IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 31, 2019, accepted September 8, 2019, date of publication September 23, 2019, date of current version October 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2943212

Clustering Analysis for Automatic Certification of LMS Strategies in a University Virtual Campus

LUISA M. REGUERAS[1], (Member, IEEE), MARÍA JESÚS VERDÚ[1], (Senior Member, IEEE), JUAN-PABLO DE CASTRO[1], AND ELENA VERDÚ[2]

[1]Higher Technical School of Telecommunications Engineering (ETSIT), Universidad de Valladolid, 47011 Valladolid, Spain
[2]School of Engineering and Technology, Universidad Internacional de La Rioja, 26006 Logroño, Spain

Corresponding author: María Jesús Verdú (marver@tel.uva.es)

ABSTRACT In recent years, the use of Learning Management Systems (LMS) has grown considerably. This has had a strong effect on the learning process, particularly in higher education. Most universities incorporate LMS as a complement to face-to-face classes in order to improve the student learning process. However, not all teachers use LMS in the same way and universities lack the tools to measure and quantify their use effectively. This study proposes a method to automatically classify and certify teacher competence in LMS from the LMS data. Objective knowledge of actual LMS use will help the university and its faculty to make strategic decisions. The information produced will be used to support teachers and institutions in the classification and design of courses by showing the different LMS usage patterns of teachers and students. In this study, we processed the structure of 3,303 courses and two million interactive events to obtain a classification model based on LMS usage patterns in blended learning. Three clustering methods were compared to find which one was best suited to our problem. The resulting model is clearly related to different course archetypes that can be used to describe the actual use of LMS. We also performed analyses of prediction accuracy and of course typologies across course attributes (academic disciplines and level and academic performance indicators). The results of this study will be used as the basis for an automatic expert system that automatically certifies teacher competence in LMS as evidenced in each course.

INDEX TERMS Clustering methods, data mining, learning systems, machine learning.

I. INTRODUCTION

Initially, LMS were designed for distance learning, but currently many universities and colleges use them as a complement to face-to-face classes. More and more colleges, schools and universities have incorporated these systems to enhance student learning [1]. In higher education, the adoption of LMS is a universal phenomenon with high rates of use and student satisfaction [2].

There are many factors that support the use of LMS for educational purposes [3], [4]. These systems provide flexibility in terms of time and space, support advanced interactivity between students and teachers and facilitate the reusability of resources. They also allow teachers to distribute learning materials, create and manage thematic debates and bulletin boards, survey and assess students, integrate on-line resources, create collaborative glossaries and manage grades.

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

However, not all teachers perceive these benefits [5]. There may be concerns among some faculty members regarding the adoption of LMS such as complexity, time, management and logistics [6]. As a result, in most cases, teachers simply use basic LMS features such as uploading files and posting their syllabus [3], [7]. If they were aware of different LMS usage strategies and their advantages were verified, they would be able to use them more often and more efficiently.

Most face-to-face educational institutions are encouraging their teachers to use LMS. In many cases, teachers are completely free to decide on their LMS strategy, resulting in multiple ways of blending. In this context, directors of educational institutions would like to know how teachers and students use LMS in order to automatically establish and categorize different classes and profiles of use. Consequently, they would be able to apply actionable intelligence to improve student success [8], enhance education planning and develop precise prediction models to prevent dropout [9]. Given the high use of LMS, it is important to take full advantage of

their potential. LMS (such as Moodle, Blackboard or Canvas) capture and store each user's online behavior data in the system (both teachers' and students') by generating large and varied data sets providing the opportunity to discover valuable information. Based on this data, it would be possible to analyze and interpret the activity and use of LMS in every course. Educational Data Mining (EDM) enables us to discover useful information from LMS to help educational institutions enhance the management of their resources.

Moreover, many national teacher accreditation systems are based on teaching quality programs that evaluate competence in technology-based learning. As a result, teachers are asking for LMS-use certificates from university institutions. Currently this task is done manually by experts or with overly basic metrics based on presence/absence of LMS activity. It would be interesting to be able to automatize this certification process.

To sum up, the adoption of analytics in higher education is a transformative process that is of immense value [10], spanning the full scope and range of activity and affecting administration, research, teaching and learning.

In this study, 3,303 courses taught at a Spanish public university were analyzed regarding the use of institutional LMS by both teachers and students. This study is part of a larger one that pursues a dual objective: 1) to classify courses, detect best practices of LMS use and give recommendations to improve academic results; and 2) to offer educational institutions a tool in order to certificate the use of LMS. Thus, the specific aim of the study described in this paper is to answer the following research questions:

- RQ1: What typologies of LMS usage can be automatically identified when the LMS is used as a complement to face-to-face learning?
- RQ2: How do clustering methods detect courses that present different LMS usage patterns?
- RQ3: What are the instructional characteristics of the different course clusters?
- RQ4: Is there any relationship between the clusters and academic performance?
- RQ5: Can the classification rules extracted for an academic year predict the structure for the next one?

The remainder of this paper is organized as follows. In Section II, a review of the related work is presented. Section III describes the context and the data mining process implemented in this study, along with the methods used. In Section IV, the results of this study and their analyses are presented. Finally, Section V contains the outcomes and insights regarding future work.

II. ACADEMIC ANALYTICS AND EDUCATIONAL DATA MINING IN LMS

Academic Analytics and Educational Data Mining have emerged in the field of education to add real value to the increasing volume of data on instruction, assessment and student effort [11], [12]. Academic Analytics and Educational Data Mining for higher education are providing new opportunities for the faculty; since they can improve learning, transform the university, increase organizational productivity and effectiveness and help to understand the institution's successes and challenges, etc. [10].

Academic Analytics is the process of evaluating and analyzing organizational data sets received from university systems in order to report and improve decision-making [13]. It combines selecting large institutional data sets, statistical analysis, and predictive modelling to create intelligence upon which teachers, or faculty administrators, can improve decision making and academic success. Academic analytics has the potential to improve teaching, learning, and student success and might become a valuable tool in institutional improvement and accountability.

On the other hand, Educational Data Mining arose as a paradigm designed to develop new methods and algorithms in order to explore educational data and to discover valuable hidden patterns that could be used to make decisions and predictions in educational systems [14]–[16].

LMS logs can provide useful insights into student online behavior, since these data sets are captured in real time and reflect aspects of users' behavior that contain very valuable information for educational institutions. The integration of LMS with face-to-face instruction presents unique challenges to implement analytics since they can be combined in myriad ways that are difficult to define and analyze [17].

In recent years, there have been important reviews regarding EDM techniques and educational methods used in LMS [18]–[22]. These papers show how most work focuses on modelling student behavior and predicting student performance. There are a lot of studies regarding clustering students but very few regarding clustering or grouping courses. Moreover, qualitative interpretation of clusters by explaining the use of LMS by teachers and students, as well as supporting institutions and teachers in the development of data-driven course planning are also important fields of research.

