Journal of Applied Data Sciences
Vol. 6, No. 1, January 2025, pp. 508-525

Student Engagement in E-Learning During Crisis: An Unsupervised Machine Learning and Exploratory Data Analysis Approach

Rachid Ait Daoud[1,*], Abdellah Amine[2], Karim Abouelmehdi[3], Ayoub Razouk[4]

[1,2]*Department of Mathematics and Computer Science, Higher School of Technology_Fkih Ben Salah, Sultan Moulay Slimane University, Morocco*

[3]*ELITES laboratory, Faculty of Sciences, Chouaib Doukkali University, El Jadida, Morocco*

[4]*Economics and Management, Polydisciplinary Faculty of Beni Mellal, Sultan Moulay Slimane University, Beni Mellal, Morocco*

(Received: August 25, 2024; Revised: October 6, 2024; Accepted: November 23, 2024; Available online: December 30, 2024)

Abstract

The lockdown caused by COVID-19 has forced educational institutions to rapidly adopt e-learning, which has revealed many significant challenges related to student engagement. Following the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, the present work aims to provide teachers and university administrators with a framework based on unsupervised machine learning and exploratory data analysis to identify engagement levels and understand the potential reasons for low engagement. Various data sources, including Microsoft Teams logs, demographic, and educational data, were merged to create a comprehensive dataset with the most relevant and useful measures for the success of our approach. This study was structured around three main research questions to achieve our goal. First, we sought to identify the most effective Microsoft Teams measures for identifying students' engagement levels. Then, our analysis focused on comparing different clustering models (two-level, three-level, and four-level models) to determine which one is most accurate in identifying low-engaged students. Finally, we examined the demographic and educational factors influencing low student engagement. The results revealed that: by applying the Sequential Forward Selection (SFS) technique, ScreenShareTime, VideoTime, NbrViewedVideos, Recency, and AvgTeamsSessionDay are the most relevant Microsoft Teams engagement metrics, improving the silhouette width from 0.37 to 0.70 when using these selected measurements. The four-level clustering model (Low, Medium, High, and Super) proved most effective in identifying low-engaged students. Analysis of factors showed that low engagement is primarily related to limited living conditions, with 66% of low-engaged students having low incomes. In addition, 50% do not use online services and 62% of low-engaged students took more than three years to reach their final year, indicating pre-existing academic difficulties. These findings provide educational institutions with valuable insights to enhance student engagement in distance learning, particularly during crisis periods such as the COVID-19 pandemic.

Keywords: Educational Data Mining, CRISP-DM Methodology, Log Files, Feature Selection Method, K-means, E-learning, Engagement Metrics

1. Introduction

One of the educational alternatives adopted by many educational institutions in the world to ensure continuity of learning, during the COVID-19 confinement period, is the use of e-learning services, in which a broad array of learning materials is delivered electronically to remote learners using Internet technologies [1]. However, adopting e-learning as a learning solution allows students to learn anywhere without time or space limitations [2], [3]. To make this experience successful in Morocco, several measures have been taken by the Ministry of National Education (MEN), Vocational Training, Higher Education, and Scientific Research.

In higher education, each university has used its own solutions (Moodle, Teams, Zoom, Google Classroom, Blog, YouTube Channel, etc.) to replace face-to-face courses with distance learning. Indeed, since the decision to suspend classes on 13 March 2020, many efforts have been made by the Kingdom's various universities. More than 111,000 digital and audiovisual resources and numerous teaching aids have been produced and published on their platforms. In this article, we examined the case of the Multidisciplinary Faculty of Beni Mellal (FMBM), part of the Sultan Moulay

***Corresponding author: Rachid Ait Daoud (r.aitdaoud@usms.ma)**
DOI: https://doi.org/10.47738/jads.v6i1.458

Slimane University in Morocco. Following the suspension of studies, FMBM adopted Microsoft Teams as its online learning platform for the second semester. It created 9,524 Microsoft Teams accounts, including 9,368 for students, 143 for professors, and 13 for MFBM administrative staff. Furthermore, 282 virtual classrooms were established and over 1,940 digital teaching resources were uploaded by educators between March 16 and May 17. Teachers and students showed a positive response towards Microsoft Teams. From the first week following the decision to suspend classes, an increasing trend was observed in the number of active students and video conferences organized by teachers, until it stabilized and peaked at the end of March. The number of video conferences reached 61 videoconferences per day. Notably, during the first two weeks of teachers' closure, the number of active students increased to 4,170 per day or 44.51% of the total. However, these encouraging figures raise an important and interesting question: to what extent are students engaging in this mode of learning? and are all students engaged to the same level? It is necessary to explore these aspects to better understand how students adapt to this new form of learning and what factors influence it?

A preliminary analysis of the evolution of e-learning reveals encouraging results for Master's degree programs, with participation rates in virtual classes reaching 97% and sometimes even 100%. However, undergraduate students, representing 80% of the total student body, show significantly lower participation rates, remaining limited between 43% and 55%, which directly reduces the interactivity and engagement of these students in e-learning. To better understand this decline, the faculty administration must have a thorough understanding of all students and analyze the usage behavior of the Microsoft Teams platform for each student. Moreover, the process of analyzing student behavior is an important key to determining the level of their interaction with the e-learning platform and their motivations [4], [5]. In general, the flexibility provided by distance learning platforms makes it challenging to identify students' behavior [6]. Another complexity lies in the inability to interpret the large amount of data generated and stored when using the Microsoft TEAMS platform, as this data cannot directly indicate students' engagement behavior without further analysis processes. Furthermore, since not all students use the Microsoft TEAMS platform in the same way (some attend live sessions, others retrieve recordings, or only download course materials), student behaviors can vary from one student to another [7].

To address these constraints, this study primarily aims to develop and implement an analytical framework based on data exploration from the log files generated by Microsoft Teams and on unsupervised machine learning techniques to identify different groups of students according to their levels of engagement. From this analysis, we will propose innovative strategies, aligned with the emerging trends in online learning, to enhance student engagement and ensure course continuity during emergencies. The remainder of this paper is structured as follows: Section 2 reviews recent research on e-learning during COVID-19. Section 3 presents the different steps of our educational data mining framework, designed to assess student engagement in e-learning, while Section 4 discusses the results obtained. Finally, conclusions and some ideas for future research are presented in Section 5.

2. Literature Review

The closure of educational institutions during the COVID-19 pandemic has led to an unprecedented shift towards distance learning globally, generating significant interest from the scientific community. This sudden shift in educational practices has led to numerous studies to assess the effectiveness and impact of this mode of education in the unique context of lockdown measures.

The literature review reveals that a major part of the studies focuses on examining the role of e-learning in COVID-19, the benefits of technology in education, and proposing alternative approaches to maintain learning continuity during this period. Srinivasan [8] proposed adopting the ZOOM tool as an alternative solution. The author confirmed the usefulness and effectiveness of using this tool to conduct e-tutorials in anatomy during the COVID-19 pandemic, while Downes et al. [9] created a multi-level questionnaire to study the impact of online platforms on the acquisition of knowledge and skills of neurosurgeons, and according to the responses they found that online neurosurgical education was useful. Favale et al. [10] studied the fruition and the performance of the online teaching system (based on the BigBlueButton framework) deployed by PoliTO University to support all the second-semester classes. The authors used the log files to understand how the students access the teaching material. Chick et al. [11] presented several innovative solutions based on technology (online practice questions, teleconferencing in place of in-person lectures, and flipped classroom model) which can be applied to mitigate the loss of learning exposure for surgical residents during the lockdown session. Kristiana et al. [12] conducted a study of 176 students from the Faculty of Education at Nusa Nipa University to assess their perceptions of Microsoft Teams as an online learning tool during the COVID-19 pandemic. The results of their research showed that the platform was perceived positively by students and that it was an asset for the continuity of education in a lockdown context.

