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

INTELLECTUAL OUTPUT #1

Student Profile for
Enhancing Tutoring Engineering

Student Profile
for Enhancing
Engineering Tutoring

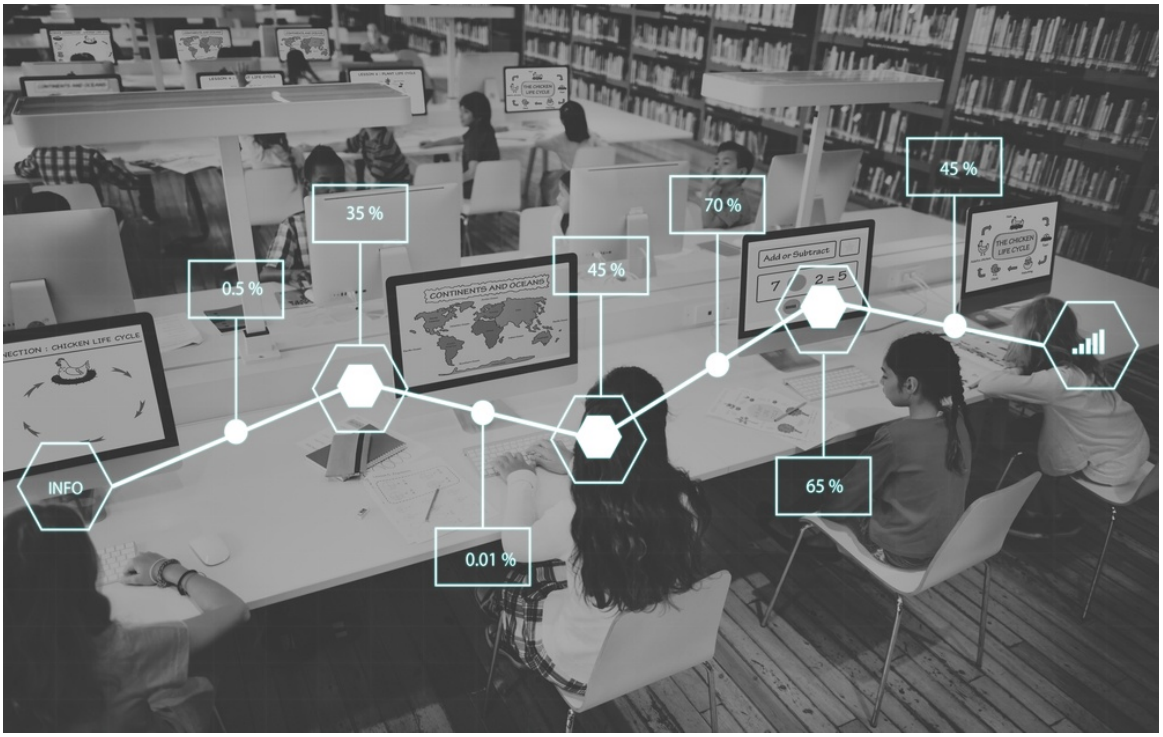
ERASMUS + KA2 / KA203

Data Mining Tool for Academic Data Exploitation

Literature review and first architecture proposal

M. Barbu (Coordinator), R. Vilanova, J. Lopez Vicario, M.J. Varanda, P. Alves, M. Podpora, M.A. Prada, A. Morán, A. Torrebruno and S. Marin, R. Tocu

June 2017



ISBN 978-972-745-228-6

ERASMUS + KA2 / KA203

INTELLECTUAL OUTPUT #1
June 2017

Data Mining Tool for Academic Data Exploitation

Literature review and first architecture proposal

M. Barbu (Coordinator)

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

R. Vilanova, J. Lopez Vicario

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

M.J. Varanda, P. Alves

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

M. Podpora

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

M.A. Prada, A. Morán

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

A. Torrebruno

Scuole di Ingegneria
Politecnico di Milano
Milano, Italy

S. Marin, R. Tocu

Department of Pedagogical Personnel Training
"Dunarea de Jos" University of Galati
Domneasca 47, 800008
Galati, Romania

Final Version

ERASMUS + KA2/KA203

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

Disclaimer: The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Table of Contents

1 Executive Summary

Using data for making decisions is not new; companies use complex computations on customer data for business intelligence or analytics. Business intelligence techniques can discern historical patterns and trends from data and can create models that predict future trends and patterns. Analytics, broadly defined, comprises applied techniques from computer science, mathematics, and statistics for extracting usable information from very large datasets.

Data itself is not new. Data has always been generated and used to inform decision-making. However, most of this was structured and organised, through regular data collections, surveys, etc. What is new, with the invention and dominance of the Internet and the expansion of digital systems across all sectors, is the amount of unstructured data we are generating. This is what we call the digital footprint: the traces that individuals leave behind as they interact with their increasingly digital world. Data analytics is the process where data is collected and analysed in order to identify patterns, make predictions, and inform business decisions. Our capacity to perform increasingly sophisticated analytics is changing the way we make predictions and decisions, with huge potential to improve competitive intelligence. These examples suggest that the actions from data mining and analytics are always automatic, but that is less often the case.

Educational Data Mining (EDM) and Learning Analytics (LA) have the potential to make visible data that have heretofore gone unseen, unnoticed, and therefore unactionable. To help further the fields and gain value from their practical applications, the recommendations are that educators and administrators:

- Develop a culture of using data for making instructional decisions;
- Involve IT departments in planning for data collection and use;
- Be smart data consumers who ask critical questions about commercial offerings and create demand for the most useful features and uses;
- Start with focused areas where data will help, show success, and then expand to new areas;

- Communicate with students and parents about where data come from and how the data are used;
- Help align state policies with technical requirements for online learning systems.

This report documents the first steps conducted within the SPEET[1] ERASMUS+ project. It describes the conceptualization of a practical tool for the application of EDM/LA techniques to currently available academic data. The document is also intended to contextualise the use of Big Data within the academic sector, with special emphasis on the role that student profiles and student clustering do have in support tutoring actions.

The report describes the promise of educational data mining (seeking patterns in data across many student actions), learning analytics (applying predictive models that provide actionable information), and visual data analytics (interactive displays of analyzed data) and how they might serve the future of personalized learning and the development and continuous improvement of adaptive systems. How might they operate in an adaptive learning system? What inputs and outputs are to be expected? In the next sections, these questions are addressed by giving a system-level view of how data mining and analytics could improve teaching and learning by creating feedback loops.

Finally, the proposal of the key elements that conform a software application that is intended to give support to this academic data analysis is presented. Three different key elements are presented: data, algorithms and application architecture. From one side we should have a minimum data available. The corresponding relational data base structure is presented. This basic data can always be complemented with other available data that may help to decide or/and to explain decisions. Classification algorithms are reviewed and is presented how they can be used for the generation of the student clustering problem. A convenient software architecture will act as an umbrella that connects the previous two parts.

The document is intended to be useful for a first understanding of academic data analysis. What we can get and what we do need to do. This is the first of a series of reports that taken all together will provide a complete and consistent view towards the inclusion of data mining as a helping hand in the tutoring action.

[1] Student Profile for Enhancing Tutoring Engineering (`www.speet-project.eu`)

2 Current Landscape

Data has always been a significant asset for institutions, and has been used to inform their day-to-day operational decisions as well as longer-term business and strategic decisions. On a strategic scale, data is used to inform senior management's business planning and overall strategy for their institutions. Student enrolment data, both historical and projected, as well as estates data, will influence the plans institutions make to build new buildings or refit current buildings to meet projected need. Financial data influences strategic decisions on expanding or reducing particular faculties or services provided.

From a more purely educational point of view, the available academic data can be collected, linked together and analyzed to provide insights into student behaviours and identify patterns to potentially predict future outcomes. In this section, usually available data will be described as well as its potential use for the benefit of students. The use of academic data for supporting tutoring action is where we will put the focus on. On that respect, the so called Intelligent Tutoring Systems (ITS) are also presented and main characteristics revisited.

2.1 Student Big Data vs. Big Data for Students

Since the definition of big data is still developing, we will start with our use of the term. In [Lan01] big data is described with a collection of "v" words (see Figure 1), referring to (1) the increasing size of data (volume), (2) the increasing rate at which it is produced and analyzed (velocity), and (3) its increasing range of sources, formats, and representations (variety). To this, other authors have added veracity to encompass the widely differing qualities of data sources, with significant differences in the coverage, accuracy, and timeliness of data.

2.1.1 Higher Education Institutions Data

Higher education institutions are not an exception and the use of analytics in education has grown in recent years for four primary reasons [SB12]: a substantial increase in data quantity, improved data formats, advances in computing, and increased sophistication of tools available for analytics. In recent years, the sophistication and ease of use of tools for analysing data make it

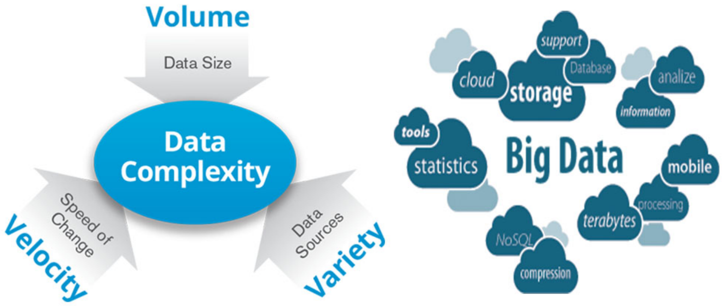


Figure 1. Big Data as Volume-Velocity-Variety

possible for an increasing range of researchers to apply data mining methodology without needing extensive experience in computer programming. Many of these tools are adapted from the business intelligence field. Higher education institutions have always operated in an information-rich landscape, generating and collecting vast amounts of data each day. A coarse classification of the types of data that higher education institutions deal with every day:

- Student record data: A typical student record may include the details of the student's name, age, address, ethnicity, socioeconomic status, school, level results, course undertaken, modules studied, examination results, degree awarded and degree classification;
- Staff data: Institutions hold data on their staff, including the number of people employed full-time and part-time, the number at each level and within each faculty, and staff equal opportunity data;
- Admissions and applications data: These records will include details of the number of students who applied to the institution, the acceptance rate, and any widening participation data such as ethnicity and socioeconomic status;
- Financial data: Universities hold data on their finances, including income streams, expenditure, and predicted profits and/or losses, held both at an institutional level and by faculty and school;
- Alumni data: The university will hold data on its alumni including graduate destinations (i.e. employment or further study), current address and contact details. This is becoming increasingly important as institutions look to diversify their income streams;
- Course data: Includes data on students enrolled in each course and per module;

- Facilities data: Includes data on the number and type of rooms across the campus (lecture theatres, classrooms, computer labs, science laboratories), room capacity, equipment, accommodation, facilities and retail.

2.1.2 Goals for Big Data in Education

In commercial fields, business and organizations are deploying sophisticated analytic techniques to evaluate rich data sources, identify patterns within the data and exploit these patterns in decision making. Recently researchers and developers from the educational community started exploring the potential adoption of analogous techniques for gaining insight into online learners activities. Even the list of goals and objectives that can be pursued with the application of analytics to students big data can be very long - see more detailed (scientific) presentation in the next section - it is possible to categorize the goals in terms of the students benefits as follows:

- Improve Student Results: The overall goal of big data within the educational system should be to improve student results. During his or her student life however, every student generates a unique data trail. This data trail can be analysed in real-time to deliver an optimal learning environment for the student as well to gain a better understanding in the individual behaviour of the students. In addition, big data can help to create groups of students that prosper due to the selection of who is in a group. Students often work in groups where the students are not complementary to each other. With the help of appropriate algorithms it will be possible to determine the strengths and weaknesses of each individual student. This will create stronger groups that will allow students to have a steeper learning curve and deliver better group results;

- Create Mass-customized Programs: All this data will help to create a customized program for each individual student. This will be created with blended learning: a combination of online and offline learning. It will give students the opportunity to develop their own personalized program. Providing mass customization in education is a challenge, but thanks to algorithms it becomes possible to track and assess each individual student. We already see this happening in the Massive Open Online Courses that are developed around the world now;

- Improve the Learning Experience in Real-time: Each student learns differently and the way a student learns affects the final grade of course. Some students learn very efficiently while others may be extremely inefficient. When the course materials are available online, it can be monitored how a student learns. This information can be used to provide a customized program to the student or provide real-time feedback to become more efficient in learning and thus improve their results. When students are monitored in real-time, it can help to improve the digital textbooks and course outlines that are used by the students. Algorithms can monitor how the students read the texts. Which parts are difficult to understand, which parts are easy and which parts are unclear. Based on how often a text is read, how long it takes to read a text, how many questions are asked around that topic, how many links are clicked for more information etc. If this information is provided in real-time, authors can change their textbooks to meet the needs of the students thereby improving the overall results. Even more, Big Data can give insights in how each student learns at an individualized level;

- Reduce Dropouts, Increase Results: All the previous reasoning will improve the student results and perhaps also reduce dropout rates at universities or colleges. Dropouts are expensive for educational institutes as well as for society. Using predictive analytics on all the data that is collected can give educational institute insights in future student outcomes. These predictions can be used to change a particular program if bad results are predicted or even run scenario analysis on a program before it is started. Universities and colleges will become more efficient in developing a program that will increase results thereby minimizing trial-and-error.