Table 1 summarizes the methodologies and results of several studies on the characterization of courses according to the level of LMS usage. It shows the methods and features used, as well as the findings and classes identified in the analysis. Different methods are used to detect hidden usage patterns in LMS: statistical methods, visual information and data mining techniques (for example, classification and clustering), where the choice depends on the objectives of the analysis [23]. Most studies use logs for data collection. However, Iwasaki *et al.* [24] use questionnaires that teachers have to fill out. Then, they identify three categories of courses (knowledge construction, knowledge transmission and mixed) by applying descriptive statistics to the results of questionnaires. Finally, they propose methods for supporting teachers in the development of course plans that encourage active learning.

Frintz [25], Rhoe *et al.* [26] and Park and Jo [27] do not use any data-mining techniques either, but instead they calculate descriptive statistics (such as median, mean, quartiles or maximum) to characterize courses and to define the most used features.

TABLE 1. Summary of studies regarding the characterization of courses.

Study	Methods	Features	Results
Iwasaki et. al [24]	Descriptive statistics	Questionnaire for teachers	Three classes: knowledge construction, knowledge transmission and mixed
Frintz [25]	Descriptive statistics	Number of items grouped by three categories: content, interactive tools (forums, chats, wikis, blogs, etc.) and assessment (quizzes, exams, gradebook, etc.)	Four quartiles ICDQx - Institutional course design quartile
Rhoe et al. [26]	Descriptive statistics	Tool use (yes-no): announcements, items, grades, folders, files, assignments, web links, plagiarism detection, discussion boards, tests	Definition of the most frequently used features
Park & Jo [27]	Descriptive statistics	Three general indicators (login frequency, members, average login frequency) and ten activity-based indicators (announcements, links, lecture notes, resources, Q&A, discussion, quiz, group work, wikis and assignment submission)	Definition of the most used features
Whitmer et al. [28]	K-Means clustering	Percentage of time spent in each tool (normalized by course enrolment and length): assessment, announcements, gradebook, discussion board, content, assignments	Five classes: supplemental, complementary, social, evaluative and holistic
Park et al. [29]	Latent Class Analysis (LCA)	Activity items: announcements, links, lecture notes, resources, Q&As, discussion forums, quiz items, group works, wikis, assignments	Four classes: inactive or immature, communication or collaboration, delivery or discussion and sharing or submission
Valsamidis et al. [30]	K-Means clustering	Time spent in each tool (normalized by course enrolment and length): assessment, announcements, gradebook, discussion board, content, assignments	Two classes: low and high activity
Jo et al. [31], [32]	Gaussian Mixture Model, K-Means clustering and Hierarchical clustering	Three general indicators (members, login frequencies, activity items) and ten activity-based indicators (resources, notices, Q&A, lecture notes, task submissions, group work, links, discussion forum postings, quiz, wikis)	Four classes: forum-based, quiz-based, wiki-based and resource-based online instruction

On the other hand, Whitmer *et al.* [28] apply k-means clustering to a data set of 18,810 courses (after filtering the initial 70,000 courses from 927 institutions), identifying five course design patterns used by teachers (supplemental, complementary, social, evaluative and holistic). This study shows that there is real diversity in the way in which LMS is used, in contrast to other research that tends to view course design as something gradual characterized by an incremental level of use [25], [30]. Diversity in LMS usage patterns is also supported by other authors [27], [29], [31]. For example, Park *et al.* [29] use Latent Class Analysis (LCA) to extract common activity features of 616 higher education courses. They identify four classes of blended-learning courses based on different use patterns more than on incremental use. In addition, Jo *et al.* [31] apply three different clustering techniques to 2,639 higher education courses and find four clusters which are considerably uncompensated, each with different use strategies (forum-based, quiz-based, wiki-based and resource-based).

These studies use data mining to establish the clusters. However, they have some limitations. For example, Valsamidis *et al.* [30] analyze a very small number of courses (only 39). Jo *et al.* [31], [32] are focused on comparing the results obtained through different clustering methods and do not go on to analyze their characteristics and implications. Moreover, these studies do not analyze whether the classification rules for an academic year predict the behavior for the following one, or what the instructional characteristics of the different course clusters are, or their relation to the students' academic performance. On the other hand, Park and Jo [27] do an interesting analysis of activity patterns across course attributes (undergraduate vs. graduate, colleges and selective vs. mandatory). However, they do not use data mining to group the courses and only consider the top-five most frequently used activity items.

III. METHODOLOGY

This study proposes a method to classify the courses taught at a Spanish public university according to LMS usage patterns. In the following sections, we describe the methods used as well as research context and ethics.

A. METHODS

In this study, we followed the steps and methodology of the KDD (Knowledge Discovery in Databases) process shown in Fig. 1.

A MySQL database engine was used for data aggregation from the educational environment, and R for data preprocessing and transformation as well as for the data mining process. We used R as the data mining tool because it is a free software environment that provides a wide variety of statistical and graphical techniques and is widely used by statisticians and researchers [33], [34]. We used the following R packages: 'caret 6.0-81', 'arules 1.6-1', 'poLCA 1.4.1' and 'rpart 4.1-13'. The specific methods used in the different steps of the methodology are described in the following sections.

1) DATA COLLECTION AND SELECTION

A preliminary phase in the data mining process is data collection and preparation. Data was collected from the Moodle

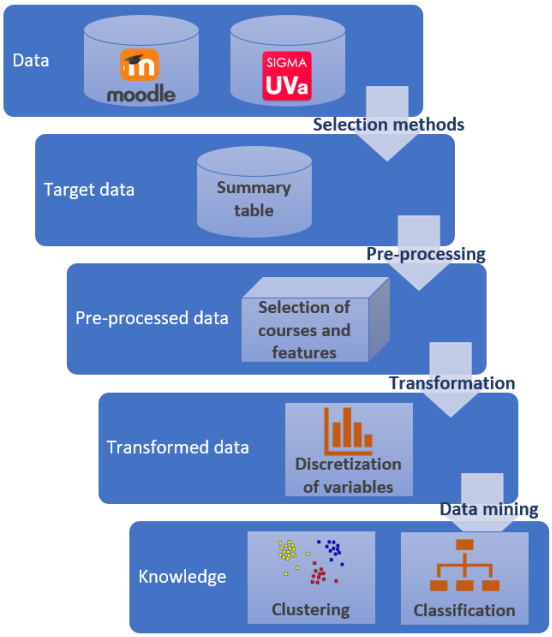

FIGURE 1. Methodology sketch.

database and the Academic Management System database (called SIGMA) for the academic year 2015-2016. The data from Moodle included approximately 2 million records of log-data on teaching and learning activities of all participants (teachers and students) for each course. The data from SIGMA included course-related administrative information (graduate vs. undergraduate, academic center, etc.) and the grades obtained by students in each course. To pre-process and link the Moodle and SIGMA data, we used SQL scripts and created a summary table of the courses with aggregated information about participants, activity and performance indicators. During this process, non-linked courses (that is, courses missing in one of the two databases), courses without students and courses without activity were eliminated, leaving a total of 3,303 pre-selected courses.

