E-Journal of Humanities, Arts and Social Sciences (EHASS)
ISSN – Online 2720-7722 | Print 2821-8949
Volume 4 Issue 12 November 2023 Special Issue pp 4-19
Available online at: https://noyam.org/journals/ehass/
DOI: https://doi.org/10.38159/ehass.20234122

Web-based Clustering Application for Determining and Understanding Student Engagement Levels in Virtual Learning Environments

Eli Nimy[1] & Moeketsi Mosia[1]
[1] Centre for Teaching, Learning, and Programme Development, Sol Plaatje University, Kimberley, South Africa.

ABSTRACT

The increasing use of virtual learning environments (VLEs) in recent years has transformed teaching and learning methods. Universities now combine VLEs with traditional classrooms to accommodate hybrid teaching and learning approaches. However, student engagement with VLEs varies, and universities lack the tools to effectively determine and analyse VLE engagement. Consequently, data-driven decision-making regarding VLE usage remains a challenge for universities. This study thus proposed a user-friendly web-based application, using a R shiny framework, to determine and understand student engagement levels in VLEs. In this study, two clustering methods, K-means and Gaussian Mixture Model (GMM) were compared, to identify the most effective method for the proposed application. The results indicated that GMM outperforms K-means by generating more accurate and comprehensive groupings of student engagement levels. One key advantage of the GMM method is its ability to capture uncertainty and provide probabilities of student membership in each level of engagement, which enhances its usefulness for decision-making. Furthermore, the GMM method achieves these outcomes efficiently, saving valuable learning time. This research holds significant implications for education by providing valuable guidance for the development of Educational Data Mining (EDM) applications. Universities can leverage these applications to gain deep insights into VLE usage and enhance their understanding of student engagement. By adopting this web-based application, educators and administrators can make informed decisions and tailor interventions to optimize student learning experiences within VLEs.

Correspondence
Eli Nimy
Email: eli.nimy@spu.ac.za

Publication History
Received 25th August, 2023
Accepted 12th October, 2023
Published online:
30th November, 2023

Keywords: *Virtual Learning Environments, Student Engagement, Clustering*

INTRODUCTION

Instructional methods (IM) within higher educational institutions vary, evolve, and change over time.[1] These shifts in IM can be attributed to various factors, including alterations in academic curricula, modifications in learning design strategies, and changes in academic educators, among others.[2] Such changes have the potential to introduce or reinforce new learning patterns among students, as their level of engagement with academic content, resources, and tools may need to fluctuate. A prime example of this dynamic can be observed during the COVID-19 pandemic.[3]

[1] Leonard Tetzlaff, Florian Schmiedek, and Garvin Brod, "Developing Personalized Education: A Dynamic Framework," *Educational Psychology Review* 33 (2021): 863–82.
[2] Rasmitadila Rasmitadila et al., "The Perceptions of Primary School Teachers of Online Learning during the COVID-19 Pandemic Period," *Journal of Ethnic and Cultural Studies* 7, no. 2 (2020): 90–109.
[3] Rasmitadila et al., "The Perceptions of Primary School Teachers of Online Learning during the COVID-19 Pandemic Period"; Daryl Zandvliet, "Towards Effective Learning Analytics for Higher Education: Returning Meaningful Dashboards to Teachers" (Vrije Universiteit, 2020).

During the peak of the pandemic, many non-essential professionals were required to work from home.[4] Consequently, higher educational institutions had to transition from traditional in-person learning to online learning.[5] This transition necessitated students to rely more heavily on VLE than ever before.[6] UNESCO estimated that approximately 1.5 billion students worldwide were unable to attend universities or schools during this period, with over 91% of students being affected by nationwide closures.[7] Consequently, both educators and students had to become more accustomed to using VLEs and other applications such as Microsoft Teams, Zoom, and Google Meet.[8]

A VLE is a web-based platform designed for educational purposes.[9] VLEs enable educators to disseminate learning materials, conduct student surveys and assessments, create collaborative glossaries, and manage grades, among other functions.[10] In addition to these educational services, VLEs also collect and store the online behavior data of each user, including both educators and students, resulting in extensive and diverse datasets.[11] This opens opportunities for analyzing and interpreting student behavior patterns within virtual learning environments.[12] One significant challenge associated with VLEs is the characterization of new student behavior patterns that emerge due to changes in instructional methods.[13]

Prior research has employed EDM techniques, specifically unsupervised learning algorithms, to characterize student behavior patterns within VLEs.[14] However, there have been limited efforts to apply unsupervised learning algorithms to web-based applications to characterize student learning behavior amidst evolving instructional methods. In this context, unsupervised learning algorithms refer to the use of clustering methods for knowledge discovery.[15] Clustering methods represent a collection of techniques that group observations in a way that similarities exist within the same group, while differences distinguish them from observations in other groups.[16] Conversely, EDM emerged as a response to the necessity of analyzing large and diverse datasets derived from educational systems.[17]

The use of clustering methods within EDM to characterize student learning behavior has demonstrated its effectiveness in discerning student engagement levels, providing personalized interventions for students, and offering insights to enhance the efficacy of IM within VLEs.[18] Consequently, these methods are increasingly being adopted in studies aiming to characterize student behavioral patterns, specifically in terms of engagement levels, within VLEs.

