variables that were considered, their degree and the specific definition of success [28]. It is also reported that it is more difficult to correctly classify average-performing students. Although this fact was not analyzed in our case, it seems to agree with the results obtained in clustering, as commented above. In [26], binary classification results on a course-wise basis showed a highest average of 0.60 ± 0.17, using only demographic variables. In [29], the results of classification in five categories of university students, based again on explanatory variables, showed prediction rates around 60%, although classification was better for students labeled as good or very good. In [27], the accuracies reported were higher for 2 classes (over 90%) and lower for 4 classes (over 60%). So, in general, it seems that higher accuracies are obtained for sufficiently aggregated classes [30]. In [30], which only used previous scores, the prediction of graduation performance showed an accuracy of up to 83.65% and that the pre-enrollment performance and the scores in the first and second year courses were the most relevant factors in a degree of Information Technology. Furthermore, they discovered that the performance of students tends to stay the same throughout the degree, and so students remain in the same classes.

3) DROP-OUT PREDICTION

The proposed drop-out prediction was initially tested by the university that developed the tool. Model accuracy on the test set reached 90.9%, with a sensitivity of 97.0% and a specificity of 75.4%, which encouraged the application of the proposed predictor [47]. Afterwards, data from two universities were analyzed with the proposed approach [53]. Common patterns were observed for these institutions: specifically, it was found that access age has a negative impact on graduation (i.e., younger students are more likely to graduate), whereas admission score and the performance of the student in the first semester (the weighted average score and the number of ECTS obtained) have a positive impact on graduation. The rest of variables did not consistently show a remarkable impact on dropout.

With regard to drop-out prediction, previous works in the literature reported accuracies in higher education that range from 0.7 to 0.8 [57]. Furthermore, the features related to the incidence of dropout have also been studied in the literature. The results obtained in [32] are interesting, due to the similarity of its target and its available variables, although the study focuses on the Spanish higher education system previous to the Bologna Process (instead of several European post-Bologna higher education systems) and does not constrain itself to engineering degrees. Some of their conclusions match those obtained in our work, i.e., that early performance (specifically pre-enrollment academic performance) is a good predictor and that age is positively associated with dropping out. Other findings by [32] were that students are more likely to drop out in the first year of studies and that the lack of parental higher education and displacement, as well as being male in the case of engineering, are positively associated with dropping out.

B. DEVELOPED TOOL

In this section, the tool that resulted from the application of the steps presented in section IV-A is described. A technical description and use guide is presented in [58].

The SPEET tool has been developed to be user-friendly and easily accessible, so that other faculties and schools outside of the project's consortium are able to easily perform a preliminary analysis of their students based on their own data after they are organized accordingly. Three templates (.csv format), which correspond to the structure described in Table 1, are provided at the website to allow other institutions to arrange their data: SubjectsPerformance.csv, Students.csv and Degrees.csv.

If data is not organized as indicated, some errors are shown by the tool along with suggestions to fix the detected problems. Once data is uploaded, the user can select two possible options: either *Clustering, Dropout and Classification* or *Coordinated views*. In the Clustering, Dropout and Classification case, five different results are provided (see Figure 4):

- Performance Clusters, projected onto a 2-dimensional representation.
- Scores Histograms, with which users can interact by choosing the reference variable (Students or Subject).

FIGURE 4. Example of execution output of the Clustering, Drop out and Classification Tool [58].

- Categorical Study, with which users can interact by choosing both the categorical dimension and the normalization setting.
- Classification Analysis, where users can observe the classification results obtained considering early performance.
- Dropout Analysis: Graduation Prediction Model, summarizing graphically both the impact in graduation of each variable, their significance and the total accuracy.

As observed, results presented by the tool are aligned with those commented at previous sections. In this particular example (see Figure 4), we are providing the results the user can obtain from the Clustering, Drop-out and Classification tool. For the ease of interpretation, we have added numbers at the figure in this paper to guide the reader. In (1), the user can easily observe that three clusters of students are generated based on their performance. In (2) and (3), the tool shows the way average scores and categorical variables are distributed at each cluster. In (4), the classification result provided by the tool for the uploaded data is shown. As observed, a classification accuracy ranging from 85% to 91% is obtained based on the amount of information used for classification. In this example, a high accuracy level with only first course information could mean that the first year of that degree is very important and performance during this year clearly determines the student's outcome. Finally, an accuracy of drop-out prediction equal to 85% is observed at (5). Besides that, the tool provides the significance level and impact on graduation for the different variables used for this prediction. On one hand, the most significant variables in this example are the weighted average score and the total number of credits obtained by the student at the first semester of the first year (Wei Avg Eval 11 and Total Credits 11 variables, respectively). On the other hand, one can also observe that these two variables have a positive impact on graduation probability.