Moreover, before analyzing and classifying the courses, it was necessary to identify and select the cases of interesting information, in order to maximize efficiency and validity [35]. We selected all cases that met some predetermined criterion of importance, that is, the courses with at least five students enrolled (since this is the minimum number of students required for an optional course to be taught, according to the Academic Management Regulation of the University of Valladolid). Finally, 3,046 courses were selected. Although there were many studies where courses with low use of LMS were eliminated [28], [29] we decided to leave them in to check if the analysis itself could detect them.

Once the courses had been filtered, we transformed and selected the variables to conduct a sound analysis of Moodle usage patterns. A Moodle course can integrate both resources and activities. Resources are items that teachers can use to support learning, such as files, links, labels, pages and folders; whereas activities are elements that allow students to interact with one another and/or with teachers (for example, forums, assignments, quizzes, glossaries, workshops and wikis) [36]. Besides, it is possible to configure tools such as the event calendar and the gradebook for management purposes.

According to this classification, 14 variables were selected (see Table 2). The first two variables are related to resources, the next nine to activities and the last three to management (gradebook and calendar). Instead of considering the different types of resources separately, we grouped data on all of them. Regarding activities, we only selected the three activities with a more extensive use in the virtual campus (forums, assignments and quizzes) and we grouped the rest in another variable, due to their limited use.

TABLE 2. Description of variables in the summary table.

Variable	Description	Role
Resources	Number of resources	Teacher
ResourceViews	Number of resource views or downloads	Student
Forums	Number of discussion forums	Teacher
ForumNews	Number of teachers' forum posts	Teacher
ForumInteractions	Number of students' forum views and posts	Student
Assigns	Number of assignments	Teacher
AssignSubmissions	Number of assignment submissions	Student
Quizzes	Number of quizzes	Teacher
QuizSubmissions	Number of quiz submissions	Student
OtherActivities	Number of other activities	Teacher
OtherActivitySubmissions	Number of other activity submissions	Student
GradeItems	Number of manual gradebook items	Teacher
GradeFeedbacks	Number of feedbacks	Teacher
CalendarEvents	Number of manual calendar events	Teacher

Table 2 shows the description of variables as well as who carries out the corresponding action. 'Resources' counts the digital course materials (html files, pdf or word documents, etc.) uploaded by teachers; while 'ResourceViews' registers students' actions to view or download a resource. 'Forums' and 'ForumNews' account for forums and posts added by teachers, respectively; while 'ForumInteractions' records the students' activity in discussion forums (posts and views). The following six items ('Assigns', 'AssignSubmissions', 'Quizzes', 'QuizSubmissions', 'OtherActivities' and 'OtherActivitySubmissions') register the number of activities of each type created by teachers and, then, the number of student participations or submissions. The following two indicate some measure of the use of Moodle assessment elements. 'GradeItems' is the number of items added manually by teachers to the gradebook; while 'GradeFeedbacks' is the volume of personalized feedback entered by teachers in the activity module or directly in the gradebook, all of which demonstrates a direct use of this tool. Finally, 'CalendarEvents' considers those calendar events that were entered manually by teachers (that is, events automatically entered in

the calendar, such as submission due dates, are excluded from the total). All activity-related variables are normalized to the number of students enrolled on the course.

2) DATA PRE-PROCESSING AND TRANSFORMATION

From the course summary table, a pre-selection process was carried out by selecting features of interest. Selecting the right features or attributes is an important task when pre-processing data. Data can contain attributes that are highly correlated with one another, and many methods work better if these attributes are removed [37]. Moreover, constant and almost constant features across samples (zero and near-zero variance predictors, respectively) also might cause failure or the fit to be unstable [38]. Thus, we used the R function 'findCorrelation' to identify which redundant features from dataset could be removed, and the R function 'nearZeroVar' to remove both zero and near-zero variance predictors. Then, the variables in Table 2 showing high correlation and/or low variability of values were removed, as explained in the results section.

Once the attributes were selected, the data was discretized to significantly reduce the number of possible values of the variables. Discretizing the attributes that will feed the learning system contributes to reducing learning time and could improve the accuracy, interpretation and comprehensibility of results [39]. Discretization divides the numerical data into categorical classes that are easier to understand.

There are different supervised and unsupervised methods for transforming continuous attributes into discrete ones [40]. In this study, we used an unsupervised method (k-means clustering) with three intervals and labels (low, medium and high) for all variables, like Romero *et al.* [41]. This is an unsupervised univariate discretization algorithm that applies the k-means clustering method to one-dimensional continuous data. This type of discretization greatly reduces the complexity of the data and makes the analysis more resistant to outliers and extreme values [42]. Therefore, we selected the function 'discretize' of the R package 'arules' [43] with the discretization method cluster (k-means clustering) and three breaks.

3) DATA MINING

Different data mining techniques were applied to obtain knowledge from the data and to be able to classify the courses. Specifically, we applied three clustering methods to our data set to compare them and choose the most suitable one for our analysis. We tested Latent Class Analysis (LCA) and two more traditional methods: K-means and hierarchical clustering, since they are often used because they are easy to understand and visualize [31].

LCA is a statistical method used in factor models, clustering and regression models for testing theories regarding analysis of multivariate categorical data [44]. In this method, classes are identified and created from unobserved categorical variables that divide a population into mutually exclusive and exhaustive latent classes. Class membership of individuals is unknown but can be inferred from a set of observed variables [45]. LCA has been used in different fields. Collins and Lanza [46] show several examples applied to social, behavioral and health sciences.

We used the R package 'poLCA' for the estimation of latent class models [47], and the 'kmeans' and 'hclust' functions for performing k-means and hierarchical clustering, respectively.

The first step to building clusters is to decide on the number of classes. Different studies [48], [49] suggest the use of BIC (Bayesian Information Criterion) as a good indicator of the number of latent classes in LCA. Moreover, silhouette (a direct method) and gap statistic (a statistical testing method) are good methods to determine the optimal number of clusters for k-means and hierarchical clustering [50].

Although at least 30 clustering quality indexes have been proposed in the literature, not all of them are applicable to every case [51]. We measured the performance of each method by two typically employed quantities: homogeneity and heterogeneity [52]–[54]. Homogeneity is a measure of the variation of the observations within each cluster, while heterogeneity gives an idea of the separation of clusters. Typically, the objective is to obtain clusters with low variability within clusters and a high degree of separation between them. We used the average distance between clusters as a heterogeneity measure and the average distance within clusters as a homogeneity measure [50].

Finally, once the clustering method had been chosen and the obtained classes had been interpreted and labeled, we were able to use these results to classify courses for other academic years. We selected the R package 'rpart' [55] to build a decision tree with the minimum prediction error. This decision tree was applied to data for the following year (2016-2017) to make predictions with the R function 'predict'. Then, we repeated the clustering analysis with these new data to check if the decision tree generated for one academic year could accurately predict the behavior for other years.

