IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Analysis of students' behaviour through user clustering in online learning settings, based on Self Organizing Maps neural networks

SOLEDAD DELGADO[1], FEDERICO MORÁN[2], JOSÉ CARLOS SAN JOSÉ[3], and DANIEL BURGOS[3]

[1]Universidad Politécnica de Madrid, Departamento de Sistemas Informáticos, Madrid, 28031, SPAIN
[2]Universidad Complutense de Madrid , Departamento de Bioquímica y Biología Molecular , Madrid, 28040, SPAIN
[3]Research Institute for Innovation & Technology in Education (UNIR iTED), Universidad Internacional de La Rioja (UNIR), 26006 Logroño, SPAIN

Corresponding authors: Soledad Delgado (e-mail: mariasoledad.delgado@upm.es), Daniel Burgos (e-mail: daniel.burgos@unir.net).

The authors thank Research Institute for Innovation & Technology in Education (UNIR iTED) and UNESCO Chair on eLearning at Universidad Internacional de La Rioja (UNIR), Universidad Politécnica de Madrid (UPM) and Universidad Complutense de Madrid (UCM), for their support to this joint research. F.M. acknowledges financial support from Grant CTQ2017-87864-C2-2-P, from the Ministerio de Economia y Competitividad (MINECO), Spain

ABSTRACT An accurate analysis of user behaviour in online learning environments is a useful means of early follow up of students, so that they can be better supported to improve their performance and achieve the expected competences. However, that task becomes challenging due to the massive data that learning management systems store and categorise. With the COVID-19 pandemic still on-going, face-to-face learning settings have migrate into online and blended ones, meaning an increase of online students and teachers in need for a tailored and effective support to their needs. A novel unsupervised clustering technique based on the Self-Organizing Map (SOM) artificial neural network model is used in this research to analyse 1,709,189 records of online students enrolled from 2015 to 2019 at Universidad Internacional de La Rioja (UNIR), a fully online Higher Education institution. SOM performs a precise and diverse user clustering based on those records. Results highlight that specific clusters are linked to the intake average profile at the university, with a clear relation between user interaction and a higher performance. Further, results show that, out of a targeted desk research compared to the analysis in this paper, face-to-face and online settings are connected through the methodological approach beyond the technology-based environment, which presents a similar behaviour in both contexts.

INDEX TERMS Artificial Neural Networks, Data science applications in education, Distance education and online learning, Pattern analysis, Self-Organizing Map (SOM), Student behaviour, Unsupervised learning.

I. INTRODUCTION

From a technological point of view, online training is practical and easy to implement. Online learning management systems (LMSs) are widely accessible, as are open-source tools (e.g., Sakai and Moodle). Thus, any higher education institution can install these systems in an agile manner and begin uploading content and registering users. At the methodological and pedagogical levels, it is somewhat more complicated. These contexts involve developing content and implementing a face-to-face model or a remote, purely online model – or a blended model that includes face-to-face and online learning [1], [2]. The means of delivery is key for tailoring the methodology, interaction, and performance.

The current context, with the pandemic still underway and vaccines gradually rolling out, is characterized by severe mobility restrictions by country and region, which has in a few months driven the migration to online learning that has been slowly evolving over the last 20 years. However, all the complexity of an academic structure and community cannot immediately be transferred from traditional teaching models to the online model [3].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3115024, IEEE Access

Adaptation requires rigorous analysis and meticulous planning. Therefore, it is necessary to identify systems that support the successful design and implementation of online learning and teaching processes.

Data science and artificial intelligence can help to scale the size of study samples and refine subsequent profiles and actions [4], [5], and analysing student behaviours and attitudes can improve the educational experience, either in face-to-face, blended or online settings. A deep analysis of what the student does (behaviour) and not only of what the student declares to do (e.g. evaluation questionnaires) provides a similar approach and performance in any context [6]. The ability to group these behaviours to identify patterns that allow personalized actions that drive academic performance provides excellent support in online contexts, especially with a large number of students, groups, or courses[7].

In the analysis of student behaviour, a set of key questions focuses on the productivity and performance of students: do these outcomes differ greatly in the online and in-person context? If so, how much and in what form [8], [9]? The training process is designed to enable the acquisition and development of a series of core competencies and competences specific to the subject of study, regardless of the means of study or teaching [10], [11]. Literature review reveals that students that are consistent in time, with regular learning patterns and a serious approach to deadlines and competence achievement succeed. Further, the student must satisfy the criteria set out in a specific rubric to demonstrate proficiency and graduate, and that rubric must be identical for the same curriculum in any context, whether face-to-face, online, or any other. In these frameworks a rubric remains unaltered since their criteria must be all met for the very same curriculum, no matter the setting [12], [13]. On the other hand, the correlation of activity patterns with final grades, and with pass or even dropout rates, would aid in the design and implementation of tailored action plans by groups that fit those patterns to reaffirm attitudes, reinforce knowledge, or reroute indicators that do not show the expected evolution [14].

The high dimension and the volume of elements of the dataset as well as the objectives proposed in this study require the use of data mining tools to extract common behaviour patterns from the activities carried out by students on the online platform. Furthermore, support is required to subsequently analyse the data's relationship with the final grade obtained by the students in the courses undertaken. To obtain a knowledge model of the behaviour patterns observed in the values of the variables that define the dataset and the relationships between them, several data mining techniques can be used. Data mining techniques exhibit two main forms: supervised and unsupervised. Supervised techniques can be applied when the dataset consists of vectors of independent variables with an associated dependent variable (or objective function), and the aim is to identify the relationship that exists between them. On the other hand, unsupervised techniques aim to find out meaningful patterns that underlie the dataset, without the presence of labels or objective function. In the study that we address in this paper we try to identify hidden relationships in a dataset made up of independent variables, without an objective function. In this context, to extract the common behaviour patterns in the dataset, clustering is the most appropriate technique. Clustering is an unsupervised technique that groups the vectors in the dataset according to a criterion of distance or similarity, such as the Euclidean distance. The vectors grouped in a cluster usually share common properties, and the subsequent analysis of the clusters makes it easier to discover the properties that characterize them and to thus obtain a description of a complex high-dimensional dataset.