In the Coordinated Views section, the tool analyzes the whole data uploaded in the session. The statistical unit is a single student-subject interaction (subject score). The distribution of subject scores is visualized across different variables using coordinated histograms and barplots. The user can interact with Coordinated Views in different ways, letting users filter out the subject scores that are not interesting for their analysis. Its operation is shown in Figure 5, where red numbers have been added to help with the explanation. For instance, a user can filter by a categorical variable, by clicking on one or more barplot columns. That is the case of (1) and (2), where the user has selected only the scores corresponding to men studying Mechanical Engineering. It is also possible to filter by a numerical variable, through the selection of a range in a histogram, such as in (3) and (6), where subjects of the third and fourth year and students with an admission score higher than 6.5 are respectively selected. Each single filter can be reset from its corresponding visual, while all filters can be reset at the same time through a link at the bottom of the page (4). The number of filtered units is also reported there. After applying a filter, the visualization is updated automatically across all its panels. The user can observe the distribution of interest, such as the scores obtained by the students under these conditions (5) and even hover over a barplot column to check its corresponding relative frequency. It is also possible to analyze the distribution the average score with respect to an explanatory variable, which can be selected from a dropdown (7). This feature could let a user find correlations.

VI. DISCUSSION

A. USEFULNESS OF THE TOOL

To evaluate the usefulness of the developed tool, it is necessary to analyze whether the stated goals are met. The aim was, in short, the development of a web-based tool for tutoring support, under the following constraints:

- Target users are not expected to have a strong background in data science and therefore the use and reports from the tool should be intuitive.
- Target users are expected to be transnational, so data structure should be as general as possible to guarantee that users are able to easily gather those data from institutions of different countries.
- The expected analyses should help users to find and study profiles of students with regard to their performance at the degree level and their risk of dropout.

To analyze usability and intuitiveness, we first need to identify the greatest obstacles a user might face when using the platform. It can be argued that the most important one is data gathering, which is unavoidable. Complexity of the files required by the platform is minimum, since their format can be created and edited almost universally, and only variables that are generally available in the universities are considered. However, even with this stress on simplicity, the task of the data manager remains critical for data collection.

Another difficulty would be the interpretability of results. The tool makes an extensive use of information visualization not only for the explanatory analysis but also to report the classification and drop-out prediction results. The coordinated view enables an interactive visual analysis, placing the user as the center of a loop of knowledge discovery. The other visualizations mainly use color as the feature channel, so that different categories stand out (in clustering and histograms) or to provide a natural interpretation for the classification analysis or the significance level of features in drop-out analysis (red-low and green-high). In drop-out, the mathematical symbol is also used to convey positive or negative impact. Except for the case of the low-dimensional representation for clustering (a scatterplot where information is provided by the relative distances among points), standard 2D bar chart displays are used. For those reasons, we argue that the user interface is simple enough to be used readily or after short training.

Finally, system operation under unexpected errors is a common source of frustration among users. For that reason, the platform is designed to provide guidance in the presence of usual errors. Indeed, 45 informative messages have been