B. CONTEXT

The study took place at the University of Valladolid, a Spanish public university in the city of Valladolid, with a campus in another three cities in Castilla-y-León (Palencia, Segovia and Soria). This institution, which was established in the 13th Century, has 25 colleges and offers more than 3,000 face-to-face undergraduate and graduate courses in different academic disciplines. It has more than 2,000 teachers and approximately 32,000 students on the enrolment each academic year. This institution has its own virtual campus, based on Moodle LMS, which has been used as a support to face-to-face classes since 2009. Moodle allows teachers to upload and share materials, hold online discussions and chats, create quizzes and surveys, propose and evaluate assignments, record and manage grades and integrate other interactive online activities [36]. All courses taught at the University of Valladolid have a corresponding course in Moodle, on which both teachers and students are enrolled. However, it is each teacher's decision how to use this platform, resulting in a

use which differs in manner and intensity. In this context, the institution is interested in classifying the courses according to LMS usage by using an expert system that could replace manual evaluation of teachers' on-line competence.

C. RESEARCH ETHICS

In the field of learning and academic analytics, the main challenge regarding ethical issues has been related to the ownership of the data and student privacy issues [56]. In this study, after combining the two databases (SIGMA and Moodle) by using course identifiers, which could reveal information on the teachers' identity, the course identifiers were re-codified to minimize possible ethical issues. Moreover, since the unit of analysis was the course, and not the student, no potential problem of student identification was involved in this study. In any case, student anonymity was always preserved by removing all personal identifiers from the data. Moreover, we did not collect any sensitive data such as racial origin, religious beliefs or data concerning health (according to the Spanish Law of Personal Data Protection).

IV. RESULTS AND DISCUSSION

In this section, we present and discuss the main results obtained in the study from the data and methods described in previous sections.

A. DATA PRE-PROCESSING AND TRANSFORMATION

After applying the methods for data pre-processing described in Section III-A2, nine variables were selected: 'Resources', 'ResourceViews', 'Forums', 'ForumNews', 'ForumInteractions', 'Assigns', 'AssignSubmissions', 'GradeItems' and 'GradeFeedbacks'.

Table 3 shows a descriptive analysis of the nine variables of interest. In this table, we can see how there are some heavily skewed variables with too many zeros, which indicate that most courses do not incorporate the corresponding activity.

TABLE 3. Descriptive statistics of features of interest.

Variables	Courses with non-zero values	Max	Mean	SD
Resources	97.1%	329	29.6	30.3
ResourceViews	97.1%	413.7	36.1	31.6
Forums	97.7%	34	1.6	1.8
ForumNews	68.6%	193	7.0	11.2
ForumInteractions	69.5%	120.8	4.7	8.7
Assigns	44.1%	61	2.2	4.3
AssignSubmissions	45.9%	48.1	1.7	3.1
GradeItems	47.8%	153	2.8	6.3
GradeFeedbacks	13.6%	33	0.4	1.6

Once selected, the features of interest were discretized using k-means cut-off thresholds as aforementioned.

B. CLUSTERING OF COURSES

Before building the clusters, we had to establish the optimal number of classes for the three clustering methods, as explained in Section III-A3. We calculated the BIC values for LCA and a model with six classes was selected since it obtained the lowest BIC value. We applied silhouette and gap statistic and we obtained six as the optimal number of clusters for k-means and hierarchical clustering.

We compared the three methods by using the homogeneity and heterogeneity measures described in Section III-A3. Table 4 shows the results obtained for the three clustering methods. K-means provided the best value for homogeneity (the lowest value is the best), that is, it offers the most homogeneous clusters; while LCA presented the best result for heterogeneity (the highest value is the best), that is, it is the most effective in differencing clusters, as was expected.

TABLE 4. Comparative analysis of clustering methods.

Method	Homogeneity	Heterogeneity
K-means	0.6548	2.0887
Hierarchical clustering	0.7565	2.0287
LCA	0.8097	2.2049

Since the objective method did not give a clear winner and different cluster methodologies would result in different class interpretation, we decided to combine the previous performance analysis with a subjective one in order to select just one clustering method. Therefore, we applied the three different methods to obtain three six-class groupings and, then, we studied and compared their possible interpretations in terms of course blended learning strategies.

After applying LCA for six classes, we obtained the results shown in Fig. 2. From this figure, six different course typologies can be established: Class 1 corresponds to courses with low use of Moodle or Inactive courses (type I or Inactive). Class 2 are courses with some content and a considerable use of assignments (type S or Submission). Class 3 corresponds to courses with a lot of content but with very little student interaction; these courses focus on classroom-based teaching but include online elements such as slides and links to resources. In this case, Moodle is used as a Web-repository, where teachers upload the material for their classes (type R or Repository). Class 4 matches courses with high interaction through discussion boards and teacher-student communication, showing a profile of communicative courses (type C or Communicative).

Class 5 looks like Class 2, although it has a greater use of assignments and evaluative elements such as gradebook manual items; thus, they are courses with some content and a considerable use of evaluative elements (type E or Evaluative). Finally, Class 6 corresponds to courses with a considerable use of Moodle tools, and with a balanced use of assignments, content, discussions and evaluative elements (type B or Balanced). Table 5 summarizes the description of course typologies.

K-means and hierarchical clustering offered some similar classes to those of LCA. However, they did not obtain any class similar to LCA Class 5, which shows an interesting

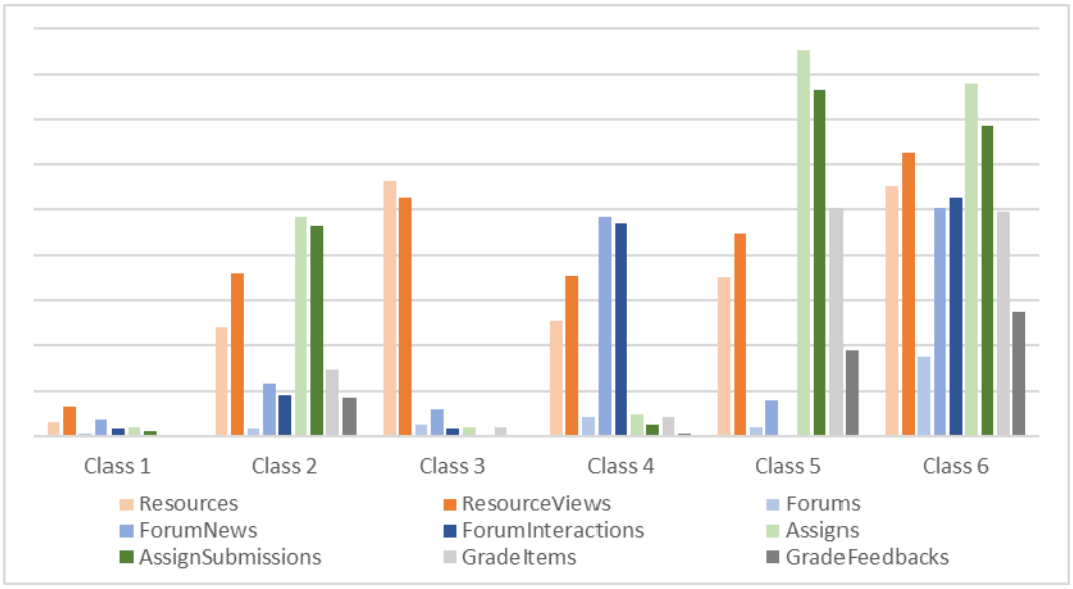

FIGURE 2. Description of the six latent classes.