Over the years, a large number of clustering methods have been proposed, which address the problems related to the identification of patterns in a dataset through different shapes, densities, and sizes. These methods can be classified into partition-based, hierarchical, and density-based clustering methods. Partition-based clustering methods divide the n-dimensional space of the dataset into a predetermined number of groups, where samples belonging to the same cluster are similar to each other and samples from different clusters are farther apart. The most widely used partition-based clustering algorithm is *k*-means [15]. This algorithm obtains a simplified model of the dataset by calculating *k* clusters, each modelled by a single representative vector of each group. Although the *k*-means algorithm is one of the most used due to its simplicity, it has some drawbacks, such as the imposition of a spherical shape of the clusters, which is not always adequate to describe cluster structures with arbitrary or complex shapes. Furthermore, since the representative vectors share the n-dimensional nature of the dataset vectors, *k*-means does not facilitate their visualization, which complicates the subsequent analysis of each cluster. Another commonly used clustering technique is hierarchical clustering [16]. Hierarchical clustering algorithms generate a representation of the dataset in the form of a tree, or dendrogram, which facilitates the analysis of the data structure, although the visualization and interpretation of the dendrogram can be complex for large datasets. One of the problems posed by both partition-based clustering algorithms (with a single centroid per cluster) and hierarchical algorithms is the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3115024, IEEE Access

discrimination of groups of samples that have arbitrary shapes, especially when these shapes are non-spherical. In these cases, density-based methods are usually more appropriate [17], since they are based on the identification of high-density regions surrounded by low-density regions. However, its main drawbacks include the difficulty in adjusting some of the parameters, its sensitivity to density drops to identify the boundaries of the clusters, as well as the weakness in the characterization of the intrinsic structure of the data [18].

In recent years, clustering techniques based on deep learning have been proposed, which have improved the accuracy of clustering for certain benchmark datasets [19]. These methods differ from traditional clustering algorithms in that they group data samples by finding complex patterns rather than using simple predefined metrics like intra-cluster Euclidean distance. One of the most popular deep learning-based clustering techniques is deep autoencoder [20]. An autoencoder network provides a non-linear mapping through learning an encoder and a decoder. The encoder performs a dimensionality reduction of the original dataset. The encoder output feeds the decoder input, which produces an output with the same dimension as the original encoder input. The goal of deep autoencoder training is the reconstruction at the output of the network of the given input. Once the network has been trained, the encoder output constitutes a compression of the original vector information into a vector with a smaller dimension. The autoencoder itself does not perform the clustering of the original dataset, but rather a transformation of it, being necessary the use of some additional algorithm to identify groups in the transformed dataset. In the context of deep clustering, feed-forward networks trained for clustering have also been proposed [19]. These are deep networks with many hidden layers, an aspect that makes it difficult to configure their architecture (number of hidden layers and number of neurons per layer). Furthermore, they require well-designed clustering loss to avoid learning a corrupted feature representation. Finally, the Generative Adversarial Networks (GAN) [21] and the Variational Autoencoder (VAE) [22] incorporate the generation of new samples from the clusters obtained, although they exhibit high computational complexity. In general, clustering techniques based on deep learning make a transformation of the input space by expressing the original dataset through new features (usually with a smaller dimension than the original) that assume a non-linear combination of the original data features. In the context of data mining, this transformation complicates or prevents the analysis and identification of the original dataset components that characterize each of the groups achieved in the clustering, so its use is not appropriate when this functionality is required.

Self-Organizing Maps (SOM) is a type of unsupervised neural network that projects a high-dimensional input space onto a low-dimensional output space [23], [24]. This network model consists of one output layer composed of neurons, each one with a synaptic vector of the same nature and dimension as the input dataset, organized in a regular grid of rows and columns that establishes neighbourhood connections between neurons. The SOM training algorithm adjusts the reduced set of synaptic vectors so that they become prototype vectors representing the input space. In this sense, SOM performs a vector quantization. In addition, it also produces a vector projection, by mapping the high-dimensional input data vectors to the SOM output grid, usually two-dimensional. While SOM is a powerful tool for high-dimensional data analysis, this network model has some limitations or drawbacks. One of them is the configuration of the network architecture, that is, the number of neurons and their arrangement in the output layer grid, characteristics that remain static throughout the entire network life cycle. The number of neurons in the network directly affects the generalizability of the SOM, therefore it is a critical aspect. In the 1990s, new architectures of self-organizing maps emerged to address this problem, such as Growing Neural Gas (GNG) [25], Growing Cell Structures (GCS) [26], Growing Hierarchical Bregman SOM (GHBSOM) [27], or Growing Self-Organizing Map (GSOM) [28], among others. In general, these models create a basic initial network, composed of few neurons connected to each other, and during the training process they incorporate new neurons and neighbourhood connections to the architecture of the output layer of the network. Some of them, such as GCS and GNG, also incorporate the removal of neurons and neighbourhood connections, being able to divide the grid of the output layer of the network into several sub-meshes, each one representing a subspace of the n-dimensional input space. Despite the advantages provided by these dynamic SOM models, it should be noted as a disadvantage the existence of a greater number of training parameters with more complexity in their configuration (e.g., the criterion for insertion of new neurons, the threshold for neuron removal, or the final size of the network). Furthermore, some of these models (such as GNG) do not ensure the two-dimensional nature of the output layer grid, so they cannot be used to visualize the knowledge acquired by the network.

To use SOM in clustering processes, the challenge is to determine the groups of neurons that identify natural clusters in the original dataset. In this regard, some

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3115024, IEEE Access

approximations based on distance matrices have been proposed [29]–[32], although they have some drawbacks, such as the high sensitivity of certain cluster shapes or the imposition of spherical or ellipsoid clusters.

In the context of identifying online student behaviour patterns from events stored in log files, different works have been published to address this challenge using some of the clustering techniques previously exposed [33]. One of the most relevant characteristics of these works is the way in which the data representation is approached, that is, how the different types of events generated in the online platform are counted throughout the duration of the course. The most common strategy is to count every event absolutely throughout the entire course [34], [35]. In these cases, the clustering results usually identify between 2 and 4 groups of behaviour that simply discriminate levels of global interaction of students with the learning platform. The most notable drawback of this type of data representation is that it is not possible to detect different patterns of event generation depending on the period of the course in which it occurs. This could be important in the early detection of students who may drop out of the course. In some studies this problem is addressed by identifying daily, weekly or monthly behaviour patterns, but it is only applicable when the courses analysed have the same duration, since the comparison must be made between patterns that occur in the same period of time [36]. Another issue to highlight is the size of the dataset used in this type of studies, which in many cases comprises only one or a few subjects taught in a single period [34], which may include a bias related to the characteristics of the course or subject (the type of activities configured on the platform, the number and type of evaluations, etc.). Another bias that can affect the type of patterns detected is related to the use of datasets generated from students who explicitly agree to be part of the study [37], since these experiments do not ensure that they cover the full range of types of enrolled students.