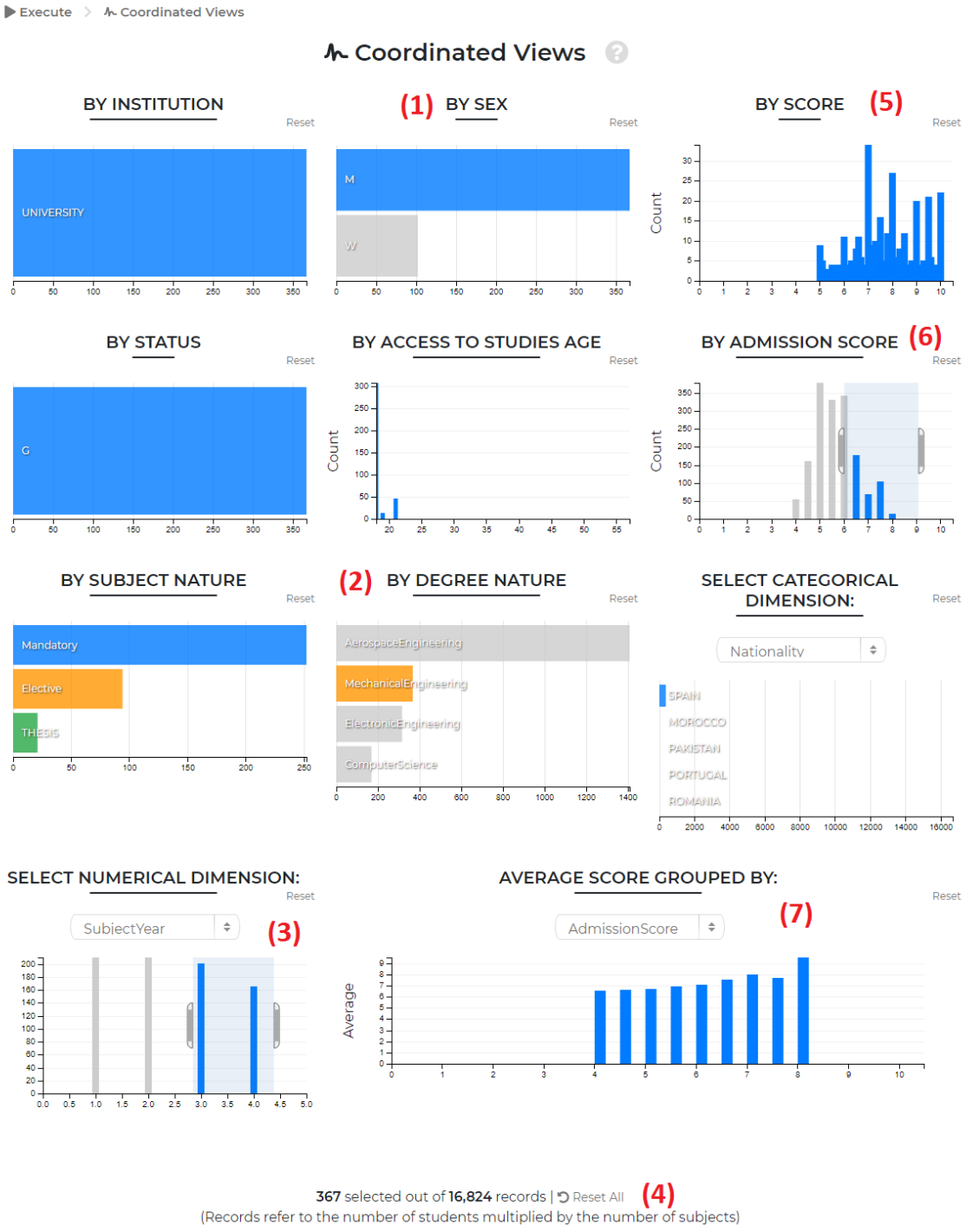

FIGURE 5. Example of execution output of the coordinated views tool.

defined, most of them related to common problems during preprocessing (see [58]).

With respect to the potential of the tool for its application to data from institutions of different countries, most features considered in the tool are expected to be stored in any database of a university administration: demographic variables, such as gender, age or nationality, and explanatory variables of degrees and subject. Local disparity is more likely to be related with length, which can consider different units (although in the European Higher Education Space the European Credit Transfer and Accumulation System is used as an standard), admission score (which might not be required in certain countries) and scores (which in some countries are rank-based instead of absolute). Nevertheless, the proposed

data structure was found to be reasonable for all the partners of the project, which comprised institutions from 5 different countries of the European Union, with different sizes and degrees taught in 7 languages.

With regard to the usefulness of the provided analyses, it must be noted that the tool provides three perspectives about the global student performance, along with an interactive visualization designed for self-guided exploration of hypotheses. As described in Section V-A, preliminary analyses showed value to obtain insights about student performance.

At the final stage of the project, the proposed tool was presented, in the context of a training workshop, to 9 people belonging to the institutions of the project consortium but who were not involved in the development of the project. The roles of the staff involved in this event were selected to cover different profiles of potential end users, such as University managers, data managers or tutors. The aim was not constrained to assess the proposed platform, but rather to analyze the potential and limitations of academic data analysis when dealing with the specific administrative and technical situations of each institution. However, the participants were able to analyze the data from their own institution and provided some feedback. The most relevant impressions about the proposed platform were that clustering/classification tools are useful to understand patterns of different students' profiles, whereas the coordinated views can reinforce the analysis thanks to their dynamic behavior and intuitiveness. Besides, it was agreed that dropout analysis is a necessary step to solve a common problem found at Universities and that one of the applications of the tool could be to give advice to students in risky situations.

B. SUPPORT IN TUTORING ACTION

The practical applications of the SPEET platform in tutoring are constrained to the identification and analysis of students' profiles. Tutors can use the tool in a first stage of exploratory analysis to identify, in a short period of time, the most salient particularities of student profiles in a certain degree. Otherwise, tutors should require much more effort to reach to these conclusions.