2.2 Intelligent Tutoring Systems

From several decades till nowadays the educational tasks are supported by computers and specialized software. Namely, systems based on Artificial Intelligence have been evolved and they are usually called Intelligent Computer-Assisted Instruction systems, Intelligent Tutoring Systems or more lately Interactive Learning Environments [Lel00].

The main objective of an ITS was to support the learning process specially on problem solving. It was assumed that the student should acquire the required knowledge before using the tutoring system. With the growth of the information technologies the ITSs provide learning material in electronic form. In this way, the student has theoretical support and problem solving in the same package [Bru00].

By definition, an Intelligent Tutoring System (ITS) is a system that uses Artificial Intelligence (AI) and tries to imitate the human tutor tasks and capabilities.

The underlying idea is to generate information and sets of exercises depending on the student's individual needs. Since those needs can change or can be very different from student to student, the system should be able to adapt its contents and methodologies every time there is a context changing.

This is done by having a "model" of each student performance which is dynamically maintained and used to drive instructions. Also when designing the system, the knowledge items and the interface are defined for all students but the sequence of instructions is determined by the ITS itself using specific information. This gives the system further ability to individualise the teaching techniques tuning the learning tasks for each student [Nwa90].

This specific information can be gathered using LA and EDM techniques. In [SWST14] the authors agree that techniques like learning analytics should be used to collect and analyze data from learners in order to profile them and ameliorate their experience inside ITS.

More recently, there are other interesting research works concerned with recommender tools.

E-learning recommender systems have become popular in educational institutions [LWM$^+$15] and they are an evolution of the traditional e-learning systems. The recommender systems aim to assist learners to choose the courses, subjects and learning materials and activities that they need to support their learning process. Zaiana [Zai02] proposed an approach to build a software agent that uses data mining techniques such as association rule mining to construct a model that represents online user behaviors, and uses this model to suggest activities or shortcuts. The suggestions created automatically by the system will guide learners to better navigate through online material. This will allow finding more assertive information in a shorter time by using the recommended shortcuts.

The architecture of an e-learning recommender system usually consists of three parts [LWM$^+$15]:

1. using Web analysis techniques to collect learners' profiles and identify their personalized demands;
2. collecting the metadata of learning objectives to identify the features;
3. acquiring related pedagogical knowledge to evaluate the matching degree between learners and learning objectives.

In [ALP16] the authors propose a recommender tool that analyzes student interactions and visually explains the collaboration circumstances to provoke the self-reflection and promote the sense making about eLearning collaboration. The tool presents a visual explanatory decision tree that graphically highlights student collaboration circumstances and helps to understand the reasoning followed by the tool when prescribing a recommendation.

To conclude, in order to construct ITS or recommender systems it is very important to get and analyze data about student activities and attitudes.

3 Privacy and Data Protection Considerations

The rise of concern over student privacy has strong implications for how new EDM approaches can be integrated into wide-reaching applications as well as the amount of funding available to public and private entities wishing to innovate in this space. The increase in EDM usage has raised public awareness of how much data is being collected about students. The applications and companies that collect and use student data are coming under scrutiny, as parents, advocates, and public officials grow concerned over student privacy. Large, well-trusted institutions have been targeted for using student data in undesirable ways.

3.1 Privacy

Privacy is chiefly a question of access. Unlike anonymity or confidentiality, people's interest in privacy is about controlling the access of others to themselves [ST12]. How to safeguard a student's privacy is a particularly complex question because of their vulnerability. Children are incapable of protecting their own interests through negotiation for informed consent because they are likely to misunderstand risks or be coerced into participating [ST12]. This need to protect has led to the formation of student privacy advocacy groups and driven the adoption of legislation. The restrictions required to comply with this legislation and maintain good public opinion have a significant impact on the adoption of data-based solutions in education.

3.2 Student Data Protection and the EU Data Protection Directive

Personal data protection has been regulated in the EU for a long time, applying a comprehensive prescriptive legal approach which focuses on the population as a whole. For the reasons that will be outlined in the following section, it is nevertheless not sufficiently equipped to deal with the pitfalls of big data in education. At present, the most important EU legal instrument on personal data protection is the 1995 Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data (DPD). Recognising the important role vendors play in processing personal data, the DPD distinguishes between first parties and vendors through the introduction of "data controllers" and "data processors" (art. 2(d)-(e)).

Within this structure, a school acts as a data controller if it decides on (a) outsourcing of student data processing; (b) delegating all or part of the processing activities to an external organisation; and (c) determining the ultimate purpose of the processing. A vendor acts as a data processor if it merely supplies the means and the platform, acting on behalf of the school[2]. The DPD has two key drawbacks in protecting student privacy and personal data in the context of "big data education". First, the DPD does not protect student data from re-identification. The DPD's definition of personal data is: "any information relating to an identified or identifiable natural person (data subject)" (art. 2(a)). If the data is anonymised or aggregated and an individual cannot be identified from the remaining data, it ceases to be personal data, and the provisions of the DPD no longer apply.

When talking about big data, it is questionable whether the personal/non-personal data distinction remains viable and whether anonymisation and aggregation remain effective in protecting users against tracking and profiling [MRP+14]. Even if identifiers, such as names and ID numbers, have been removed, one can use background knowledge and cross-correlation with other databases in order to re-identify student data records [NS08]. Therefore, it could be that when student data is anonymised or aggregated the provisions of the DPD will not apply, but the risk of identifying the student - or more precisely: re-identifying - still remains. Second, setting consent as the DPD's main legal guide may be ineffective. A key principle in the DPD is the need to obtain personal unambiguous consent before data can be processed (art. 2(h)). Before big data, parents could roughly gauge the expected uses of their children's personal data and weigh the benefits and the costs at the time they provided their consent. Today, the ability to make extensive, often unexpected, secondary uses of student data makes it simply too complicated for the average parent to make fine grained choices for every new situation [KKO12]. Moreover, in many instances vendors do not offer users the option of choosing which data they agree to share and for which purposes, thus users are forced to accept or deny the service as a whole. Consequently, parents could end up unintentionally excluding their children from services necessary for their education just because they are unable or unwilling to parse out complex data policy statements [PJ14].

The Directive does not address the fact that opting-out is hardly a feasible alternative for users in the educational context, since most parents do not have the privilege of changing their children's schools based on the applicable privacy policy. Therefore, student privacy should not be a binary concept that

[2] Article 29 Working Party, 2012, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2012/wp196_en.pdf

is either on or off and parents should be given the option of choosing which data they agree to share and for which specific purposes, without having to disengage their children from "big data education".

3.3 Student Data Protection and the EU General Data Protection Regulation

EU data protection law has undergone a long-awaited, rigorous and comprehensive revision. After long discussions in the various committees, on 16 April 2016, the EU Parliament formally approved the General Data Protection Regulation (GDPR or Regulation) and it is set to go into effect in May 2018 in all EU member states. The GDPR was adopted by the European Commission "to strengthen online privacy rights and boost Europe's digital economy", recognising that "technological progress and globalisation have profoundly changed the way our data is collected, accessed and used" (European Commission, 2012).

4 Educational Data Mining and Learning Analytics

During the last decade, analytics and data mining, understood as methodologies that extract useful and actionable information from large datasets, have transformed one field of scientific inquiry after another. When applied to education, these methodologies are referred to as learning analytics (LA) and educational data mining (EDM). LA can be defined as the measurement, collection, analysis and reporting of data about learners and their contexts, for the purposes of understanding and optimising learning and the environments in which it occurs. On the other hand, as an area of scientific inquiry, EDM is concerned with the analysis of large-scale educational data, with a focus on automated methods.

There is considerable thematic overlap between EDM and LA. In particular, both communities share a common interest in data-intensive approaches to education research, and share the goal of enhancing educational practice. At the same time, there are several interesting differences, with one viewpoint on the differences given in [SB12]. In that work, it was argued that there are key areas of difference between the communities, including a preference for automated paradigms of data analysis (EDM) versus making human judgment central (LA), a reductionist focus (EDM) versus a holistic focus (LA), and a comparatively greater focus on automated adaptation (EDM) versus supporting human intervention (LA).

LA and EDM are both emerging fields that have a lot in common, although there are differences in their origins and applications. LA is a multidisciplinary field that involves Machine Learning, artificial intelligence, information retrieval, statistics, and visualization. Additionally, it contains the Technology Enhanced Learning areas of research such as EDM, recommender systems, and personalized adaptive learning.

The combination of LA, as a new research discipline with a high potential to impact the existing models of education, and EDM, a novice growing research area to apply Data Mining methods on educational data, leads to new insights on learners' behavior, interactions, and learning paths, as well as to improve the technology enhanced learning methods in a data-driven way. In this regard, LA and EDM can offer opportunities and great potentials to increase our understanding about learning processes so as to optimize learning through educational systems.

4.1 Educational Data Mining

Applications of EDM methods comprise several steps. Initially, a design is planned, i.e., the main aim of the study and the required data are identified. Afterwards, the data is extracted from the appropriate educational environment. Data will need to be pre-processed, since it may come from several sources or have different formats and levels of hierarchy. Models or patterns are obtained from applying EDM methods, which have to be interpreted. If the conclusions suggest applying changes to the teaching/learning process or are not conclusive[3], the analysis is performed again after modifying the teaching/learning process or the study design. The basic steps to test a learning/teaching process-related hypothesis are the same as those explained for EDM: an iterative process in which data is extracted from an educational environment and pre-processed before applying computational/quantitative methods in order to support stakeholders (instructors, course managers, etc.) when making decisions.

A wide range of EDM methods have emerged through the last several years. Some are roughly similar to those seen in the use of data mining in other domains, whereas others are unique to EDM.

EDM [SM12] is related to the application of computational algorithms for pattern detection in educational data sets [RV13]. It is an interdisciplinary area that combines topics of computer science, statistics and education. Research tends to be more focused on the description and comparison of the computational techniques used for the analysis, which are generally well-known unsupervised and supervised learning algorithms. Furthermore, EDM generally aims to provide automated discovery of student models or prediction of learning outcomes, i.e., its main target is the technical challenge [Fer12].

In [RVPB10] EDM is also introduced as a method for: "developing, researching, and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist".

In [WB13] the authors state that EDM helps to operationalize constructs that are not obvious, to investigate several kinds of relationships among that

[3] Main reasons may be: the problem has not been adequately addressed, the raw data is small or not suitable, or the selected methods are not powerful enough

constructs, to model and understand student learning trajectories, and to identify the trust level of the results. In these ways, EDM will help to support the richness, concreteness, and usefulness of research in metacognition, motivation, and self-regulated learning.

4.2 Learning Analytics

LA [CDST12] is, in turn, focused on the application of data analysis to understand and optimize the educational process [RV13], [Fer12], i.e., it integrates both the technical and social/pedagogical dimensions to solve educational challenges. As a result, the feedback to educational stakeholders (instructors, administration, etc.) is key and its goal is to provide useful results and reports that can leverage human judgment [SB12]. Pedagogy can be implicitly contained in the data or explicitly addressed in the goals and objectives [GD12].

In [SPM16] the authors say that besides academic records there are other kinds of information that can be used. For instance, every time a student interacts with their university, going to the library, entering to the virtual learning platform, submitting assessments online, etc., they leave behind a digital footprint. LA is the process of using this data to improve learning and teaching.

In this sense, LA can be seen as a tool for quality assurance and quality improvement, for boosting retention rates, for assessing and acting upon differential outcomes among student population. Moreover it can be used for development and introduction of adaptive learning because it can help to identify student at risk before it is too late to intervene.

It can also help students by themselves when they have to select future studies and career choices. They become more reflective learners. Unfortunately, there are still lots of institutions that have not implemented LA.

LA is different from academic analytics, the latter only uses administrative data like personal information, programs, subjects and grades.

A more complete student information system containing data about students such as their prior qualifications, socioeconomical status, ethnic group, module selections, grades obtained to date, number of attempts, years to finish the course and so on, can be used to predict academic performance. This data can also be supplemented by attendance monitoring systems and other systems that record campus visits, or their presence in locations like lecture halls, libraries and refectories.

In [MTS15] the authors state that learner-centered approaches to higher education require that instructors have insight into their student's characteristics before they start the course. For that, they also agree that all the information about students should be given to them: demographics, enrollment history and general academic performance.

In [VGO+15] LA is also considered as a great help to discover the hidden information and patterns from raw data collected from educational environments. This is one of the reasons motivating the raising awareness on LA, and stimulating the vital strengthening of its connections with data-driven research fields like Data Mining and Machine Learning.

Summarizing, the understanding of LA (as well as Teaching Attitudes) may be useful in helping the students pursue their career opportunities in opposition to finding the final consequence of the change in Learning/Teaching Attitude. Linking consequences (e.g. "student failed to complete 2nd semester in mathematics") with leaving academia is far more easy than spotting a change in a Learning Attitude (e.g. "loss of motivation in 1st semester due to extremely boring lectures of a mathematician"). However, spotting a change in a Learning Attitude would give the possibility to help that particular student to recover and reinforce. In order to detect changes in these attitudes and to prevent unsuccessful situations, LA techniques should be used.