In the study presented in this work, we address the identification and analysis of behaviour patterns of students in online subjects, based on events stored in log files. We propose a representation of the patterns collected by the generation of events in different periods of the subject and that can be applied to subjects with a variable range of duration. In addition, the dataset is made up of 1,709,189 records of online students enrolled from 2015 to 2019, which means different subjects developed in 4 different academic years. The clustering methodology used in this study is the one proposed by Delgado et al. [38] based on the Self-Organizing Map (SOM) artificial neural network model. From the existing unsupervised clustering techniques, this novel approach has demonstrated important advantages in visualizing, preserving and analysing complex knowledge [38], [39]. This methodology consists of two phases. In the first phase, the SOM network size that best adapts to the dataset is identified. This first phase addresses the previously described problem of the SOM model related to determining the number of neurons that best represents the analysed dataset. In the second phase, the clustering of the SOM prototype vectors is carried out, where several prototypes can be grouped into a single cluster, providing cluster structures with complex or arbitrary shapes and different densities. Adjusting the parameters of this clustering methodology to determine the optimal size of the network as well as the number of clusters to analyse is very simple, depending only on the number of patterns in the dataset. Furthermore, by using a methodology based on SOM, it has been possible to take advantage of the ability of this model to project a high-dimensional data space into a low-dimensional one (usually two-dimensional) that can be used to generate graphs to visualize an analyse the intrinsic knowledge of the data.

II. MATHERIALS AND METHODS

An overview of the context of the e-learning data and the analysis performed in the study is shown in Fig. 1. The detailed description of each part is given below.

A. DATASETS

1) DESCRIPTION OF THE ONLINE PLATFORM

The data used for this research were collected and extracted from two servers with the Sakai LMS installed at Universidad Internacional de La Rioja (UNIR), a young Spanish university, 100% online, with premises in Spain and Latin-America, over 45.000 annual students and 2.500 faculty members. Versions 10 and 11 were used at the time of data extraction.

Sakai is open-source educational software that manages online learning. It allows the distribution of different subjects, or courses, with their corresponding teaching units and offers a wide variety of features and configurations. In addition to the distribution of content, forums, and deliverable tasks, questionnaires and grades can be managed, among many other options. The system also allows content providers to control and store data related to students' progress. For example, the system records the number of times a student has accessed the subject forum and whether the student has read or posted messages. Similarly, it allows users to measure access to the content provided by the teacher and attendance in online classes.

2) SOURCE DATASETS

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3115024, IEEE Access

The raw data extracted from the LMS described above reflect the efforts of the professors and students who participated in online subjects from 2015 to 2019 in one of the faculties of the university, the School of Engineering and Technology (ESIT). In this context, 'subject' refers to a teaching unit that must be completed to earn academic credit. All available subjects were downloaded, some at undergraduate and others at master's level, and of varying duration.

3) DESCRIPTION OF THE ORIGINAL FILES

The data were distributed across several files. The first, called DataSet.csv, with 1,709,189 entries, contains the log of all events performed by the users of the LMS, linked to each subject. These events include attending online classes, participating in forums, performing tasks, and completing questionnaires. The file consisted of the following fields that were used for analysis:

- Registration ID, user_id – a unique hash per user present in other files.
- site_id – subject in which the event occurred.
- edition_id – subject's edition (subject content can change over time, with different editions published).
- curriculum_id – curriculum of the subject.
- subjetct_id –identifier of the subject.
- event – action the user performed
- section – tool or area that received the action (content area, tasks, calendar, online class, etc.).
- occurrence – number of times an even occurred on the day.
- timestamp – day on which the action was performed.

Once we captured all the actions that had been performed in the Sakai LMS during the four years under study, we generated another file that contains the roles of LMS users, including teachers, administrators, and students. This file, called DataSetRoles.csv, contains 95,874 records and the following fields:

- Id – unique record identifier.
- user_id – user identifier, with an encrypted hash and that relates to other files.
- role – user's role.
- site_id – subject where the user fulfils their role.

DataSetCourses.csv, with 1,108 records, contains the fields with the duration of the subjects. The extracted fields are 'site_id', which as in the other files identifies the subject; 'begin', which marks the start date; and 'end', which marks the day of completion of the subject.

The raw extraction was completed with the generation of the file ScoresOutput.csv, whose data were extracted from the university's management system and which contains the students' results by subject. The file consists of 19,863 tuples. The file includes the following fields:

- user_id – encrypted hash student ID.
- site_id – code that identifies the subject evaluated.
- continuous evaluation – grade obtained following the formula and weight in the scored works during continuous evaluation.
- coordination examination – grade resulting from applying weight and regulations in ordinary call for face-to-face examination.
- external examination – grade resulting from applying weight and regulations in extraordinary call for face-to-face examination.
- FinalNote – final grade, considering continuous evaluation, and ordinary and extraordinary face-to-face examinations.
- ordinary – grade obtained by student if they passed ordinary examinations.
- extraordinary – grade obtained by student during extraordinary registration period.

4) ANONYMIZATION

In the exported files, an encryption hash algorithm was applied to sensitive data such as student identified data. This makes these values anonymous, distorting sensitive data by minimizing the risk of reidentifying sensitive fields without altering the results of the analyses.

5) PREPROCESSING

The content of the files DataSet (event per user, subject, and day), DataSetRoles (role per user and subject), and DataSetCourses (subject) were processed to obtain the activity record for each student-course. In the DataSetRoles file, 16 types of roles were identified. Crossing the Dataset and DataSetRoles files, only those events produced by students were selected, generating the DataSetStudents file as a result, with a total of 1,385,682 records. In the DataSetCourses file, 1,108 subjects were identified, some with inconsistent or poorly established start and end dates or with an end date after the date on which the data for this study were obtained, so these subjects were filtered, resulting in 799 subjects correctly configured. Within this group of subjects, a range of duration between 12 and 441 days was observed.

To analyse the activity patterns of the students on the platform in a set of subjects with homogeneous length, only those with an approximate duration between three and four months were selected. As a result of this filter, 534 subjects were obtained with a duration between 90 and 138 days. Given that each record in the DataSetStudents file indicated the number of times that a student had generated an event in a subject on a specific day and that the subjects considered had a variable duration in days, to unify the sizes of the activity vectors per student-course, the duration of each

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3115024, IEEE Access

subject was divided into 10 periods (PER). In the DataSetStudents file, 28 types of events were identified, of which the 10 that were considered most relevant were selected (see Table 1). Those 10 are the most significant indicators in the methodological framework of the online setting.

TABLE 1
TYPES OF EVENTS SELECTED FOR THE ANALYSIS

Event	Description
EV1	Submit homework
EV2	Create resource
EV3	Read resource
EV4	Create new post in forum
EV5	Read forum post
EV6	Reply to forum post
EV7	Read page
EV8	Submit exam
EV9	Read external page from web content
EV10	BasicLTI tool released

Crossing the information from the DataSetStudents files and the 534 subjects with a duration range between 90 and 138 days, the activity vectors of each student in each subject were generated. Each activity vector consisted of 100 values that counted the number of times a student had produced each of the 10 types of events in the 10 periods into which the duration of the subject was divided. As a result of this processing, 13,292 vectors of student-course activity were obtained.