Indeed, the preliminary experiments were driven by some questions that presumably tutors would pose during this first profiling stage, such as:

- Can we separate students at different groups based on their performance behavior?
- Can we observe clear students' profiles at these groups based on categorical variables such as age, admission score, sex, previous studies?
- Can we observe if early performance determines the behavior of students at one degree?
- Can group separation be explained by the way categorical variables are distributed?
- Can we formulate hypotheses about the relationship between explanatory variables and performance?
- Does any score distribution grouped by an explanatory variable show an evident trend?

The results obtained with academic data from partner organizations provided positive answers to those questions, linked to the perspective of tutoring activity. Even if that is not the case for other scenarios, we argue that the findings would still be interesting for tutoring, because they would provide insight about the students' profiles.

However, the subsequent process that tutors would follow to promote the students' successful completion of university studies is by no means trivial or automatic. The SPEET tool only provides global hints about students' profiles and performance and cannot identify problems of specific students. Student data are anonymized, and any in-depth analysis of the situation would need to consider new information that is not included in the tool. For that reason, once this exploration stage is finished, tutors should rely on their experience to define a protocol of action to address the profiles associated to failure and dropout. This course of action will probably require gathering more information to understand the inner causes of low performance and will result in prevention or intervention strategies. This pedagogical work is beyond the scope of this paper, which focuses on the technological aspects.

VII. CONCLUSION, LIMITATIONS AND FUTURE WORK

This paper presents a web-based software tool for student profiling, providing support to tutoring staff without a data scientist background. The presented tool is focused on the analysis and forecasting of students' performance, in terms of the observable scores and of the completion of studies. The study has focused in Engineering Bachelor degree programs currently running at higher education institutions from 5 different countries of the European Union, with different sizes and degrees taught in 7 languages. For those reasons, the considered variables are those commonly found in the administrative records of the students (student's explanatory variables, student's performance and information about subjects and degrees) and analyses are aimed to provide a global, degree-wide view of performance, instead of course-wise.

To achieve these purposes, a variety of exploratory and predictive analyses have been proposed:

- Clustering and visualization in a low-dimensional representation to help end user to understand the data structure.
- The coordinated visualization of histograms, which can be automatically updated on the basis of custom filters set interactively, to facilitate the validation of hypotheses about a set of students.
- Classification of already graduated students to three performance levels using exploratory variables and early performance information, to understand the degree of course-dependency of students' behavior at different degrees.
- Risk of dropout, computed on the basis of first semester performance and selected explanatory variables, since

the analysis of the impact of the student's variables and early performance in the graduation probability can lead to a better understanding of the causes of dropout.

A preliminary application of the proposed algorithms to data of the 6 higher education institutions led to the following observations:

- Exploratory Analysis: The usefulness of coordinated views for a self-guided exploration of hypotheses was qualitatively assessed through the confirmation of assumptions suggested by the teaching staff.
- Clustering: After applying the proposed clustering to all the degrees, it was possible to find three clusters of students with regard to their performance for all the institutions, which are generally well separated.
- Classification: It was possible to to satisfactorily classify a new student into the obtained clusters but accuracy depended both on the degree type and the institution.
- Drop-out prediction: The proposed drop-out prediction model showed a high accuracy, with better sensitivity than specificity. Common patterns were observed: specifically, it was found that access age has a negative impact on graduation whereas admission score and the performance of the student in the first semester have a positive impact on graduation.

The successful results of the proposed methods in preliminary experiments motivated their implementation in the SPEET tool, which is conceived to provide easy upload and processing of the academic data, different perspectives on the information, and interactivity for self-guided exploration. The reliance on visualization and the preprocessing support messages make the tool easy to use. The data structure has been kept simple enough to be applicable to diverse institutions. The positive feedback about the SPEET tool by managing and tutoring staff of the 6 universities of the project let us conclude that it offers remarkable insight about the features related to student performance, which can become potentially useful for an initial stage of student profiling.

As an immediate continuation work, it is expected to extend the experiences with these tools to other knowledge domains to find how students' profiles in humanities and sciences differ from engineering students analyzed in this work. It would also be useful to add further information about classroom attendance and results at the course level, obtained from learning management systems, to the analysis. Including these variables is a pre-requisite to study, for instance, the effects of teaching methodologies. Nevertheless, an agreement on methodologies and learning management tools is needed to guarantee enough consistency, so experiments might be constrained to only one degree.

The tool should also be used in routine and massive conditions to gather feedback from tutoring offices. Feedback should lead to its improvement but also to motivate its extension to elaborate more comprehensive analyses. It would also be interesting to design a structured approach such as a laboratory user study to validate the usefulness of the proposed platform and the appropriateness of the design choices and abstractions for the target user population.