4.3 Visual Analytics

Visual data analysis blends highly advanced computational methods with sophisticated graphics engines to tap the extraordinary ability of humans to see patterns and structure in even the most complex visual presentations. Currently applied to massive, heterogeneous, and dynamic datasets, such as those generated in studies of astrophysical, fluidic, biological, and other complex processes, the techniques have become sophisticated enough to allow the interactive manipulation of variables in real time. New research is now beginning to apply these sorts of tools to the social sciences and humanities as well, and the techniques offer considerable promise in helping us understand complex social processes like learning, political and organizational change, and the diffusion of knowledge.

Beyond simply mining data, there is an increasing awareness that good visualization methods can support both analysts and practitioners in deriving meaning from data [Duv11]; [BI14]; [TSK+14]; [VDK+13]. In [JLSS10] it described the promise of visual data analysis (in the four to five year time frame) for teaching undergraduates to model complex processes. Visual data analysis also may help expand our understanding of learning because of its ability to support the search for patterns. It may be applied, for example, to illustrate the relationship among the variables that influence informal learning and to "see" the social networking processes at work in the formation of learning communities.

The aim of this section is to introduce some of the general tools and methods for visual analytics, which enable building interactive visual interfaces for gaining knowledge and insight from data as well as communicating important implications for learning to students and teachers.

- Tableau[4]: presents a family of products for interactive data analysis and visualization. Although the primary focus of the Tableau toolset is support for business intelligence, it has been commonly applied in educational settings to analyze student data, provide actionable insights, enhance teaching practices, and streamline educational reporting. The main advantage of Tableau is that no programming knowledge is needed to analyze large amounts of data from various sources, making a range of visualizations easily available to a wider community. However, Tableau's functionality is limited to this; it does not support predictive analytics or relational data mining. Moreover, Tableau, as a commercial tool, is not extendable and does not support integration with other software platforms;
- Data Driven Documents (D3.js)[5]: is a JavaScript library that allows manipulation of data-driven documents, enabling researchers and practitioners to build complex, interactive data visualizations. D3.js has several benefits: It allows considerable flexibility in building a range of kinds of data visualization, does not require installation, supports code reuse, and is free and open source. However, there are challenges to wider adoption for educational research purposes. As a technology, D3.js requires extensive programming knowledge and has compatibility issues as well as some performance limitations for larger data sets. Finally, it does not provide any means to hide data from users of visualizations, requiring data preprocessing to ensure privacy and data security.

Beyond D3.js, many other programmatic data visualization tools exist, aimed at providing different ways to present data visually and build interactive dashboards. Some of the commonly used tools include Chart.js, Raw,

[4] www.tableau.com

[5] www.d3js.org

JavaScript InfoVis Tool kit, jpGraph, and Google Visualization API[6]. These tools offer broadly similar functionality to D3.js but have been less frequently used by EDM and LA researchers.

4.4 Resources and Goals

It is necessary to analyze both the specific data sources available in educational settings as well as the problems that data analysis pretends to solve. Data analysis can be applied to any type of information obtained in the context of education [RV13]:

- Administrative data, including demographic data, academic performance, etc.;
- Data obtained from the interaction of students with an educational information system, such as the results to exercises and quizzes etc.;
- Data obtained from questionnaires, field observations or controlled experiments.

The hierarchical structure of data, its context, its granularity and the time range of available data must also be considered [RV13]. In traditional education frameworks, information about the student path and performance is stored along with administrative data in databases. However, nowadays, there is a great proliferation of computer-based educational environments, which produce additional data with higher granularity, related to specific activities or events, resulting from the interaction of students with educators and software. Among computer-based educational systems, the following can be distinguished [RV13]:

- Test and quiz systems, which measure the performance of a student with regard to some concepts or subjects. It is assumed that the system is only limited to the processing and storage of answers, scores and subsequent statistics;
- Learning Management Systems, which might comprise the functionalities included in test and quiz systems and additionally support and record other course-delivery tasks (administration, documentation, tracking and reporting);

[6] (see "The 38 best tools for data visualization" at www.creativebloq.com/design-tools/data-visualization-712402)

- Intelligent tutoring systems, which record all student-educator interaction with the ultimate goal of modeling student behavior and customizing instruction;
- Others, including educational games, content and interaction support systems (e.g., wikis, forums, etc.)

The application of LA to the available data requires a non-trivial stage of preprocessing to perform the appropriate data integration, selection and transformation.

The potential goals of LA are diverse, including monitoring, prediction, tutoring, personalization, recommendation, etc. [CDST12], [KE15]. They can be categorized according to the targeted stakeholders:

- Students
  - Identifying tutoring needs;
  - Receiving effective advice or feedback;
  - Receiving customized content adapted to their needs.
- Educators
  - Student profiling;
  - Predicting student performance;
  - Predicting student dropout;
  - Improving content organization;
  - Receiving suggestions for instructional actions;
  - Providing feedback educators' interventions;
  - Determining and retaining the students' level of interest.
- Administrators
  - Improving curriculums, schedules and procedures;
  - Identifying useful frameworks, methods and tools;
  - Domain modeling;
  - Improving efficiency and marketing/admission measures;

In [MTS15] the authors state that learner-centered approaches to higher education require that instructors have insight into their student's characteristics before they start the course. For that, they also agree that all the information about students should be given to them: demographics, enrollment history and general academic performance.

The aggregation of great quantities of data can raise ethical concerns and the rights to privacy, informed consent and transparency must be taken into account for the design of the procedures for data collection, aggregation and analysis. As a result, data often needs to be anonymized so that the data subjects are unidentifiable and security safeguards might be considered. Some authors propose to follow a checklist to guarantee ethical and law-abiding implementations [DG16].

4.5 Methods for Learning Analytics and Educational Data Mining

This section addresses the methods that can be applied for the analysis of educational data, highlighting those oriented to model student profiles and to predict student performance and dropout with the final goal of enhancing tutoring. Nevertheless, other previous works also apply EDM to improve assessment or facilitate student support and feedback.

Several surveys review the techniques that have been used in different applications of EDM [PE14], [PA14], [RV13], [RV10].

The unsupervised learning techniques [HTF09] can be used to improve the understanding about the data structure, either at an exploratory data analysis stage or iteratively (and interactively) to guide exploration [RV10], [SGVD12]. For instance, clustering methods such as k-Means [XW05], which group instances according to their similarity, can be used to obtain student clusters that can help discovering and understanding student profiles. Dimensionality reduction algorithms, which transform the original data set to a low-dimensional representation (a representation with fewer variables) that preserves as much information as possible according to certain criteria [LV07], are generally used to enable information visualization. Visual data mining takes advantage of the human reasoning abilities for insight discovery and interpretation [Kei02]. An example of the application of information visualization to educational data mining is the analysis of students' behavior and their interaction with learning environments [RV10]. Additionally, association rule mining techniques, which find joint values of the variables that appear frequently together in a data set [HTF09], can also be used in educational environments to produce recommendations and customize learning.

On the other hand, supervised learning methods [Bis06] can be used to enable profiling [IW14] or academic performance and dropout prediction [YOT14]. Supervised learning methods use labeled training data, i.e., examples of input vectors along with their desired output values. If the output is continuous, it is a regression problem. When the outputs are discrete categories, it is called classification. Due to its importance for the project goals, the classification techniques are described in more detail.

Collaborative filtering techniques can also be used to predict student performance [RPI17]. These techniques, generally used in recommender systems, predict missing values in a sparse matrix from other rows [SK09].

Finally, social network analysis is used in the cases where the objects of interest are the relationships between individuals, instead of individual attributes or properties [RV10].

4.6 Advanced Classification Methods

One of the most significant tasks in EDM is classification, e.g., to identify to which category/profile a new student belongs so that tutoring actions can be carried out. The classification task requires a training data set, labeled according to pedagogical principles. Among the plethora of classification algorithms [HTF09], [DHS01], some very well-known techniques can be highlighted. Although the description of the algorithms in the rest of the section assumes that the goal is binary classification, it is normally possible to generalize the algorithms to multi-class problems.

Logistic regression is a generalized linear model that models the conditional distribution of the target y given the input variables x as Bernoulli, because the dependent variable is binary. For that purpose, unlike linear regression, it uses the logistic/sigmoid function in the hypothesis [HL89]. Due to its simplicity, this method has been widely used in the area of EDM and LA [RPI17]. Naive Bayes is also a linear classifier, but it is a generative model, in contrast to logistic regression, which is a discriminative model. For that reason, it models the joint probability of the input variables from the training data and uses Bayes rules to estimate $p(y|x)$. This probabilistic technique assumes that feature values are completely independent of each other [DP97]. Again, its use in educational analytics is common [YOT14], [RPI17].

Support vector machines construct a hyperplane that separates, with a maximum margin, two classes. The hyperplane only depends on a small subset of samples that are found near to it (support vectors) [CV95]. It can perform nonlinear classification by using nonlinear kernel functions in the dot products. This allows fitting the maximum-margin hyperplane in a high-dimensional transformed feature space. There are also examples of its application to LA in the literature [IW14], [RPI17].

A different strategy to statistical and geometrical classifiers are decision trees, which create a predictive model by learning simple decision rules inferred from the data features. In this direction, the use of random forests to detect student dropout has been already studied [RPI17]. Random forests combine many decision trees through majority vote. Each tree is obtained from a subset of the training data set obtained through bootstrapping, and using a random subset of the features [Bre01]. This procedure might lead to better model performance, decreasing the variance of the model without increasing the bias.

Furthermore, a random forest can be seen as a particular case of ensemble method. Ensemble classifiers combine the prediction of several base estimators (not limited to decision trees) in order to improve the performance or robustness of single classifiers [Die00]. Common types of ensembles are obtained through bootstrap aggregating (or bagging) or boosting. Bagging methods choose the training instances using a uniform distribution, whereas boosting methods choose the instances to concentrate the efforts on those that have been not learned appropriately.

5 Student Profiles and Tutoring action

EDM and LA are used to research and build models in several areas that can influence online learning systems. One area is user modelling, which encompasses what a learner knows, what a learner's behaviour and motivation are, what the user experience is like, and how satisfied users are with online learning. At the simplest level, analytics can detect when a student in an online course is going astray and nudge him or her on to a course correction. At the most complex, they hold promise of detecting boredom from patterns of key clicks and redirecting the student's attention. Because these data are gathered in real time, there is a real possibility of continuous improvement via multiple feedback loops that operate at different time scales immediate to the student for the next problem, daily to the teacher for the next days teaching, monthly to the principal for judging progress, and annually to the district and state administrators for overall school improvement.

The same kinds of data that inform user or learner models can be used to profile users. Profiling as used here means grouping similar users into categories using salient characteristics. These categories then can be used to offer experiences to groups of users or to make recommendations to the users and adaptations to how a system performs.

5.1 Teaching Learning Attitudes and Methods

Choosing the right career scenario and consistent pursuit of its track is important and beneficial for every citizen of the global community. Sadly, there are students, who apparently have reconsidered their career and decided to leave the studies. However, the number of students who decide to quit can and should be lowered.

Some of the universities are aware of the problem and to find the answer, various methodologies are chosen. The most common way of determining the reason for leaving academia is by a simple questionnaire. Such a document usually unveils reasons that, according to the university authorities, are suspected to be relevant. Online resources of the biggest universities in Poland were inspected, as well as 100 Google search results (for such a document), to get the basic view of the issue - from the universities' point of view. A total

Table 1. Occurrence of specific questions in analyzed questionnaires.

No	Rank	Question	Occurrence
1	1	Disappointed with contents (subjects) of the studies (other than expectations)	4
2	2	Disappointed with the quality of Workshops/Lectures or methodology	3
3	2	Problems with mastering material and/or not completing the semester	3
4	2	Lack of career prospects after graduation	3
5	2	Inadequate premises of university or equipment of teaching rooms	3
6	2	Financial problems	3
7	2	Changing professional interests or life plans	3
8	3	Changing University / Faculty / Discipline	2
9	3	Low level of education	2
10	3	Lack of acceptance of other students	2
11	3	Family or health problems	2
12	3	Bad service in the dean's office /administration, discourtesy or tardiness	2
13	3	Too difficult subjects/requirements	2
14	3	Other	2
15	4	Finding a better didactic offer	1
16	4	Mistreatment by teachers	1
17	4	Discrimination (e.g. Due to cultural differences, national origin, gender or other)	1
18	4	Lack of help and support from the University	1
19	4	Not sufficient adaptations for persons with disabilities	1
20	4	Difficult subjects right from the 1st year	1
21	4	Functioning of the dormitories	1
22	4	Leaving country	1
23	4	Working while studying	1
24	4	Recruited just to get the status of a student	1
25	4	Studying just temporarily because I did not get to my 1st choice studies	1
26	4	Low salaries after graduation	1
27	4	Poor organization of teaching hours (timetable)	1

number of (only) 5 such documents and 1 report have been found. Table 1 shows the occurrence of specific questions in all analysed questionnaires.

Although the questionnaires are designed based on the universities experience regarding students that decided to quit their studies, the questions are usually chosen empirically - not as a result of any research - so the options that are suggested to the student might not be quintessential.

