# Analyzing MathE Platform through Clustering Algorithms*

Beatriz Flamia Azevedo[1,2][0000−0002−8527−7409], Yahia Amoura[1][0000−0002−8811−0823], Ana Maria A. C. Rocha[2][0000−0001−8679−2886], Florbela P. Fernandes[1][0000−0001−9542−4460], Maria F. Pacheco[1,3][0000−0001−7915−0391], and Ana I. Pereira[1,2][0000−0003−3803−2043]

[1] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, Bragança - 5300-253, Portugal
[2] ALGORITMI Center, University of Minho, Campus de Azurém, 4800-058 Guimarães, Portugal
[3] Center for Research & Development in Mathematics and Applications CIDMA, University of Aveiro, Aveiro, Portugal
{beatrizflamia, yahia, fflor, pacheco, apereira}@ipb.pt, arocha@dps.uminho.pt

**Abstract.** University lecturers have been encouraged to adopt innovative methodologies and teaching tools in order to implement an interactive and appealing educational environment. The MathE platform was created with the main goal of providing students and teachers with a new perspective on mathematical teaching and learning in a dynamic and appealing way, relying on digital interactive technologies that enable customized study. The MathE platform has been online since 2019, having since been used by many students and professors around the world. However, the necessity for some improvements on the platform has been identified, in order to make it more interactive and able to meet the needs of students in a customized way. Based on previous studies, it is known that one of the urgent needs is the reorganization of the available resources into more than two levels (basic and advanced), as it currently is. Thus, this paper investigates, through the application of two clustering methodologies, the optimal number of levels of difficulty to reorganize the resources in the MathE platform. Hierarchical Clustering and three Bio-inspired Automatic Clustering Algorithms were applied to the database, which is composed of questions answered by the students on the platform. The results of both methodologies point out six as the optimal number of levels of difficulty to group the resources offered by the platform.

**Keywords:** E-learning · Data analysis · Educational technology · Machine learning · Clustering

* This work has been supported by FCT Fundação para a Ciência e Tecnologia within the R&D Units Project Scope UIDB/00319/2020, UIDB/05757/2020 and Erasmus Plus KA2 within the project 2021-1-PT01-KA220-HED-000023288. Beatriz Flamia Azevedo is supported by FCT Grant Reference SFRH/BD/07427/2021.

## 1 Introduction

The development of Information and Communication Technologies (ICT) is reflected in the dynamics of the educational fields. This advancement has facilitated and made everyday tasks more accessible, effective, and faster to perform [20]. The ICT are directly involved in the development of teaching and learning processes by supporting innovative pedagogical actions and providing new learning spaces. In this way, it is possible to transform the classical classroom into a virtual one by eliminating the existing space-time barriers [7]. Among the ICT-based teaching methods is e-learning; e-learning platforms such as MathE are intensively involved in higher education teaching and learning practices through the support of online classes. They offer many advantages in terms of communication, students interaction, group development, and greater access to knowledge [6], as well as providing students access to a wide spectrum of information in a multitude of ways. The results of the use of an e-learning platform are reflected in the improvement of students' skills, self-motivation, engagement, and attitude towards educational content. This teaching approach is currently proliferating since the COVID-19 pandemic situation; its independence regarding location, time, effort, and cost makes it the most suitable option for student learning and assessment [16]. This type of pedagogy has particularities that distinguish it from other teaching modalities. Some researchers recognize it as a progression of distance education [5, 25]. According to others, it represents a novelty that differs significantly from face-to-face teaching [18].

Promoting an e-learning method requires different types of resources, in particular digital and technological resources. Among the available digital tools are videos, teaching platforms, video conferences, podcasts, social networks, as well as many other resources [26]. The technological resources represent hardware tools including the desktop computer, tablet, smartphone, among others [17].

E-learning offers a range of particularities such as stimulating the development of dialogue and group work [4], strengthening interprofessional relationships among learners [10], promoting collaboration between the participants themselves, allowing the achievement of joint goals in the development of different tasks [13], making synchronous and asynchronous communication easier [22], and allowing learning from anywhere where there is an available internet connection [24]. E-learning also encourages the acquisition of digital competences by the students [11], allowing the adjustment to their personal rhythm [14], increasing their interest and motivation towards learning, as they are able to adapt to their specific learning style [1], giving everyone an unlimited number of learning resources [21]; E-learning also facilitates the monitoring of student activity by the teacher [2].

A solid education in Mathematics, in particular, has great importance both in the areas of exact sciences as well as human and biological sciences. However, Mathematics is frequently the subject of disorientation and complaints from students at all levels of their educational journey. One way to change the students' pragmatic view of mathematics is through interactive learning platforms. Under this scenario, the MathE platform emerges as a digital, innovative, dynamic, and

intelligent tool for teaching and learning mathematics. The MathE platform will be better described in Sect. 2.

This paper aims to analyze some of the data collected by the MathE platform over the 3 years the platform has been online. In particular, with this research it is expected to reach conclusions about the best way to reorganize the resources available on the platform into different levels of difficulty. For this, unsupervised learning techniques, namely clustering, will be used for data analysis.

The rest of the paper is divided as follows. In Sect. 2, the concepts of the MathE collaborative learning platform are described. Section 3 explains the methodology applied in this work, which are Hierarchical and Partitioning Clustering techniques. The dataset used is described in Sect. 4 and the obtained results and their discussion are presented in Sec. 5. Finally, Sect. 6 concludes the work and sets forward-looking guidelines for the future of the platform.

## 2 The MathE Platform

MathE is a collaborative e-learning platform that aims to provide users with greater mathematical skills in higher education by creating a virtual space for learning and exchange. Like other e-learning platforms in mathematics, this platform represents a remarkable transition from the classical Learning Management Systems (LMS) to an interactive Intelligent Tutoring System (ITS). MathE is distinguished by its dynamic teaching environment involving both teachers and students or also external contributors and learners in the field of Mathematics. Furthermore, MathE is a non-commercial tool, being completely free and available 24 hours a day, for all individuals interested in improving their knowledge and understanding of Mathematics.