Yuet-Ming et al. [13] suggested a three-step virtual classroom education approach based on Zoom and Blockboard online platforms to support nursing educators in online theoretical hand hygiene enhancement. The authors have confirmed that this approach provides real-time interaction between educators and nursing students. Oyeniran et al. [14] proposed a new framework that uses certain open-source computer and Android applications (Google Classroom, Zoom, Whats App, Blog). These are grouped in a zone called zone e-learning, which represents the point of interaction between students and their teachers. The authors also cited some advantages and disadvantages of this framework. In [15], the effectiveness of the e-learning method (based on the Moodle platform) in the teaching of mathematics was identified. By applying an ad hoc questionnaire, the authors have shown that the e-learning method has a positive influence on aspects related to motivation, participation, interactivity, results, and grades compared to the traditional expository method.

Although distance learning platforms generate large amounts of data that can be used to assess the online teaching experience [16] and reveal hidden insights through learning analytics, the majority of these studies have not sufficiently exploited these data resources to assess student engagement and interactivity in this new learning mode, which is no longer considered a complementary learning mode but has become the only solution to ensure the continuity of education during the COVID-19 crisis. The limitations of these previous studies [8], [11], [12], [13], [14] are reflected in the fact that they mainly focus on suggestions for distance learning solutions during the lockdown, examining only students' perceptions, without assessing their interactivity and engagement level in e-learning. On the other hand, some of these studies [9] and [15] rely on questionnaire responses, which may be insufficient and inaccurate in identifying students' learning behavior [17]. Since log file data are more reliable and authentic than those collected through questionnaires, our work enriches the existing literature in two significant ways. First, we integrate two data sources, the first is the log file data generated by the TEAMS platform, while the second data source comes from the university application APOGEE (Application for the organization and management of students and teachers) which provides demographic and educational data about students. This allows us to extract many engagement metrics and indicators that we use as behavioral variables in our clustering model, maximizing their ability to distinguish between different groups of students. Second, our approach is distinguished from previous studies by adopting a comprehensive methodology dedicated to data mining. While previous research mainly relied on traditional statistical techniques such as descriptive statistics, Student's t-test, Cohen's d-test, and biserial correlation [9], [15] to analyze data from questionnaires, the present work exploits the potential of educational data mining to analyze and extract the knowledge hidden behind a large amount of data, and more specifically we use the clustering technique (K-means), which is an unsupervised machine learning algorithm, to deeply analyze student behaviors and divide them into an appropriate number of homogeneous groups, so that students in the same group will have the same learning behaviors. This helps to quickly and accurately identify different student profiles in terms of interactivity and engagement level. Educational data mining is a variant of data mining, specifically used to discover knowledge from data that comes from educational context (includes information about teachers, students, programs, and courses) [18]. It takes log files and data from various education systems and transforms them into actionable insights to better understand students and their learning environments [19]. Although there are many studies on the effectiveness of distance learning and student engagement, few of them address this mode of teaching in emergencies. Given that pre-planned and carefully designed distance learning courses differ significantly from emergency distance learning courses [20], more studies based on exploring educational data extracted from the COVID-19 experience are needed. A systematic and applicable tool is required to implement the data mining analysis in the online educational context. Hence, in this study, the Cross-Industry Standard Process for data mining (CRISP-DM) [21] is used to present all the steps necessary for the understanding and success of our approach. This choice is motivated by the fact that the latter has currently become the reference model for the Knowledge Discovery in Databases process [22].

3. Research Methodology

As mentioned earlier, this study applies the CRISP-DM standard process model, which is known for its flexibility and applicability. CRISP-DM is a data mining process model that breaks down the life cycle of a data mining project into six phases ranging from understanding the business problem to deployment and production [23]. In this section, we detail these phases essential for building our clustering model, with the main objective of identifying the students' values in terms of interactivity and engagement.

3.1. Business Understanding

It is important to note that while our initial analysis revealed high participation rates (97-100%) among Master's degree students, this study focuses specifically on 5,000 undergraduate students who logged into the platform at least once during the study period. This decision was driven by two critical factors. First, undergraduate students, who comprise 80% of the total student body, exhibited significantly lower participation rates, ranging from 43% to 55%. Second, given that these undergraduates represent the majority of the faculty's student population, understanding and addressing their engagement issues is paramount for overall educational improvement. As previously stated, our issue is related to the difficulty of understanding the low participation rate of undergraduate students in virtual classes. This challenge stems from the inability to identify student's behavior using traditional analysis methods due to the abundant amounts of data generated by the Microsoft TEAMS tool. According to this brief description of the issue, we set both educational and data mining objectives for our study. From an educational perspective, our primary aim is to strengthen students' engagement and enhance their participation in distance learning by finding solutions to their difficulties, ensuring continued learning for all students during the COVID-19 confinement period. We seek to segment students into several clusters based on their levels of engagement in e-learning, which will enable the identification of differences and similarities in their learning behavior to make appropriate decisions and deploy targeted strategies. Additionally, we aim to identify the engagement metrics that contribute significantly to determining student engagement when using the Microsoft TEAMS platform as an alternative solution, while also understanding the demographic and educational factors influencing low levels of engagement. Regarding data mining objectives, our main focus is to exploit the large amount of data from activity log files generated by the Microsoft Teams service. This data is used to develop a clustering model adapted to the education sector, with the primary purpose of identifying different types of student profiles and analyzing the value of each profile to assess student interactivity and engagement. A secondary objective involves identifying and selecting the features (metrics) necessary for the success of our clustering model. In this study, the success criterion for data mining is the quality of the clustering model that will be created, which will be evaluated using the average silhouette method.

3.2. Data Understanding

The data utilized in this study were collected from the MFBM, Sultan Moulay Slimane University. As mentioned previously, the faculty adopted the Microsoft TEAMS service as an alternative solution to ensure continuity of education during the COVID-19 pandemic. When students log into this service, all their actions are recorded and stored as log files. Additionally, the MFBM also has an information system (APOGEE) for student affairs management, which provides demographic and educational data on students.

3.2.1. Initial Data Collection

The data sources used in this study combine the APOGEE database and Log-files data from the Reporting Office 365 Service. The APOGEE database provides comprehensive student information, including contact details, demographic and educational data such as name, gender, age, address, city, country, student profession, parents' profession, degree, course type, semester, pedagogical registration in modules, results, and grades. Additionally, the Log-files data and Reporting Office 365 Service capture all TEAMS sessions opened by students, storing the complete history of events and activities performed when accessing the platform. These logs encompass various interactions including Teams session started, session status, conversations, and chats, videoconference participation, files accessed, files downloaded, watching the video conferencing recording, feedback on video conferencing (like or dislike), device information, and IP address. This provides information on students' interactivity in virtual classrooms, including their interactions with shared course materials and participation in scheduled video conferences.

ISSN 2723-6471

3.2.2. Data Description and Exploration

The dataset contains contact information, browsing habits, and learning behavior of 5,000 students who accessed the Microsoft Teams platform. The events and activities performed by these students during the period from March 16, 2020, to May 31, 2020, have been utilized. Student activities in Microsoft Teams are categorized as follows: Activities related to session creation (Teams Session Started), course activity (Files viewed, Files downloaded, and Files uploaded), and class recording activities (Microsoft Stream) which include downloading and viewing videos, as well as liking and unliking mentions. To clarify the information collected, table 1 describes the types of data used, their attributes, and the source of the data. There are five categories of data: demographic data, educational data, session data, user activity data, and virtual classroom data. Each category includes specific attributes that are collected from different sources.