From the student-course vectors obtained, the distribution of the values of the number of events per period and type of event was analysed (Fig. 2).

6) DATA STANDARDIZATION

In most of the events, the first and second quartiles had a value of 0. Furthermore, EV5 ('Read forum post') presented the largest data deviation in all periods. Since the analysis of these data was carried out using a methodology based on SOM networks that use Euclidean distances, to balance the weight of all the features in the clustering process, the values of the activity vectors were standardized using z-score: $v\_standardized=(v-\mu)/\sigma$, where v is the original value (number of times an event was generated in a period), μ is the mean value of the event-period to which v belongs, and σ is the standard deviation. In this way, the values of the student-subject activity vectors were expressed as the number of deviations from their mean.

7) STUDENT'S FINAL GRADE DATASET

To perform the analysis on the types of activity patterns and their relationship with the student's grade, the final scores of the students in the subjects were retrieved from the platform. In the extraction of the final grades, students whose records did not have consistency to calibrate the final grade were discarded (e.g., students with low activity on the platform and who probably did not take any exams). Of the 13,292 vectors of student-course activity, it was possible to obtain the final grade of 10,473, that is, 78.8% of the original vectors used to train the SOM networks.

B. SOM CLUSTERING

The SOM network is a model based on unsupervised learning that produces a discrete and ordered representation of the dataset. A SOM is made up of a set of neurons, each with a synaptic vector with the same dimension and nature as the training dataset. The synaptic vectors of the neurons of the SOM network are prototype vectors that, once the network has been trained, represent areas of the input space. In this sense, all those points of the input space that are closer to the synaptic vector of a neuron than to the rest constitute what is known as the Voronoi region of the neuron. Thus, each vector of the training dataset is captured by the Voronoi region of one of the neurons in the SOM, which is the one that represents it (usually known as best matching unit, or *bmu*).

Furthermore, the neurons of the SOM map are organized in a two-dimensional (2D) hexagonal or rectangular neighbourhood grid. Based on these characteristics, the SOM algorithm addresses two objectives simultaneously. First, it aims to adjust the synaptic vectors of the neurons so that each one represents groups of similar vectors in the dataset, usually known as vectorial quantization. Second, it aims to achieve that the neighbouring neurons in the 2D map represent nearby areas of the n-dimensional input space and vice versa, so that the nearby input space areas are represented by neighbouring neurons. This is known as topology preservation in the context of the vector projection of the n-dimensional input space to the 2D space defined by the SOM map.

The SOM-based clustering methodology proposed by [38] consists of two phases. The first phase addresses the decision to determine the SOM network size (in terms of number of neurons) that best adapts to the dataset, also taking into account the non-deterministic nature of the training algorithm, using in this process the topology preservation indices of the Kaski–Lagus error function [40] and the topographic function [41]. To that end, several SOM networks with different sizes are trained, all of them with square architecture (i.e., the same number of rows and columns of neurons), where the number of vectors in the dataset is the only factor to establish the range of SOM sizes to train. To address the non-deterministic nature of the training algorithm, 20 SOMs are trained for each size in the range, and the one that produces the lowest Kaski–Lagus error ($\varepsilon_{k\text{-}l}$) is selected. The $\varepsilon_{k\text{-}l}$ function combines the quantization error of each sample of the data set to the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3115024, IEEE Access

closest synaptic vector, with an index that measures the continuity of the mapping of the data set to the SOM grid. Thus, it captures the degree of topology preservation of the trained network. This function is sensitive to the size of the network so it is only suitable to compare the degree of topology preservation of SOMs with the same number of neurons [38]. From the best networks obtained for each size, the one with the lowest value of the topographic function ($\Phi_A^P(0)$) is finally chosen. The topographic function is calculated by adding the average number of leftover neighbourhood connections per neuron, denoted as $\Phi_A^P(-1)$, with the average number of missing neighbourhood connections per neuron, denoted as $\Phi_A^P(1)$. The computation of both values is based on the induced connectivity matrix and the SOM connectivity matrix. The induced connectivity matrix defines how the neurons of the SOM should be connected, based on the inherent knowledge of the input dataset, while the SOM connectivity matrix denotes the existing neighbourhood connections between the neurons of the output layer of the network [41]. The first phase of the SOM-based clustering methodology achieves the objective of vector quantization of the dataset through a SOM that represents a simplified model of the input space, as it is composed of a number of neurons significantly lower than the number of vectors in the dataset.

The second phase of the methodology analyses the potential clusters in the data by clustering the synaptic vectors of the SOM obtained in the first phase, instead of using the dataset itself. For this purpose, a range of the number of clusters to be analysed is established, and the synaptic vectors of the SOM obtained in the first phase are grouped for each value of the range. To determine the SOM clustering that best suits to the potential groups in the dataset, the connectivity indices CONNIndex [42] and Davies–Bouldin Index (DBI) are used [43]. The CONNIndex tends to favour groupings with fewer clusters than DBI [39], so it is appropriate to evaluate the results of both indexes for the dataset used.

C. SOM 3D MAPS

The vector quantization and the topology preservation characteristics of SOM networks cause the *n*-dimensional input space to be projected into an ordered 2D one that can be used as a base plane for producing 2D graphics that display information learned by the SOM and therefore information from the input space itself [44]. The 2D map of the neurons arranged in a grid can also be used to build a three-dimensional (3D) map, assigning to each neuron some numerical value obtained from the dataset but not used in SOM training. This value assigned to each neuron would represent the third dimension of the 3D map. In this study, the final grade obtained by the students in each subject has been used.

The criteria for assigning a value to each neuron can vary, since a neuron can map different vectors in the dataset. In this paper, six 3D maps are proposed: (i) MAX, which assigns the maximum grade of the student-course activity vectors mapped in the neuron; (ii) MIN, which assigns the minimum grade; (iii) MEAN, which assigns the grade average value; (iv) Q1, which assigns the value of the first quartile; (v) Q2, which assigns the value of the median or second quartile; and (vi) Q3, which assigns the value of the third quartile. If a neuron does not map any student-course activity vector, the score associated with the activity vector closest to its synaptic vector is assigned. By generating these six 3D grade maps and delimiting the boundaries between clusters obtained by the SOM network (as described in section II.B), it is possible to visually assess whether any type of activity pattern presents a specific grade distribution, for example, a high pass or failure rate.