MathE relies on an essential set of resources presented in the form of lessons, exercises, quizzes, videos, and other materials. These resources encourage students to study and practice mathematics outside the classic classroom rhythm, without the need for a teacher to be present.

Currently, there are 99 teachers and 1161 students from different nationalities enrolled on the platform: Portuguese, Brazilian, Turkish, Tunisian, Greek, German, Kazakh, Italian, Russian, Lithuanian, Irish, Spanish, Dutch and Romanian. In its current stage, the platform is organized into three main sections: **Student's Assessment** (composed of multiple-choice questions divided into topics, with two difficulty levels (basic and advanced), which were previously defined by a professor member of the platform); **MathE Library** (composed of valuable and diversified materials related to the topics and subtopics covered by the platform, such as videos, lessons, exercises, training tests and other formats); and **Community of Practice** (provides a virtual place where teachers and students have the opportunity to interact in order to fulfill their common goals, thus consolidating a strong network community). More details about each section are described in [2, 3], and can also be found in the Platform Website (mathe.pixel-online.org).

MathE includes fifteen topics in Mathematics, among the ones that are in the classic core of graduate courses: Analytic Geometry, Complex Numbers, Set Theory, Differential Equations, Differentiation (including 3 subtopics: Derivatives, Partial Differentiation, Implicit Differentiation and Chain Rule), Fundamental Mathematics (2 subtopics: Elementary Geometry and Expressions and Equations), Graph Theory, Integration (3 subtopics: Integration Techniques, Double Integration and Definite Integrals), Linear Algebra (5 subtopics: Matrices and Determinants, Eigenvalues and Eigenvectors, Linear Systems, Vector Spaces Linear Transformations, Others), Optimization (2 subtopics: Linear Optimization and Nonlinear Optimization), Probability, Real Functions of a Single Variable - RFSV (2 subtopics: Limits and Continuity and Domain, Image and Graphics), Real Functions of Several Variables - RFSV (1 subtopic: Limits, Continuity, Domain and Image) and Statistics, as presented in Fig. 1. However, it is essential to mention that the platform's content is constantly being updated, and other topics and subtopics may be created whenever necessary.

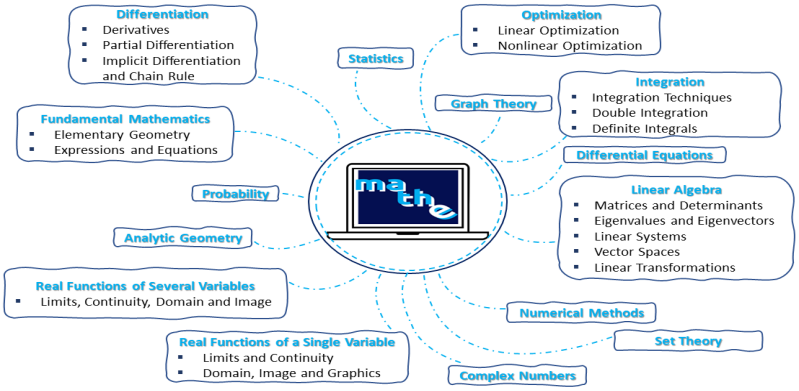



Fig. 1: Topics and subtopics currently available on the MathE Platform.

This paper is focused on the Student's Assessment section of MathE Platform. In this section of the platform, the students can train and test their skills in the *Self Need Assessment* (SNA) and *Final Assessment* (SFA) subsections, respectively. The Self Need Assessment section aims to provide the students with some training assessment to test if a particular topic that he/she enrolled in is already known and understood: suppose that the students or the teachers believe that their understanding needs to be deepened. In this case, the student can choose to answer to another training assessment to measure his/her level of confidence to perform a final assessment. Each training assessment will be randomly generated from the assessments database composed of questions/answers. In this way, the same student will be able to answer different training assess-

ments on the same topic. When answering a training assessment, the students will have immediate access to the obtained mark: the test will randomly select seven questions from a given set, and after the student submits the test, the mark will appear automatically, allowing self-assessment. On the other hand, the purpose of the Final Assessment section is to evaluate the student performance after practicing with training assessment questions (and all the related resouces available in MathE platform). In the Final Assessment section, the teacher can select the questions and the assessment will be available for the students at a chosen moment, defined by the teacher. In this case, the student will submit the test and receive feedback on the following day; the teacher will have access to the results at the end of the test, one day before the students [3].

As already mentioned, the questions available on the MathE platform are divided into two levels of difficulty (basic and advanced). The classification into basic or advanced is done by a professor registered on the platform. However, previous studies [3] concluded that two levels are insufficient for separating the available content. Therefore, to further meet the needs of users of the MathE platform in a more suitable and personalized way, it is undergoing profound changes that will make it even more interactive and customized. For this, a digital intelligence system is being developed and, in the near future, the questions will be addressed to students in a personalized way and not randomly, as is currently the case. One of the first needs that were identified is to reorganize the available resources. Thus, this work intends to investigate an optimal number of levels of difficulty to reorganize the questions available in MathE. For this, the data from the questions belonging to SNA, answered in the last 3 years in which the platform was online, were analyzed by hierarchical clustering and automatic clustering techniques to define the number of levels of difficulty that are defined by the algorithms as the optimal number of clusters.