In summary, IM within Higher Educational Institutions are not static but rather dynamic, reflecting the varying levels of student engagement within VLEs.[19] Recognizing this dynamism in IM and student engagement on VLEs underscores the necessity for solutions capable of capturing this dynamic nature. Therefore, this research aims to develop a web-based application that employs cluster analysis to capture the fluctuations in

---

[4] Pradeep Sahu, "Closure of Universities Due to Coronavirus Disease 2019 (COVID-19): Impact on Education and Mental Health of Students and Academic Staff," *Cureus* 12, no. 4 (2020).
[5] Sumitra Pokhrel and Roshan Chhetri, "A Literature Review on Impact of COVID-19 Pandemic on Teaching and Learning," *Higher Education for the Future* 8, no. 1 (2021): 133–41.
[6] Zandvliet, "Towards Effective Learning Analytics for Higher Education: Returning Meaningful Dashboards to Teachers."
[7] Pokhrel and Chhetri, "A Literature Review on Impact of COVID-19 Pandemic on Teaching and Learning"; Zandvliet, "Towards Effective Learning Analytics for Higher Education: Returning Meaningful Dashboards to Teachers."
[8] B Gaikar Vilas et al., "An Impact of Covid-19 on Virtual Learning: The Innovative Study on Undergraduate Students of Mumbai Metropolitan Region," *Academy of Strategic Management Journal* 20 (2021): 1–19.
[9] Kamalesh Palani, Paul Stynes, and Pramod Pathak, "Clustering Techniques to Identify Low-Engagement Student Levels.," in *CSEDU (2)*, 2021, 248–57.
[10] Palani, Stynes, and Pathak, "Clustering Techniques to Identify Low-Engagement Student Levels."
[11] Gabriella Casalino, Giovanna Castellano, and Corrado Mencar, "Incremental and Adaptive Fuzzy Clustering for Virtual Learning Environments Data Analysis," in *2019 23rd International Conference Information Visualisation (IV)* (IEEE, 2019), 382–87.
[12] Luisa M Regueras et al., "Clustering Analysis for Automatic Certification of LMS Strategies in a University Virtual Campus," *IEEE Access* 7 (2019): 137680–90.
[13] Palani, Stynes, and Pathak, "Clustering Techniques to Identify Low-Engagement Student Levels."; Kun Liang et al., "Online Behavior Analysis-Based Student Profile for Intelligent E-Learning," *Journal of Electrical and Computer Engineering* 2017 (2017); Ashish Dutt et al., "Clustering Algorithms Applied in Educational Data Mining," *International Journal of Information and Electronics Engineering* 5, no. 2 (2015): 112.
[14] Palani, Stynes, and Pathak, "Clustering Techniques to Identify Low-Engagement Student Levels."; Liang et al., "Online Behavior Analysis-Based Student Profile for Intelligent E-Learning."
[15] Kelvin P. Murphy, "Introduction," in *Machine Learning: A Probabilistic Perspective*, 1st Edition (London, England: MIT Press, 2012), 1–2.
[16] Murphy, "Introduction."
[17] Casalino, Castellano, and Mencar, "Incremental and Adaptive Fuzzy Clustering for Virtual Learning Environments Data Analysis."
[18] Palani, Stynes, and Pathak, "Clustering Techniques to Identify Low-Engagement Student Levels."; Liang et al., "Online Behavior Analysis-Based Student Profile for Intelligent E-Learning"; Zehra Bilici and Durmuş Özdemir, "Data Mining Studies in Education: Literature Review for the Years 2014-2020," *Bayburt Eğitim Fakültesi Dergisi* 17, no. 33 (2022): 342–76.
[19] Casalino, Castellano, and Mencar, "Incremental and Adaptive Fuzzy Clustering for Virtual Learning Environments Data Analysis."

student engagement levels based on individual student behavior in VLEs. The objectives of this study align with this aim and include:

1. Identifying a clustering method for use in a web-based application, determined through considerations of clustering time, Silhouette coefficient, Calinski-Harabasz, and Davies Bouldin indexes.
2. Determining and understanding student engagement levels in a virtual learning environment using the most effective clustering method.
3. Recognizing IM and student characteristics associated with the identified student engagement levels.
4. Developing a clustering web-based application that adapts to the dynamic nature of virtual learning environment data.