On the other hand, a preliminary research was conducted, in which students of last years were asked about the honesty of reasons given by other students in such questionnaires, and the results are intriguing: in their opinion only about 40% of students would give the real reason for leaving university. The

Table 2. Top 5 questions, which - according to students - are chosen to hide the real reason for leaving university

No	Question	Score (1..5)
1	Family or health problems	3.90
2	Financial problems	3.62
3	Lack of career prospects after graduation	3.59
4	Leaving country	3.50
5	Other	3.33

Table 3. Top 5 questions, which - according to students - remain untold (and another reason is chosen untruthfully

No	Question	Score (1..5)
1	Too difficult subjects/requirements	4.09
2	Problems with mastering material and/or not completing semester	4.00
3	Working while studying	3.95
4	Difficult subjects right from the 1st year	3.95
5	Lack of career prospects after graduation	3.84

respondents were also told to choose the most common "fake" reasons for quitting, as well as the real reasons that remain untold. The top 5 of these two categories are listed in Tables 2 and 3.

In order to find the real reasons for leaving academia, in order to help students profile their careers, much broader research has to be conducted. The most fundamental issue seems to be to understand student's situation, and the circumstances that led to it. And above all, to determine at which point in his/her career a student has lost the motivation and/or capacity to continue learning. After graduating, young generation should not only have specific knowledge and skills that they learn during their education period, but they should also have a broader point of view and, most of all, they should have the ability to learn. The PISA (Programme for International Student Assessment) 2012 report [PIS13] explicitly points out that in the time of ongoing economic crisis the urgency of development of citizens skills has increased, and that the investment in developing skills (not only in the education system, but also in the workplace) is fundamental for the economic recovery and for future growth. For this reason, students not only should be able to recognize the sense and the need for getting new knowledge and skills, but they also should be capable of setting new objectives, adapting their learning approach, and to face and overcome any problems and challenges in their lives. Nowadays, everyday life and every job require the ability to adapt to changes, the ability to learn throughout the life.

According to the literature [PIS04], the loss of motivation happens more frequently in situations where a student has lost their potential to master fundamental concepts and skills, usually due to feeling alienated and disengaged from the context of learning. For this reason, not only the cognitive outcomes of education are important, but also: their belief in their capabilities, their motivation and their engagement.

These aspects lead to an assumption that the problem of loss of career opportunities should not be investigated in the day of leaving the university, but much earlier. Even capturing the timespan of difficulties in learning specific subject might be insufficient. It seems that the very beginning of the breakdown may be associated with the change of student's interest in learning for that particular subject - a change in their learning attitude.

Although the terms "learning attitude" and "attitude" seem to be self-explanatory and their definition can be found in many publications, according to [Ric96] and [GJ63], attitudes research often is limited to drawing the most obvious conclusions. Nevertheless, Learning Attitudes are (and should) be defined in literature - the PISA report 2003 [PIS04] authors describe four aspects of students' attitudes to learning:

1. student's "general attitude towards school";
2. student's "beliefs about themselves as learners", (although some literature [Ric96] differentiates attitudes from beliefs);
3. student's anxiety in a particular subject or discipline;
4. student's learning strategies.

Whereas in [TM04] authors evaluated 49 factors and proposed six categories, five of which consider students Learning Attitudes:

1. "confidence - confidence and self-concept of their performance";
2. "anxiety - feelings of anxiety and consequences of these feelings";
3. "value - students' beliefs on the usefulness, relevance and worth of" a specific subject "in their life now and in the future";

4. "enjoyment - the degree to which students enjoy" that specific subject;

5. motivation - interest in a specific subject and desire to use that knowledge in the future.

These five categories do not completely overlap with PISA report's list of attitude. For that reason, it might be crucial to consider a new set of attitudes, updated due to the availability of new technology in global community (as well as modern teaching methodologies), which could be addressed to investigate the reasons for leaving academia.

Keeping in mind the questionnaires referred to in table 1, as well as the above-mentioned Learning Attitudes, the following list of Learning Attitudes may be proposed:

1. student's general attitude towards learning (i.e. willing to make an effort to get new knowledge or skills);

2. student's beliefs about their learning capability and confidence (i.e. self-confidence of being able to master new knowledge or skills);

3. student's anxiety in a particular subject or discipline (i.e. emotional distress, tension or uneasiness, in some cases caused by apprehension of possible future failure);

4. value (as in [TM04]);

5. student's motivation in learning that particular subject or discipline (including inter alia: personal career objectives, student's interest in a specific field, successful teaching methodologies);

6. enjoyment (as in [TM04]);

7. student's discipline-oriented attitude towards learning and self-learning (i.e. student's curiosity and determination to search for the knowledge and skills, to make active learning efforts, to search for the optimal learning strategies), including a broader view than just a single subject;

8. student's attitude towards the economical situation of country or families in that country (in relation to other countries, taking into consideration possible migration).

Of course, this set of propositions of Learning Attitudes should be revised by conducting a more detailed study, and it should be used for determining the real drop-out reasons (e.g. "loss of personal motivation in learning mathematics on 1st semester") instead of the resulting drop-out reasons (e.g. "failing to complete the 2nd semester of mathematics").

Understanding the Learning Attitudes of students should underlie reasoning and argumentation on the explanation of the drop-out rate.

On the other hand, not only students are the ones who should be influenced to change attitude. The literature also describes Teaching Attitudes that characterize teachers, to show specific aspects that influence the learning process. The paper [dSBE97] lists the following Teaching Attitudes that affect negatively the learning process:

1. teachers' "lack of confidence about subject content";
2. teachers acting just as "information providers";
3. teachers tending to "separate theory and practice";
4. teachers avoiding new and innovative teaching methodologies;
5. teachers' "lack of coherence" between the interactivity of their teaching methodologies (in their subjective opinion) and the real interactivity level of particular lessons;
6. teachers, when interacting with academically weaker students, tend to have lower expectations and this "generates poor teaching practices",
7. teachers' actual working conditions. Paper [Ric96] indicates also that some of the social issues that are visible in the literature should be taken into consideration:
8. teacher's attitudes "related to democratic and authoritarian attitudes" [Ric96], [Rok61];
9. teachers' attitudes resulting from their gender, ethnicity and race [EGB95];

   The analysis of drop-out rate, as well as any success indicator of the learning process, should also include relation to teaching methodologies used in the classroom. For this reason, it might be worth to examine:
10. teacher's attitudes towards "practical knowledge and experiential learning" and "orientation and structure of practical knowledge, cognitive style,

and reflections on research activity" [Elb83] in a particular subject, understood also in terms of teaching methodology.

As for the purpose of dropout rate analysis, the following set of Teaching Attitudes can be proposed (and should be verified):

1. teacher's general attitude towards teaching (including job satisfaction and working conditions);
2. teacher's attitude towards diverging teaching methodologies (including new and innovative methodologies, interactive methodologies, using variety of methodologies, searching for optimal methodology for a specific group or for academically weaker students);
3. teacher's attitude towards linking theory to practice (i.e. showing the usefulness of the subject in real life applications, combining the practical knowledge with the used teaching methodologies);
4. teacher's attitude towards the subject being taught (including the subject foundations, applicability, as well as the state-of-the-art research in the field);
5. teacher's attitude towards students' motivation, understanding and valuing the subject (i.e. igniting students' curiosity, motivating them, showing the value of the subject in their careers);
6. teacher's attitudes towards social issues of students (hopefully neutral - inter alia: gender, ethnicity, nationality, race, etc.).

Although Teaching Attitudes do have influence on students' outcomes (including the decisions of leaving the university), it seems to be more difficult to assess if there is a connection between information included in the student's university records and particular teachers. If yes, than it would be clearly possible to determine some of the Teaching Attitudes influencing the dropout rate.

It is also important to keep in mind, that there are other factors influencing students' decisions, some of which concern for example:

1. the influence and attitudes of the most important people in student's life, family and friends [Ren94];

2. community's involvement and attitudes [MTS15].

The social context of the student is related with the education level of the people he lives with, health habits and financial support. As it was seen in Table 1 these parameters can have a great influence in student performance. However, these factors seem to be difficult to be inferred from the student's university records.

This is the biggest challenge of our project: how to infer from academic information the factors that are responsible for unsuccessful students.

For that techniques of LA like statistics and data mining must be used over several kinds of digital information that are available in the higher educational institution.

5.2 On Defining Different Profiles and Identified Label Flags

This first intellectual output of the project proposes to capture major paradigm changes, from official statistics reported about students at universities in European states, in working with students and new approaches in university education, notably in the field of IT implementation and valorization of intelligent systems for data processing from official statistics reported about students at universities in European states and early identification of students at risk of dropout. Although the goal is very clear, the approach to achieve student profiles in such a situation raises several questions and problems arising from the difficulty of the challenge assumed by the project partners:

- the official data reported by universities are quantitative, numerical, allowing the statistical processing and obtaining relevant feedback on the following indicators:
  1. Number of students;
  2. Student Explanatory Information;
  3. Degree Information;
  4. Student Performance Information.
- the phenomenon of dropout from university studies has multiple causes and requires a factorial analysis with at least two major categories of factors: internal factors related to the student's personality and its level of bio-psycho-social development and external factors related to the socio-economic, cultural and educational environment in which the student develops;

- the official data reported by universities about students partly reflects the two categories of factors, but enough to be able to identify in the first instance the students with educational risk of dropout, information which once obtained then calls on the attention of the teachers and the management of the university to initiate some tutorial actions, counseling and failure avoidance;
- tutoring and counseling will later complete the student profile by obtaining qualitative data about the student with dropout risk, information generated by tools such as questionnaire, interview, checklist, structured essay. The data collected will allow personalization of the profile and identification of other causes of socio-emotional and attitude-behavioral nature not found in official data statistically reported by universities.

The issue of the project is complex and the hypotheses generated and to be validated during the project are outlined around the following coordinates:

- What are the dominant causes and the factors favoring the risk of student dropout?
- What are the characteristics of a student who is at risk of dropout?
- Can we outline some generally valid student profiles with an educational risk which reflects the typology and the patterns found in our universities?
- Can we identify students with educational risk of dropout by using data processing based on the data reported officially by universities?
- What is the structure of such a database and what other processing opportunities can be created for this?
- What are the desirable and most effective tutorial actions that follow identifying students with dropout risk?

Answers to these questions are not easy, and research and innovation work involves an integrated and complementary approach between the partner teams in the project to study the literature, legislative documents, the European Council official reports, official student reports, to have quantitative and qualitative investigation based on interviews, questionnaires, focus groups etc. with students and academic staff. The approach will thus allow to capture the official guidelines and their correlation with the student reality reflected in the perception of the young people, of the academic staff, actions that will favor the collection of sufficient data to accurately outline the student profiles with dropout risk and to propose an innovative and functional IT tool.

The proposed IT solution integrates the main functions of such a system, to provide strong support to project managers and team members and, on the other hand, to highlight the positive impact of information systems developed using web technologies in project management and subsequently on the success of businesses. Thus, the objectives associated with the information system designed and implemented refers, primarily, to the facilitation of communication between team members and the project manager, eliminating the geographical barriers and the extension of access area, to monitoring and continuous control of their project development through team, costs, time and risk management.[7]

5.2.1 On Definying Psycho-Pedagogical Profiles of Students

University education, through its specificity, facilitates the access of various categories of students, both from the point of view of academic abilities and psychological, socio-cultural or economic background. This diversity implies the assumption by the higher education institution of the responsibility to create an appropriate environment for the formation and development of students' academic and professional competences. In this respect, the development of typologies, psycho-pedagogical profiles of students can provide universities with the necessary support to better understand student behavior and attitudes towards learning.

Any scientific approach that aims to develop a human typology is based on identifying the main characteristics of the target group and the factors that influence them. Factors that interfere positively or negatively with academic results are subjects of major interest to university educational programs aimed at increasing/maximizing academic performance, learning experiences of students and reducing dropout rates.

Numerous categories of variables influencing academic success can be identified in the literature, with most researches in the field highlighting the importance of four major levels of their analysis: individual/personal level, institutional/organizational level, level regarding the national educational policies in higher education and economic level. In general, the individual level empha-

[7] Lupasc, A., 2016, Project Management Information System Based on Web Technologies, accept publicare ICI Proceedings, Conference EduWorld 2016, The European Proceedings of Social and Behavioural Sciences EpSBS, e-ISSN: 2357-1330, Published by the Future Academy

sizes the importance of personal, psychosocial characteristics; the institutional level is described by the specifics of the higher education institution (organizational culture, learning facilities, socio-cultural activities etc.); the level of national educational policies refers to the particularities of the university education system from the perspective of access, learning, evaluation, structure; the economic level being represented by the financial aspects derived both from the study years path and from the professional alternatives offered to students and graduates. It is important to note that these factors are interdependent, influencing each other and acting on the student's academic behavior, both directly and indirectly [HKT15], [VKJ$^+$15].

The psychological theories and research undertaken to date emphasize the particular role that individual psychological characteristics have on the student's performance [Ban82]; [BE00]; [BH05]; [Tin92]. What is the level of student's academic abilities? Can he/she face the demands of the university environment?

Most studies have demonstrated the existence of a relationship between the student's educational path, high school and university performances, with a higher dropout rate for students with lower secondary school outcomes [LG08].

Successful completion of university studies involves, besides cognitive training, also motivational processes, self-esteem, self-efficacy and adaptability. A low level regarding these issues have a negative impact on academic achievements and increasse the probability of failure and dropout [BE00], [Ban86], [CHG01], [TCH09]. Although the intrinsic motivation is predictively associated with academic success, however, lack of motivation is negatively associated with student performance. An interesting aspect is related to admission criteria, students' performance analysis indicating a higher rate of academic abandonment in case of study programs with more permissive criteria than those with high requirements, due to the motivational mechanisms involved and the level of effort made.