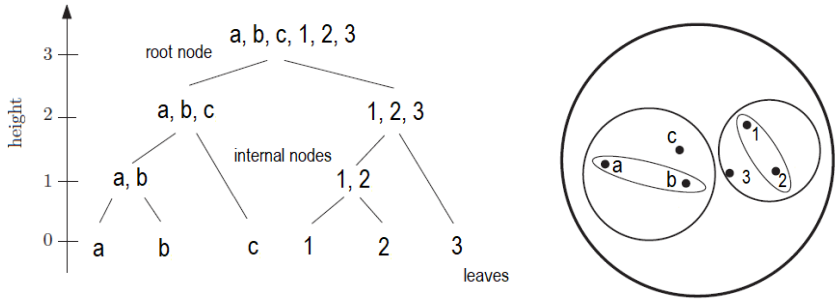
## 3 Methodology

Clustering is one of the most widely used methods for unsupervised learning. It is used in datasets where there is no defined association between input and output. Thus, clustering algorithms consist of performing the task of grouping a set of elements with similarities in the same group and those with dissimilarities in other groups [27]. The methodology used in this work refers to two types of clustering: Hierarchical Clustering and Partitioning Clustering.

### 3.1 Hierarchical Clustering

Hierarchical clustering is an unsupervised technique for performing exploratory data analysis. This technique consists of building a binary merge tree, starting from the data elements stored in the leaves and proceeding by merging the closest "sub-sets" two by two until reaching the root of the tree, which contains all the elements of a dataset, denoted as $X$ [23]. The graphical representation of this binary merge tree is called dendrogram. Basically, a dendrogram consists of many

$U$-shaped lines that connect data points in a hierarchical tree. The height of each $U$ represents the distance between the two data points being connected. To draw a dendrogram, we can draw an internal node $s(X')$ composed by the subset $X' \subseteq X$ at height $h(X') = |X'|$. Thereafter, the edges between this node $s(X')$ and its two sibling nodes $s(X_1)$ and $s(X_2)$ with $X' = X_1 \cup X_2$ (and $X_1 \cap X_2 \neq \oslash$) are drawn. Considering this, each defined subset of $X$ can be interpreted as a cluster. Fig. 2a illustrates a generic representation of a dendrogram, whereas Fig. 2b is the equivalent Venn diagram. Note that the set $X \equiv \{a, b, c, 1, 2, 3\}$ is the root note, and the subset $\{a, b, c\}, \{1, 2, 3\}, \{a, b\}, \{a, 2\}$ are the internal nodes. At the end, are the leaves, in this case, represented by the subsets $\{a\}, \{b\}, \{c\}, \{1\}, \{2\}, \{3\}$.




(a) Dendrogram visualization (b) Venn diagram visualization

Fig. 2: Dendrogram and Venn diagram representation.

In this approach, the Matlab® `dendrogram` function [19] was used to generate the Hierarchical cluster and, consequently, the dendrogram representation. In this particular function, two important informations must be considered:

- If there are 30 or fewer data points in the original dataset, then each leaf in the dendrogram corresponds to one data point.
- If there are more than 30 data points, then the dendrogram collapses lower branches so that there are 30 leaf nodes. As a result, some leaves in the plot correspond to more than one data point.

### 3.2 Partitioning Clustering

Partitioning clustering decomposes a dataset into a set of disjoint clusters. Considering a dataset of $X_m$ points, a partitioning method constructs $K$ ($X_m \geq K$) partitions of the data, with each partition representing a cluster $C$. That is, it classifies the data into $K$ groups by satisfying the following requirements: (1) each group contains at least one point, and (2) each point belongs to exactly one group [12].

In real-world data clustering analysis problems, identifying the number of clusters and, consequently, the appropriate partitioning of the dataset is quite a difficult task. An unappropriated selection of the number of clusters results in poor performance since, in traditional clustering algorithms, the results often depend on the initial starting points [9]. In this context, automatic data clustering techniques that combine clustering and optimization techniques have helped to overcome these challenges and have also offered several improvements in the clustering methods. The automatic clustering process consists of solving an optimization problem, aiming to minimize the similarity within a cluster and maximize the dissimilarity between the clusters.

In this work, the Davies–Bouldin index (DB) [8] will be used as a clustering similarity and dissimilarity measure that will define the number of cluster centroids, which is the number of groups into which the dataset will be divided. DB index is based on a ratio of intra-cluster and inter-cluster distances. It is used to validate cluster quality and also to determine the optimal number of clusters. Consider that cluster $C$ has members $X_1, X_2, ..., X_m$. The goal is to define a general cluster separation measure, $S_i$ and $M_{ij}$, which allows computing the average similarity of each cluster to its most similar cluster. The lower the average similarity, the better the clusters are separated and the better the clustering results. To better explain how to get the Davies-Bouldin index, four steps are considered [8].

In the first step, it is necessary to evaluate the average distance between each observation within the cluster and its centroid, that is the dispersion parameter $S_i$, also known as intra-cluster distance, given by Equation (1),

$$S_i = \left\{ \frac{1}{T_i} \sum_{j=1}^{T_i} |X_j - A_i|^q \right\}^{\frac{1}{q}} \tag{1}$$

where, for a particular cluster $i$, $T_i$ is the number of vectors (observations), $A_i$ is its centroid and $X_j$ is the $j$th (observation) vector.

The second step aims to evaluate the distance between the centroids $A_i$ and $A_j$, given by Equation (2), also known as inter-cluster distance. In this case, $a_{ki}$ is the $k$th component of the $n$-dimensional vector $a_i$, which is the centroid of cluster $i$, and $N$ is the total number of clusters. It is worth mentioning that $M_{ij}$ is the Minkowski metric of the centroids which characterize clusters $i$ and $j$ and $p = 2$ means the Euclidean distance.

$$M_{ij} = \left\{ \sum_{k=1}^{N} |a_{ki} - a_{kj}|^p \} \right\}^{\frac{1}{p}} = ||A_i - A_j||_p \tag{2}$$

In the third step, the similarity between clusters, $R_{ij}$, is computed as the sum of two intra-cluster dispersions divided by the separation measure, given by Equation (3), that is the within-to-between cluster distance ratio for the $i$th and $j$th clusters.

$$R_{ij} = \frac{S_i + S_j}{M_{ij}} \text{ for } i, j = 1, ..., N \tag{3}$$

Finally, the last step calculates the DB index, that is, the average of the similarity measure of each cluster with the cluster most similar to it (see Equation (4)). $R_i$ is the maximum of $R_{ij}$ $i \neq j$, so, the maximum value of $R_{ij}$ represents the worst-case within-to-between cluster ratio for cluster $i$. Thus, the optimal clustering solution has the smallest Davies-Bouldin index value.

$$DB = \frac{1}{N} \sum_{i=1}^{N} R_i. \tag{4}$$

Considering the definition of the DB index, a minimization problem can be defined, whose objective function is the DB index value. Thus, metaheuristics can be used in order to solve this problem as an evolutionary bio-inspired algorithm.