LITERATURE REVIEW

Educational Data Mining in Web-based Educational Platforms

Web-based educational platforms were initially created for online learning, but presently, many colleges and universities are integrating them as a supplementary tool for in-person instruction.[20] These platforms are swiftly becoming integrated into higher education to improve student learning in diverse formats, including E-learning, VLEs, Massive Open Online Courses (MOOCs), and Learning Management Systems (LMS).[21] Learning Management Systems, in particular, prioritize the development of VLEs for educational purposes, so web-based platforms like Moodle, Blackboard, and Canvas can be regarded as VLEs as well.[22]

Numerous factors advocate the adoption of VLEs for educational purposes. VLEs offer flexibility concerning both time and space, promote resource reusability, and facilitate enhanced interaction between educators and students.[23] Additionally, VLE platforms enable functions like content management, curriculum mapping and planning, learner engagement and administration, communication, and collaboration, as well as real-time interaction between educators and students.[24] Among the array of services provided by VLE platforms, they also accumulate extensive and diverse data, including system logs documenting student activities within the platform (such as browsing time, login times, and click counts).[25] This data collection extends to personal information like user profiles and academic performance.[26] The escalating volume of data generated by VLEs frequently necessitates the extraction of valuable insights from this vast dataset.[27]

EDM represents a knowledge discovery approach designed to transform raw data from VLEs into valuable insights. Its primary objective is to assist Higher Educational Institutions in resource management enhancement, optimization of learning processes, and the refinement of instructional methods, including monitoring, evaluation, and personalization of teaching procedures.[28] EDM places a strong emphasis on developing data mining algorithms that can delve into educational data, uncover hidden patterns, and use these discoveries to make predictions and informed decisions within educational settings.[29] Applications of EDM encompass activities such as data clustering in education, the creation of e-learning systems, and predicting student dropouts and performance within VLEs. Among these applications, clustering stands out as the most widely employed.[30] EDM assumes a critical role within HEIs, evolving in response to the necessity of extracting value from the data generated on web-based educational platforms.

---

[20] Regueras et al., "Clustering Analysis for Automatic Certification of LMS Strategies in a University Virtual Campus."
[21] Palani, Stynes, and Pathak, "Clustering Techniques to Identify Low-Engagement Student Levels."
[22] Eli Nimy, Moeketsi Mosia, and Colin Chibaya, "Identifying At-Risk Students for Early Intervention—A Probabilistic Machine Learning Approach," *Applied Sciences* 13, no. 6 (2023): 3869.
[23] Regueras et al., "Clustering Analysis for Automatic Certification of LMS Strategies in a University Virtual Campus."
[24] Regueras et al., "Clustering Analysis for Automatic Certification of LMS Strategies in a University Virtual Campus."
[25] Zandvliet, "Towards Effective Learning Analytics for Higher Education: Returning Meaningful Dashboards to Teachers."
[26] Zandvliet, "Towards Effective Learning Analytics for Higher Education: Returning Meaningful Dashboards to Teachers."
[27] Casalino, Castellano, and Mencar, "Incremental and Adaptive Fuzzy Clustering for Virtual Learning Environments Data Analysis."
[28] Zandvliet, "Towards Effective Learning Analytics for Higher Education: Returning Meaningful Dashboards to Teachers"; Regueras et al., "Clustering Analysis for Automatic Certification of LMS Strategies in a University Virtual Campus"; Dutt et al., "Clustering Algorithms Applied in Educational Data Mining"; Bilici and Özdemir, "Data Mining Studies in Education: Literature Review for the Years 2014-2020."
[29] Bilici and Özdemir, "Data Mining Studies in Education: Literature Review for the Years 2014-2020."
[30] Marcos Wander Rodrigues, Seiji Isotani, and Luiz Enrique Zarate, "Educational Data Mining: A Review of Evaluation Process in the e-Learning," *Telematics and Informatics* 35, no. 6 (2018): 1701–17.

The Use of Clustering for Student Engagement Levels

Student learning behaviors within VLEs are primarily characterized by indicators linked to actions, specifically, information regarding students' activities in VLEs.[31] These action-based indicators are often presented in summarized formats, encompassing metrics such as clicks per session, the number of file downloads, session duration, login frequencies, artifact production quantities, and time allocated to specific tasks.[32] One can determine student engagement levels by examining these action-based indicators of interest.[33] Among these indicators, a frequently employed measure for gauging student engagement levels in VLEs is clicks per session or the aggregated sum of clicks.[34] The task of identifying student engagement levels within a VLE is commonly treated as an unsupervised machine learning task.[35] In unsupervised machine learning, the data lacks labels and contains only input information.[36]

Due to remarkable shifts in student learning patterns or the expansion of educational data, the composition of data within VLEs undergoes changes.[37] Typically, it becomes challenging to determine the precise level of engagement within VLEs.[38] The established standards for characterizing engagement levels, as exemplified by the COVID-19 pandemic, often lose their relevance. Consequently, the primary dataset available for constructing classification models to assess student engagement levels is the input data, which comprises of action-based indicators.[39] In such scenarios, the most suitable approach involves the application of clustering methods.[40] These methods excel in situations where only input data is accessible, as they aim to identify natural groupings through similarity metrics.[41]