REFERENCES

[1] J. J. Vossensteyn, A. Kottmann, B. W. Jongbloed, F. Kaiser, L. Cremonini, B. Stensaker, E. Hovdhaugen, and S. Wollscheid, "Dropout and completion in higher education in Europe," Eur. Union, Brussels, Belgium, Tech. Rep. NC-04-15-779-EN-N, 2015.

[2] C. Lang, G. Siemens, A. Wise, and D. Gasevic, *Handbook of Learning Analytics*. Beaumont, AB, Canada: SOLAR, Society Learning Analytics Research, 2017, doi: 10.18608/hla17.

[3] A. Peña-Ayala, "Learning analytics: A glance of evolution, status, and trends according to a proposed taxonomy," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 8, no. 3, p. e1243, May 2018.

[4] R. M. M. F. Luis, M. Llamas-Nistal, and M. J. F. Iglesias, "Enhancing learners' experience in e-learning based scenarios using intelligent tutoring systems and learning analytics: First results from a perception survey," in *Proc. 12th Iberian Conf. Inf. Syst. Technol. (CISTI)*, Jun. 2017, pp. 1–4.

[5] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telematics Informat.*, vol. 37, pp. 13–49, Apr. 2019.

[6] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15991–16005, 2017.

[7] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 6, pp. 601–618, Nov. 2010.

[8] O. Viberg, M. Hatakka, O. Bälter, and A. Mavroudi, "The current landscape of learning analytics in higher education," *Comput. Hum. Behav.*, vol. 89, pp. 98–110, Dec. 2018.

[9] R. Ferguson, "Learning analytics: Drivers, developments and challenges," *Int. J. Technol. Enhanced Learn.*, vol. 4, nos. 5–6, pp. 304–317, 2012.

[10] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education," *EDUCAUSE Rev.*, vol. 46, no. 5, pp. 31–40, 2011.

[11] G. Siemens and D. Gasevic, "Guest editorial–learning and knowledge analytics," *Educ. Technol. Soc.*, vol. 15, no. 3, pp. 1–3, 2012.

[12] J. T. Avella, M. Kebritchi, S. G. Nunn, and T. Kanai, "Learning analytics methods, benefits, and challenges in higher education: A systematic literature review," *Online Learn.*, vol. 20, no. 2, pp. 13–29, 2016.

[13] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 3, no. 1, pp. 12–27, 2013.

[14] R. S. J. D. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *J. Edu. Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.

[15] C. Vieira, P. Parsons, and V. Byrd, "Visual learning analytics of educational data: A systematic literature review and research agenda," *Comput. Edu.*, vol. 122, pp. 119–135, Jul. 2018.

[16] C. C. Gray and D. Perkins, "Utilizing early engagement and machine learning to predict student outcomes," *Comput. Edu.*, vol. 131, pp. 22–32, Apr. 2019.

[17] A. Ortigosa, R. M. Carro, J. Bravo-Agapito, D. Lizcano, J. J. Alcolea, and O. Blanco, "From lab to production: Lessons learnt and real-life challenges of an early student-dropout prevention system," *IEEE Trans. Learn. Technol.*, vol. 12, no. 2, pp. 264–277, Apr. 2019.

[18] R. Ferguson, L. P. Macfadyen, D. Clow, B. Tynan, S. Alexander, and S. Dawson, "Setting learning analytics in context: Overcoming the barriers to large-scale adoption," *J. Learn. Anal.*, vol. 1, no. 3, pp. 120–144, Sep. 2014.

[19] R. Vilanova, M. Dominguez, J. Vicario, M. Prada, M. Barbu, M. Varanda, P. Alves, M. Podpora, U. Spagnolini, and A. Paganoni, "Data-driven tool for monitoring of students performance," *IFAC-PapersOnLine*, vol. 52, no. 9, pp. 190–195, 2019.

[20] M. Barbu, R. Vilanova, J. Lopez Vicario, M. J. Varanda, P. Alves, M. Podpdora, M. A. Prada, A. Morán, A. Torreburno, S. Marin, and R. Tocu, "Data mining tool for academic data exploitation: Literature review and first architecture proposal," Erasmus+ KA2 / KA203 project SPEET—Student Profile for Enhancing Engineering Tutoring, Instituto Politécnico de Bragança, Bragança, Portugal, Tech. Rep. SPEET-IO1, 2017.

[21] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, "Visual analytics: Definition, process, and challenges," in *Information Visualization*. Berlin, Germany: Springer, 2008, pp. 154–175.

[22] Y. Chen, Q. Chen, M. Zhao, S. Boyer, K. Veeramachaneni, and H. Qu, "DropoutSeer: Visualizing learning patterns in massive open online courses for dropout reasoning and prediction," in *Proc. IEEE Conf. Vis. Anal. Sci. Technol. (VAST)*, Oct. 2016, pp. 111–120.