Therefore, in order to compare the results obtained through different approaches, three bio-inspired evolutionary algorithms are used in this work: Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Differential Evolution (DE). Thus, the main difference between the so-called automatic algorithms that will be used in this paper is the optimization process to define the DB index, since each one of them employs a different bio-inspired optimization algorithm: GA, PSO or DE. More information about these algorithms can be found at [15, 28, 29].

## 4 Dataset

In this paper, all the questions answered on the section SNA of the MathE platform were analyzed to identify patterns based on the type of student answers to each question, whether correct or incorrect. Thus, the data collected considers information of 6942 answers distributed among 766 questions of 15 topics. These answers were provided by 285 students of different nationalities, over 3 years, in which the platform is online. It is important to highlight that the questions and the topics are constantly being added to the platform, so naturally some topics have more questions answered than others. Table 1 describes the dataset.

The *Topic* column describes all the MathE topics available on the platform. Next, the *Question Available* column describes the number of questions available on the platform for each topic. Thereafter the *Question Answered* column gives the number of different questions answered in each topic. The last three columns refer to the type of answers provided by the students: the first column shows the number of correct answers per topic, followed by the column of the incorrect ones, and the last column corresponds to the sum of the correct and incorrect answers, which is equal to the total number of answers per topic.

To investigate the optimal number of levels of difficulty into which the questions will be divided, the probability of correct answers for each of the 766 questions was evaluated. That is, for each question the number of correct answers

Table 1: MathE dataset

| **Topic** | **Questions Available** | **Questions Answered** | **Correct Answers** | **Incorrect Answers** | **Total Answers** |
|---|---|---|---|---|---|
| Linear Algebra | 211 | 199 | 1741 | 1955 | **3696** |
| Fund. Math | 91 | 84 | 365 | 396 | **761** |
| Graph Theory | 49 | 34 | 29 | 19 | **48** |
| Differentiation | 144 | 96 | 193 | 397 | **590** |
| Integration | 127 | 54 | 67 | 94 | **161** |
| Analytic Geometry | 40 | 40 | 156 | 183 | **339** |
| Complex Numbers | 41 | 37 | 231 | 277 | **508** |
| Dif. Equation | 41 | 30 | 56 | 40 | **96** |
| Statistic | 41 | 26 | 155 | 175 | **330** |
| R. F. Single Variable | 52 | 25 | 28 | 46 | **74** |
| Probability | 46 | 34 | 32 | 54 | **86** |
| Optimization | 96 | 25 | 11 | 26 | **37** |
| R. F. Several Variable | 58 | 15 | 5 | 13 | **18** |
| Set Theory | 40 | 26 | 26 | 16 | **42** |
| Numerical Methods | 42 | 41 | 73 | 83 | **156** |
| **Total** | **1119** | **766** | **3168** | **3774** | **6942** |

divided by the total number that this question was answered. This information was then considered as the input variable of the clustering algorithm, in a first scenario.

However, it is known that the information contained in the probability variable of a question that was, for example, answered 20 times, is different from a question that was answered only 2 times.

Thus, to achieve greater reliability in the results, it was decided to divide the complete dataset into smaller sets, according to the number of questions answered, resulting in 4 datasets, as described below:

- **dataset** 1: questions answered 1 time, at least.
- **dataset** 2: questions answered 5 times, at least.
- **dataset** 3: questions answered 10 times, at least.
- **dataset** 4: questions answered 15 times, at least.

Note that datasets 2, 3, and 4 are subsets of dataset 1. The dimension of each dataset, according to the topics, is presented in Table 2.

## 5 Results and Discussion

This section presents a general analysis of the data described in Sect. 4. Thereafter, the results from Hierarchical and Partitioning Clustering Algorithms are presented and discussed.

Table 2: Number of different questions per dataset

| Topic | Dataset 1 | Dataset 2 | Dataset 3 | Dataset 4 |
|---|---|---|---|---|
| Linear Algebra | 199 | 154 | 125 | 70 |
| Fund. Math | 84 | 63 | 29 | 8 |
| Graph Theory | 34 | 0 | 0 | 0 |
| Differentiation | 96 | 61 | 20 | 0 |
| Integration | 54 | 12 | 1 | 0 |
| Analytic Geometry | 40 | 38 | 14 | 0 |
| Complex Numbers | 37 | 33 | 30 | 4 |
| Dif. Equation | 30 | 6 | 0 | 0 |
| Statistic | 26 | 25 | 22 | 1 |
| R.F.Single Variable | 25 | 8 | 0 | 0 |
| Probability | 34 | 3 | 0 | 0 |
| Opmization | 25 | 0 | 0 | 0 |
| R.F.Several Variable | 15 | 0 | 0 | 0 |
| Set Theory | 26 | 0 | 0 | 0 |
| Num. Methods | 41 | 14 | 1 | 0 |
| **Total** | **766** | **417** | **242** | **83** |

### 5.1 General Analysis of the Data

From Table 1, more specifically in the *Total Answers* column, it is possible to observe that the Linear Algebra topic is the most used topic in the platform. Approximately 53% of the total questions answered correspond to this topic. However, this fact is not surprising, considering that Linear Algebra is a subject present in almost all curricula of higher education courses that include mathematics. After Linear Algebra, the most requested topic is Fundamentals of Mathematics, which corresponds to 10% of the total answered questions. Fundamentals of Mathematics includes questions about the essential background for higher education and, in turn, has also substantial demand on the platform. The other topics have a lower rate of use, however, from the data, it is possible to conclude that all topics are consistently being exploited.