METHODOLOGY

To determine and understand student engagement levels and create a web-based clustering application, the Knowledge Discovery in Database (KDD) methodology is adopted. The methodology involves the following steps: (a) data selection and understanding; (b) data pre-processing and transformation; (c) modelling; (d) evaluation; and (d) web-based application development. KDD is a widely employed methodology within.[42]

Data Selection and Understanding

This study relies on the Open University Learning Analytics Dataset (OULAD), which was developed by Kuzilek, Hlosta, and Zdrahal to facilitate research in EDM.[43] What sets this dataset apart from other educational datasets is its inclusion of demographic information combined with aggregated clickstream data, detailing student interactions within the VLE. This unique combination enables the analysis of student behavior as reflected in their actions within the VLE. The dataset encompasses a total of 22 distinct modules and covers 32,593 students for the years 2013 and 2014. Kuzilek, Hlosta, and Zdrahal meticulously constructed this dataset in compliance with the ethical and privacy guidelines of the Open University.[44] They rigorously anonymized the data to eliminate any personally identifiable information concerning the students. The structure and tables of the OULAD dataset are depicted in Figure 1 (a) and (b) respectively.

[31] Zandvliet, "Towards Effective Learning Analytics for Higher Education: Returning Meaningful Dashboards to Teachers."
[32] Zandvliet, "Towards Effective Learning Analytics for Higher Education: Returning Meaningful Dashboards to Teachers."
[33] Palani, Stynes, and Pathak, "Clustering Techniques to Identify Low-Engagement Student Levels."
[34] Palani, Stynes, and Pathak, "Clustering Techniques to Identify Low-Engagement Student Levels."; Jihyun Park et al., "Detecting Changes in Student Behavior from Clickstream Data," in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 2017, 21–30.
[35] Abdallah Moubayed et al., "Student Engagement Level in an E-Learning Environment: Clustering Using k-Means," *American Journal of Distance Education* 34, no. 2 (2020): 137–56.
[36] Murphy, "Introduction."
[37] Casalino, Castellano, and Mencar, "Incremental and Adaptive Fuzzy Clustering for Virtual Learning Environments Data Analysis."
[38] Moubayed et al., "Student Engagement Level in an E-Learning Environment: Clustering Using k-Means."
[39] Moubayed et al., "Student Engagement Level in an E-Learning Environment: Clustering Using k-Means."
[40] Moubayed et al., "Student Engagement Level in an E-Learning Environment: Clustering Using k-Means."
[41] Murphy, "Introduction."
[42] Palani, Stynes, and Pathak, "Clustering Techniques to Identify Low-Engagement Student Levels."; Regueras et al., "Clustering Analysis for Automatic Certification of LMS Strategies in a University Virtual Campus."
[43] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal, "Open University Learning Analytics Dataset," *Scientific Data* 4, no. 1 (2017): 1–8.
[44] Kuzilek, Hlosta, and Zdrahal, "Open University Learning Analytics Dataset."

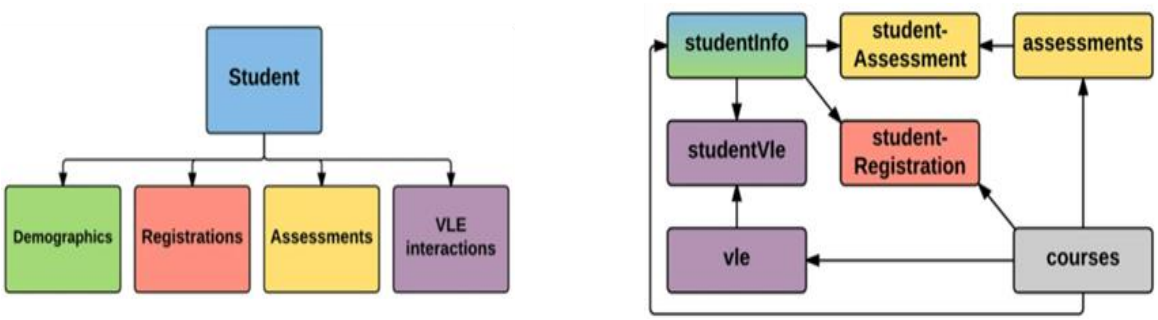

Figure 1. overall OULAD (left) structure and OULAD tables (right).[45]

The dataset is categorized into three distinct groups: student demographics, performance, and learning behavior, further subdivided into seven tables (Figure 1, on the right). As illustrated in Figure 1 (on the left), the dataset is centered around students, making it well-suited for this study, as a part of the research objective is to determine and understand student engagement levels. To assess student engagement levels, the focal metric is drawn from Table 1, specifically, the cumulative clicks recorded in the student Virtual Learning Environment (VLE). This metric represents the frequency of a student's interactions with the instructional materials presented within the VLE. Table 1 offers an overview of the dataset, including the associated quantity of observations.