[23] J. Zimmerman, K. H. Brodersen, H. R. Heinimann, and J. M. Buhmann, "A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance," *J. Edu. Data Mining*, vol. 7, no. 3, pp. 151–176, 2015.
[24] V. Tinto, "Student attrition and retention," in *The Encyclopedia of Higher Education*, vol. 3. Oxford, U.K.: Pergamon Press, 1992, pp. 1697–1709.
[25] K. E. Arnold and M. D. Pistilli, "Course signals at purdue: Using learning analytics to increase student success," in *Proc. 2nd Int. Conf. Learn. Analytics Knowl. (LAK)*, 2012, pp. 267–270.
[26] P. Strecht, L. Cruz, C. Soares, J. Mendes-Moreira, and R. Abreu, "A comparative study of classification and regression algorithms for modelling students' academic performance," in *Proc. 8th Int. Conf. Educ. Data Mining*, Madrid, Spain, Jun. 2015, pp. 392–395.
[27] N. T. Nghe, P. Janecek, and P. Haddawy, "A comparative analysis of techniques for predicting academic performance," in *Proc. 37th Annu. Frontiers Edu. Conf.-Global Eng., Knowl. Borders, Opportunities Passports*, Oct. 2007, pp. 7–12.
[28] E. Pathros Ibarra Garcia and P. Medina Mora, "Model prediction of academic performance for first year students," in *Proc. 10th Mex. Int. Conf. Artif. Intell.*, Dec. 2011, pp. 169–174.
[29] D. Kabakchieva, "Predicting student performance by using data mining methods for classification," *Cybern. Inf. Technol.*, vol. 13, no. 1, pp. 61–72, 2013.
[30] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Comput. Edu.*, vol. 113, pp. 177–194, Oct. 2017.
[31] A. U. Khasanah, "A comparative study to predict student's performance using educational data mining techniques," in *Proc. IOP Conf. Ser., Mater. Sci. Eng.*, 2017, vol. 215, no. 1, Art. no. 012036.
[32] G. Lassibille and L. N. Gómez, "Why do higher education students drop out? Evidence from Spain," *Edu. Econ.*, vol. 16, no. 1, pp. 89–105, 2008.
[33] C. Aina, "Parental background and university dropout in Italy," *Higher Edu.*, vol. 65, no. 4, pp. 437–456, Apr. 2013.
[34] *Regulation (EU) 2016/679 of 27 April 2016 on the Protection of Natural Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)*, Eur. Parliament Council Eur. Union, Brussels, Belgium, 2016.
[35] D. Ifenthaler and C. Widanapathirana, "Development and validation of a learning analytics framework: Two case studies using support vector machines," *Technol., Knowl. Learn.*, vol. 19, nos. 1–2, pp. 221–240, Jul. 2014.
[36] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," *Sci. Data*, vol. 4, 2017, Art. no. 170171, doi: 10.1038/sdata.2017.171.
[37] S. Rovira, E. Puertas, and L. Igual, "Data-driven system to predict academic grades and dropout," *PLoS ONE*, vol. 12, no. 2, Feb. 2017, Art. no. e0171207.
[38] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.
[39] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab., Statist.*, L. LeCam and J. Neyman, Eds. Berkeley, CA, USA: Univ. California Press, 1967, pp. 281–297.
[40] R. Vilanova, J. Vicario, M. Prada, M. Barbu, M. Dominguez, M. J. Pereira, M. Popdora, U. Spagnolini, P. Alves, and A. Paganoni, "SPEET: Software tools for academic data analysis," in *Proc. EDULEARN Int. Conf. Edu. New Learn. Technol.*, 2018, pp. 1–10.
[41] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: A comparative review," Tilburg Centre Creative Comput., Tilburg Univ., Tilburg, The Netherlands, Tech. Rep. TiCC TR 2009-005, 2009.
[42] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.
[43] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[44] M. Prada, A. Domínguez, A. Morán, R. Vilanova, J. L. Vicario, M. J. Varanda, P. Alves, M. Podpdora, M. Barbu, A. Torreburno, U. Spagnolini, and A. Paganoni, "Data mining tool for academic data exploitation: Graphical data analysis and visualization," Erasmus+ KA2 / KA203 project SPEET—Student Profile for Enhancing Engineering Tutoring, Instituto Politécnico de Bragança, Bragança, Portugal, Tech. Rep. SPEET-IO3, 2018.
[45] M. Domínguez, R. Vilanova, M. Prada, J. Vicario, M. Barbu, M. J. Pereira, M. Podpora, U. Spagnolini, P. Alves, and A. Paganoni, "SPEET: Visual data analysis of engineering students performance from academic data," in *Proc. Learn. Anal. Summer Inst. Spain*, 2018, pp. 50–61.
[46] I. Díaz Blanco, A. A. Cuadrado Vega, D. Pérez López, M. Domínguez González, S. Alonso Castro, and M. Á. Prada Medrano, "Energy analytics in public buildings using interactive histograms," *Energy Buildings*, vol. 134, pp. 94–104, Jan. 2017.
[47] J. Lopez 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, "Data mining tool for academic data exploitation: Selection of most suitable algorithms," Erasmus+ KA2 / KA203 project SPEET—Student Profile for Enhancing Engineering Tutoring, Instituto Politécnico de Bragança, Bragança, Portugal, Tech. Rep. SPEET-IO2, 2018.
[48] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[49] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[50] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer-Verlag, 2001.
[51] L. Fontana and A. M. Paganoni, "Analysis of dropout in engineering BSc using logistic mixed-effect models," in *XLIX Scientific Meeting of the Italian Statistical Society*. London, U.K.: Pearson, 2018.
[52] C. E. McCulloch and J. M. Neuhaus, "Generalized linear mixed models," in *Encyclopedia of Biostatistics*, vol. 4. Hoboken, NJ, USA: Wiley, 2005.
[53] M. Barbu, R. Vilanova, J. Lopez Vicario, M. J. Varanda, P. Alves, M. Podpdora, A. Kawala-Janik, M. A. Prada, Domínguez, U. Spagnolini, and L. Fontana, "Data mining tool for academic data exploitation: Publication report on engineering students profiles," Erasmus+ KA2 / KA203 Project SPEET—Student Profile for Enhancing Engineering, Instituto Politécnico de Bragança, Bragança, Portugal, Tech. Rep. SPEET-IO4, 2019.
[54] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. Newton, MA, USA: O'Reilly Media, Inc, 2012.
[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[56] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 data-driven documents," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 12, pp. 2301–2309, Dec. 2011.
[57] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting students drop out: A case study," in *Proc. Int. Work. Group Educ. Data Mining*, 2009, pp. 1–10.
[58] 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, "Data mining tool for academic data exploitation: Webtool description and usage," Erasmus+ KA2 / KA203 project SPEET—Student Profile for Enhancing Engineering Tutoring," Instituto Politécnico de Bragança, Bragança, Portugal, Tech. Rep. SPEET-IO5, 2019.