III. RESULTS AND DISCUSSION

A. SOM CLUSTERING RESULTS: ANALYSIS OF ACTIVITY PATTERNS

According to the protocol established by [38], to determine the SOM network size that best adapted to the training dataset, it was necessary to analyse the size range from 10 x 10 to 44 x 44. However, given the high number of vectors in the training dataset (13,292), a maximum size of 38 x 38 was established, which represented slightly more than 10% of neurons with respect to the size of the dataset. For each size in the range 10 x 10 to 44 x 44, 20 SOMs were trained; 580 networks were trained in this phase. From each group of 20 SOMs of the same size, the one with the lowest Kaski–Lagus error value ($\varepsilon_{k\text{-}l}$) was selected. Subsequently, among the best SOMs obtained in the previous step, the one with the lowest value of the topographic function ($\Phi_A^P(0) = \Phi_A^P(1) + \Phi_A^P(-1)$) was chosen.

Fig. 3 visualizes the value of the topographic function of these 29 SOM networks, where the SOM with the size 37 x 37 obtained the best value (hereinafter referred to as KOH37x37). Between sizes 37 and 38 the functions $\Phi_A^P(-1)$ and $\Phi_A^P(1)$ are crossed. In this scenario, the topographic function typically increases in value as the number of neurons in the SOM increases because the average number of leftover neighbourhood connections per neuron increases ($\Phi_A^P(-1)$), while the decrease in the average number of missing connections slows down ($\Phi_A^P(1)$). This situation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3115024, IEEE Access

also justifies not analysing the total range of original SOM sizes, initially established up to 44 x 44.

Using the KOH37x37, the range of clusters [2...15] was analysed in the second phase of the methodology. The complete range from 2 to 44 clusters proposed by [38] was not explored, again due to the large number of vectors in the training dataset. The limit of 15 clusters was considered a reasonable number in the search for typologies of activity patterns. As a result of this phase, 14 different groupings of KOH37x37 neurons were obtained (see Fig. 4) and the $CONN_{Index}$ and DBI values were calculated for each of them (Table 2).

$CONN_{Index}$ highlighted 2 clusters as the best result (0.64), followed by 4 clusters (0.48). On the other hand, according to the DBI index, the best clustering was 12 (1.425), followed by 13 (1.426). Given that the objective was to identify the student-course activity patterns with the finest precision, the DBI results were chosen, as they offered better results for a greater number of clusters compared to those obtained by $CONN_{Index}$. The network of 12 clusters (Fig. 4) shows a greater dispersion of the grouped neurons (e.g., the yellow cluster appears in 7 separate groups of neurons) compared to the network of 13 clusters, which achieves a more compact distribution of neurons per cluster. Since both networks offer similar DBI values, the one with 13 clusters was finally selected (hereinafter referred to as KOH37x37_13clusters).

TABLE 2
$CONN_{INDEX}$ AND DBI VALUES (CLUSTERED SOM37X37)

Number of clusters	$CONN_{Index}$	DBI
2	**0.64**	1.55
3	0.43	1.60
4	0.48	1.63
5	0.43	1.57
6	0.47	1.54
7	0.33	1.46
8	0.46	1.52
9	0.24	1.51
10	0.32	1.56
11	0.39	1.72
12	0.36	**1.43**
13	0.31	**1.43**
14	0.35	1.55
15	0.35	1.55

To identify the characteristics of the activity patterns in KOH37x37_13clusters, the average synaptic vector of each cluster was calculated to locate the most significant events. Fig. 5 summarizes the values of the average prototype vectors per cluster, where those features with values close to 1 or higher, that is, those that represent significant deviations from the mean of the event in the period (PER), have been marked with red squares.

Clusters C1 and C2 showed high peaks in the event 'BasicLTI tool released' (EV10) in the 10 periods, taking average values around 3 in the first 4 periods (Fig. 5). The main difference between clusters C1 and C2 was found in the EV10 event values in the last 6 periods, being higher than 6 in cluster C1 and lower than 3 in cluster C2. In addition, while in cluster C2 the rest of the events in all periods presented values close to 0 and even below it, in cluster C1 in periods PER7 and PER8, the events 'submit homework' (EV1) and 'create resource' (EV2) took slightly higher values (0.7 and 1.6 in PER7 and 1.5 and 1.2 in PER8).

Clusters C3 and C4 were characterized by high peaks in the event 'read external page from web content' (EV9) in the 10 periods, although cluster C3 had average values in the range [4–9] in the last 9 periods, while in cluster C2 these values remained below 3. Furthermore, in cluster C4, the rest of the events in all periods presented low values close to 0, while in cluster C3 in periods PER6, PER7, PER9, and PER10, some of the events took values between 1 and 1.5 (see values boxed in red in the tables of clusters C3 and C4 in Fig. 5).

Cluster C5 represented patterns that stood out for high activity in forums in most of the periods: 'create new post in forum' (EV4), 'read forum post' (EV5), and 'reply to forum post' (EV6). In addition, it took values slightly higher than 1 in the event 'read external page from web content' (EV9) in the periods from PER2 to PER9. In the rest of the events, C5 generally presented values close to 0, except in EV2 and EV3 in the PER1, PER5, PER6, and PER9 periods, in which it took values greater than 1.

The behaviour pattern of cluster C6 marked high activity significantly above the average in all periods of the course in events related to resources ('create resource', EV2, and 'read resource', EV3) and forums ('create new post in forum', EV4; 'read forum post', EV5; and 'reply to forum post', EV6), maintaining a generation of the rest of the events similar to the average, except in some periods, in which it produced some values slightly higher than 1.

Cluster C7 was characterized by the activity of reading resources ('read resource', EV3) and pages ('read page', EV7) slightly above the average (between 1 and 2) throughout the entire course, maintaining a generation of the rest of the events similar to the average.

The pattern of the C8 cluster was distinguished by moderate activity (between 1 and 2) in events related to resources ('create resource' EV2, and 'read resource', EV3), and from moderate to high in forums ('create new post in forum', EV4; 'read forum post', EV5; and 'reply to forum post', EV6), especially in the final periods of the course. It also showed moderate activity in the reading of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3115024, IEEE Access

pages ('read page', EV7) and maintained values similar to the mean in the rest of the events.

Clusters C9 and C10 highlighted activity in forums ('create new post in forum', EV4; 'read forum post', EV5; and 'reply to forum post', EV6) above the average, but only in the first third of the course in cluster C9 and only in the second third in cluster C10, maintaining values similar to the average in the rest of the periods.

Cluster C11 identified a normal behaviour pattern for all events throughout the course, except in the first period, in which it stood out for the activities of 'submit homework' (EV1) and 'create resource' (EV2).

Cluster C12 characterized patterns with moderate-high activity in forum events ('create new post in forum', EV4; 'read forum post', EV5; and 'reply to forum post', EV6) and in resource reading ('read resource', EV3) in the third quartile of the course, maintaining values similar to the average in the rest of the periods.

Finally, cluster C13 represented the normal behaviour pattern, with values around 0 in all events and periods of the course.