The academic expectations of the student towards university studies and the university environment are also very important, and a high level of self-efficacy is related to academic success [Ban86]; [CHG01]. Instead, unrealistic expectations about academic outcomes (and/or social) caused by various factors such as incorrect or incomplete information about the curriculum, incorrect estimates of one's own ability to cope with the academic environment may lead to lower performance or dropout.

Another important factor is the socio-economic status of the student's family ([TQ03], [TCH09]. A supportive family that encourages the development of communication skills, autonomy and respect for learning can favor the academic performance. The financial resources available during the studies influence the student's ability to integrate into the academic community, sometimes their absence or their low level determining either the interruption of their studies, the dropout or a lower involvement in academic activity (often students must have a job resulting that the time allocated to studies diminishes significantly). Parents' level of training is relevant to the extent that they can constitute informational sources (cognitive, social and cultural) and affective-motivational to facilitate student adaptation and academic integration. Although the female student's academic outcomes are generally considered to be superior to male students (both in terms of rate of advancement and performance level), these differences are based on several variables to be considered: the social-economic status, the specifics of the study program, family support. The study by [SD12] highlights the existence of a dropout or domain change rate greater if the student is a minority in terms of the specificity of the field.

Regarding the influence of ethnicity/nationality on academic results, studies have shown a higher dropout rate and lower performance in the case of minorities [HJ01]; [RB10]; [TCH09]. The basis for these results is often the cultural differences in education and training or incorrect information on the specifics of the study program.

The impact of socio-economic factors, belonging to a certain category of gender or ethnicity/nationality on academic success is not uniquely determined, and no pattern of direct influence is identified. Rather, we can talk about a mediated influence, an interaction between these factors and other conditions that mark the success or academic failure.

Tinto's perspective [Tin87] on institutional functioning and development underlines the importance of establishing and communicating specific expectations to create a productive learning environment, highlighting the ongoing interactions between each student and the institutional structures of the university. The organizational characteristics of the higher education institution (size, structure, educational resources, facilities for students) can have a strong effect on students' psychosocial and economic adaptation and integration, and therefore on dropout rates.

The institutional factor can influence academic success and the dropout or failure rate through strategies implemented at institutional level to involve the students in the development of a culture appropriate to learning, adaptation and psychosocial and educational integration (clear presentation of pre-requirements for a study program, clear formulation of the expectation regarding the curriculum, granting support in choosing the study program according to the individual characteristics of the student).

One of the factors with the greatest impact on the success or dropout rate of studies is the economic one. In the current socio-economic context, participation in learning activities is influenced by the cost-benefit relation. Costs include issues related to school fees, participation in various activities, giving up work etc. Thus, institutional policies related to highlighting short, medium and long-term benefits, financial support for certain social categories, graduates' employability become extremely important.

By analyzing the impact of these factors on the student's academic performance, we can identify predictors of student success or failure, study dropout or willingness to continue. In this way prevention and intervention strategies can be developed and implemented to reduce or even eliminate the negative effects of the involved elements.

5.2.2 On Definying the Profiles of Students with a Dropout Risk

Previous studies and research have highlighted the causes and factors involved in the risk of dropping out of university studies, many of which have been identified since the first year of study. Thus, students considered to be with dropout risk in the higher education system are:

- Students from families with low socio-economic status: parents do not have higher education; family income is low;
- Students with a low level of previous school performance;
- Students belonging to minorities or disadvantaged social categories;
- Students with poor results from first-year assessment (a special case is represented by students who have high performance before entering the higher education system and have low academic outcomes);
- Students enrolled in several study programs.

The literature ([HM11], [CO14]) highlights various typologies of students at risk of dropping out of university studies with various (statistically and empirically) different criteria, including:

- self-confidence;
- choice of the study program - the factors that influenced the student's choice (parents, friends, school, own choice);
- career success - interest regarding the career and the status in the chosen study program;
- school results - both previous ones and those obtained during university studies.

The psycho-pedagogical profiles of students with dropout risk can be elaborated by complex analysis of the data obtained and those of demographic nature. The following typologies/profiles can therefore be proposed:

- undecided - are those students who are still unsure about their chosen education/professional path or career they want to follow. They are students with low self-confidence, whose decision about the field of study has been strongly influenced by external factors, not by their career interests ("I want a professional status/career success, but I do not yet know in which domain"). In their case, academic performance is often low (previous school results may be high);
- exploratory - are students who go from one study program to another in search of the right field. They have common features with the "undecided" category regarding uncertainty about the choice of the study program. These students want academic performance and professional success, they often know their own abilities and interests, having difficulties in selecting the right study program. Often they enroll in more specializations at the same time, or they pass from one specialization to another (sometimes the study areas are extremely different). Abandonment of a study program can take place through definitive withdrawn, by changing the higher education institution or by achieving very low performance;
- uncommitted - students who are not interested in academic progress or career success. They are characterized by a low self-confidence (even if apparently, their behavior suggests the opposite), significant influences from external factors in choosing the study program or even the educational institution. Their interests are not related to academic performance but are centered on other aspects of their lives.

To these profiles, which present a clear pattern of the risk of dropping out of academic studies, can be added others which, under certain conditions, may present behaviors of this kind:

- scholars - is the category of students who have/want academic performance, but for one reason or another they give up. These reasons may be financial, medical, social (membership of a social minority) or intellectual (the program of study chosen does not meet the expectations, they are pursuing a different direction/professional training than the offered one etc.).

To conclude, several student profiles can be found and we feel that they are common to all the partner institutions. The main challenge now is to relate these profiles to the academic data. In the next chapter a tool to implement this relation will be proposed. It will allow to infer student profiles and to identify the most risk groups.

6 Software application

6.1 Proposal for data set format

It must be remarked that the final aim of this activity is to determine and categorise the different profiles for engineering students across Europe, under the assumption that students' performance can be classified according to their behaviour while conducting their studies. The academic records of those students are stored on the academic offices of our Engineering Schools/Faculties and these records do not only include the performance of the student on the different subjects of the degree but also collateral information (geographical info, previous studies, age, etc.). This information could be used to help characterise the student by means of data science techniques and, as a result, help tutors to better understand their students and improve counselling actions.

One of the characteristics of the proposed approach is its transnational nature, since the fact of obtaining (or not) the same student classifications and profiles will help identify the common characteristics on engineers coming from different EU institutions. The differences on a country/institution basis will also be exposed and lead to deeper analysis.

Due to this transnational nature, it is necessary to choose the appropriate variables and representation to cover the differences in course organization at a country level. Additionally, the data set must include students' demographic data while complying with privacy regulations of the European Union. As a result, the proposed data set uses variables obtained from the administrative records of the students, such as demographic data, courses taken or academic performance.

Figure 2 shows the initial, minimum core data set, proposed to perform the analysis. It is also possible to augment the data set with other potentially useful additional data sources, e.g., the regional/metropolitan socioeconomic indicators provided by organizations such as the Organisation for Economic Cooperation and Development.

Examples of similar data set formats 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 ed-

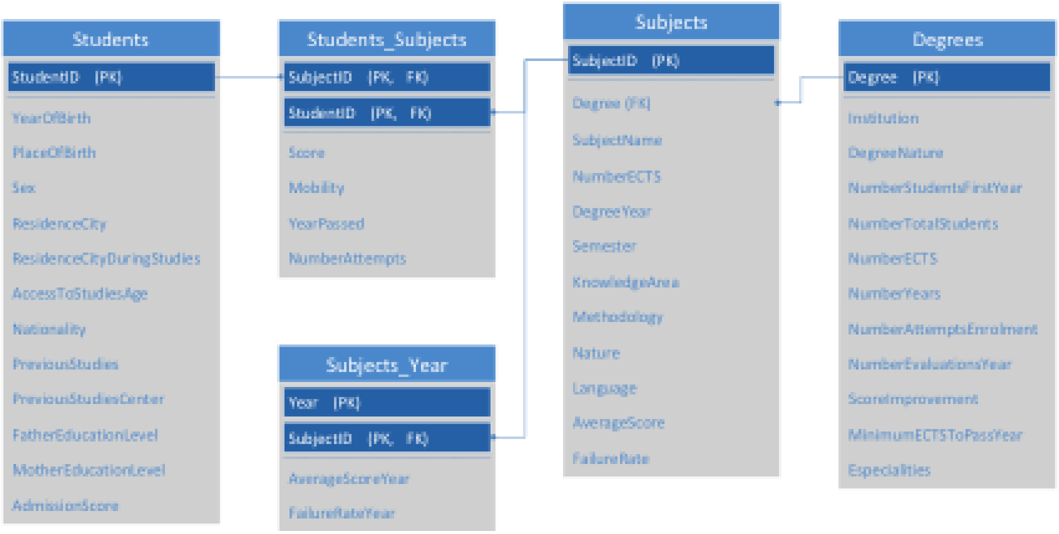


Figure 2. Relational model of the proposed structure of the dataset

ucation level), or current and past information about study units (study area, enrolment mode, delivery method, average grade) are considered in [IW14] for student profile modeling. The publicly available Open University Dataset [KHZ16] also includes this information along with data from the interactions of the students with a virtual learning environment. In [RPI17], only the final grades of the courses are used to predict dropout and performance.

6.2 Proposed software architecture

This section describes the software architecture needed to deploy the supporting IT tools for tutoring. These tools should help determine the profile one student complies, once key labels for the different profiles have been obtained, so that the tutor knows how to provide the student with the appropriate suggestions in order to increase its performance and satisfaction with the studies.

These IT tools will be deployed as web applications, since their benefits with regard to accessibility, interoperability and easier maintenance make them more suitable for this purpose.

From an architectural point of view, the applications might follow a Model-View-Controller (MVC) pattern, which divides an application into three interconnected parts in order to separate internal representations of information from the ways that information is presented to and accepted from the user (see Figure 3).

In this pattern, the models represent knowledge, i.e., they will include the student data and the results of the methods that we applied for the analysis of educational data. All this information will be stored in a database whose structure will be based on the relational model described in the previous section. On the other hand, the view is a (visual) representation of the model that can highlight certain attributes of the model and suppress others. A view is attached to its model (or model part) and gets the data necessary for the presentation from the model by asking questions. Dynamic web pages with support for querying, interaction and different types of visualizations can be used as the views of our architecture. Finally, the controller is the link between the user and the system that provides the views and presents the data to the user. It provides means for user output by presenting the user with menus or other means of giving commands and data.

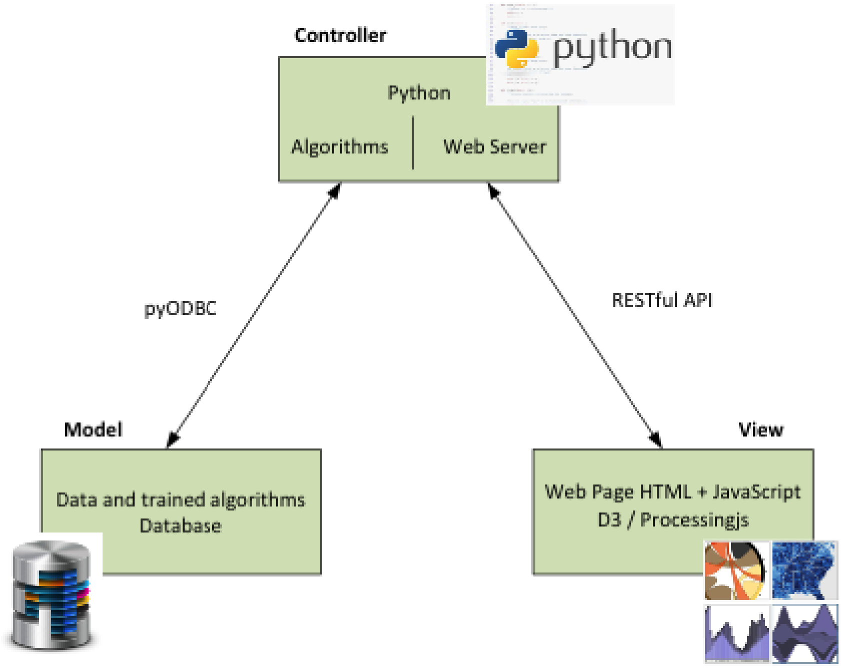


Figure 3. Architecture for the software tool to be developed

Additionally, the architecture might also follow the Representational State Transfer (REST) style. This approach specifies constraints to a web service that restrict how the server may process and respond to client requests so that desirable non-functional properties are achieved, such as performance, scalability, simplicity, etc. RESTful web services are stateless and their data and functionality are considered resources. These resources are accessed through a uniform interface using the HTTP methods: PUT, GET, POST, and DELETE.

From an implementation perspective, a suitable technology for the server-side components is Python, because of its simplicity and the availability of

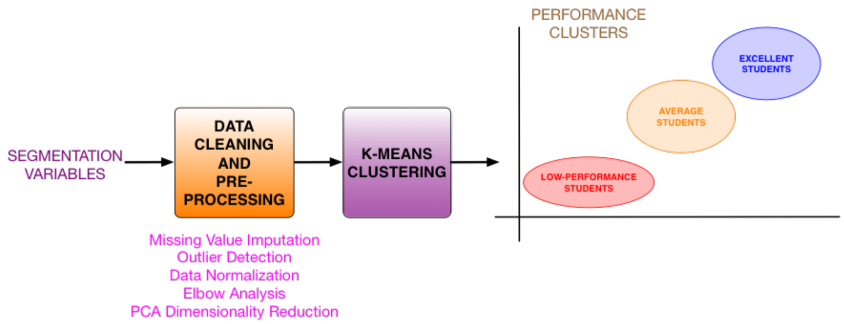


Figure 4. Clustering procedure for student's data analysis and classification

widely-supported libraries for data analysis (Pandas, Scikit-learn, TensorFlow, etc.) and web frameworks (Django, Flask, etc.). Data will be stored in a relational database which can be interfaced directly by the framework or through some object-relational mapping.