MIGUEL ÁNGEL PRADA was born in 1981. He received the bachelor's degree in computer engineering and the Ph.D. degree from the Universidad de León, León, Spain, in 2003 and 2009, respectively.

He has worked at Aalto University, Finland, and the Universidad de León, Spain. He is currently an Associate Professor with the Department of Electrical and Systems Engineering, Universidad de León. He is the author of 19 articles in recognized scientific journals and over 45 papers in international conferences. His research interests include machine learning, visualization, industrial cybersecurity, and innovation of education in control engineering.

MANUEL DOMÍNGUEZ (Member, IEEE) was born in 1956. He received the M.S. degree in industrial engineering and the Ph.D. degree from the Universidad de Oviedo, Gijón, Spain, in 2003.

He is currently a Professor of automatic control with the Department of Electrical and Systems Engineering, School of Engineering, Universidad de León, León Spain. He is author of more than 100 publications in international conferences and journals, as well as several book chapters. His main research interests are industrial remote laboratories, the development of innovative tools for education in automatic control, remote monitoring, automation, and industrial cybersecurity. He is a member of IFAC. He is also a member of the Spanish Committee on Automatic control.

JOSE LOPEZ VICARIO (Member, IEEE) received the bachelor's degree in electrical engineering and the Ph.D. degree *(cum laude)* from the Universitat Politècnica de Catalunya (UPC), Barcelona, in 2002 and 2006, respectively, and the M.B.A. degree from the IESE Business School, Universidad de Navarra.

He is currently an Associate Professor with the Universitat Autònoma de Barcelona (UAB). He has a wide expertise on signal processing and machine learning, publishing 39 articles in recognized international journals, more than 50 papers in conferences and supervising five Ph.D. Thesis (more than four in progress). He received the 2005/06 best Ph.D. prize in Information Technologies and Communications by the UPC and holds the Advanced Research accreditation issued by Catalan Government.