Figure 3 compares the performance of the students by topics through the percentage of correct answers and incorrect ones, constituted by the data from columns *Correct Answers*, *Incorrect Answers* and *Total Answers* of Table 1. Although these values are complementary, presenting them in confrontation allows a better evaluation of the results.

Thus, from Fig. 3, in practically all topics, represented by 1 to 15, at least 30% of the questions were answered correctly. The topics Set Theory (14), Graph Theory (3), and Differential Equation (8) had the highest percentage of correct answers, 62%, 60%, and 58% respectively. It is important to highlight in these three topics the highest rate of 50%, which means that in this topic the rate of correct answers is higher than the rate of incorrect ones. On the other hand, Optimization (12) and Real Function of Several Variable (13) had the lowest rate

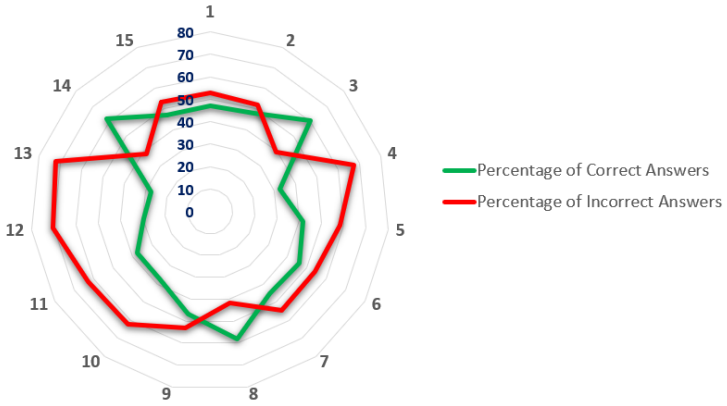


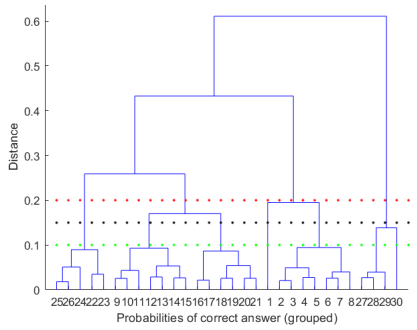

Fig. 3: Percentage of correct and incorrect answered questions, by topics.

of correct answers, it is 30% and 28%, respectively. Moreover, it is important to mention that these two topics had the fewest questions answered, as can be seen in the *Total Answers* column of Table 1.

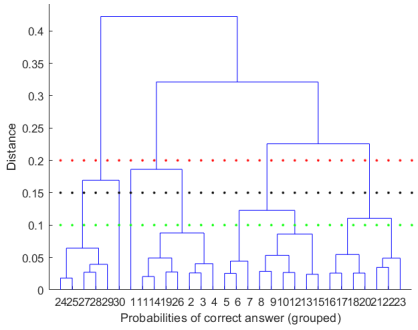
In order to determine the optimal number of partitions of the dataset, two different clustering methodologies were performed: Hierarchical Clustering and Partitioning Clustering. The results are presented bellow, and they were obtained using an Intel(R) i5(R) CPU @1.60 GHz with 8 GB of RAM using Matlab 2019a ® software [19].

### 5.2 Hierarchical Clustering Results

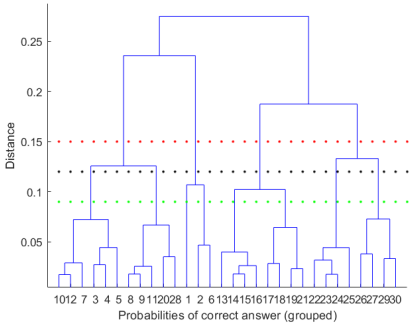
The data from all datasets was evaluated by Hierarchical Clustering techniques, as presented in the Sect. 3. Figure 4 presents the results in the form of Dendrogram for the datasets 1–4. Figure 4a shows the Dendrogram generated by all questions answered (dataset 1); the results of the dataset that includes only the questions that were answered at least 5 times (dataset 2) are depicted in Fig. 4b; Figure 4c refers to questions answered at least 10 times (dataset 3); and the results of questions answered at least 20 times (dataset 4) are presented in Fig. 4d.



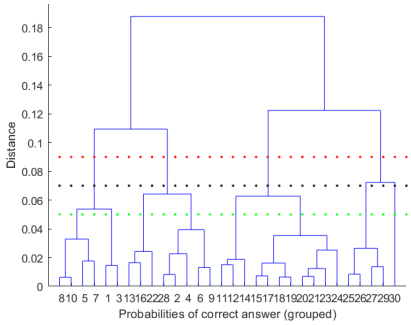

(a) All questions answered.




(b) Questions answered at least 5 times.




(c) Questions answered at least 10 times.




(d) Questions answered at least 20 times.

Fig. 4: Dendrogram results.

When comparing the four dendrograms, it is noticed that the distances between the clusters become smaller and smaller as the datasets become more restrictive. That is, dataset 1 is less homogeneous (similar) than dataset 2 and so on. This makes perfect sense, given the divisors applied to generate each of the datasets used.

In each dendrogram it can be seen precisely the possible divisions of the clusters, which depend on the chosen similarity measure (distance) between the groups. In this work, this division also represents how many levels of difficulty the platform questions will be divided into. In this way, when choosing horizontal cut lines with small distances, there will be more levels of difficulty and, consequently, very similar questions in terms of difficulty within each cluster. In contrast, larger distances lead to more heterogeneity concerning the difficulty level in each generated cluster. The prominent question is,"What is the best value for the cut line?". From the previous work [3], it is known that two difficulty levels, meaning 2 clusters, are not enough, so we are interested in delimiting a horizontal line that includes 3 clusters (levels) or more.