Table 1. The data used and their attributions

Data type	Attributions	Data source
Demographic data	IdStudent, FirstName, LastName, Gender, Age, Address, City, EmailStudent, Phone, StudentProfession, ParentsProfession	APOGEE
Educational data	IdStudent, CollegeYear, Degree, CourseType, Semester, ModuleName	APOGEE
Session data	EmailStudent, SessionStartDate, Device, ClientIP, UserAgent	Log-files
User activity data	CreationTime, EmailStudent, Operation, RecordType, EventSource, Workload, VirtualClassromName, SourceFileName, SourceFileExtension, ItemType	Log-files
Virtual Classrooms data	EmailStudent, VirtualClassromParticipated, AudioTime, VideoTime.	Reporting Office 365 Service

From the data in the log files, some symbolic variables (such as the date of the first activity in TEAMS, the date of the last participation of a student in a TEAMS activity, as well as the courses downloaded and consulted) must be exploited to create relevant numerical variables. This includes total number of sessions, daily and weekly session averages, recency, duration of Microsoft Teams usage, total files viewed and downloaded, and total number of virtual classes the student has attended.

3.3. Data Preparation Phase (design and feed of the data warehouse)

Data acquisition, data cleaning, building new metrics, and integrating data into the data warehouse are the main tasks of the data preparation phase. This phase represents the most important and time-consuming step in the design of decision-making systems, accounting for an estimated 60-80% of the effort and time spent on a data mining project [24]. The Extraction, Transformation, and Loading (ETL) tools are used in this phase to select and merge data from all sources. Subsequently, the records and attributes selection operation must be performed. The selection of attributes largely depends on the desired data mining objectives. For example, in our case, the scope of the study will be limited to students who have performed at least one activity in TEAMS between March 16 (the starting date of e-learning) and May 31, 2020. Concerning attributes, we will focus on those more effective in describing students' learning behavior and interactivity. Filters should also be defined to exclude students who have not opened a TEAMS session (students who have not activated their account) during the study period. In this study, we utilized the ETL tool from the KNIME ANALYTICS platform to develop our data integration process, drawing from data sources such as APOGEE, log files, and Reporting Office 365 Service, to construct our data mart. A summary and explanation of this process are presented in table 2. The design of this data store was particularly focused on the student to extract only the metrics necessary for the success of our clustering model. For this, the fact table must contain as much information as possible (measures) to identify those that contribute more to the analysis of students learning behavior. Therefore, 15 measures were calculated in the fact table. We excluded highly correlated measures to ensure better diversity and representativeness of information in the clusters. This also reduces data dimensionality and enhances the performance and interpretability of the clustering model. For instance, "Audio time" representing the total time a student engaged in audio activities, and "VirtualClassromParticipated" indicating the count of virtual classes attended, exhibited a strong correlation ($r = 0.902$, p-value <0.0).

Table 2. The actions used in the creation and feeding of the data mart.

Loading Process	Source	Target	Actions
Dim_student_loading	Apogee.Student Apogee.PedagRegist Reporting.Office365	dim_student	Update/Insert
Dim_session_loading	Log-files	dim_TeamSession	Insert
Dim_coursesActivity_loading	Log-files	dim_CourseMateriels	Update/Insert
Dim_ClassesRecording Activity_loading	Log-files	dim_RecordingVirtualClasses	Update/Insert
Dim_time_loading	Log-files	dim_time	Update/Insert
Fact_LearningBehavior_loading	Joining dimensions	Fact_LearningBehavior	Update/Insert

Table 3 presents the finalized metrics used to assess engagement levels. High values in all metrics, except for "Recency," indicate positive student engagement. For the Recency metric, a lower value reflects more recent student sessions.

Table 3. Engagement metrics description

Metric category	Metric name	Description
Session	Recency	Refers to the number of days from the last session date to the last day of study (May 31, 2020). (The lower the recency value, the more recent the student's open sessions.)
	AllSession	Refers to the total number of sessions opened by the student during the period analyzed.
	AvgSessionDay	Refers to the average number of open sessions per day.
Course Material	NumbrFileDownloaded	Refers to the total number of course materials downloaded
Classes Recording (Videos)	NbrViewedVideos	Refers to the total number of recorded videos (virtual classroom recording) watched.
Virtual Classes Room	NumbrVirtualClassromParticipated	Refers to the number of virtual classrooms a student participated in during the period analyzed.
	VideoTime	Refers to the total duration of video content (minutes) viewed by the student during the analyzed period.
	ScreenSharingTime	The total duration of screen (minutes) sharing by students.

It is interesting to note that during the integration process, we encountered and addressed several specific challenges when merging data from our three main sources (APOGEE, log files, and Reporting Office 365 Service). One of the primary challenges was ensuring data consistency across the three sources, as each had its own format and standards. For instance, the APOGEE database utilized a different timestamp format compared to our log files, which complicated the synchronization of events. We also encountered data quality issues, including missing values in Teams student attendance due to connectivity issues and the existence of students not enrolled in their course modules at the APOGEE system level. Another challenge is the consistency of student identification as each system uses different identifiers, for example email addresses in Microsoft Teams and APOGEE number in the faculty system APOGEE. In addition, legal constraints related to data protection have added a layer of complexity to our work. In order to ensure a minimum level of protection of personal data, we have notably removed sensitive personal information (telephone numbers, addresses, precise dates of birth) while retaining the aggregated demographic data necessary for the analysis of engagement patterns and replacing direct identifiers (names, first names, emails) with unique coded identifiers. These challenges were systematically addressed through our ETL pipeline, implementing specific data cleaning and transformation rules for each data source while ensuring the consistency and reliability of the integrated dataset for

ISSN 2723-6471

subsequent analysis. Table 4 shows the descriptive statistics for their maximum, minimum, average, and standard deviation values.

Table 4. The descriptive statistics of engagement metrics.

Metrics	Maximum	Minimum	Average	Standard deviation
Recency (days)	57	0.0	10.051	14.9941
AllSession	78	1	18.6066	15.3066
AvgSessionDay	5.2	1	1.1504	0.3402
NumbrFileDownloaded	34	0.0	4.1824	6.312
NbrViewedVideos	24	0.0	0.8615	2.706
NumbrVirtualClassromParticipated	73	0.0	10.4418	13.1677
VideoTime (minutes)	17 108	0.0	685.8769	1 983.4934
ScreensharingTime (minutes)	7 164	0.0	123.5033	546.5221

3.4. Modeling Phase

This phase aims to design and implement a clustering model based on the K-means algorithm to identify the different clusters of students according to their engagement levels. Each cluster represents a distinct group of students with similar engagement levels. Regarding clustering quality, each group must be distinct in all engagement metrics to avoid unnecessary overlaps and obtain the optimal result.

Students behavioral analysis in learning management systems, and assessment of student engagement, in particular, have received growing attention during the last decade [25], [26], [27], [28], [29]. Most studies on student engagement levels in e-learning adopt a three-level [25], [29] or five-level model [26], [27], [28]. Concerning the metrics used to measure engagement levels, the majority of these studies focus on behavioral and cognitive, as they offer a more measurable perspective on engagement. To this end, as shown in table 3, this study analyzes four categories of engagement metrics: session metrics, course material metrics, virtual class participation metrics, and metrics from virtual classroom session recordings, each comprising various indicators.