PAULO ALEXANDRE VARA ALVES received the master's degree in multimedia technology from the University of Porto, Portugal, in 2001, and the Ph.D. degree in technology and information systems from the University of Minho, Portugal, in 2008. He is a Coordinator Professor with the Department of Computer Science, Technology and Management School, Polytechnic Institute of Bragança. He is a Vice-President of Pedagogical Council of the School of Technology and Management and e-learning coordinator at the Polytechnic Institute of Bragança. He is an integrated member of the Research Centre in Digitalization and Intelligent Robotics. His research interests include multimedia, e-learning, web development, learning analytics, academic analytics, machine learning, big data analytics, data science, and business intelligence.

MARIAN BARBU (Senior Member, IEEE) was born in Galaţi, Romania, in 1978. He received the B.S., M.S., and Ph.D. degrees from ''Dunărea de Jos'' University of Galaţi, in 2001, 2002, and 2006, respectively, all in control engineering.

He is currently a Full Professor with ''Dunarea de Jos'' University of Galaţi. In 2005, he was a Young Researcher with the Technical University of Crete, Greece, and from 2015 to 2016 and in 2017, he was an Invited Professor at Universitat Autonoma de Barcelona, Spain. His research interest is focused on modeling and control of energy and environmental systems, having results published in top journals such as: *ISA Transactions; Water Research; Chemical Engineering Journal; Journal of Cleaner Production; Industrial and Engineering Chemistry Research*. He is a member of the IFAC TC8.3. Modeling and Control of Environmental Systems.

MICHAL PODPORA received the B.Sc. and M.Sc. degrees in computer engineering and the Ph.D. degree in control engineering and robotics from the Opole University of Technology, Opole, Poland, in 2002, 2004, and 2012, respectively.

Since 2010, he has been a Research Department Manager and a Project Manager of research projects for industry. He is currently employed as a Humanoid Robot Developer with the Research Department, Weegree Sp. z o.o S.K., Poland, and as an Assistant Professor with the Department of Computer Science, Opole University of Technology. His main research interests include cybernetics, artificial intelligence in robotics, machine vision, cognitive systems, education, big data, smart infrastructure, embedded systems, IoT, cybersecurity, and forensic science.

UMBERTO SPAGNOLINI (Senior Member, IEEE) is currently a Faculty member with Politecnico di Milano, since 1990, where he is a Professor of statistical signal processing. His research in statistical signal processing covers remote sensing and communication systems with more than 300 papers on peer-reviewed journals/conferences and patents. He is author of the book *Statistical Signal Processing in Engineering* (J. Wiley, 2017). The specific areas of interest include channel estimation and space-time processing for single/multi-user wireless communication systems, cooperative and distributed inference methods including V2X systems, mmWave communication systems, parameter estimation/tracking, and wavefield interpolation for remote sensing (UWB radar and oil exploration). He was a recipient/co-recipient of Best Paper Awards from EAGE on geophysical signal processing methods (1991, 1998), and IEEE on array processing (ICASSP 2006) and distributed synchronization for wireless sensor networks (SPAWC 2007, WRECOM 2007). He served as part of IEEE Editorial boards as well as member in technical program committees of several conferences for all the areas of interests.

MARIA J. VARANDA PEREIRA was born in Braga, in November 1971. She received the M.Sc. and Ph.D. degrees in computer science from the University of Minho, in 1996 and 2003, respectively.

She is integrated member of the Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança and collaborator member of the Language Processing group in the Algoritmi Research Center, University of Minho. She is currently a Coordinator Professor with the Technology and Management School, Polytechnic Institute of Bragança, and she is a vice-president of the Technology and Management School. As a Computer-Science Researcher from 25 years ago, she is interested and usually supervises Master and Ph.D. students in the following areas: domain specific software development, visualization tools, human-computer interaction, QA systems, data science, machine learning, learning analytics, generation of virtual learning spaces and computer-assisted education. She is coauthor of 20 articles in journals and 77 articles at international conferences, more than half of which are indexed. She is responsible for or participates in several research projects with international partners from countries such as: Poland, Italy, Spain, Romania, Serbia, Slovenia, Argentina, and Brazil.

RAMON VILANOVA (Member, IEEE) graduated in the Universitat Autònoma de Barcelona in 1991. He received the Ph.D. degree from Universitat Autònoma de Barcelona in 1996.

He is currently a Full Professor of automatic control and systems engineering with the School of Engineering, Universitat Autònoma de Barcelona, where develops educational task teaching subjects of signals and systems, automatic control and technology of automated systems. His research interests include methods of tuning of PID regulators, systems with uncertainty, analysis of control systems with several degrees of freedom, application to environmental systems and development of methodologies for design of machine-man interfaces. He is author of several book chapters and has more than 100 publications in international congresses/journals. He is a member of IFAC and IEEE-IES. He is also member of the Technical Committee on Factory Automation.