TABLE 5. Description of course typologies.

Typology	Description
Type I or Inactive	Low use of Moodle
Type S or Submission	Some content and considerable use of assignments
Type R or Repository	A lot of content and low student interaction
Type C or Communicative	High interaction teacher-students
Type E or Evaluative	Some content and considerable use of evaluative elements
Type B or Balanced	Considerable and balanced use of Moodle tools

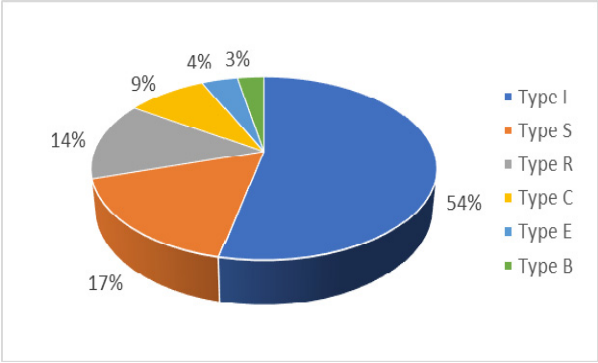

FIGURE 3. Distribution of the six LCA typologies of courses.

component of formative assessment (with more grade items and feedback than the other classes, except Balanced type) and a clearer prevalence of interactive tasks against static resources than LCA Class 2. On the contrary, both K-means and hierarchical clustering found a class of courses that was missing in LCA results. It is a class that includes courses with many resource views but only a few resources. This could be due to very large or compressed files, for example, which could include all course materials and, therefore, should be accessed a lot by students during the learning process. With LCA, those courses are integrated into Class 3 (Repository type) or in Class 1 (Inactive type). Since LCA clustering allowed us to make a richer interpretation of teacher strategies, we chose it as clustering method for our study.

Generally, the LCA classes are not divided by a higher or lower use of LMS, but by the different ways of using it; these results are similar to the findings of other studies [27]–[31].

In Fig. 3, it can be seen how the distribution of the courses is not homogeneous. The results of this study indicated that classes were considerably imbalanced and that most courses had low use of virtual campus (type I).

These results are consistent with the findings of [27], [31]. However, unlike the studies analyzed in Section II there is a greater variability of course typologies.

To test the prediction accuracy of the model, we built the decision tree. Then, it was applied to data for the following year to check if the decision tree generated for one academic year could accurately predict the behavior for other years. Very good results were obtained, with a prediction accuracy of 0.9325.

C. ANALYSIS OF CLASSES

Further analysis can be done from the classification of the courses, attending to different aspects: (1) academic discipline or field of knowledge, (2) undergraduate versus graduate, (3) the number of enrolled students and (4) academic performance.

1) FIELD OF KNOWLEDGE

Fig. 4 shows how the six classes are distributed in the different fields of knowledge. We can observe how the disciplines

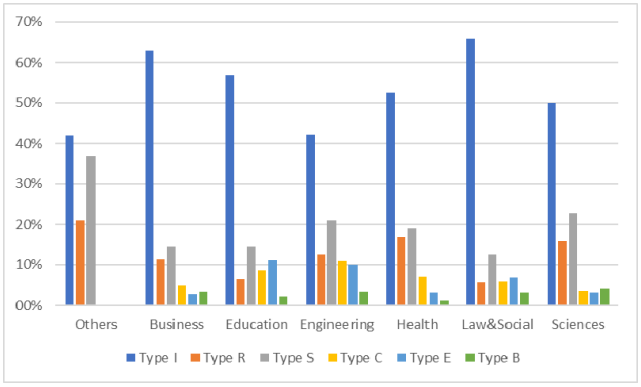

FIGURE 4. Course typology distribution by academic discipline.

of Business, Law and Social are characterized by a lower use of virtual campus (type I predominates with more than 60%); whereas Sciences and Engineering have a greater distribution of type S than the other disciplines. Moreover, typology B is the least used (it does not exceed 5%), which shows that there is still a lot of work to be done to incorporate an LMS into teaching.

A more detailed analysis allows us to see if these two variables are independent. We use the chi-square test to compare the relationships between the two nominal variables, course typology and field of knowledge, and to see if they are independent or not. By conducting the chi-square test, we obtain a high chi-squared value and a very small p-value significance level (201.37, $p < 0.001$), which provides evidence to suggest that field of knowledge and course typology have a significant relationship. Teachers from different fields of knowledge tend to have different preferences for typologies of courses. Fig 5. represents a mosaic plot, where we can see that typology I is preferably used by Law and Social (fully saturated blue color); whereas Engineering preferably uses types C, E, R and S, Sciences types R and S, Education type E and Health type R.

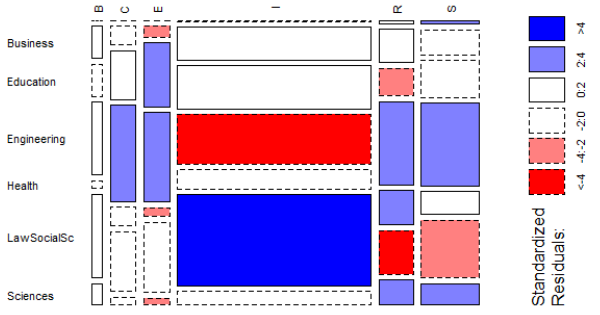

FIGURE 5. Correlation mosaic between typology and academic discipline.

2) UNDERGRADUATE VERSUS GRADUATE

Undergraduate and graduate courses pursue different academic objectives. We use the chi-square test to compare the relationship between the two nominal variables, course typology and academic level (undergraduate or graduate), and to see if they are independent or not. Conducting the chi-square test, we obtain a high chi-squared value and a very small p-value significance level (46.54, $p < 0.001$), which provides evidence to suggest that academic level and course typology have a significant relationship.

Specifically, Fig. 6 represents a mosaic plot, in which we can see that Inactive typology is used preferably by postgraduate teachers (blue color). For undergraduate courses there is no significant correlation with any class.