Using the engagement measures presented in table 3, this study aims to answer the following three research questions:

Q1: Do all the metrics presented in table 3 contribute to developing an effective clustering model? If not, which metrics are most helpful in identifying consistent and meaningful student profiles in terms of engagement in Microsoft Teams?

Q2: What clustering model helps educational managers identify unengaged students accurately?

Q3: Are engagement metrics extracted from Microsoft Teams log files enough to understand low levels of student engagement, or do we need to integrate other demographic and educational data?

To find all the hidden student profiles in the data warehouse created during the data preparation phase, our model must be able to discover them autonomously. This is because the learning style of each student is not known in advance. To achieve this, the data mining algorithms adapted to this kind of problem are those of unsupervised learning because those of supervised learning would not be suitable for our specific case due to the lack of pre-labeled engagement patterns. More precisely, these are clustering techniques that discover the inherent groupings in the data. [30], [31]. There are several clustering algorithms. The clustering algorithm applied in this paper is the K-means algorithm. It is one of the most popular unsupervised clustering algorithms for pattern clustering [32] due to its ability to process large amounts of data quickly and efficiently, and has been widely used in the literature in the field of e-learning [25], [33], [34]. This work proposed the k-means algorithm because of its simplicity and recognized efficiency in this field and the most important is the ease of interpretation of its results. The centroids of the clusters clearly represent the average characteristics of each group, which is particularly important in our educational context where understanding and explaining engagement patterns is essential.

Based on some attributes this algorithm clusters n vectors with high similarity into k partitions, where k < n. It first selects k random initial centroids. Then, each vector in the dataset is assigned to the closest centroid using Euclidean distance and recalculates the new centroids as means of the assigned data vectors. This process is repeated many times until the stopping criteria are met. Each vector of the space corresponds to a student whose attributes are the metrics mentioned in table 3. Then, each student is positioned to the closest centroid. Once the groups of students have been formed according to their learning style, the position of the centers is recalculated and the operation is repeated until a stable state is reached. Briefly, the student's engagement metrics are input into the k-means clustering algorithm, which groups the students into clusters according to their engagement patterns. These clusters are then analyzed to identify groups of students who are not engaged and might require extra assistance. In the k-means algorithm, the number of clusters is selected randomly, which introduces a potential for unreliability in the clustering result if the assumed number of clusters is incorrect. Additionally, the choice of features (metrics) employed during the clustering process can also influence the quality of the resulting clusters. While some features are essential for effective clustering, others can lead to erroneous results [35]. However, to efficiently choose the optimal number of clusters and the metrics that best uncover clusters, we have combined the silhouette and feature selection methods. These two methods are provided and explained in the next phase "Evaluation phase".

Evaluation Phase

The experimental evaluation of our models represents the last step of the clustering process. To evaluate the efficiency of our clustering model - that is, its capacity to properly discover hidden patterns - we selected the silhouette method, a widely-used technique that measures both how similar an object is to its own cluster (cohesion) and how different it is from other clusters (separation).

The silhouette method, introduced by Rousseeuw [36], works by analyzing two key aspects for each data point. First, it examines how well the point fits within its assigned cluster (measured by 'ap') by calculating the average distance between this point and all other points in the same cluster. A smaller value of 'ap' indicates better placement. Second, it evaluates how well separated the point is from neighboring clusters (measured by 'bp') by calculating the average distance between the point and all points in the nearest neighboring cluster, where a larger value of 'bp' indicates better separation.

These measurements are combined into a single score (Sp) using the formula:

$$S_p = \frac{b_p - a_p}{\max(a_p, b_p)} \tag{1}$$

The resulting silhouette score (Sw), calculated as the average of all individual Sp values, ranges from -1 to 1, where a score near 1 indicates well-defined clusters where points are far from neighboring clusters, a score near 0 suggests overlapping clusters with unclear boundaries, and a negative score indicates potential misclassification, where points might be assigned to the wrong clusters.

Regarding the task of feature selection, the Sequential Forward Selection (SFS) technique was used in this paper to identify the metrics that effectively contribute to building a powerful clustering model capable of generating more distinctive groups in terms of learning style. This technique is one of the most popular wrapper methods. It's an iterative approach that starts with having no feature selected and gradually adds features selected using some criterion function. In each iteration, the best feature that greatly enhances the model is added to the feature set [37].

However, the integration of these two techniques (Silhouette and Sequential Forward Selection) will help to properly identify the best number (K) of student groups with varying engagement levels, and the clustering task becomes more efficient and targeted because only consistent, impartial, and explanatory features will be used. The basic idea involves running the K-means algorithm with various K values (k=2,3,...7). For each value of K, we apply the SFS method to identify the subsets of metrics that improve the clustering model the most. Therefore, we have set a threshold on average silhouette width S_w, and only a subset of metrics that offer a value greater or equal to 0.60 will be taken into account for comparison to identify the most suitable model in an e-learning context. Figure 1 depicts the algorithm for selecting the best subset of metrics.

Algorithm SELECT BEST METRICS

1. Start with empty Metrics set $M_0 = \emptyset$ and $N = 0$
2. Specify a silhouette score threshold SilhouetteThershold = 0.60
3. for K = 2 to 7
4. Run clustering algorithm $\underset{m \in M - M_N}{K - means} \; (M_N \cup \{m\})$
5. Select the next best metric m* = arg $\underset{m \in M - M_N}{max} \; S_w(M_N \cup \{m\})$
6. if $S_w(M_N \cup \{m *\}) \geq$ SilhouetteThreshold
7. Update $M_{N+1} = M_N \cup \{m *\}; N = N + 1$
8. Go to 4

Figure 1. Algorithm SELECT BEST METRICS

4. Results and Discussion

Incorporating all engagement metrics into the clustering process, table 5 shows the clustering quality for different clustering models (k=2, k=3, ...k=7) based on the average silhouette coefficient. The silhouette values show that clustering quality is poor for all values of K because a low silhouette value generally indicates that the clusters are poorly defined and poorly separated. It is, therefore, necessary to make a metric selection to examine those that improve the model the most.

Table 5. Silhouette coefficient with all metrics

Clustering Model	Average Silhouette Width Sw using all metrics
K = 2	0.36
K = 3	0.378
K = 4	0.348
K = 5	0.342
K = 6	0.317
K = 7	0.272

Table 6 presents the results of the metric selection algorithm illustrated in the previous section. The metrics are ranked according to their importance for clustering using the average silhouette width Sw. Therefore, only a subset of metrics offering a silhouette value greater than or equal to 0.60 will be considered. The distinctive characteristics of each clustering model, including the number of clusters k, the number of selected metrics, the metric name, and the Sw value, are presented.