Table 1. OULAD Description

Category	Table	Table Description	Observations
Demographic	Student Information	Contains demographic information about the students together with their results	32 593
Demographic	Student Registration	Contains information about the time when the student registered for the module presentation	32 593
Performance	Student Assessment	Contains the results of students' assessments	173 912
Performance	Courses	Contains the list of all available modules and their presentations	22
Performance	Assessments	Contains information about assessments in module presentations.	206
Learning Behavior	Student VLE	Contains information about each student's interactions with the materials in the VLE	10 655 280
Learning Behavior	VLE	Contains information about the available materials in the VLE	6 364

Source: Open University Learning Analytics dataset

Given that this research is structured as an unsupervised machine learning task, the training dataset will be denoted as $D = \{x_i\}_{i=1}^{N}$, where x represents the sum of clicks (input data) extracted from the student VLE table, N signifies the count of training instances, and D constitutes the training set. Within D, each observation indicates how frequently a student interacted with VLE materials (such as quizzes, forums, URLs, etc.). In the context of this study, the student engagement level(s) corresponds to cluster(s). The training data is presented in Table 2.

Table 2. Training data

Variable	Table	Variable Description	Observations
sum_click	Student VLE	Aggregated sum of clicks. Sum of clicks represents the number of times a student interacts with a VLE material in a day.	26 074

Source: Open University Learning Analytics dataset

[45] Kuzilek, Hlosta, and Zdrahal, "Open University Learning Analytics Dataset."

Data Pre-processing and Transformation

Data pre-processing and transformation represent crucial initial phases conducted prior to data modeling.[46] These procedures yield clean data, potentially improving the performance of machine learning models.[47] The action-based indicator of interest (sum of clicks) underwent the following data pre-processing and transformation steps. Initially, missing values and outliers were identified and addressed. Subsequently, the data underwent standardization to ensure that all observations were on a consistent scale. Standardization is mathematically defined as follows:

$$x_{std} = \frac{x - \mu_x}{\sigma_x} \quad (1)$$

Here, x denotes the observations of the indicator, while μ_x and σ_x represents the sample mean and standard deviation, respectively.

Modelling

Two distinct clustering methods, namely the GMM and K-means, were chosen for the clustering of the pre-processed and transformed input data x (sum of clicks).

Firstly, K-means, a conventional clustering method often employed in EDM, was included due to its simplicity in visualization and interpretation.[48] Its widespread use in the EDM community signifies its utility as a baseline for comparison and a straightforward means to characterize clusters. In contrast, the GMM was introduced as it represents a probabilistic model belonging to the soft clustering approach, which is the counterpart of hard clustering, to which K-means belongs.[49] In hard clustering, each data point belongs to exactly one cluster, while in soft clustering, like GMM, data points can belong to multiple clusters with associated probabilities. By incorporating probabilistic modelling, GMM offers a more nuanced understanding of complex data structures, which can be particularly advantageous when dealing with intricate and overlapping patterns, thereby enhancing the depth of the analysis.

To determine the optimal number of clusters, which directly translates to the number of engagement levels in this study, well-established methods were relied upon. The application of the elbow method for K-means aligns with recommendations from previous research solidifying its selection as a clustering option.[50] Additionally, the Bayesian Information Criterion (BIC) emerged as a robust measure for determining the optimal number of classes within GMMs, further reinforcing the choice of GMM for this analysis.[51]

Elbow Method

The Elbow method is employed to identify the optimal number of clusters in K-means through data visualization.[52] It involves locating a point on a plot where the distortion value experiences the most significant decline, resembling an elbow or bend. This point serves as an indicator for determining the appropriate number of clusters.[53] The distortion value is mathematically defined as follows:

$$D(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2 \quad (2)$$

Here, x_i represents a data point belonging to the cluster C_k, and μ_k is the mean value of the data points assigned to the cluster C_k.

Bayesian Information Criterion (BIC)

The Bayesian Information Criterion is an analytical method used to assess the goodness-of-fit of statistical models when compared to each other, given a specific dataset.[54] It also quantifies the model's ability to

[46] Regueras et al., "Clustering Analysis for Automatic Certification of LMS Strategies in a University Virtual Campus."
[47] Nimy, Mosia, and Chibaya, "Identifying At-Risk Students for Early Intervention—A Probabilistic Machine Learning Approach."
[48] Regueras et al., "Clustering Analysis for Automatic Certification of LMS Strategies in a University Virtual Campus."
[49] Regueras et al., "Clustering Analysis for Automatic Certification of LMS Strategies in a University Virtual Campus."
[50] Eva Patel and Dharmender Singh Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model," *Procedia Computer Science* 171 (2020): 158–67; Lalitha Agnihotri et al., "Mining Login Data for Actionable Student Insight.," *International Educational Data Mining Society*, 2015.
[51] Patel and Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model."
[52] Patel and Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model."
[53] P. Dangeti, "Unsupervised Learning," in *Statistics for Machine Learning*, 1st Edition (Birmingham, United Kingdom: Packt Publishing Ltd, 2017), 313–14.
[54] Patel and Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model."

generalize and accurately represent future data generated by the same process that produced the current data.[55] Generally, models with lower BIC scores are preferred. The BIC score is calculated as follows:

$$BIC = -2\log(\hat{L}) + \log(N)d \quad (3)$$

Here, N represents the number of data points, d signifies the number of parameters, and $\hat{L}$ is the maximum likelihood of the model.