If 7 or 8 clusters are used (green dashed line), the clusters will be composed of few questions, but of high similarity. As we are dealing with people (students and teachers), and the requirements, preferences, needs, and characteristics would make a lot of difference, it is interesting to have a little heterogeneity between the groups. Therefore, it is not interesting to restrict the dissimilarities of the elements of each cluster so much.

Another possibility is 3 - 4 levels (red dashed line). However, knowing that 2 is very little, 3 or 4 may not make much difference, and the problems presented in [3] may remain. We need a middle ground. Thus, 6 is a good split possibility. With 6 clusters (black dashed line) it is expected to be able to maintain a balance between the similarities and dissimilarities of the questions at each level.

However, the analyzes obtained so far have a high content of partiality of the authors. For this reason, it was decided to analyze the same datasets by another clustering technique, in this case Partitioning Clustering, through Bio-inspired automatic clustering techniques, whose results are presented below.

### 5.3 Partitioning Clustering Results

The four datasets mentioned were also evaluated by the Partitioning Clustering. In this case, three Bio-inspired clustering techniques were used to define the optimal number of clusters automatically. Consequently, the main difference in the definition of the number of clusters is in the algorithm used to minimize the DB index, that is GA, PSO and DE, as presented in Sect. 3. Besides,since these algorithms are stochastic, the results may vary from one iteration to another, requiring more than one execution of the algorithm. Moreover, it is interesting to compare the results of the different bio-inspired algorithms.

For all bio-inspired algorithms, the common parameters used were: maximum number of clusters equal to 10; initial population equal to 100, maximum number of iterations equal to 250, which was also the stopping criterion considered. For the GA, a rate of 0.8 was considered for selection and crossover, and 0.3 for a mutation. On the other hand, for PSO, the chosen rates were: global learning coefficient equal to 2, personal learning coefficient equal to 1.5, inertia weight equal to 1 and inertia weight damping equal to 0.99. Finally, for DE, the rates are equal to 0.2 for crossover and the scaling bound factor varies between $[0.2, 0.8]$. Each algorithm was performed 30 times for each dataset, and the smaller DB index obtained was defined as the optimal solution.

Table 3 presents the results of each algorithm, in terms of DB index and the optimal number of clusters (No. Clusters), for each dataset.

As can be seen, the smallest DB index was obtained by the DE approach on the dataset 1 resulting in an index of 0.4525, with 6 clusters. Moreover, the number 6 appears at least once for each dataset considered. And, for a general analysis, 7 out of the 12 tests performed (each algorithm on each dataset) indicated 6 as the value of the clusters, which is in line with the results and conclusions obtained by the hierarchical clusters. Considering this, the DE approach on the dataset 1 was chosen as the optimal solution. The detailed results of this approach are presented in Table 4, in terms of centroid coordinator, probability

Table 3: Clustering bio-inspired algorithm results

| Algorithm | Results | Dataset 1 | Dataset 2 | Dataset 3 | Dataset 4 |
|---|---|---|---|---|---|
| **GA** | DB index | 0.4547 | 0.4638 | 0.4981 | 0.4624 |
| | No. Clusters | 3 | 5 | 6 | 6 |
| **PSO** | DB index | 0.4546 | 0.4757 | 0.6131 | $\approx 0$ |
| | No. Clusters | 6 | 6 | 6 | 2 |
| **DE** | DB index | 0.4525 | 0.4753 | 0.4686 | 0.4468 |
| | No. Clusters | 6 | 6 | 9 | 7 |

intervals (Prob. Inter.), intra-cluster distance (Intra C. Dist.) and inter-cluster distance. Note that, each cluster is defined as $C_k$, where $k \in [1, 6]$.

Table 4: Detailed results of the optimal DE solution

| Results | | C1 | C2 | C3 | C4 | C5 | C6 |
|---|---|---|---|---|---|---|---|
| **Centroids** | | 0.0000 | 0.8338 | 1.0000 | 0.4876 | 0.6664 | 0.2787 |
| **Prob. Inter.** | | [0 , 0.13] | [0.14 , 0.37] | [0.38 , 0.57] | [0.58 , 0.75] | [0.76 , 0.94] | [0.95, 1.00] |
| **Intra C. Dist.** | | 0.026 | 0.0358 | 0.0000 | 0.0511 | 0.0485 | 0.0634 |
| **Inter Cluster Distance** | **C1** | 0 | | | | | |
| | **C2** | 0.8338 | 0 | | | | |
| | **C3** | 1.0000 | 0.1662 | 0 | | | |
| | **C4** | 0.4876 | 0.3463 | 0.5124 | 0 | | |
| | **C5** | 0.6664 | 0.1675 | 0.3336 | 0.1788 | 0 | |
| | **C6** | 0.2787 | 0.5551 | 0.7213 | 0.2033 | 0.3876 | 0 |

Note that the centroids have only one coordinate as they are on a straight line, since the clusters were defined from a single variable. The values in *Prob. Inter.* define the probabilistic intervals that delimit each of the clusters and, consequently, the interval of each level of difficulty, based on the probability of a student hitting a question.

Finally, Fig. 5 illustrates the optimal solution, given by the dataset 1 with the DE algorithm. In this case, each level of difficulty is presented by a different color. And the axis $x$ represents the probability of a question being answered correctly while the axis $y$ represents each topic available on MathE platform. So, the level 1 are the easiest questions, while the level 6 are the most difficult questions, based on the probability of correct answers.

So far, only questions per topic have been analyzed. However, in some topics, there are also subtopics, so the need arose to verify that the results of the topics can be verified for the subtopics. For this, the topic of Linear Algebra, which has 6 subtopics and has the largest number of questions answered, was chosen for further analysis. Thus, the same metrics were applied to the definition of datasets, and the parameters of the algorithms were reproduced exclusively for the questions that compose the subtopics that correspond to the Linear Algebra topic.