From a cross-analysis amongst patterns, the results reveal that the majority pattern (C13) reflects neutral student behaviour. Of the 10 activities measured, student type C13 does not stand out; this is in line with C4, with a slight uptick when consulting external websites (EV9); C7, with a little more intensity in reading resources (EV3) and pages (EV7); and for C12 with resource reading (EV3) and the creation, reading, and response to forum posts (EV4, EV5, EV6). We can observe similar behaviours in C2, C9, C10, and C11. Therefore, 8 clusters are defined by the normality and homogeneity of behaviour.

In addition, clusters C1, C3, C5, C6, and C8 stand out for one or two activities: C1 for the use of live online classes (LTI connection, EV10), C3 for reading external resources (EV9), and C5 and C8 for interaction with other users, specifically for creating and responding to forum posts (EV4, EV6). Likewise, C6 focuses on reading resources (EV3) and interaction on the forums (EV4, EV5, and EV6).

B. SOM 3D MAPS (FINAL GRADE)

Before generating the 3D maps, the percentage of student-course activity vectors mapped in each cluster was calculated, both with the dataset of 13,292 vectors used to train the KOH37x37 network and with the dataset of 10,473 vectors labelled with final grade (see Table 3). In general, the distribution of patterns by cluster of the labelled dataset was similar to the distribution obtained with the training dataset, except in cluster C1, which was slightly lower. Having vectors labelled with a final grade distributed by all the clusters ensured the correct generation of the 3D maps.

TABLE 3
PERCENTAGE OF STUDENT-COURSE VECTORS MAPPED IN EACH CLUSTER

Cluster	[a]Unlabelled	[b]Labelled
C1	0.54	0.12
C2	2.54	2.35
C3	0.19	0.24
C4	1.94	2.35
C5	0.13	0.16
C6	0.17	0.21
C7	2.83	3.33
C8	0.50	0.56
C9	2.20	2.60
C10	1.61	1.92
C11	3.25	3.65
C12	1.71	2.09
C13	82.39	80.42

[a]Using the 13,292 vectors from the training dataset (unlabelled); [b]Using the 10,473 vectors from the dataset labelled with final grade;

Using the 10,473 vectors labelled with the final grade of the student-course, the six labelled 3D maps of the KOH37x37_13clusters network were generated, as described in section II.C (Fig. 6). Fig. 7 shows these maps in a top view, where the boundaries of the 13 clusters have been delimited.

To explore the distribution of the final grades shown by the 3D maps in detail, the 10,473 vectors were classified and the percentage of student-course vectors with a final grade in the ranges [0–1) ... [9–10] (Fig. 8) was calculated for each cluster. In addition, the distribution of student-course vectors with failure in the final grade was also obtained for each cluster (Table 4). It should be noted that in the C3 and C8 clusters, all the grades corresponded to 'passed', while the C13 cluster accumulated most of the failures.

Given that the C13 cluster is the one that grouped the largest number of neurons, the distribution of suspended student-grade vectors was analysed in five ranges (Fig. 9).

TABLE 4
PERCENTAGE OF STUDENT-COURSE VECTORS WITH FAILURE IN THE FINAL GRADE DISTRIBUTED IN EACH CLUSTER WITH RESPECT TO THE TOTAL OF FAILURES

Cluster	Student-course vectors (%)
C1	0.32
C2	1.6
C3	0
C4	0.64
C5	0.16
C6	0.16
C7	1.44
C8	0
C9	0.64
C10	0.16
C11	0.96
C12	0.32
C13	93.61

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3115024, IEEE Access

Although the student-course vectors with a failed grade were distributed throughout the entire area of cluster C13, there was a region in the upper central zone in which the highest concentration of them was located, as shown in Fig. 9 in the range [0–5) and in Fig. 6 and Fig. 7 in the valley displayed in all the 3D maps, so a detailed analysis of this area was conducted. Cluster C13 characterized the student-course vectors with normal behaviour with respect to the mean; that is, the value of its 100 features was at a maximum distance of approximately 1 standard deviation with respect to the mean of the feature.

To delimit the special area of cluster C13, the number of features with a positive value in each of the 972 synaptic vectors of the neurons in this cluster was counted, and a variation was obtained from 0 (the 100 components with a negative value) to 83 (only 17 features with a negative value). In particular, 6, 12, and 15 neurons were identified with 0, 1, and 2 features with positive values, respectively. All these neurons were located in the compact region in the upper central area of the C13 cluster. Neurons with synaptic vectors that included only three positive features were found in scattered positions throughout different areas of the C13 cluster, so the study area was limited to the 33 previously identified neurons (Fig. 10).

The average synaptic vector was calculated for the 33 neurons in the study area, as well as for the 939 remaining neurons of the C13 cluster (Fig. 11). The neurons in the study area clearly modelled a below-average normal behaviour pattern, while the rest covered the above-average normal behaviour pattern. Of the 586 student-course vectors with a fail grade mapped in cluster C13, 181 were located in the 33 neurons of the study area; that is, 3.4% of the neurons in cluster C13 mapped 30.89% of the student-course vectors with failed qualification of this cluster.

In cluster C13, 7,837 student-course vectors with passing grades were mapped, of which only 2.78% were mapped in the 33 neurons of the study area. Even though the neurons in the study area presented an activity pattern with values below the average in practically all their features, it must be considered that the platform does not collect all the work a student can do in a subject (such as study hours). In addition, in this study, only 10 of the 28 events generated by the students on the platform were used, so the rest of the events not considered could make the difference in behaviour between the students with pass and fail grades in this special study area.

IV. CONCLUSIONS

Cluster analysis is a useful tool when working with thousands of students in online settings. This need becomes crucial because of the compulsory new context brought by the COVID-19 pandemic that muscles the migration of face-to-face settings into online and blended settings, in no time. The virtualisation of learning and teaching processes from face-to-face settings can apply some of the lessons learnt by online universities, even before the pandemic, where this research is set. This situation makes school teachers and university professors work in a different setting than usual, with other performance indicators and potential students' behaviour. Further, a clear understanding of how the student behaves and how that behaviour is connected to specific actions, means an excellent instrument for improving performance and competence achievement. In this study, there are a number of clear readings on specific clusters: a) Cluster C5 shows the most gainful result, highlighting the benefit of interacting with other users – creating messages in the forum and replying to existing ones (EV4, EV6); b) Cluster C8 shows equal support for interaction, although with a lower median (8.50); c) Cluster C13, the majority, does not stand out for any peak or valley behaviour, showing a normality that causes its users to regularly pass.