Table 6. Results of select best metrics algorithm

	K=2			***K=3***	
Nr. of metrics	Best metrics	SW	Nr. of metrics	Best metrics	SW
1	ScreenShareTime(Minutes)	0.961	1	ScreenShareTime(Minutes)	0.923
2	VideoTime(Minutes)	0.942	2	AvgTeamsSessionDay	0.918
3	NbrViewedVideos	0.914	3	VideoTime (Minutes)	0.905
4	AvgTeamsSessionDay	0.867	4	NbrViewedVideos	0.883
5	Recency	0.706	5	Recency	0.715
	K=4			***K=5***	
Nr. of metrics	Best metrics	SW	Nr. of metrics	Best metrics	SW

1	ScreenShareTime(Minutes)	0.903	1	NbrViewedVideos	0.930
2	NbrViewedVideos	0.871	2	ScreenShareTime(Minutes)	0.763
3	VideoTime (Minutes)	0.818	3	AvgTeamsSessionDay	0.710
4	AvgTeamsSessionDay	0.716			
5	Recency	0.709			
	K=6			***K=7***	
Nr. of metrics	Best metrics	SW	Nr. of metrics	Best metrics	SW
1	NbrViewedVideos	0.930	1	NbrViewedVideos	0.930
2	Recency	0.728	2	Recency	0.711

The results presented in table 6 and the graph in figure 2 confirm that including all engagement metrics during the clustering process can only obscure the discovery of the cluster structure. Conversely, the exclusion of irrelevant metrics can improve the quality of clustering. However, as demonstrated in table 6, the two-level (k=2), three-level (k=3), and four-level (k=4) models offer the best performance in terms of cluster separation when using subsets of metrics containing only five metrics. While other models (k=5, k=6, k=7) are only effective when using limited metrics, it is best to focus on models with enough metrics to ensure a deep understanding of student engagement behavior.

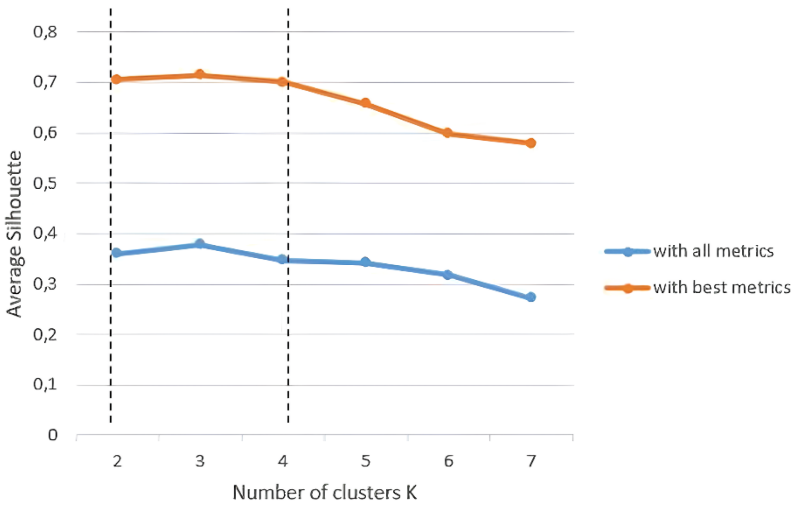

Figure 2. Impact of metrics on average Silhouette: Comparative analysis by number of clusters.

Furthermore, it is important to highlight that ScreenShareTime and VideoTime are powerful metrics that contribute significantly to the clustering process. This is confirmed by the similarity in the importance ranking of these two metrics, where we find them ranking among the top three most important engagement metrics in the three models. On the other hand, it appears that the metrics indicating the number of virtual classrooms a student participated in, the total course materials downloaded, and the total sessions opened by the student seem less significant and do not fully contribute to identifying student engagement levels. One reason for this could be that students opening sessions on Microsoft Teams to join virtual classes or access course materials may not necessarily indicate their level of engagement, as students may close their sessions before the class ends, leading to situations where students have attended numerous virtual classes with an average rate not exceeding five minutes per class. The same is true for the number of course materials downloaded because the number of downloads may not accurately reflect students' motivation, persistence, or critical thinking, which are essential components of engagement [38]. In these situations, it is better not to consider useless metrics in the process of identifying engagement levels, as they obscure finding more valuable results for clustering. Therefore, ScreenShareTime, VideoTime, NbrViewedVideos, Recency, and AvgTeamsSessionDay are the potential metrics that our models will use to measure engagement levels.

Table 7 presents a summary of the results obtained from the three clustering models. It depicts, for each model, the descriptive statistics of the clusters, including the sample size, the mean value of each metric representing the centroid of each cluster, and the last row of each model showing the engagement level pattern. By observing table 4, if the average value of a cluster metric exceeds the total average value, then an upward arrow is shown; otherwise, a downward arrow is shown. Except "Recency", a metric with an upward arrow means that in terms of that metric, the level of student engagement is satisfactory. But in the case of the "Recency" metric, an upward arrow indicates that the value of the latter is high, and according to its definition given in table 3 (Recency: Refers to the number of days from the last session date to the last day of study (May 31, 2020)), this indicates that the student did not use the platform for a long time.

Table 7. Clustering results of three models

Two-level clustering model K=2				
Cluster	Cluster_0	Cluster_1		
Simple size	375 (82.41%)	80 (17.58%)		
Mean(AvgTeamsSessionDay) M1	1.185	1.098		
Mean(NbrViewedVideos) M2	0.976	0.333		
Mean(VideoTime (Minutes)) M3	757.829	353.654		
Mean(ScreenShareTime (Minutes)) M4	135.059	70.148		
Mean(Recency) M5	3.746	39.160		
Pattern	M1↑,M2↑,M3↑,M4↑, M5↓	M1↓,M2↓,M3↓, M4↓, M5↑		
Three-level clustering model K=3				
Cluster	Cluster_0	Cluster_1	Cluster_2	
Simple size	367 (80.65%)	80 (17.58%)	8 (1.75%)	
Mean(AvgTeamsSessionDay) M1	1.187	1.099	1.109	
Mean(NbrViewedVideos) M2	0.984	0.338	0.5	
Mean(VideoTime (Minutes)) M3	646.308	290.388	6456	
Mean(ScreenShareTime (Minutes)) M4	65.834	30.725	3.696,875	
Mean(Recency) M5	3.785	39.175	6.25	
Pattern	M1↑,M2↑, M3↑,M4↑, M5↓	M1↓, M2↓, M3↓, M4↓, M5↑	M1↓, M2↓, M3↑, M4↑, M5↓	
Four-level clustering model K=4				
Cluster	Cluster_0	Cluster_1	Cluster_2	Cluster_3
Simple size	76(16.70%)	50(10.98%)	8 (1.75%)	321(70.54%)
Mean(AvgTeamsSessionDay) M1	1.170	1.058	1.109	1.189
Mean(NbrViewedVideos) M2	0.908	0.340	0.500	0.941
Mean(VideoTime (Minutes)) M3	459.934	70.720	6456.00	691.386
Mean(ScreenShareTime (Minutes)) M4	65.461	6.360	3696.875	66.436
Mean(Recency) M5	19.513	46.840	6.250	2.174
Pattern	M1↑, M2↑, M3↓, M4↓, M5↑	M1↓, M2↓, M3↓, M4↓, M5↑	M1↓, M2↓, M3↑, M4↑, M5↓	M1↑, M2↑, M3↑, M4↓, M5↓

4.1. Two-Level Clustering Model

According to table 7, looking at the results of the two-level model, most of the students are in cluster_0. This cluster represents a significant group of students with high indicators for metrics such as average sessions per day (M1), number of videos viewed (M2), video time (M3), and screen sharing time (M4), with low average recency (M5). This suggests a high level of recent engagement and active use of Microsoft Teams. Without a doubt, these 374 students are the most engaged and the most interactive with e-learning, defined as students with high engagement levels. By directly observing the "Pattern" row, students in cluster_1 are considered as unengaged students since the engagement metrics values are smaller compared with the total average values. This suggests a lower level of engagement and less frequent activity on the platform recently. Also, it can be seen that the metric that determines the average number of sessions per day is similar between these two clusters, which means that this metric does not contribute to identifying the students' engagement level, and on the other hand, this indicates that if the number of sessions opened by the student is high, this does not necessarily mean that the level engagement of this student's will also be high. This analysis further confirms the conclusion we have already reached is that the number of sessions opened by students does not reflect their real engagement level.