Interactivity and visualization in the client-side can be provided by JavaScript libraries, such as D3.js or P5.js, that take advantages of the features supported by HTML 5. The aim, in any case, is that these rich Internet applications that provide tutoring support use web standards.

6.3 Preliminary classification methods

As a first iteration, SPEET project will apply clustering and classification methods to start to obtain students' performance patterns. The objective is to obtain initial results that will allow us to identify the kind of students existing at the different degrees/universities along with the main attributes that characterize such students. To do this, a static approach will be followed by analyzing performance results of students that already have finished their degrees. Next, we present the three steps considered to perform this analysis.

6.3.1 Clustering

In Figure 4, we present the clustering procedure. Concerning the Segmentation Variables, these are the variable associated to students' performance, i.e., attributes included at subjects table (see data set format at previous section).

The idea behind this procedure is to organize students at different groups (clusters) based on their performance results. To do so, a classical clustering approach will be adopted, k-means, based on gathering in a cluster those elements with the highest similarity. The goal is to obtain 3 or 4 clusters as

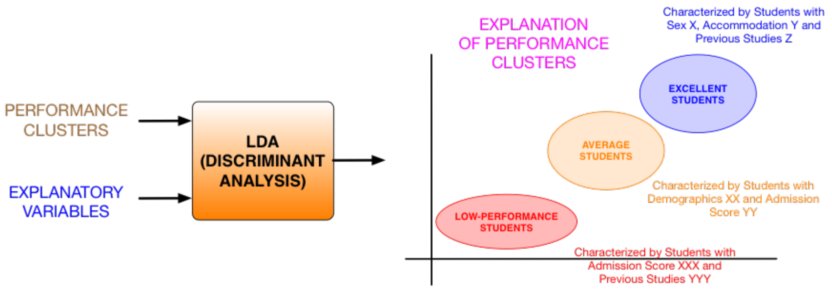

PERFORMANCE CLUSTERS
EXPLANATORY VARIABLES
LDA (DISCRIMINANT ANALYSIS)
EXPLANATION OF PERFORMANCE CLUSTERS
Characterized by Students with Sex X, Accommodation Y and Previous Studies Z
EXCELLENT STUDENTS
AVERAGE STUDENTS
Characterized by Students with Demographics XX and Admission Score YY
LOW-PERFORMANCE STUDENTS
Characterized by Students with Admission Score XXX and Previous Studies YYY

Figure 5. Clustering procedure for student's data analysis and classification

observed in the figure. As for the Data cleaning and pre-processing block, this block is aimed at prepare the data. Data coming from academic management databases could not be complete. Besides, data coming from different universities should be uniformed to allow for the exploitation of the same data mining algorithms.

6.3.2 Clustering explanation

Once students are properly organized in clusters, the goal is to try to understand why these students have these performance results. In other words, the goal is to answer the question: "Are students belonging to a given cluster characterized by a set of explanatory variables?". These explanatory variables are those belonging to students table (see data set format at previous section). Figure 5, we present the clustering explanation procedure. As observed, a Linear Discrimination Analysis (LDA) algorithm will be adopted due to the ability the algorithm have to obtain the exogenous attributes driving the elements belonging to a cluster.

6.3.3 New student classification

The final step of this static approach is to classify a new student entering the system. In other words, to assign a performance cluster to a new student based on the explanatory variables of this student. The goal here is to try to anticipate performance results of this student to provide proper tutoring aid. This procedure is presented at Figure 6, where LDA is adopted again due to the ability this algorithm provides to perform classification tasks.

6.4 Advanced classification methods for grouped data

The main rationale behind SPEET proposal is the observation that students performance can be classified according to their behavior while conducting their studies, that is determine and categorize the different profiles for engi-

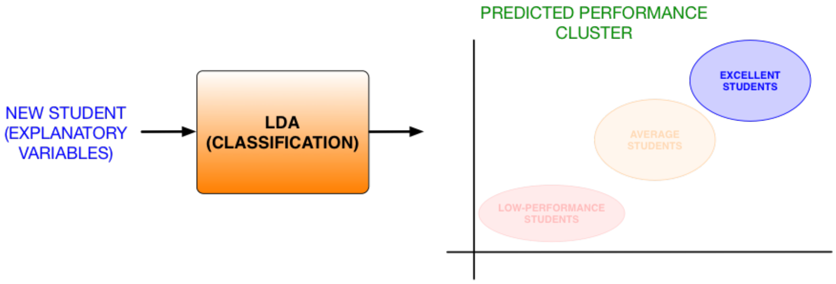


Figure 6. Clustering procedure for student's data analysis and classification

neering students across Europe. In order to complement the static analysis presented in the previous section, the project will also perform an advanced classification approach as a second iteration. To do it, this document presents the most advanced methods of analysis on data that are grouped according to one or more classification factors (students within engineering faculties, engineering faculties within universities, universities within countries and so on).

6.4.1 Generalized linear mixed-effects models (GLME)

Many common statistical models can be expressed as linear models that incorporate both fixed effects, which are parameters associated with an entire population or with certain repeatable levels of experimental factors, and random effects, which are associated with individual experimental units drawn at random from a population. A model with both fixed effects and random effects is called mixed-effects model. By associating common random effects to observations sharing the same level of a classification factor, mixed-effects models flexibly represent the covariance structure induced by the grouping of the data.

The generalized linear mixed-effects model (GLME model) is a further extension that admits for example a response which is binary in nature (logistic model like: $0 = studentgraduatedontime$, $1 = studentwhodelaysgraduation$) or multi-category (multinomial observations that includes nominal responses like "students that finish degree on time", "students blocked on a certain set of subjects", "students that leave degree earlier", ...).

6.4.2 Classification and regression trees (CART)

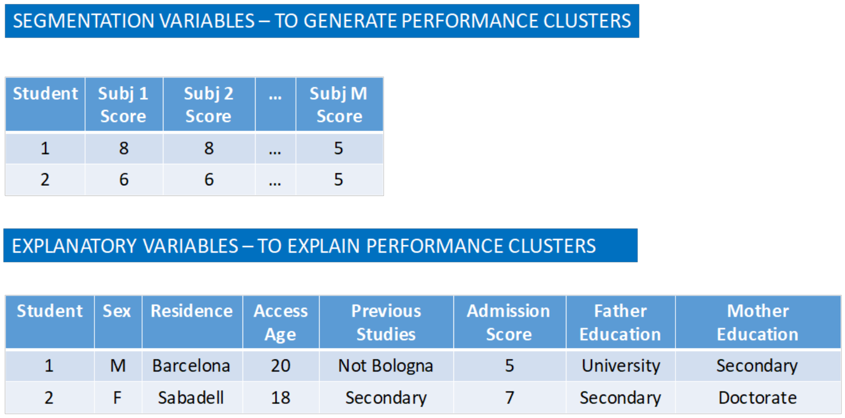
Classification and regression trees are machine-learning methods for constructing prediction models from data. The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree. Classification trees are designed for dependent variables that take a finite number of unordered values, with prediction error measured in terms of misclassification cost. Regression trees are for dependent variables that take continuous or ordered discrete values, with prediction error typically measured by the mean sum of squared difference between the observed and predicted values.

6.4.3 Resulting estimation method (RE-EM Tree)

The methodology that combines the structure of mixed-effects models for longitudinal and clustered data with the flexibility of tree-based estimation methods, is the resulting estimation method, called the RE-EM tree. The RE-EM tree is less sensitive to parametric assumptions and provides improved predictive power compared to linear models with random effects and regression trees without random effects. Traditional mixed-effects models, such as the linear mixed-effects model, neither the random effects nor the fixed effects are known, so there is an alternation between estimating the regression tree, assuming that the estimates of the random effects are correct, and estimating the random effects, assuming that the regression tree is correct. This alternation between the estimation of different parameters is reminiscent of the EM algorithm, as used by [LW82]; for this reason, the resulting estimator is called Random Effects/EM Tree, or RE-EM Tree.

6.4.4 Support Vector Machines

Support Vector Machine (SVM) is a supervised machine-learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, each data item is plotted as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, a classification is identified by finding the hyper-plane that differentiate the two classes in the best way. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.



SEGMENTATION VARIABLES – TO GENERATE PERFORMANCE CLUSTERS

Student	Subj 1 Score	Subj 2 Score	...	Subj M Score
1	8	8	...	5
2	6	6	...	5

EXPLANATORY VARIABLES – TO EXPLAIN PERFORMANCE CLUSTERS

Student	Sex	Residence	Access Age	Previous Studies	Admission Score	Father Education	Mother Education
1	M	Barcelona	20	Not Bologna	5	University	Secondary
2	F	Sabadell	18	Secondary	7	Secondary	Doctorate

Figure 7. Segmentation and explanatory variables

6.5 Preliminary Results

6.5.1 Introduction

In order to explore available data and obtain some preliminary patterns, some analysis have been performed by taking into account the following considerations:

- Bologna-based engineering degrees;
- Data from students that have finished their degrees;
- Only data sets from Universitat Autonoma de Barcelona (UAB) and Instituto Politécnico de Bragança (IPB);
- Only obligatory subjects.

Besides, the amount of segmentation variables (attributes to generate profiles) and explanatory variables (attributes used to explain the obtained profiles) have been reduced as shown in Figure 7.

6.5.2 Data Mining Algorithm

The adopted algorithm is the k-means based clustering approach presented in the previous section, where students have been grouped in three clusters (Excellent students, Average students and Low-Performance students). Concerning the clustering explanation part, preliminary tests have been based on the analysis of the histograms of the different explanatory variables at each cluster (see Figure 8). For instance, one analysis is based on observing the percentage of females and males at each cluster to obtain the predominant sex.

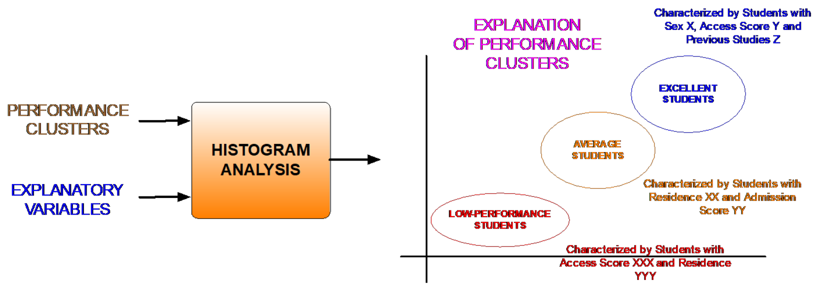

PERFORMANCE CLUSTERS
EXPLANATORY VARIABLES
HISTOGRAM ANALYSIS
EXPLANATION OF PERFORMANCE CLUSTERS
Characterized by Students with Sex X, Access Score Y and Previous Studies Z
EXCELLENT STUDENTS
AVERAGE STUDENTS
Characterized by Students with Residence XX and Admission Score YY
LOW-PERFORMANCE STUDENTS
Characterized by Students with Access Score XXX and Residence YYY

Figure 8. Clustering procedure for student's data analysis and classification

It is worth noting that the first experiments showed the absence of relation between sex and cluster belonging.

6.5.3 UAB Results

In this case, the Telecommunication Systems Engineering degree was analyzed. Bologna degrees are quite new at UAB and, for this reason, the amount of available data is limited. In this case, only 25 students are available and the number of considered subjects is 30. By applying the k-means algorithm, three clusters are obtained as shown in Figure 9. Notice that a PCA dimensionality reduction is applied, so the dimensions are reduced to two.

In order to verify whether performance clusters are correctly generated, the average scores at the different clusters are calculated. On one hand, by taking all the students belonging to a cluster, the average of all the subjects of each student is computed (average of Student' score). With the averages of all students' scores, the histogram from Figure 10 is obtained showing how averages are higher for students belonging to Excellent Students cluster.

Secondly, by considering a given cluster, all the scores for each subject is averaged (average of Subject' score). This is presented in the Figure 11. Again, differences between clusters are observed but one can appreciate wider histograms and some overlapping. This is due to the existence of high and low challenging subjects, where the scores of high or low performance students could be affected by the difficulty level of the subject.

Finally, the access score of students (score obtained to access to the selected degree) is analyzed along with the access age. As presented in Figure 12, excellent students tend to be younger and have a higher access score. However, this analysis provides limited results as more data would be needed.