Table 5 presents the results for the 4 datasets, where the first dataset is composed of questions answered at least 1 time, in the second are the questions answered at least

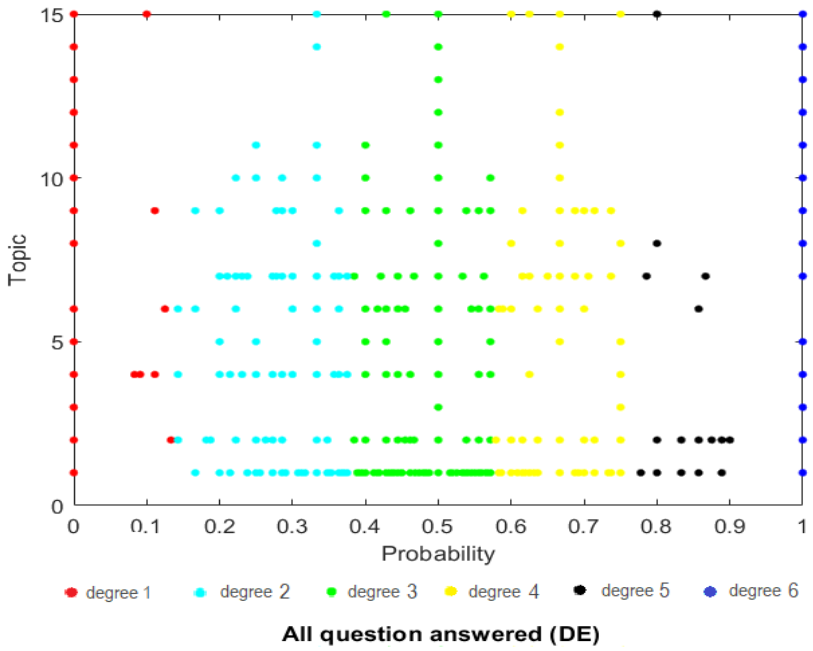



Fig. 5: Clustering optimal solution.

5 time, followed by the datasets composed of questions answered 10 times and 20 times, respectively. Again, a trend towards 6 clusters was observed. As can be seen, for the 12 tests performed, in 5 of them, the number 6 was pointed out as the optimal solution. In this case, the optimal solution, represented by the smallest DB index, was obtained by the PSO algorithm in dataset 1, being 0.4237 the DB index value.

Table 5: Linear Algebra Clustering Results

| **Algorithm** | **Results** | **Dataset 1** | **Dataset 2** | **Dataset 3** | **Dataset 4** |
|---|---|---|---|---|---|
| **GA** | DB index | 0.4282 | 0.4363 | 0.4244 | 0.4322 |
| | No. Clusters | 6 | 5 | 5 | 7 |
| **PSO** | DB index | 0.4237 | 0.4652 | 0.4614 | 0.4479 |
| | No. Clusters | 6 | 6 | 6 | 2 |
| **DE** | DB index | 0.4246 | 0.4731 | 0.4409 | 0.4258 |
| | No. Clusters | 7 | 6 | 8 | 7 |

## 6 Conclusions and Future Work

The MathE platform is an online educational system that aims to help students who struggle to learn college mathematics as well as students who want to deepen their knowledge of a multitude of mathematical topics, at their own pace. The platform has the aim of offering a dynamic and engaging tool to teach and learn mathematics, relying on interactive digital technologies that enable customized study. This work extracted information from the data collected by the MathE platform in order to trace paths for the creation of an intelligent and customized management system for the platform. It is expected that in the near future the platform will be able to make use of intelligent mechanisms, based on optimization algorithms and machine learning, to make autonomous decisions, tailored according to the needs of each user. One of the decisions to be made refers to the distribution of questions according to the students' background and demands. Thus, the information collected through this research will serve as a guide to make the choice of optimal strategies to improve the performance of the platform. Hence, the information from 285 students who used the *Students Assessment Section* on the MathE platform between April 2019 and February 2022 was considered.

Currently, the resources available in the MathE platform are organized into two levels of difficulty, basic and advanced; any user of the platform with teacher profile can define the level of each question. However, in order to improve the resources that are available in the platform, making it autonomous for making certain decisions, some adjustments are necessary. This work aimed to investigate the optimal number of levels of difficulty in which the resources in the MathE platform should be reorganized. For this, the information from the questions answered over the time that the platform is online was analyzed, through the probabilities of correct answers for each question. This information was analyzed through two clustering techniques, namely hierarchical clustering and partitioning clustering.

According to the presented results, it can be concluded that both methodologies reached a consensus that 6 levels of difficulty is the optimal solution for the reorganization of the platform, both for topics and subtopics. With 6 levels of difficulty, it will be possible to work better with the implementation of algorithms that will make the MathE platform autonomous and intelligent. In addition, a more accurate division of the content tends to motivate student users even more, since in this way, they will be able to better follow the advance or retreat of their knowledge when moving through the different levels of difficulty.

Thus, considering the described results, the reorganization of the available questions remains as future work. Besides, it is also expected that future work will be developed in order to distribute the questions in an intelligent and personalized way, respecting the needs of each user.

## References


1. Ashwin, T., Guddeti, R.M.R.: Impact of inquiry interventions on students in e-learning and classroom environments using affective computing framework. User Modeling and User-Adapted Interaction **30**(5), 759–801 (2020). https://doi.org/10.1007/s11257-019-09254-3
2. Azevedo, B.F., Amoura, Y., Kantayeva, G., Pacheco, M.F., Pereira, A.I., Fernandes, F.P.: Collaborative Learning Platform Using Learning Optimized Algorithms, vol. 1488. Springer (2021). https://doi.org/10.1007/978-3-030-91885-9-52