4.2. Three-Level Clustering Model

The results of the Three-level clustering model show a similar trend to that observed for the two-level model. In Cluster_0, students demonstrate relatively high values for average sessions per day, number of videos viewed, time spent on videos, and screen sharing. The recency of actions is low, indicating recent and ongoing engagement. Students in this cluster appear very engaged in using Microsoft Teams for their distance learning activities. Moving to Cluster_1, we observe insufficient values for all engagement metrics. The recency of actions is high, indicating older use of Microsoft Teams. Students in this cluster seem less involved and show lower engagement. Finally, Cluster_2 represents the major difference between the two-level clustering model and the three-level clustering model.

Key Features: Although the number of videos viewed and sessions per day are slightly lower than the average total value, this cluster stands out with very high values for screen sharing and time spent on videos. The average recency value is low. These students use Microsoft Teams extensively, with highly recent activity and consumption of video content and frequent use of screen sharing, which may indicate intense engagement in specific learning activities or group projects. Students in this cluster show extremely high levels of engagement, particularly in terms of time spent on videos and screen sharing. For this, these eight students will be designated as Super Engaged students. Their relentless involvement and passion for online learning are an example for the entire student community. A survey was conducted with these eight students to understand why they spend much time watching videos and why they use the screen sharing feature extensively. The results showed that these students are used to working together on projects, discussing and solving exercises in groups, and making presentations requested by their teachers during online sessions. All of these collaborative activities are recorded in video form, which generates additional educational resources for these students allowing them to review the content at their own pace, take breaks, and go back to clarify misunderstood points. This automatically increases the time spent by these students on the videos compared to other students.

4.3. Four-Level Clustering Model

The clusters provided by the four-level clustering model are characterized by by distinct patterns. Cluster_0 represents a medium engagement level where students demonstrate good participation in sessions and video watching, spending sufficient time on videos with limited screen sharing interaction. Cluster_1, indicates a low engagement level standing out for its extremely low values on the first four metrics. Regarding "Recency", this cluster has the highest value (46,840), suggesting older use of the Microsoft Teams platform. Undoubtedly, students in this cluster represent the students with the lowest level of engagement. Moving to Cluster_2, which exhibits extremely high engagement level, we find the same cluster that was identified with the three-level clustering model (Cluster_1), which is characterized by its small size and very distinct behavior due to exceptionally high values for video viewing and screen sharing time. Finnaly, Cluster_3 demonstrates high and regular engagement level, containing the largest number of students with values generally above the total average value, and the lowest "Recency" value (2.174). This means that students in this cluster show high activity on Microsoft Teams, with frequent participation in sessions, regular viewing of videos, and significant screen sharing. Their recency shows the most recent activity in Microsoft Teams

After describing the characteristics of each model, it is time to carry out a comparative study to identify which one best meets our objective.

4.4. Comparison

From figure 2, we see that the three models (k=2, k=3, and k=4) have very close Silhouette coefficients and above 0.7, which generally indicates good clustering and similar performance, with relatively small differences between them. Comparing the two-level model with the three-level model, we notice that in both models there is the same group (cluster_1: 80 students) representing low engagement students. Since our main goal is to identify these students, this similarity does not provide any new relevant information. Therefore, our comparative study will focus on the three-level model and the four-level model.

When comparing the Three-level engagement model versus Four-level engagement model, several similarities and differences emerge. In terms of similarities, both models identify a very small cluster of 8 students with exceptionally high values for video watching time and screen sharing (Cluster_2 in both cases). Additionally, both models feature a large group of engaged students (Cluster_3 in the 4-level model, Cluster_0 in the 3-level model), and both contain a group with lower engagement (Cluster_1 in both cases). Regarding differences, the three-level model merged Clusters_0 and Cluster_3 from the four-level model into one large group (Cluster_0), and the number of students in the low engagement group (cluster_1) varies between the two models. To determine which model best allows teachers to accurately identify unengaged students, several factors were considered. In terms of clarity of identification, both models clearly identify a group of less engaged students (cluster_1 in both cases), which stands out from the others with insufficient values on all measures. However, when examining accuracy of identification, the four-level model provides more precise identification of unengaged students, with only 50 students in its Cluster_1, targeting a more specific group of students most in need of interventions. Concerning the risk of over- or under-identification, the Three-level model, with its cluster of 80 less engaged students, could potentially include some students who are actually moderately engaged, which could lead to over-identification of 'unengaged' students.

Therefore, the answer to the first research question "What clustering model helps educational managers to identify unengaged students accurately ?" is that the four-level model is the most effective for this specific task. With its smallest group of 50 students, it likely targets the truly unengaged students, thereby reducing the risk of false positives (over-identification), and allowing teachers to focus their efforts on a smaller group of students who need help the most.

To better understand low student engagement and identify factors that could explain this phenomenon, we extracted demographic and educational factors from the APOGEE database. The description of these factors is presented in table 8. Thus, the different engagement levels obtained by the four-level model were analyzed according to these factors. The results of this analysis are presented in table 9.

Table 8. Demographic and educational data description

Feature Type	Feature Name	Description	Categories
Demographic	Living Standard	Economic situation based on parents' profession	High Medium Limited
	Gender	Sex of the student	F (Female) M (Male)
Educational	Study Duration	Number of years to reach the final year of bachelor's degree	3 years 4 years 5 years or more
	Online Service Usage	Frequency of using faculty's online services	Frequent Regular Infrequent

Table 9. Demographic and educational factors distribution for each engagement level group (in %)

Feature	Categories	Engagement levels %			
		Low	**Medium**	**High**	**Super**
Living Standard	Good	0.0	9.2	13.7	50.0
	Limited	66.0	38.2	38.0	37.5
	Medium	34.0	52.6	48.3	12.5
Gender	F	66.0	53.9	57.9	87.5
	M	34.0	46.1	42.1	12.5
Study Duration	3 years	38.0	44.7	47.0	50.0
	4 years	38.0	28.9	31.8	50.0
	5 years or more	24.0	26.3	21.2	0.0
Online Service Usage	Frequent	20.0	10.5	16.2	25.0
	Infrequent	50.0	40.8	37.4	12.5
	Regular	30.0	48.7	46.4	62.5
Cluster size		**50(11%)**	**76 (16.7%)**	**321(70.5%)**	**8(1.8%)**

Table 8 shows the percentage of each factor category at each engagement level. These percentages reveal several key trends. Regarding Living Standard, 66% of low-engaged students have a limited standard of living, compared to only 38-38.5% at Medium and High levels, and notably, the 'Low level' group does not have any students with a good standard of living. In terms of Gender, 66% of low-engagement students are women, a rate slightly higher than that of the Medium and High groups. Although more female students are in the low engagement group, this proportion is similar in the other groups, suggesting that gender is probably not a determining factor in low engagement. Concerning Study Duration, the 'Super level' group has the largest percentage of students who completed their bachelor's degree in just 3 years (50.0%), followed by 'High level' (47.04%), 'Medium level' (44.74%), and finally 'Low level' (38.0%), which indicates that the proportion of students finishing in 3 years decreases slightly with the drop in the engagement level. In the 'Low' group, only 38% of students took 3 years (normal duration) to reach the final year of their degree, while 38% took four years, signaling a one-year delay, and 24% took five years or more, indicating a significant delay. This means that 62% of low-engagement students took longer than normal to reach their final year of degree. Regarding Online Service Usage, frequent use of online services offered by the faculty is generally low, not exceeding 25% at all engagement levels. In contrast, infrequent use is highest among 'Low' level students (50.0%), followed by 'Medium' (40.8%), 'High' (37.04%), and 'Super Engaged' (12.5%). Regarding regular use, this increases with the level of engagement, reaching 62.5% among 'Super Engaged'.