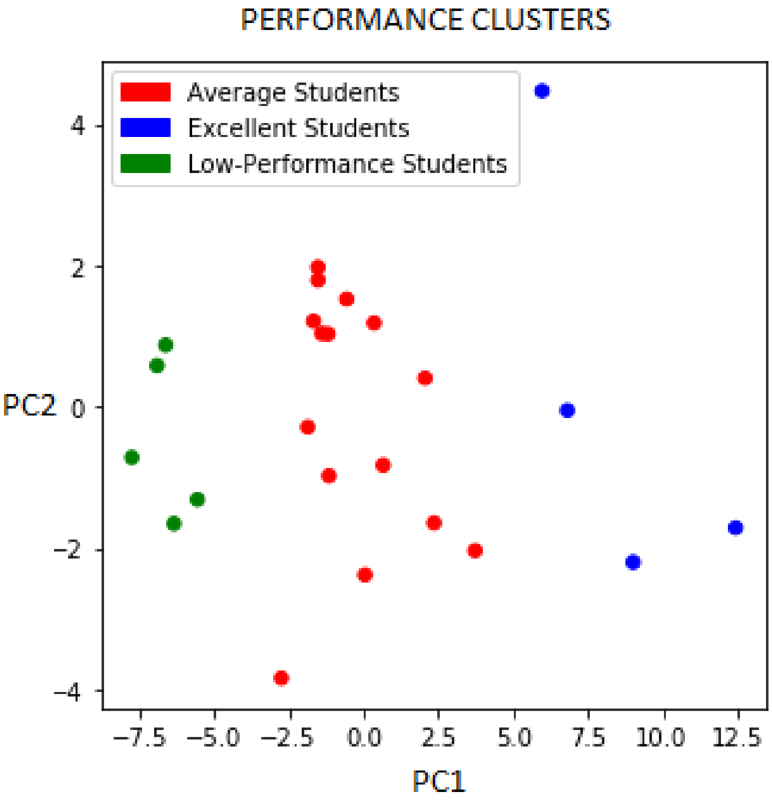


Figure 9. Performance clusters in the case of UAB

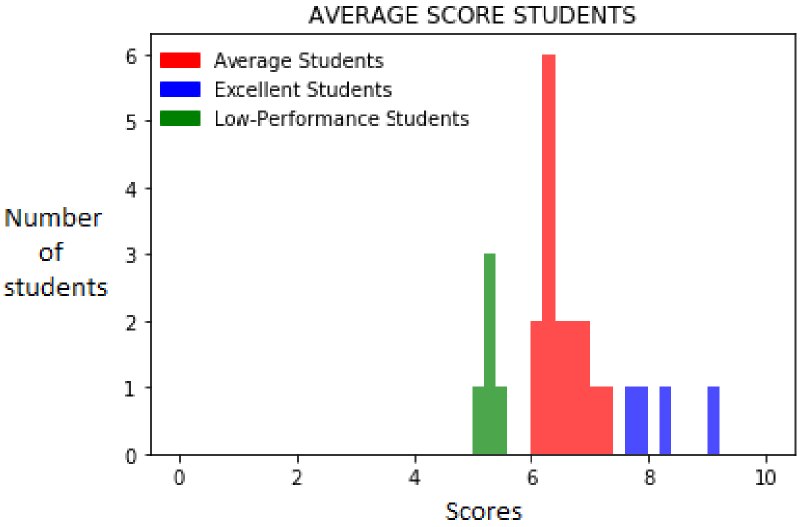


Figure 10. Average score for students in the case of UAB

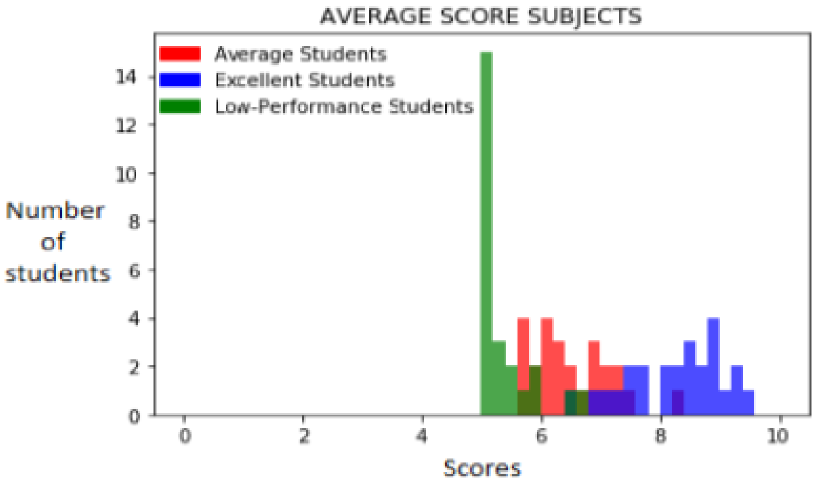


Figure 11. Average score for subjects in the case of UAB

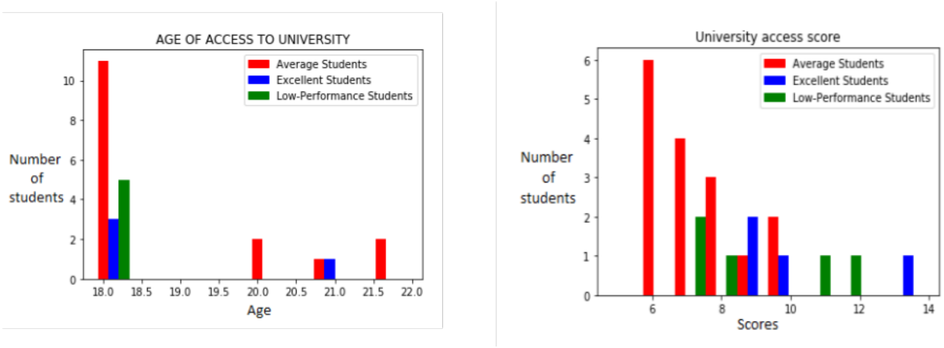


Figure 12. Results regarding the age and access score in the case of UAB

6.5.4 IPB Results

In this case, Mechanical Engineering students are considered (266 students). Again, three clusters are obtained and similar trends to those observed with UAB data are appreciated here in terms of average scores and access age (see Figure 13).

In order to expand the analysis, previous studies of the different students are analyzed. This is reflected at the pie charts presented in Figure 14. Here one can observe that excellent students tend to come from Secondary education, whereas average students have a wider set of kinds of previous studies. Low performance students' previous studies are mostly shared by Secondary and Not Bologna degrees. Not Bologna degrees are referred to those degrees before Bologna harmonization. So, this case is related to those students that were not able to finish these degrees before their elimination and were forced to move to the new Bologna degrees.

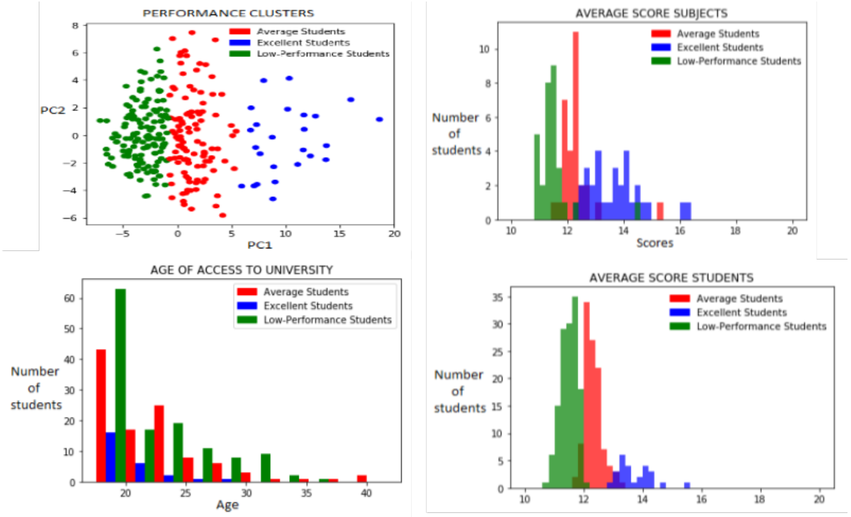


Figure 13. Results obtained in the case of IPB

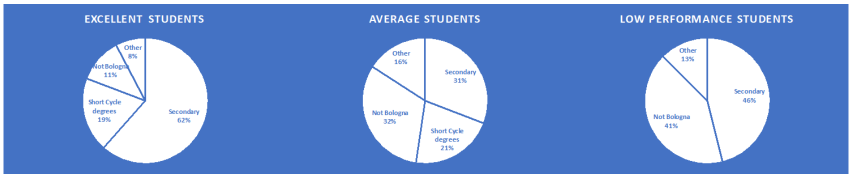


Figure 14. Results regarding the students previous studies in the case of IPB

Finally, students are analyzed in terms of the cities where they come from. As shown in the map from Figure 15 (Average students - red circles, Excellent students - blue circles), excellent students tend to come from a broader region. The hypothesis is that excellent students tend to go to those universities where they believe they can obtain a better curriculum, whereas the rest of students tend to go to universities closer to their residences.

6.5.5 Preliminary Results Conclusions

Some preliminary analysis has been performed with a reduced set of data and simplified versions of the data mining algorithms considered for the project. As observed, results are quite aligned to what teaching professionals would expect based on their experience. So, this shows that preliminary analysis is correct and that the exploitation of this kind of algorithms could significantly simplify the analysis of students' patterns. Next tasks in the project will be focused on expand these analysis and provide more sophisticated tools.

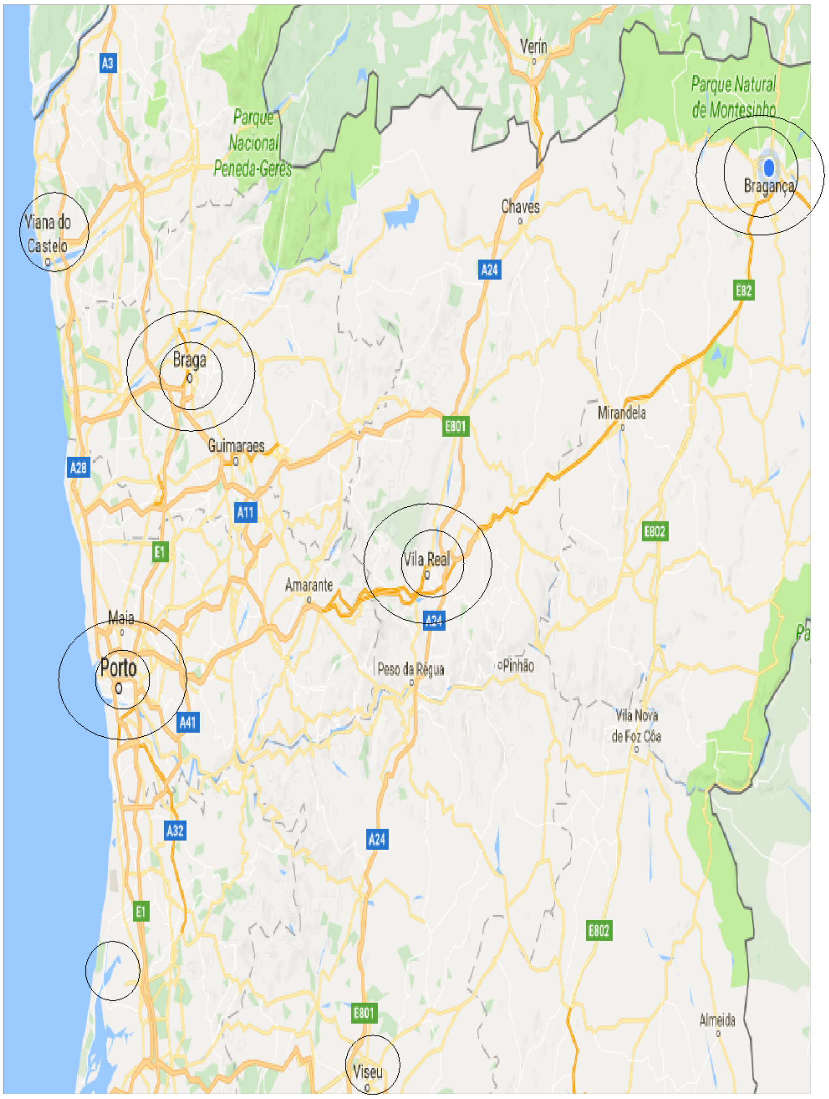


Figure 15. Results regarding the students location in the case of IPB

7 Summary

The use of IT in teaching and learning have multiple approaches, ranging from the simplest, such as wireless internet, email, web pages to the most sophisticated online tutorials, databases, e-portal, electronic platforms or intelligent learning systems. Whatever the concrete form, these applications intensify the student-centered learning and turn the passive-to-active learning environment, allowing a combination between the guided and mediated learning, between learning in campus and distance learning. In this way, the educational action is continued in other social environments, enhancing the learning time, diversifying the workspaces, and adapting to the personal learning style of any student.

IT thus demonstrates its pedagogical valences both inside the university space by helping teachers create an interactive environment in the classroom and outside the university space by extending and enhancing students' learning processes and fixing/completing the acquired information through direct interaction in the classroom. These new learning methods allow students to access information and use them during and in their chosen space, in any media format, whether digital, printed or multimedia. In the same time the institutions can take advantage of the huge amount of data they gain regarding the students activity and use it to improve the students experience on the university.

This report documents the first steps conducted within the SPEET[8] ERASMUS+ project. It describes the conceptualization of a practical tool for the application of EDM/LA techniques to currently available academic data. The document is also intended to contextualise the use of Big Data within the academic sector, with special emphasis on the role that student profiles and student clustering do have in support tutoring actions.