At this point, we can reflect on the comparison between behaviour that is meaningful and behaviour that stands out. The significance of cluster C13 implies the regularity of activities and results, without any highlightable peak or valley and therefore without any behaviour that stands out. This pattern fits the entry profiles at Universidad Internacional de La Rioja (UNIR), where the average age is around 40 years, and includes qualified professionals who are employed and seeking to complete unfinished degree programmes or pursuing second or more advanced degrees. This profile is characterized by being focused and responsible with regards to activities and university standards, which includes compliance with deadlines and academic requirements. Without promoting a conspicuous behaviour, it shows the profitability of consistency in mostly self-study processes.

Additionally, these results show a similarity between online and face-to-face learning–teaching environments. Through an initial literature review, introduced in [6-14], we concluded that students that are responsible and show a clear commitment to the academic programme and to the role of learner are the ones who obtain the highest performance, in general. To this extent, we observed that the comparison between the attitudes, results, and impressions were similar in both contexts, without significant differences due to the environment. This conclusion becomes relevant in terms of methodology planning since it provides a timely input about the

relevance of educational methodology over ICT support and technology-based environments.

For future research, one approach to explore is the various clusters in combination with events and periods. That is, for example, in cluster C6, students were doing more than studying and visiting the pages, but they raise the values in periods between 2 and 9 that can be linked to exams. In cluster C12, the rise of EV3, EV4, EV5, and EV6 events occurs in periods 7 and 8. Another example is cluster C1, which includes more active students who view classes more times from period 5. This cross analysis between clusters, events and periods will help understanding the timeline in the student behaviour and how the university can adapt resources and support actions to specific groups.

In addition, a new recording and collection of user-tracking information is in place, specifically focused on the dates where the COVID-19 pandemic has been spread. Further, an analysis with time-of-pandemic data will provide a broader focus and reflection on user behavior presented in this paper.

REFERENCES

[1] R. Huang *et al.*, "Emergence of the Online-Merge-Offline (OMO) Learning Wave in the Post-COVID-19 Era: A Pilot Study," *Sustainability*, vol. 13, no. 6. p. 3512, 2021.

[2] M. Hughes and C. Hagie, "The Positive and Challenging Aspects of Learning Online and in Traditional Face-to-Face Classrooms: A Student Perspective," *J. Spec. Educ. Technol.*, 2005.

[3] D. Burgos, A. Tlili, and A. Tabacco, "Education in a Crisis Context: Summary, Insights and Future," in *Radical Solutions for Education in a Crisis Context*, A. Burgos, D.; Tlili, A.; Tabacco, Ed. 2021, pp. 3–10.

[4] F. Simanca, R. González, L. Rodriguez, and D. Burgos, "Personalized tutoring model through the application of Learning Analytics phases," *IEEE Lat. Am. Trans.*, vol. 18, no. 1, pp. 7–15, 2020.

[5] F. Simanca, R. Gonzalez Crespo, L. Rodríguez-Baena, and D. Burgos, "Identifying Students at Risk of Failing a Subject by Using Learning Analytics for Subsequent Customised Tutoring," *Appl. Sci.*, vol. 9, no. 3, p. 448, 2019.

[6] A. Horspool and C. Lange, "Applying the scholarship of teaching and learning: student perceptions, behaviours and success online and face-to-face," *Assess. Eval. High. Educ.*, vol. 37, no. 1, pp. 73–88, Feb. 2012.

[7] D. Burgos, "A Predictive System Informed by Students' Similar Behaviour," *Sustainability*, vol. 12, no. 2, p. 706, 2020.

[8] P. Euzent, T. Martin, P. Moskal, and P. Moskal, "Assessing Student Performance and Perceptions in Lecture Capture vs. Face-to-Face Course Delivery," *J. Inf. Technol. Educ.*, vol. 10, no. 1, pp. 295–307, Jan. 2011.

[9] D. K. Larson and C. H. Sung, "Comparing student performance: Online versus blended versus face-to-face," *J. Asynchronous Learn. Netw.*, vol. 13, no. 1, pp. 31–42, 2009.

[10] A. Lopez, S. A. Gómez, D. Martín, and D. Burgos, "A Framework for a Semiautomatic Competence Valuation," in *Radical Solutions and eLearning: Practical Innovations and Online Educational Technology*, D. Burgos, Ed. Singapore: Springer Singapore, 2020, pp. 215–236.

[11] D. Xu and S. S. Jaggars, "The impact of online learning on students' course outcomes: Evidence from a large community and technical college system," *Econ. Educ. Rev.*, vol. 37, pp. 46–57, 2013.

[12] A. Driscoll, K. Jicha, A. N. Hunt, L. Tichavsky, and G. Thompson, "Can Online Courses Deliver In-class Results?: A Comparison of Student Performance and Satisfaction in an Online versus a Face-to-face Introductory Sociology Course," *Teach. Sociol.*, vol. 40, no. 4, pp. 312–331, 2012.

[13] N. Kemp and R. Grieve, "Face-to-face or face-to-screen? Undergraduates' opinions and test performance in classroom vs. online learning," *Front. Psychol.*, vol. 5, p. 1278, Nov. 2014.

[14] M. Utari, B. Warsito, and R. Kusumaningrum, "Implementation of Data Mining for Drop-Out Prediction using Random Forest Method," in *2020 8th International Conference on Information and Communication Technology (ICoICT)*, 2020, pp. 1–5.

[15] S. Z. Selim and M. A. Ismail, "K-means-type algorithms: a generalized convergence theorem and characterization of local optimality.," *IEEE Trans Pattern Anal Mach Intell.*, vol. 6, no. 1, pp. 81–7, 1984.

[16] F. Nielsen, "Hierarchical Clustering," in *Introduction to HPC with MPI for Data Science*, Springer, 2016, pp. 195–211.

[17] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 1, no. 3, pp. 231–240, May 2011.

[18] G. Xu, Y. Zong, and Z. Yang, *Applied data mining*. CRC Press, 2013.

[19] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long, "A Survey of Clustering with Deep Learning: From the Perspective of Network Architecture," *IEEE Access*, vol. 6, pp. 39501–39514, Jul. 2018.

[20] W. Wang, Y. Huang, Y. Wang, and L. Wang, "Generalized autoencoder: A neural network

framework for dimensionality reduction," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 496–503, Sep. 2014.

[21] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative Adversarial Networks: An Overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.

[22] K. L. Lim, X. Jiang, and C. Yi, "Deep Clustering with Variational Autoencoder," *IEEE Signal Process. Lett.*, vol. 27, pp. 231–235, 2020.

[23] T. Kohonen, "Essentials of the self-organizing map," *Neural Networks*, vol. 37, pp. 52–65, Jan. 2013.

[24] T. Kohonen, *Self-Organizing Maps*, 3rd ed. Springer-Verlag Berlin Heidelberg, 2001.

[25] B. Fritzke, "A Growing Neural Gas Learns Topologies," *Adv. Neural Inf. Process. Syst.*, vol. 7, pp. 625–632, 1995.