Based on these observations, several potential factors for low engagement emerge. Limited standard of living is the most striking factor, as students with limited resources may have difficulty accessing adequate equipment or a stable internet connection. Infrequent use of online services suggests a lack of familiarity or interest in online services and digital tools in general. Study duration is also significant, as 62% of low-engagement students took longer than expected (3 years) to reach their final year, highlighting the relationship between students' failure to reach the final year within the normal duration and low engagement levels. This may be due to a significant percentage of students who do not achieve the required academic performance to complete their studies regularly also do not make sufficient effort in distance learning.

As noted above, adopting the four-level model is one that allows faculty leaders to accurately identify unengaged students. After analyzing these factors, several recommendations are proposed to improve student engagement. For immediate interventions, faculty should offer material support or access to equipment for students with a limited standard of living, organize practical workshops on the use of Microsoft Teams, especially for those who rarely use online services, and conduct a qualitative study to understand the specific reasons for delay in studies and its link to low engagement. For long-term strategic recommendations, the focus should be on developing partnerships with companies in the information technology sector for sustainable access to technologies, establishing a system based on the analysis of students' impressions of their academic and social experience, implementing periodic engagement surveys to track long-term trends, and conducting in-depth qualitative studies to understand the specific reasons for delay in studies and its link to low engagement.

The implementation of these recommendations should follow a phased approach. In the short term (first year), the focus should be on deploying immediate support measures and establishing the monitoring system setup to quickly address current engagement challenges. The medium-term phase (2-3 years) would concentrate on building institutional infrastructure and developing capacity through systematic training and resource development. Finally, the long-term vision (beyond 3 years) involves fully integrating these engagement strategies into institutional policies and practices, ensuring their sustainability and continuous evolution based on collected data and feedback. This staged implementation ensures both immediate impact and sustainable long-term effectiveness of the engagement support system.

Apart from the effectiveness demonstrated above, the proposed framework also maintains certain limitations. Our methodological framework focuses on online learning as an exclusive modality, and this specialization limits its applicability to hybrid learning environments where students can alternate between online and face-to-face modes. Additionally, the absence of quiz data and detailed discussion participation information creates a limitation, as this lack of data restricts our ability to perform comprehensive engagement analysis based on important metrics such as quiz performance and the quality and frequency of student interactions in discussions.

5. Conclusion and Future Works

The necessity to replace face-to-face teaching with distance learning during health emergencies such as the COVID-19 lockdown has revealed significant challenges. Preliminary analyses of engagement rates show differences among students in the same field, and that some students do not fully engage in this new mode of learning. The present work has investigated three research questions with the final aim of providing teachers and university administrators with a framework based on an unsupervised machine learning algorithm (k-means) aimed at identifying different levels of engagement and understanding the potential reasons for low engagement. The first research question focuses on identifying which Microsoft Teams engagement metrics are most relevant in the process of identifying engagement levels. The results show that including all engagement metrics during the clustering process can only obscure the discovery of the cluster structure. According to the "Sequential Forward Selection (SFS)" technique ScreenShareTime, VideoTime, NbrViewedVideos , Recency, and AvgTeamsSessionDay emerge as the most representative metrics for assessing student engagement levels. These metrics demonstrate the highest efficacy in distinguishing between different engagement levels among the student population. The second question is "What clustering model helps educational managers to identify unengaged students accurately?". Three distinct clustering models were analyzed: two-level, three-level, and four-level models. Experimental results demonstrated that all three models exhibit very similar Silhouette coefficients, exceeding 0.7. However, the four-level model outperforms the others in terms of accurately identifying low-engagement students without over-identifying this category. The third question examines whether demographic and educational data extracted from the faculty information system (APOGEE) can help university administrators understand the reasons for low engagement and identify the factors that influence it. The analysis of the four levels of engagement (Low, Medium, High, Super) based on factors Living Standard, Gender, Study Duration, and Online Service Usage revealed that low student engagement with Microsoft Teams is primarily associated with a limited standard of living, with 66% of unengaged students having a low income. This lack of engagement is also linked to unfamiliarity with digital tools, as 50% of unengaged students use online services infrequently, and to academic difficulties, with 62% of unengaged students taking longer than the normal duration to reach their final year.

To deepen and extend this study on student engagement, several research avenues focused on artificial intelligence can be envisaged. A significant future work involves combining Natural Language Processing (NLP) techniques with predictive methods such as neural networks or random forests to analyze sentiments and emotions expressed by students during their learning process. This combination could provide insights into engagement and enable early intervention for students at risk of disengagement. Another idea to explore is the implementation of anomaly detection algorithms to rapidly identify unusual changes in a student's engagement behaviors, thus allowing for prompt intervention. A further future research direction is to utilize reinforcement learning techniques to optimize intervention strategies aimed at improving engagement and evaluating their effectiveness in real time.

While our study found a strong link between low income and student disengagement, but a deeper exploration of exploration of the effects of demographic and living conditions factors effects on engagement is necessary. Future research should qualitatively examine how factors like internet quality, device access, quiet study spaces, and other environmental elements influence student engagement. Another promising research direction would be a comparative study of student engagement across different learning platforms, such as Microsoft Teams, Zoom, and Moodle. This would help understand how platform-specific features and design influence engagement patterns.

6. Declarations

6.1. Author Contributions

Conceptualization: R.A.D., A.A., K.A., and A.R.; Methodology: R.A.D.; Software: R.A.D., A.R.; Validation: R.A.D., A.A., and K.A.; Formal Analysis: R.A.D., A.A., and K.A.; Investigation: R.A.D.; K.A.; Resources: R.A.D., A.A. and A.R.; Data Curation: R.A.D.; Writing Original Draft Preparation: R.A.D., A.A., and K.A.; Writing Review and Editing: R.A.D., A.A., and K.A.; Visualization: R.A.D, A.R. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. G. Ruiz, M. J. Mintzer, and R. M. Leipzig, "The Impact of E-Learning in Medical Education," *Academic Medicine*, vol. 81, no. 3, pp. 207–212, 2006

[2] I. P. W. Ariawan, "Forms and Field Trials of a Digital Evaluation Tool: Integrating F-S Model, WP Method, and Balinese Local Wisdom for Effective E-Learning," *Journal of Applied Data Sciences*, vol. 5, no. 2, pp. 441–454, May 2024

[3] A. Akmal, "Predicting Dropout on E-learning Using Machine Learning," *Journal of Applied Data Sciences*, vol. 1, no. 1, pp. 29–34, Sep. 2020

[4] K. A. Dahleez, A. A. El-Saleh, A. M. Al Alawi, and F. Abdelmuniem Abdelfattah, "Higher education student engagement in times of pandemic: the role of e-learning system usability and teacher behavior," *International Journal of Educational Management*, vol. 35, no. 6, pp. 1312–1329, Jul. 2021

[5] E. Popescu, "Diagnosing Students' Learning Style in an Educational Hypermedia System," *IGI Global eBooks*, pp. 187–208, Jan. 2009

[6] 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 Computer Science*, vol. 163, pp. 550–559, Jan. 2019

ISSN 2723-6471

[7] R. Jeljeli, F. Farhi, S. Setoutah, and A. A. Laghouag, "Microsoft teams' acceptance for the e-learning purposes during Covid-19 outbreak: A case study of UAE," *International Journal of Data and Network Science*, vol. 6, no. 3, pp. 629–640, 2022

[8] D. K. Srinivasan, "Medical Students' Perceptions and an Anatomy Teacher's Personal Experience Using an e-Learning Platform for Tutorials During the Covid-19 Crisis," *Anatomical Sciences Education*, vol. 13, no. 3, pp. 318–319, May 2020

[9] S. R. Downes and T. Lykina, "Closing the Gap in Global Neurosurgical Education via Online Conference: A Pre-Covid Survey," *Cureus*, vol. 12, no. 5, pp. 1-12, May 2020.