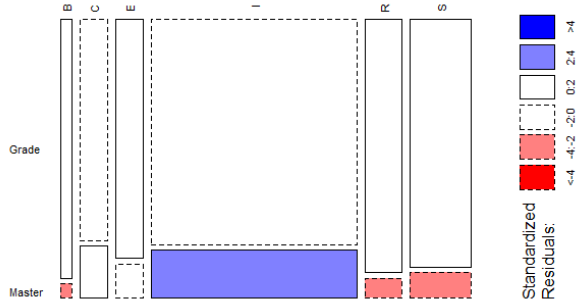

FIGURE 6. Correlation mosaic between typology and academic level.

3) NUMBER OF ENROLLED STUDENTS

Fig. 7 shows the frequency distribution according to the number of enrolled students for the six typologies of courses. We can see how the number of students does not follow a very different pattern in each class. Type C and type B are used more by courses with a higher number of students, but the difference is not significant (Kruskal-Wallis test, $p > 0.05$). Thus, the size of class does not determine the use of LMS.

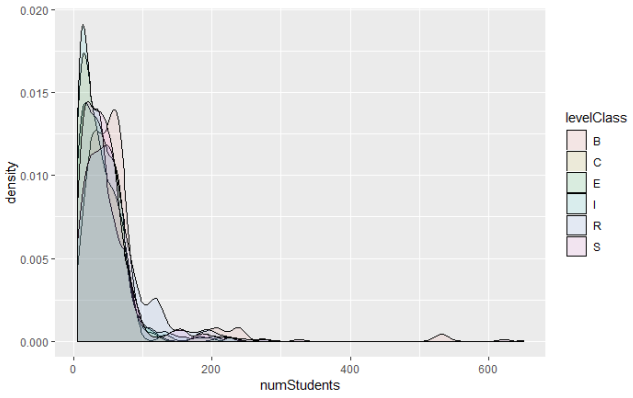

FIGURE 7. Histogram of the number of students for each course typology.

4) ACADEMIC RESULTS

Another interesting analysis is to check if the different course classes provide different academic results according to several parameters (such as performance rate, success rate and average grade), and whether these differences are significant or not. Fig. 8 shows the values for these three parameters. We can see how the Evaluative and Submission typologies obtain the best academic results, whereas Repository and Communicative types obtain the worst.

TABLE 6. Pairwise comparisons using Wilcoxon test for performance rate.

	Type C	Type E	Type B	Type I	Type R	Type S
Type C	-	-	-	-	-	-
Type E	0.00025***	-	-	-	-	-
Type B	0.05837	0.70379	-	-	-	-
Type I	2.2e-10***	1.00000	0.70379	-	-	-
Type R	0.34465	0.00582**	0.59153	3.5e-08***	-	-
Type S	3.8e-12***	1.00000	0.19187	0.34465	4.0e-10***	-

***p<0.001
**p<0.01

TABLE 7. Pairwise comparisons using Wilcoxon test for success rate.

	Type C	Type E	Type B	Type I	Type R	Type S
Type C	-	-	-	-	-	-
Type E	2.9e-06***					-
Type B	0.1251	0.1664				-
Type I	7.2e-15***	1.0000	0.0912			-
Type R	0.0927	0.0024***	1.0000	1.3e-09***		-
Type S	5.1e-16***	1.0000	0.0970	0.3717	1.4e-10***	-

***p<0.001

TABLE 8. Pairwise comparisons using Wilcoxon test for average grade.

	Type C	Type E	Type B	Type I	Type R	Type S
Type C						-
Type E	8.6e-07***					-
Type B	0.056	0.177				-
Type I	3.3e-11***	0.354	0.354			-
Type R	0.339	1.7e-05***	0.341	2.0e-09***		-
Type S	1.5e-13***	0.608	0.104	0.339	1.2e-11***	-

***p<0.001

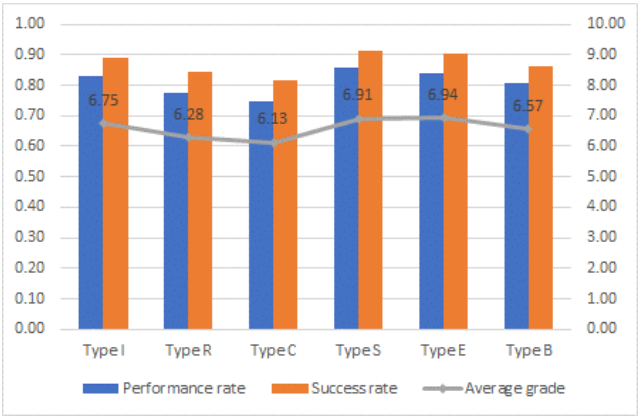

FIGURE 8. Academic results by course typology.

The performance rate for each course was calculated as the number of students passing the course divided by the number of students enrolled. We wanted to know if there was any significant difference between the performance rate in the six typologies of courses (S > E > I > B > R > C, as shown in Fig. 8). We used the Kruskal-Wallis test, since the samples do not have a normal distribution and, therefore, the ANOVA test cannot be applied. From this result ($H = 90.76$, $p < 0.005$), we can conclude that there are significant differences between the six groups in question, although we do not know which pairs of groups are different. Then, pairwise comparisons using the Wilcoxon test were made (see Table 6). The pairwise comparison shows that E-R, E-C, I-R, I-C, S-R and S-C pairs are significantly different ($p < 0.01$). Thus, the E-Evaluative type, I-Inactive type and S-Submission type are significantly better than the R-Repository type and the C-Communicative type. However, the differences in performance are too low to be used in a decision system. A deeper analysis should be done with the courses belonging to each class. Regarding the B-Balanced type, no conclusion can be obtained. Similar results are obtained for the success rate (see Table 7), calculated as the number of students passing the course divided by the number of students taking the exam. Finally, average grade also shows similar results (see Table 8).

V. CONCLUSION

This paper proposes a method of classification for academic courses in higher education in accordance with LMS usage patterns.

We have identified six typologies of LMS usage with different profiles: Inactive type, Repository type, Submission type, Communicative type, Evaluative type and Balanced type (answering RQ1). The Latent Class Analysis has been able to detect the existence of courses with a low use of Moodle, without having to filter them previously. Moreover, different clustering techniques offer different results

(answering RQ2), LCA being the one which renders more useful classes.

These six course typologies have different instructional characteristics (answering RQ3). The classes are not influenced by the volume of interaction but by the strategy of integrating LMS into face-to-face courses. Thus, the statistical analysis plus the subjective interpretation of the classes could define a practical expert system which could be suitably incorporated into the decision workflows of the university.

There is a significant relationship between the classes and the field of knowledge and between the classes and the academic level (undergraduate and graduate). The Inactive type is used preferably in graduate courses and by Law and Social teachers. Moreover, there is also a significant relationship between course typologies and academic performance (answering RQ4): The Evaluative, Inactive and Submission types have significantly better results than the Repository and Communicative types.

We have also checked that the classification rules for an academic year can predict the behavior for the following one, with great accuracy (answering RQ5). Thus, the model is highly accurate and can be used suitably as a practical classifier.

These findings allow us to design an expert system that can automatically assess and certify teacher LMS competence. The results provide a solid grounding for an objective classification. Consequently, the expert system could relieve from this task human experts who previously needed to subjectively judge LMS usage competence from scarce and partial data.