[26] B. Fritzke, "Growing cell structures—A self-organizing network for unsupervised and supervised learning," *Neural Networks*, vol. 7, no. 9, pp. 1441–1460, Jan. 1994.

[27] E. López-Rubio, E. J. Palomo, and E. Domínguez, "Bregman divergences for growing hierarchical self-organizing networks," *Int. J. Neural Syst.*, vol. 24, no. 04, p. 1450016, 2014.

[28] T. Villmann and H. U. Bauer, "Applications of the growing self-organizing map," *Neurocomputing*, vol. 21, no. 1–3, pp. 91–100, Nov. 1998.

[29] J. Vesanto and M. Sulkava, "Distance Matrix Based Clustering of the Self-Organizing Map," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2415 LNCS, pp. 951–956, 2002.

[30] D. Brugger, M. Bogdan, and W. Rosenstiel, "Automatic cluster detection in Kohonen's SOM," *IEEE Trans. Neural Networks*, vol. 19, no. 3, pp. 442–459, 2008.

[31] A. M. Newman and J. B. Cooper, "AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number.," *BMC Bioinformatics*, vol. 11, no. 1, p. 117, 2010.

[32] G. Cabanes and Y. Bennani, "Learning the Number of Clusters in Self Organizing Map," in *Self-Organizing Maps*, IntechOpen, 2010.

[33] A. Dutt, S. Aghabozrgi, M. A. B. Ismail, and H. Mahroeian, "Clustering algorithms applied in educational data mining," *Int. J. Inf. Electron. Eng.*, vol. 5, no. 2, pp. 280–291, 2015.

[34] N. B. Ahmad, U. F. Alias, N. Mohamad, and N. Yusof, "Principal Component Analysis and Self-Organizing Map Clustering for Student Browsing Behaviour Analysis," *Procedia Comput. Sci.*, vol. 163, pp. 550–559, Jan. 2019.

[35] M. Cantabella, R. Martínez-España, B. Ayuso, J. A. Yáñez, and A. Muñoz, "Analysis of student behavior in learning management systems through a Big Data framework," *Futur. Gener. Comput. Syst.*, vol. 90, pp. 262–272, Jan. 2019.

[36] L. Youngjin, "Using Self-Organizing Map and Clustering to Investigate Problem-Solving Patterns in the Massive Open Online Course: An Exploratory Study," *J. Educ. Comput. Res.*, vol. 57, no. 2, pp. 471–490, 2018.

[37] R. Cerezo, M. Sánchez-Santillán, M. P. Paule-Ruiz, and J. C. Núñez, "Students' LMS interaction patterns and their relationship with achievement: A case study in higher education," *Comput. Educ.*, vol. 96, pp. 42–54, May 2016.

[38] S. Delgado, C. Higuera, J. Calle-Espinosa, F. Morán, and F. Montero, "A SOM prototype-based cluster analysis methodology," *Expert Syst. Appl.*, vol. 88, 2017.

[39] D. Guamán, S. Delgado, and J. Perez, "Classifying Model-View-Controller Software Applications Using Self-Organizing Maps," *IEEE Access*, vol. 9, pp. 45201–45229, 2021.

[40] S. Kaski and K. Lagus, "Comparing self-organizing maps," in *Proceedings of the International Conference on Artificial Neural Networks*, 1996, pp. 809–814.

[41] T. Villmann, R. Der, M. Herrmann, and T. M. Martinetz, "Topology preservation in self-organizing feature maps: Exact definition and measurement," *IEEE Trans. Neural Networks*, vol. 8, no. 2, pp. 256–266, 1997.

[42] K. Tasdemir and E. Merényi, "A validity index for prototype-based clustering of data sets with complex cluster structures," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 41, no. 4, pp. 1039–1053, 2011.

[43] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 2, pp. 224–227, 1979.

[44] A. Ultsch, "Maps for the visualization of high-dimensional Data Spaces," in *Proceedings of the workshop on Self-Organizing Maps*, 2003, pp. 225–230.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3115024, IEEE Access

Figure captions

FIGURE 1. **Overview of the e-learning system and the proposed analysis.**

FIGURE 2. **Boxplot with the distribution of the values of the 100 features of the student-course vectors. Each period (PER) groups the 10 events in Table 1 (from left to right, in the order indicated in the table). The line that separates the light red from the dark red boxes represents the median. The light red box displays the Q2 percentile and the dark red the Q3 percentile. The vertical line arising from the Q3 percentile represents the range of values from the Q3 percentile to the 90th decile. The grey diamond displays the mean value.**

FIGURE 3. **Value of the topographic function ($\Phi_A^P(0)$) for SOMs of sizes 10 x 10...38 x 38.**

FIGURE 4. **Maps of the KOH37x37 network with neurons grouped from 2 to 15 clusters. Each map visualizes the neurons of each cluster in the same colour.**

FIGURE 5. **Average synaptic vectors per cluster of the KOH37x37_13clusters network. Each graph includes the image and the values (lower table) of the 10 events in each of the 10 periods (PER) of the average synaptic vector of the cluster. Values close to or greater than 1 have been marked in red boxes.**

FIGURE 6. **Final grade 3D maps (isometric view).**

FIGURE 7. **Final grade 3D maps (top view) including the boundaries of each of the 13 clusters of the KOH37x37_13clusters network.**

FIGURE 8. **Histograms with the percentage of student-course vectors with final grade by ranges (x-axis) in clusters C1 to C13, and with the distribution of grades without considering clusters (Total).**

FIGURE 9. **Histograms of the C13 cluster of the KOH37x37_13clusters network (each neuron, or circle, is coloured based on the number of patterns for which it is *bmu*. Neurons that appear in white identify 0 patterns. Each histogram is colour coded with a scale included at the right of each graph; note that the scale is not identical in the different maps). First row: Histograms of cluster C13 according to the final grade obtained by the student-course in the ranges [0–5) and [5–10]. Remaining rows: Histograms of cluster C13 for the final qualification ranges [0–1), [1–2), [2–3), [3–4), and [4–5). The number of student-subject vectors located in each histogram is included at the bottom of each graph.**

FIGURE 10. **The blue line in the upper central area of the map on the right (the histogram of patterns classified in cluster C13 whose final grade is less than 5) delimits the special region of cluster C13 where there is a higher concentration of failed grades. On the left is an enlargement of this area, where the blue neurons are those with synaptic vectors with 0 features with positive values, the green ones are those with 1 out of 100, and the orange ones are those with 2 out of 100.**

FIGURE 11. **Comparison of the average synaptic vectors of the 33 neurons in the study area (dark red) versus the average synaptic vector of the rest of the neurons of the C13 cluster (light red). The upper and lower bars displayed in the components of the average synaptic vector of the study area represent the standard deviation.**