K-means Clustering

K-means is a method for clustering that groups data points based on their proximity to a central point known as the cluster centroid.[56] In this study, K-means clustering is applied to analyze the observations in x and create distinct groupings referred to as student engagement levels (clusters). These clusters represent student VLE data with similar characteristics. Given that x is a numerical variable, each cluster is characterized by a centroid, which is essentially the mean of the sum of clicks within that cluster. To measure the similarity between student engagement levels, the squared Euclidean distance is employed, defined as follows:

$$d_{sq} = \sum_{i=1}^{D} (x_i - y_i)^2 \quad (4)$$

In this equation, x and y represent observations within the D-dimensional training dataset. The determination of the number of student engagement levels involved minimizing the Sum of Squared Errors (SSE), which comprises the squared error between each observation and its nearest centroid. The SSE is expressed as follows:

$$SSE = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{i,j} \left\|x_i - c_j\right\|^2 \quad (5)$$

Here, c_j represents the centroid of the j^{th} student engagement level, and $w_{i,j} = 0$ if an observation x_i does not belong to the student engagement level j, while $w_{i,j} = 1$ if x_i is part of the student engagement level j.

Gaussian Mixture Model

The Gaussian Mixture Model (GMM) is a clustering method that assigns data points to clusters in a probabilistic manner, with each cluster being characterized by a distinct Gaussian Distribution.[57] This Gaussian distribution is mathematically defined as:

$$N(X|\mu,\sum) = \frac{1}{(2\pi)^{\frac{D}{2}}\sqrt{|\sum|}} exp^{\{-\frac{(X-\mu)^T\sum^{-1}(X-\mu)}{2}\}} \quad (6)$$

In this equation, $|\sum|$ represents the determinant of the covariance matrix $\sum$, μ is a D-dimensional vector, and the shape of the Gaussian is determined by $\sum$, which is a $D \times D$ covariance matrix. Since each student engagement level is modeled as a Gaussian distribution (GD), the GMM can be represented as a linear combination of these fundamental Gaussian probability distributions, defined as:

$$p(X) = \sum_{K=1}^{k} \pi_K \; N(X|\mu_k,\sum_K) \quad (7)$$

Here, π_K denotes the mixing coefficient, which approximates the density of each Gaussian student engagement level, and K represents the number of student engagement levels in the mixture model (MM). The student engagement level within the MM is characterized by $N(X|\mu,\sum_K)$, which is the Gaussian density. Ultimately, each student engagement level K is expressed as a GD with covariance $\sum_K$, mean μ_k and π_K serving as the mixing coefficient.

[55] Patel and Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model."
[56] Patel and Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model."
[57] Patel and Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model."

Evaluation

The evaluation of the two clustering methods, namely GMM and K-means, was based on their execution time and three cluster evaluation metrics: the Calinski-Harabasz index, the Davies-Bouldin index, and the Silhouette coefficient. These cluster metrics gauge the degree of separation between clusters and the variation within each cluster,[58] while the execution time quantifies the duration it takes for the clustering method to fit the training data.

Calinski-Harabasz Index

The Calinski-Harabasz index score is a measure of the ratio between the sum of within-cluster dispersion and between-cluster dispersion for all clusters. A higher score indicates a model with well-defined clusters.[59] Mathematically, the Calinski-Harabasz index score is expressed as:

$$CH = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k - 1} \qquad (8)$$

Here, E represents a dataset of size n_E grouped into k clusters. The parameters $tr(W_k)$ and $tr(B_k)$ denote the trace of within-cluster and between-cluster dispersion matrices, respectively. These matrices are defined as follows:

$$B_k = \sum_{q=1}^{k} n_q \, (c_q - c_E)(c_q - c_E)^T \qquad (9)$$

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T \qquad (10)$$

Where n_q is the number of points in cluster q, c_q represents the center of cluster q, c_E is the center of E, and $c_q$is the set of points in cluster q.