Hence, the next step is to implement an extension for the LMS Moodle that gives the teacher an automatic assessment and certification of their actual use of the LMS and some advice and guidance on how to discover new tools and techniques for their courses.

To make the model evolve with the dynamics of the institution, feedback will be obtained proactively from the teachers. This feedback will allow us to further analyze the inner structure of the classes and their true impact on performance indicators and to adjust the model if needed. Moreover, the teachers' feedback will also provide information on why they use certain features for a particular course.

REFERENCES

[1] R. C. Kushwaha, A. Singhal, and S. K. Swain, "Learning pattern analysis: A case study of moodle learning management system," in *Recent Trends in Communication, Computing, and Electronics* (Lecture Notes in Electrical Engineering), vol. 524. Singapore: Springer, 2019, pp. 471–479.

[2] J. D. Galanek, D. C. Gierdowski, and D. C. Brooks, "ECAR study of undergraduate students and information technology 2018," EDUCAUSE Center Anal. Res., Lousville, KY, USA, Tech. Rep., 2018.

[3] A. Balderas, L. De-La-Fuente-Valentin, M. Ortega-Gomez, J. M. Dodero, and D. Burgos, "Learning management systems activity records for students' assessment of generic skills," *IEEE Access*, vol. 6, pp. 15958–15968, 2018. doi: 10.1109/ACCESS.2018.2816987.

[4] Y. Psaromiligkos, M. Orfanidou, C. Kytagias, and E. Zafeiri, "Mining log data for the analysis of learners' behaviour in Web-based learning management systems," *Oper. Res.*, vol. 11, no. 2, pp. 187–200, Aug. 2011. doi: 10.1007/s12351-008-0032-4.

[5] F. A. A. Trayek and S. Sariah, "Attitude towards the use of learning management system among university students: A case study," *Turkish Online J. Distance Educ.*, vol. 14, no. 3, pp. 91–103, 2013.

[6] M. A. Ocak, "Why are faculty members not teaching blended courses? Insights from faculty members," *Comput. Educ.*, vol. 56, no. 3, pp. 689–699, Apr. 2011. doi: 10.1016/j.compedu.2010.10.011.

[7] C. R. Graham, W. Woodfield, and J. B. Harrison, "A framework for institutional adoption and implementation of blended learning in higher education," *Internet Higher Educ.*, vol. 18, pp. 4–14, Jul. 2013. doi: 10.1016/j.iheduc.2012.09.003.

[8] J. Campbell, P. DeBlois, and D. Oblinger, "Academic analytics: A new tool for a new era," *Educause Rev.*, vol. 42, no. 4, p. 40, Jul. 2007.

[9] A. Essa and H. Ayad, "Student success system: Risk analytics and data visualization using ensembles of predictive models," in *Proc. 2nd Int. Conf. Learn. Anal. Knowl.*, May 2012, pp. 158–161.

[10] P. Long and G. Siemens, "Penetrating the fog: Analytics in learning and education," *Educause Rev.*, vol. 46, pp. 31–40, Sep./Oct. 2011.

[11] P. Baepler and C. J. Murdoch, "Academic analytics and data mining in higher education," *Int. J. Scholarship Teach. Learn.*, vol. 4, no. 2, p. 17, 2010.

[12] S. K. Mohamad and Z. Tasir, "Educational data mining: A review," *Procedia Social Behav. Sci.*, vol. 97, pp. 320–324, Nov. 2013.

[13] J. P. Campbell, P. B. DeBlois, and D. G. Oblinger, "Academic Analytics," *Educause Rev.*, vol. 42, no. 4, pp. 40–57, 2007.

[14] S. P. Algur, P. Bath, and N. Kulkarni, "Educational data mining: Classification techniques for recruitment analysis," *Int. J. Mod. Educ. Comput. Sci.*, vol. 8, no. 2, pp. 59–65, Feb. 2016.

[15] W. Klösgen and J. M. Zytkow, *Handbook of Data Mining and Knowledge Discovery*. London, U.K.: Oxford Univ. Press, 2002.

[16] J. Luan, "Data mining and knowledge management in higher education—Potential applications," presented at the Annu. Forum Assoc. Institutional Res., Toronto, ON, Canadá, 2002.

[17] A. G. Picciano, "Big data and learning analytics in blended learning environments: Benefits and concerns," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 2, no. 7, pp. 35–43, 2014.

[18] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.*, vol. 33, pp. 125–146, Jul. 2007.

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

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

[21] K. Sin and L. Muthu, "Application of big data in education data mining and learning analytics—A literature review," *ICTACT J. Soft Comput.*, vol. 5, no. 4, pp. 1035–1049, Jul. 2015.

[22] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1432–1462, 2014.

[23] M. C. S. Manzanares, R. M. Sánchez, C. I. G. Osorio, and J. F. DÃez-Pastor, "How do B-learning and learning patterns influence learning outcomes," *Frontiers Psychol.*, vol. 8, p. 745, May 2017.

[24] C. Iwasaki, T. Tanaka, and K. Kubota, "Analysis of relating the use of a learning management system to teacher epistemology and course characteristics in higher education," *Knowl. Manage. E-Learn., Int. J.*, vol. 3, no. 3, pp. 478–490, 2011.

[25] J. Frintz, "LMS Course Design As Learning Analytics Variable," in *Proc. PCLA@LAK*, Apr. 2016, pp. 15–19.

[26] J. Rhode, S. Richter, P. Gowen, T. Miller, and C. Wills, "Understanding Faculty Use of the Learning Management System," *Online Learn.*, vol. 21, no. 3, pp. 68–86, 2017.

[27] Y. Park and I.-H. Jo, "Using log variables in a learning management system to evaluate learning activity using the lens of activity theory," *Assessment Eval. Higher Educ.*, vol. 42, no. 4, pp. 531–547, 2017.

[28] J. Whitmer, N. Nuñez, T. Harfield, and D. Forteza, *Patterns in Blackboard Learn Tool Use: Five Course Design Archetypes*. Washington, D.C., USA: Blackboard, 2016.

[29] Y. Park, J. H. Yub, and I.-H. Jo, "Clustering blended learning courses by online behavior data: A case study in a Korean higher education institute," *Internet High Educ.*, vol. 29, pp. 1–11, Apr. 2016.

[30] S. Valsamidis, S. Kontogiannis, I. Kazanidis, T. Theodosiou, and A. Karakos, "A clustering methodology of Web log data for learning management systems," *J. Educ. Technol. Soc.*, vol. 15, no. 2, pp. 154–167, 2012.