3. Azevedo, B.F., Pereira, A.I., Fernandes, F.P., Pacheco, M.F.: Mathematics learning and assessment using mathe platform: A case study. Education and Information Technologies (2021). https://doi.org/10.1007/s10639-021-10669-y
4. Bakhouyi, A., Dehbi, R., Banane, M., Talea, M.: A semantic web solution for enhancing the interoperability of e-learning systems by using next generation of scorm specifications. In: International Conference on Advanced Intelligent Systems for Sustainable Development. pp. 56–67. Springer (2019). https://doi.org/10.3991/ijet.v14i11.10342
5. Beinicke, A., Bipp, T.: Evaluating training outcomes in corporate e-learning and classroom training. Vocations and learning **11**(3), 501–528 (2018). https://doi.org/doi.org/10.1007/s12186-018-9201-7
6. Benta, D., Bologa, G., Dzitac, I.: E-learning platforms in higher education. case study. Procedia Computer Science **31**, 1170–1176 (2014). https://doi.org/10.1016/j.procs.2014.05.373
7. Cabero Almenara, J., Barroso Osuna, J.M.: Los escenarios tecnológicos en realidad aumentada (ra): posibilidades educativas en estudios universitarios. Aula Abierta - Revistas Eletronicas de la Unviersidad de Oviedo (2018)
8. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-1**(2), 224–227 (1979). https://doi.org/10.1109/TPAMI.1979.4766909
9. Ezugwu, A., Shukla, A., Agbaje, M., Oyelade, O., José-García, A., Agushaka, J.: Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature. Neural Computing and Applications (2020). https://doi.org/10.1007/s00521-020-05395-4, `https://hal.archives-ouvertes.fr/hal-03217646`
10. Gunasinghe, A., Abd Hamid, J., Khatibi, A., Azam, S.F.: The adequacy of utaut-3 in interpreting academician's adoption to e-learning in higher education environments. Interactive Technology and Smart Education (2019). https://doi.org/10.1108/ITSE-05-2019-0020
11. Herodotou, C., Rienties, B., Hlosta, M., Boroowa, A., Mangafa, C., Zdrahal, Z.: The scalable implementation of predictive learning analytics at a distance learning university: Insights from a longitudinal case study. The Internet and Higher Education **45**, 100725 (2020). https://doi.org/10.1016/j.iheduc.2020.100725
12. Jin, X., Han, J.: Partitional Clustering. Springer US, Boston, MA (2010). https://doi.org/10.1007/978-0-387-30164-8-631
13. Kalpokaite, N., Radivojevic, I.: Teaching qualitative data analysis software online: a comparison of face-to-face and e-learning atlas. ti courses. International Journal of Research & Method in Education **43**(3), 296–310 (2020). https://doi.org/doi.org/10.1080/1743727X.2019.1687666
14. Kayser, I., Merz, T.: Lone wolves in distance learning?: An empirical analysis of the tendency to communicate within student groups. International Journal of Mobile and Blended Learning (IJMBL) **12**(1), 82–94 (2020). https://doi.org/10.4018/IJMBL.2020010106
15. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of ICNN'95 - International Conference on Neural Networks. vol. 4, pp. 1942–1948 vol.4 (1995). https://doi.org/10.1109/ICNN.1995.488968
16. Khlifi, Y.: An advanced authentication scheme for e-evaluation using students behaviors over e-learning platform. International Journal of Emerging Technologies in Learning **15**(4) (2020). https://doi.org/10.3991/ijet.v15i04.11571

17. Laskaris, D., Heretakis, E., Kalogiannakis, M., Ampartzaki, M.: Critical reflections on introducing e-learning within a blended education context. International Journal of Technology Enhanced Learning **11**(4), 413–440 (2019). https://doi.org/10.1504/IJTEL.2019.102550
18. Luo, N., Zhang, Y., Zhang, M.: Retaining learners by establishing harmonious relationships in e-learning environment. Interactive Learning Environments **27**(1), 118–131 (2019). https://doi.org/10.1080/10494820.2018.1506811
19. MATLAB: The mathworks inc. https://www.mathworks.com/products/matlab.html (2019a)
20. Moreno-Guerrero, A.J., Aznar-Díaz, I., Cáceres-Reche, P., Alonso-García, S.: E-learning in the teaching of mathematics: An educational experience in adult high school. Mathematics **8**(5), 840 (2020). https://doi.org/10.3390/math8050840
21. Moubayed, A., Injadat, M., Shami, A., Lutfiyya, H.: Student engagement level in an e-learning environment: Clustering using k-means. American Journal of Distance Education **34**(2), 137–156 (2020). https://doi.org/10.1080/08923647.2020.1696140
22. Mousavi, A., Mohammadi, A., Mojtahedzadeh, R., Shirazi, M., Rashidi, H.: E-learning educational atmosphere measure (eeam): A new instrument for assessing e-students' perception of educational environment. Research in Learning Technology **28** (2020). https://doi.org/10.25304/rlt.v28.2308
23. Nielsen, F.: Hierarchical Clustering, pp. 195–211. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-21903-5-8
24. Rakic, S., Tasic, N., Marjanovic, U., Softic, S., Lüftenegger, E., Turcin, I.: Student performance on an e-learning platform: Mixed method approach. International Journal of Emerging Technologies in Learning **15**(2) (2020). https://doi.org/10.3991/ijet.v15i02.11646
25. Sathiyamoorthi, V.: An intelligent system for predicting a user access to a web based e-learning system using web mining. International Journal of Information Technology and Web Engineering (IJITWE) **15**(1), 75–94 (2020). https://doi.org/10.4018/IJITWE.2020010106
26. Shakah, G., Al-Oqaily, A., Alqudah, F.: Motivation path between the difficulties and attitudes of using the e-learning systems in the jordanian universities: Aajloun university as a case study. International Journal of Emerging Technologies in Learning (iJET) **14**(19), 26–48 (2019). https://doi.org/10.3991/ijet.v14i19.10551
27. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory To Algorithms. Cambridge University Press (2014)
28. Sivanandam, S.N., Deepa, S.N.: Introduction to Genetic Algorithms. Springer, 1 edn. (2008). https://doi.org/10.1007/978-3-540-73190-0
29. Storn, R., Price, K.: Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization **11**(4), 341–359 (1997). https://doi.org/doi.org/10.1023/A:1008202821328