[10] T. Favale, F. Soro, M. Trevisan, I. Drago, and M. Mellia, "Campus traffic and e-Learning during COVID-19 pandemic," *Computer Networks*, vol. 176, no. 7, pp. 1-9, Jul. 2020

[11] R. C. Chick et al., "Using Technology to Maintain the Education of Residents During the COVID-19 Pandemic," *Journal of Surgical Education*, vol. 77, no. 4, pp. 729-732, Apr. 2020

[12] K. N. Wea and A. Dua Kuki, "Students' Perceptions of Using Microsoft Teams Application in Online Learning During the Covid-19 Pandemic," *Journal of Physics: Conference Series*, vol. 1842, no. 1, pp. 1-7, Mar. 2021.

[13] Y.-M. Ng and P. L. P. Or, "Coronavirus disease (COVID-19) prevention: Virtual classroom education for hand hygiene," *Nurse Education in Practice*, vol. 45, no. 5, pp. 1-2, May 2020

[14] Oyeniran, A. Oluwashina, Oyeniran, T. Stella, Oyeniyi, O. Joshua, Ogundele, A. Rita, Ojo, O.e. Adeolu, "E-Learning: Advancement in Nigerian Pedagogy Amid Covid-19 Pandemic," International Journal of Multidisciplinary Sciences and Advanced Technology, vol 1, no. Special Issue1 Covid-19, pp. 73–79, May 2020.

[15] A.-J. Moreno-Guerrero, I. Aznar-Díaz, P. Cáceres-Reche, and S. Alonso-García, "E-Learning in the Teaching of Mathematics: An Educational Experience in Adult High School," *Mathematics*, vol. 8, no. 5, pp. 1-16, May 2020

[16] D. Clow, "An overview of learning analytics," Teaching in Higher Education, vol. 18, no. 6, pp. 683–695, Aug. 2013

[17] I. H. Jo, D. Kim, and M. Yoon, "Constructing Proxy Variables to Measure Adult Learners' Time Management Strategies in LMS.," Educational Technology & Society, vol. 18, no. 3, pp. 214–225, Jul. 2015.

[18] K. Aulakh, Rajendra Kumar Roul, and M. Kaushal, "E-learning enhancement through Educational Data Mining with Covid-19 outbreak period in backdrop: A review," *International Journal of Educational Development*, vol. 101, no. 9, pp. 102814–102814, Sep. 2023.

[19] J. Wang, L. Hai-yan, B. Cao, and Y. Zhao, "Application of educational data mining on analysis of students' online learning behavior," 2017 2nd International Conference on Image, Vision and Computing (ICIVC), vol. 2017, no. 6, pp. 1011-1015, Jun. 2017

[20] F. Ferri, P. Grifoni, and T. Guzzo, "Online Learning and Emergency Remote Teaching: Opportunities and Challenges in Emergency Situations," Societies, vol. 10, no. 4, pp. 1-18, Nov. 2020.

[21] P. Chapman, "CRISP-DM 1.0: Step-by-step data mining guide," Semantic Scholar, 2000.

[22] M. Saltan, S. Terzi, and E. U. Küçüksille, "Backcalculation of pavement layer moduli and Poisson's ratio using data mining," *Expert Systems with Application*s, vol. 38, no. 3, pp. 2600–2608, Mar. 2011

[23] H. Wiemer, L. Drowatzky, and S. Ihlenfeldt, "Data Mining Methodology for Engineering Applications (DMME)—A Holistic Extension to the CRISP-DM Model," Applied Sciences, vol. 9, no. 12, p. 2407, 2019.

[24] T. R. Sahama and P. R. Croll, "A data warehouse architecture for clinical data warehousing," *Privacy Enhancing Technologies*, vol. 2007, no. 1, pp. 227–232, Jan. 2007.

[25] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Student Engagement Level in e-Learning Environment: Clustering Using K-means," *American Journal of Distance Education*, vol. 34, no. 2, pp. 137–156, Mar. 2020

[26] V. Kovanović, S. Joksimović, D. Gašević, J. Owers, A.-M. Scott, and A. Woodgate, "Profiling MOOC Course Returners: How Does Student Behavior Change Between Two Course Enrollments?," Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S '16, vol. 2016, no. 4, pp. 269-272, Apr. 2016.

[27] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Engaging with massive online courses," Proceedings of the 23rd international conference on World wide web - WWW '14, vol. 2014 , no. 4, pp. 687-698, Apr. 2014.

[28] A. Kamath, A. Biswas, and V. N. Balasubramanian, "A crowdsourced approach to student engagement recognition in e-learning environments," IEEE Winter Conference on Applications of Computer Vision (WACV), vol.2016, no.3, pp. 1-9 Mar. 2016.

[29] Y.-Y. Kuo, J. Luo, and J. Brielmaier, “Investigating Students’ Use of Lecture Videos in Online Courses: A Case Study for Understanding Learning Behaviors via Data Mining,” In Proceedings of the International Conference on Web Based Learning, vol. 2015, no. 11, pp. 231–237, Nov.2015

[30] S. I. Rahman, A. Shadman, T.A. Fariha, M. Ammar, M. N. Mubasshirul Haque, S. Chellappan, N. Jannatun, “Unsupervised machine learning approach for tailoring educational content to individual student weaknesses, ” High-Confidence Computing, vol 4, no. 4, pp. 1-14, Dec. 2024,

[31] S. Wang, M. Li, N. Hu, E. Zhu, J. Hu, X. Liu, J. Yin“K-Means Clustering With Incomplete Data,” IEEE Access, vol. 7, no. 4, pp. 69162–69171, 2019

[32] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” *In Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, vol. 5, no. 1, pp. 281–297, Jan. 1967.

[33] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, “Analyzing undergraduate students’ performance using educational data mining,” *Computers & Education*, vol. 113, no. 10, pp. 177–194, Oct. 2017

[34] D. Zakrzewska, “Cluster Analysis in Personalized E-Learning Systems,” *Studies in computational intelligence*, vol. 252, no. 1, pp. 229–250, Jan. 2009

[35] M. Dash and Poon Wei Koot, “Feature Selection for Clustering,” Springer eBooks, pp. 1119–1125, Jan. 2009

[36] P. J. Rousseeuw, “Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, no. 0377–0427, pp. 53–65, Nov. 1987

[37] H. Li, C.-J. Li, X.-J. Wu, and J. Sun, “Statistics-based wrapper for feature selection: An implementation on financial distress identification with support vector machine,” *Applied Soft Computing*, vol. 19, no. 6, pp. 57–67, Jun. 2014

[38] M. Cents-Boonstra, A. Lichtwarck-Aschoff, E. Denessen, N. Aelterman, and L. Haerens, “Fostering student engagement with motivating teaching: an observation study of teacher and student behaviours,” *Research Papers in Education*, vol. 36, no. 6, pp. 754–779, 2021