Davies Boudin index

The Davies-Bouldin index metric calculates the average similarity between each cluster C_i for $i = 1, \ldots, k$ and its most similar one C_j. Values closer to zero indicate better partitioning of clusters [24]. It is defined as:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max R_{ij} \;\; \{i \neq j\} \qquad (11)$$

Where R_{ij} is a similarity measure that considers the trade-off between s_i and d_{ij}, calculated as:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \qquad (12)$$

Here, d_{ij} represents the distance between cluster centroids i and j, and s_i is the average distance between the centroid of cluster i and each point in cluster i.

Silhouette Coefficient

The silhouette coefficient computes the average distance between data points to assess the density and separation of clusters. It ranges from -1 to 1, with values closer to 1 indicating appropriate cluster configuration.[60] The formula for the Silhouette coefficient is:

$$SC = \frac{b - a}{\max(a, b)} \qquad (13)$$

[58] Palani, Stynes, and Pathak, "Clustering Techniques to Identify Low-Engagement Student Levels."
[59] Tadeusz Caliński and Jerzy Harabasz, "A Dendrite Method for Cluster Analysis," *Communications in Statistics-Theory and Methods* 3, no. 1 (1974): 1–27.
[60] Peter J Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics* 20 (1987): 53–65.

Where b is the average distance between a data point and all other data points in the nearest neighboring cluster, and a is the average distance between a data point and all other data points in the same cluster.

Web-based Application Development

Following the identification of a clustering method with favorable results during the evaluation phase, a web-based application was developed using the Shiny framework. Shiny is an open-source library within the R programming language, offering a robust web framework for creating interactive web applications.[61] A Shiny application is composed of two fundamental components: a user interface (UI) and a server. The server acts as the backend of the application, housing a set of instructions for executing tasks such as data processing, cluster model creation, and data visualizations, among others. On the other hand, the UI serves as the front end, encompassing instructions for presenting results to users within a web browser.[62] The server segment of the Shiny application was programmed to provide the following interactive functionalities:

- Offer a step-by-step user guide tutorial on clustering VLE data and comprehending cluster analysis.
- Present a user information modal on application tabs.
- Enable users to filter VLE data and select the desired number of clusters (engagement levels).
- Generate visualizations and tabular outputs regarding student engagement levels based on user selections in cluster analysis.
- Generate interactive visualizations that offer insights into IM and student characteristics associated with each student engagement level.
- Allow users to produce a downloadable report delivering insights into student engagement levels derived from cluster analysis.
- Permit user(s) to update cluster analysis with new inputs.

The UI aspect of the Shiny application was designed to display results and insights generated from the cluster analysis to the user(s).

FINDINGS AND DISCUSSION

Data Pre-processing and Transformation

The sum of clicks, extracted from the student VLE table, was employed as an action-based indicator to depict student engagement. Descriptive details regarding the sum of clicks are presented in Tables 3 and 4, both before and after aggregation.

Table 3. Sum of clicks information before and after aggregation.

Sum of Clicks	Observations	Variable size	Missing Value %
Before Aggregation	10 655 280	40.6 MB	0%
After Aggregation	26 074	101.9 KB	0%

To prevent student redundancy and alleviate computational load, the sum of clicks per student since the beginning of the semester was aggregated (summed).

Table 4. Sum of clicks statistics before and after aggregation.

Sum of Clicks	Mean	SD	Min	Max
Before Aggregation	3.72	8.85	1	6 977
After Aggregation	1 518.95	1 936	1	28 615

Table 4 illustrates a consistent minimum engagement of 1 before and after aggregation, indicating that some students remained disengaged throughout the semester. The aggregated sum of clicks underwent standardization, following Equation 1, in preparation for k-means clustering during modeling.

[61] Rachma Hermawati and Imas Sukaesih Sitanggang, "Web-Based Clustering Application Using Shiny Framework and DBSCAN Algorithm for Hotspots Data in Peatland in Sumatra," *Procedia Environmental Sciences* 33 (2016): 317–23.
[62] Hermawati and Sitanggang, "Web-Based Clustering Application Using Shiny Framework and DBSCAN Algorithm for Hotspots Data in Peatland in Sumatra."

Student Engagement Clustering

Before determining the distinct engagement levels, it was essential to establish the optimal number of engagement levels for K-means and GMM. The optimal number of engagement levels for K-means and GMM is depicted in Figure 2.