[31] I.-H. Jo, J. Song, Y. Park, H. Lee, and S. Kang, "Clustering analysis of academic courses based on LMS usage levels and patterns: Gaussian mixture model, K-means and hierarchical clustering," in *Proc. 4th Int. Conf. Data Analytics*, 2015, pp. 1–8.
[32] I.-H. Jo, Y. Park, and J. Song, "Comparisons on clustering methods: Use of LMS log variables on academic courses," *Educ. Technol. Int.*, vol. 18, no. 2, pp. 159–191, 2017.
[33] Y. Zhao, "Introduction to data mining with R," presented at the Stat. Modeling Comput. Workshop Geosci., Canberra, ACT, Australia, 2015.
[34] W. N. Venables and D. M. Smith, "An introduction to R," R Foundation for Statistical Computing, Vienna, Austria, Tech. Rep., 2015.
[35] L. A. Palinkas, S. M. Horwitz, C. A. Green, J. P. Wisdom, N. Duan, and K. Hoagwood, "Purposeful sampling for qualitative data collection and analysis in mixed method implementation research," *Admin. Policy Mental Health Mental Health Services Res.*, vol. 42, pp. 533–544, Sep. 2015.
[36] J. Cole and H. Foster, *Using Moodle: Teaching with the Popular Open Source Course Management System*. 2nd ed. Newton, MA, USA: O'Reilly Media, 2007.
[37] D. C. Yu-Wei, *Machine Learning with R Cookbook*, Birmingham, U.K.: Packt Publishing, 2015.
[38] M. Khun and K. Johnson, *Applied Predictive Modeling*. Berlin, Germany: Springer, 2013.
[39] J. Catlett, "On changing continuous attributes into ordered discrete attributes," in *European Working Session on Learning* (Lecture Notes in Computer Science), vol. 482. Berlin, Germany: Springer, 1991, pp. 164–178. doi: 10.1007/BFb0017012.
[40] J. Dougherty, R. Kohavi, and R. Sahami, "Supervised and unsupervised discretization of continuous features," in *Proc. 12th Int. Conf. Mach. Learn.*, 1995, pp. 194–202.
[41] C. Romero, S. Ventura, and E. García, "Data mining in course management systems: Moodle case study and tutorial," *Comput. Educ.*, vol. 51, pp. 368–384, 2008.
[42] X. Yan and T. Zheng, "Selecting informative genes for discriminant analysis using multigene expression profiles," *BMC Genomics*, vol. 9, no. 2, p. S14, 2008.
[43] M. Hahsler, K. Hornik, and C. Buchta, "The arules R-package ecosystem: Analyzing interesting patterns from large transaction data sets," *J. Mach. Learn. Res.*, vol. 12, pp. 2021–2025, Jun. 2011.
[44] D. Rindskopf, "Latent class analysis," in *The SAGE Handbook of Quantitative Methods in Psychology*. London, U.K.: SAGE Publications, 2009, pp. 199–218.
[45] J. K. Vermunt and J. Magidson, "Latent class cluster analysis," in *Applied Latent Class Analysis*, J. A. Hagenaars A. L. McCutcheon, Eds., Cambridge, U.K.: Cambridge Univ. Press, 2002, pp. 89–106.
[46] L. M. Collins and S. T. Lanza, *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. Hoboken, NJ, USA: Wiley, 2010.
[47] D. A. Linzer and J. B. Lewis, "poLCA: An R package for polytomous variable latent class analysis," *J. Stat. Softw.*, vol. 42, no. 10, pp. 1–29, 2011.
[48] J. Hagenaars and A. McCutcheon, *Applied latent Class Analysis Models*. New York, NY, USA: Cambridge Univ. Press, 2002.
[49] K. L. Nylund, T. Asparouhov, and B. O. Muthén, "Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study," *Struct. Equation Model. Multidisciplinary J.*, vol. 14, no. 4, pp. 535–569, 2007.
[50] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, "NbClust Package: Finding the relevant number of clusters in a dataset," *J. Stat. Softw.*, vol. 61, no. 6, pp. 1–36, 2014.
[51] S. A. Marklouf and B. Yagoubi, "Data-aware scheduling strategy for scientific workflow applications in IaaS cloud computing," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 5, no. 4, pp. 75–85, 2019.
[52] D. Haughton, P. Legrand, and S. Woolford, "Review of three latent class cluster analysis packages: Latent gold, poLCA, and MCLUST," *Amer. Statistician*, vol. 63, no. 1, pp. 81–91, 2009.
[53] A. Eshghi, D. Haughton, P. Legrand, M. Skaletsky, and S. Woolford, "Identifying Groups: A Comparison of Methodologies," *J. Data Sc.*, vol. 9, no. 2, pp. 271–291, 2009.
[54] A. M. Navarro and P. Moreno-Ger, "Comparison of clustering algorithms for learning analytics with educational datasets," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 5, no. 2, pp. 9–16, 2018.
[55] T. Therneau and B. Atkinson. (2019). *An Introduction to Recursive Partitioning Using the RPART Routines*. [Online]. Available: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf
[56] S. Slade and P. Prinsloo, "Learning Analytics: Ethical Issues and Dilemmas," *Amer. Behav. Scientist*, vol. 57, no. 10, pp. 1510–1529, 2013.

LUISA M. REGUERAS (M'10) received the master's and Ph.D. degrees in telecommunications engineering from the University of Valladolid, Spain, in 1998 and 2003, respectively, where she is currently an Associate Professor with the Department of Signal Theory, Communications and Telematics Engineering.

Her research interests include new e-learning technologies, gamification, educational data mining, and expert systems.

MARÍA JESÚS VERDÚ (M'09–SM'11) received the master's and Ph.D. degrees in telecommunications engineering from the University of Valladolid, Spain, in 1996 and 1999, respectively, where she is currently an Associate Professor and the Deputy Director of the Higher Technical School of Telecommunications Engineering.

She has experience in coordinating projects in the fields of new telematic applications for the Information Society and Telecommunication Networks, especially related to e-learning. Her research interests include new e-learning technologies, gamification, educational data mining, and expert systems.

JUAN-PABLO DE CASTRO received the master's degree in telecommunications engineering from the University of Valladolid, Spain, in 1996, and the Ph.D. degree in telecommunications engineering from the Polytechnic University of Madrid, in 2000.

He was the Research Director of the Technological Centre for the Development of Telecommunications (CEDETEL), from February 2001 to June 2003. He is currently an Associate Professor with the Department of Signal Theory, Communications and Telematics Engineering, University of Valladolid. He currently acts as a Research and Development and Technology Consultant. His research interests include new e-learning technologies, gamification, educational data mining, expert systems, and spatial data infrastructures.

ELENA VERDÚ received the master's and Ph.D. degrees in telecommunications engineering from the University of Valladolid, Spain, in 1999 and 2010, respectively.

She is currently an Associate Professor with the Universidad Internacional de La Rioja (UNIR), where she is also a member of the Research Group "Data Driven Science." For more than 15 years, she has worked on research projects at both national and European levels. Her research has focused on e-learning technologies, intelligent tutoring systems, competitive learning systems, data mining, and expert systems. She is a Managing Editor of the Journal *International Journal of Interactive Multimedia and Artificial Intelligence*.