In the second chapter, the report presents the current landscape regarding the use of data mining in the case of higher education institutions, categorizing the goals in term of student benefits. The challenges that arise from using student private are presented in Chapter 3.

[8] Student Profile for Enhancing Tutoring Engineering (www.speet-project.eu)

Chapter 4 is the result of analysing the main methodologies used for analytics and data mining. In the chapter the resources for Educational Data Mining and Learning Analytics are presented, together with the general methods and tools available for visual data analysis.

Chapter 5 is devoted to analyse the traditional methods regarding the student profiling. These methods are based usually on the questionnaires filled by students and by psycho-pedagogical analysis. The students with a high dropout risk are the main focus of these analyses and teaching learning attitudes are presented in this context.

The last chapter presents the first steps made in building the software application object of this project. For that are presented: the software architecture, the structure of the data set and the classification methods that will be considered. Finally some preliminary results based on the datasets provided by Universitat Autonoma de Barcelona and Instituto Politecnico de Braganca. The obtained results showed that preliminary analysis is correct and that the exploitation of this kind of algorithms could significantly simplify the analysis of students' pattern.

References

[ALP16] A. Anaya, M. Luque, and M. Peinado. A visual recommender tool in a collaborative learning experience. Expert Systems With Applications, 45:248–259, 2016.

[Ban82] A. Bandura. Self-efficacy mechanism in human agency. American Psychologist, 37:122–147, 1982.

[Ban86] A. Bandura. Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, 1986.

[BE00] J. Bean and S.B. Eaton. Reworking the Student Departure Puzzle, chapter A psychological model of college student retention, pages 48–62. Vanderbilt University Press, 2000.

[BH05] J.M. Braxton and A.S. Hirschy. College Student Retention: Formula for Student Success, chapter Theoretical developments in the study of college student departure. American Council on Education/Praeger, 2005.

[BI14] R.S. Baker and P.S. Inventado. Educational data mining and learning analytics. In Learning Analytics, pages 61–75. Springer New York, 2014.

[Bis06] C. M. Bishop. Pattern Recognition and Machine Learning. Springer. 2006.

[Bre01] L. Breiman. Random forests. Machine Learning, 45:5–32, 2001.

[Bru00] P. Brusilovsky. ITS 2000, LNCS 1839, chapter Adaptive Hypermedia: From Intelligent Tutoring Systems to Web-Based Education, pages 1–7. Springer-Verlag, 2000.

[CDST12] M.A. Chatti, A.L. Dyckhoff, U. Schroeder, and H. Thus. A reference model for learning analytics. International Journal of Technology Enhanced Learning, 4:318–331, 2012.

[CHG01] M.M. Chemers, L. Hu, and B.F. Garcia. Academic self-efficacy and first-year college student performance and adjustment. Journal of Educational Psychology, 93:55–64, 2001.

[CO14] K.C. Cheong and B. Ong. Pre-college profiles of first year students: A typology. Procedia - Social and Behavioral Sciences, 123:450–460, 2014.

[CV95] C. Cortes and V. Vapnik. Support-vector networks. Machine Learning, 20:273–297, 1995.

[DG16] H. Drachsler and W. Greller. Privacy and analytics: it' a delicate issue a checklist for trusted learning analytics. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, pages 89–98, 2016.

[DHS01] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern classification. 2nd Edition. John Wiley & Sons, 2001.

[Die00] T.G. Dietterich. Ensemble methods in machine learning. In International workshop on multiple classifier systems, pages 1–15, 2000.

[DP97] P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. Machine Learning, 29:103–130, 1997.

[dSBE97] S. de Souza Barros and M.F. Elia. Physics teacher's attitudes: How do they affect the reality of the classroom and models for change? Technical report, 1997.

[Duv11] E. Duval. Attention please! learning analytics for visualization and recommendation. In Proceedings of the 1st International Conference on Learning Analytics and Knowledge - LAK '11. ACM Press, 2011.

[EGB95] R.G. Ehrenberg, D.D. Goldhaber, and D.J. Brewer. Do teachers' race, gender, and ethnicity matter? evidence from the national educational longitudinal study of 1988. ILR Review, 48(3):547–561, apr 1995.

[Elb83] F. Elbaz. Teacher Thinking: A Study of Practical Knowledge. Nichols Pub Co, 1983.

[Fer12] R. Ferguson. Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning, 4:304–317, 2012.

[GD12] W. Greller and H. Drachsler. Translating learning into numbers: A generic framework for learning analytics. Educational technology & society, 15:42–57, 2012.

[GJ63] J. Getzels and P. Jackson. The handbook of research on teaching,, chapter The teacher's personality and characteristics, pages 506–582. Rand McNally, Chicago, 1963.

[HJ01] S. Hu and E.P. St. John. Student persistence in a public higher education system: Understanding racial and ethnic differences. Journal of Higher Education, 72:265âĂŞ286, 2001.

[HKT15] E. Hovdhaugen, A. Kottman, and L. Thomas. Drop -out and completion in higher education in europe: Annex 1 literature review, european commission: Education and culture. Technical report, 2015.

[HL89] D.W. Hosmer and S. Lemeshow. Applied Logistic Regression. John Wiley & Sons, 1989.

[HM11] S. Hu and A.C. McCormick. An engagement-based student typology and its relationship to college outcomes. In Annual forum of the Association for Institutional Research, 2011.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Second Edition. Springer, 2009.

[IW14] D. Ifenthaler and C. Widanapathirana. Development and validation of a learning analytics framework: Two case studies using support vector machines. Technology, Knowledge and Learning, 19:221–240, 2014.

[JLSS10] L. Johnson, A. Levine, R. Smith, and S. Stone. The 2010 Horizon Report. The New Media Consortium, 2010.

[KE15] M. Khalil and M. Ebner. Learning analytics: principles and constraints. In Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, pages 1326–1336, 2015.

[Kei02] D.A. Keim. Information visualization and visual data mining. IEEE Transactions on Visualization and Computer Graphics, 8:1–8, 2002.

[KHZ16] J. Kuzilek, M. Hlosta, and Z. Zdrahal. Open university learning analytics dataset. In Data Literacy For Learning Analytics Workshop at Learning Analytics and Knowledge (LAK16), 2016.

[KKO12] D. Kay, N. Korn, and C. Oppenheim. Legal, risk and ethical aspects of analytics in higher education. techreport Vol.1 No.6, CETIS Analytics Series, ISSN 2051-9214, 2012.

[Lan01] D. Laney. 3D data management: Controlling data volume, velocity, and variety. Technical report, META Group, February 2001.

[Lel00] R. Lelouche. ITS 2000, LNCS 1839, chapter A Collection of Pedagogical Agents for Intelligent Educational Systems, pages 143–153. Springer-Verlag, 2000.

[LG08] G. Lassibille and L. Gomez. Why do higher education students dropout? evidence from spain. Education Economics, 16:89–105, 2008.

[LV07] J.A. Lee and M. Verleysen. Nonlinear Dimensionality Reduction. Springer, 2007.

[LW82] N.M. Laird and J.H. Ware. Random effects models for longitudinal data. Biometrics, 38:963–974, 1982.

[LWM$^+$15] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang. Recommender system application developments: A survey. Decision Support Systems, 74:12–32, 2015.

[MRP$^+$14] A. Monreale, S. Rinzivillo, F. Pratesi, F. Giannotti, and D. Pedreschi. Privacy-by-design in big data analytics and social mining. EPJ Data Science, 3(1), sep 2014.

[MTS15] B. Motz, J. Teague, and L. Shepard. Know thy students: Providing Aggregate Student Data to Instructors. Educause, 2015.

[NS08] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In 2008 IEEE Symposium on Security and Privacy (sp 2008). IEEE, may 2008.

[Nwa90] H.S. Nwana. Intelligent tutoring systems: an overview. Artificial Intelligence Review, 4:251–277, 1990.

[PA14] A. Pena-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. Expert systems with applications, 41:1432–1462, 2014.

[PE14] Z.K. Papamitsiou and A.A. Economides. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Educational Technology & Society, 17:49–64, 2014.

[PIS04] PISA 2003 Learning for Tomorrow's World: First Results from PISA 2003. OECD Publishing, dec 2004.

[PIS13] PISA 2012 Results: Ready to Learn: Students' Engagement, Drive and Self-Beliefs (Volume III). OECD Publishing, dec 2013.

[PJ14] J. Polonetsky and J. Jerome. Student data: Trust, transparency, and the role of consent. SSRN Electronic Journal, 2014.

[RB10] L. Reisel and I. Brekke. Minority dropout in higher education: A comparison of the united states and norway using competing risk event history analysis. European Sociological Review, 26:691–712, 2010.

[Ren94] L.I. Rendon. Validating culturally diverse students: Toward a new model of learning and student development. Innovative Higher Education, 19(1):33–51, sep 1994.

[Ric96] V. Richardson. Handbook of research on teacher education, chapter The role of attitudes and beliefs in learning to teach, pages 102–119. MacMillan Reference Books, 1996.

[Rok61] M. Rokeach. The open and closed mind: Investigations into the nature of belief systems and personality systems. Political Science Quarterly, 76:462–464, 1961.

[RPI17] S. Rovira, E. Puertas, and L. Igual. Data-driven system to predict academic grades and dropout. PLoS one, 12:1–21, 2017.

[RV10] C. Romero and S. Ventura. Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40:601–618, 2010.

[RV13] C. Romero and S. Ventura. Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3:12–27, 2013.

[RVPB10] C. Romero, S. Ventura, M. Pechenizkiy, and R. Baker. Handbook of educational data mining. Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC Press, 2010.

[SB12] G. Siemens and R.S. Baker. Learning analytics and educational data mining: towards communication and collaboration. In Proceedings of the 2nd international conference on learning analytics and knowledge, pages 252–254, 2012.

[SD12] S. Severiens and G. Dam. Leaving college: A gender comparison in male and female-dominated programs. Research in Higher Education, 53:453–470, 2012.

[SGVD12] J.L. Santos, S. Govaerts, K. Verbert, and E. Duval. Goal-oriented visualizations of activity tracking: a case study with engineering students. In Proceedings of the 2nd international conference on learning analytics and knowledge, pages 43–152, 2012.

[SK09] X. Su and T.M. Khoshgoftaar. A survey of collaborative filtering techniques. Advances in artificial intelligence, 4, 2009.

[SM12] O. Scheuer and B.M. McLaren. Encyclopedia of the Sciences of Learning, chapter Educational data mining, pages 1075–1079. Springer, 2012.

[SPM16] N. Sclater, A. Peasgood, and J. Mullan. Learning analytics in higher education, a review of uk and international practice. Technical report, 2016.

[ST12] J.E. Sieber and M. Tolich. Planning Ethically Responsible Research. SAGE PUBN, 2012.

[SWST14] C. Schatten, M. Wistuba, and L. Schmidt-Thieme. Minimal invasive integration of learning analytics services in intelligent tutoring systems. In 2014 IEEE 14th International Conference on Advanced Learning Technologies, 2014.

[TCH09] E.A. Turner, M. Chandler, and R.W. Heffer. The influence of parenting styles, achievement motivation, and self-efficacy on academic performance in college students. Journal of College Student Development, 50:337–346, 2009.

[Tin87] V. Tinto. Leaving College: Rethinking the Causes and Cures of Student Attrition. University of Chicago Press, 1987.

[Tin92] V. Tinto. The Encyclopedia of Higher Education, chapter Student attrition and retention, pages 1697–1709. Oxford/Tarrytown, 1992.

[TM04] M. Tapia and G.E. Marsh. An instrument to measure mathematics attitudes. Academic Exchange Quarterl, 8(2):16–22, 2004.

[TQ03] E. Thomas and J. Quinn. International insights into widening participation. Technical report, Staffordshire University, Institute for Access Studies, 2003.

[TSK$^+$14] A.M. Tervakari, K. Silius, J. Koro, J. Paukkeri, and O. Pirttila. Usefulness of information visualizations based on educational data. In 2014 IEEE Global Engineering Education Conference (EDUCON). IEEE, apr 2014.

[VDK$^+$13] K. Verbert, E. Duval, J. Klerkx, S. Govaerts, and J. L. Santos. Learning analytics dashboard applications. American Behavioral Scientist, 57(10):1500–1509, feb 2013.

[VGO$^+$15] M. Vahdat, A. Ghio, L. Oneto, D. Anguita, M. Funk, and M. Rauterberg. Advances in learning analytics and educational data mining. In ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2015.

[VKJ$^+$15] H. Vossensteyn, A. Kottmann, B. Jongbloed, F. Kaiser, L. Cremonini, B. Stensaker, E. Hovdhaugen, and S. Wollscheid. Dropout and completion in higher education in europe: Main report. Technical report, 2015.

[WB13] P.H. Winne and R.S.J. Baker. The potentials of educational data mining for researching metacognition, motivation and self-regulated learning. Journal of Educational Data Mining, 5, 2013.

[XW05] R. Xu and D. Wunsch. Survey of clustering algorithms. Neural Networks, IEEE Transactions on, 16:645–678, 2005.

[YOT14] E. Yukselturk, S. Ozekes, and Y.K. Turel. Predicting dropout student: an application of data mining methods in an online education program. European Journal of Open, Distance and E-learning, 17:118–133, 2014.

[Zai02] O.R. Zaiane. Building a recommender agent for e-learning systems. In Proceedings of 2002 International Conference on Computers in Education, volume 51, pages 55–59, 2002.