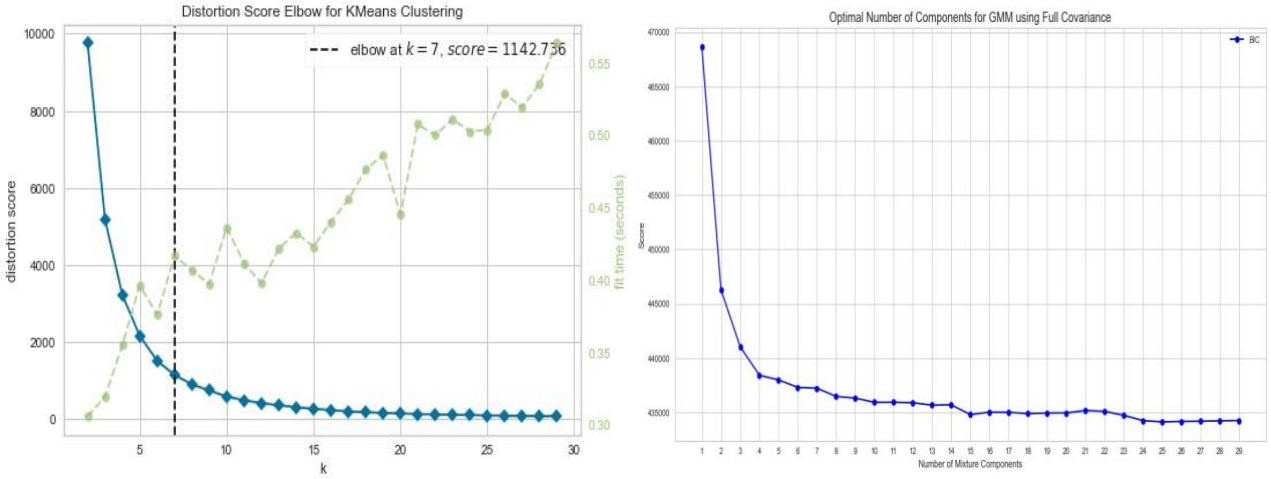

Figure 2. Elbow method for K-means (left) and BIC scores for GMMs (right)

In Figure 2 (left), the graph reveals a significant reduction in distortion until it reaches 7, after which it stabilizes at a constant distortion value. Consequently, the optimal number of engagement levels for K-means is determined to be 7. By analyzing Figure 2 (right), it becomes apparent that the optimal number of engagement levels for GMM is 25, as the lowest BIC score is associated with 25 components. Both K-means and GMM were modeled with 7 and 25 engagement levels, facilitating a fair comparison of engagement level means. These means signify the average number of times a student engages with the VLE, indicating that students within the same engagement level exhibit similar VLE engagement patterns.

The engagement level means, as presented in Table 5, have been arranged in ascending order to enhance interpretability. Consequently, lower engagement levels correspond to reduced engagement, while higher levels signify heightened engagement. The engagement level means, as derived from K-means for 7 and 25 levels, fall within the ranges of [299, 16,330] and [79, 26,602], respectively. For GMM, these values span [276, 13,153] and [49, 26,570] for 7 and 25 levels. Notably, these ranges align closely with the minimum and maximum values of Sum of Clicks (Table 4) when modeling more engagement levels for both methods.

Table 5. Engagement Level (Cluster) - Sum of Clicks.

Method	Number of Engagement Levels	Engagement Level Means
K-means	7	299, 1 187, 2 418, 4 014, 6 124, 9 444, 16 330
GMM	7	276, 1 083, 2 430, 4 138, 6 208, 8 934, 13 153
K-means	25	79, 296, 524, 765, 1 039, 1 337, 1 661, 2 028, 2 409, 2 818, 3 269, 3 774, 4 292, 4 828, 5 432, 6 189, 7 065, 7 993, 9 201, 10 617, 12 280, 14 573, 16 898, 19 859, 26 602
GMM	25	49, 199, 371, 557, 759, 996, 1 266, 1 579, 1 969, 2 381, 2 841, 3 348, 3 905, 4 523, 5 181, 5 953, 6 668, 7 519, 8 677, 9 790, 11 335, 13 251, 14 950, 18 721, 26 570

It's worth noting that GMM, with its optimal number of engagement levels substantially higher than K-means, adeptly captures a broader spectrum of engagement patterns, including both minimal and maximal engagement. Consequently, GMM offers the potential for more comprehensive groupings of VLE engagement.

Student Engagement Clustering Evaluation

The K-means and GMM methods were compared using execution time and three clustering metrics. The results for both clustering methods are presented in Table 6. Notably, the GMM method demonstrated superior performance in terms of execution time for the OULAD dataset. This suggests that GMM is notably more efficient in the task of categorizing 26,074 students into 7 and 25 engagement levels compared to K-means, which require more time. However, it's worth noting that despite the execution time of K-means being slightly longer, it excels in creating well-separated engagement levels, as evident from the higher Calinski-Harabasz index scores. When modeling 7 engagement levels, K-means proves to be particularly effective at generating well-defined engagement levels, boasting a slightly higher Silhouette Coefficient and a lower Davies-Bouldin Index score. This trend is however not consistent with 25 engagement levels, as both K-means and GMM exhibit similar Silhouette coefficients and Davies-Bouldin index scores. As demonstrated in Table 5, a higher number of engagement levels accurately captures the full range of VLE engagement, further emphasizing the importance of a method capable of efficiently establishing well-defined and separated engagement levels.

Table 6. Comparative analysis of clustering methods.

Method	Engagement Levels	Execution Time	Calinski-Harabasz	Silhouette Coefficient	Davies-Bouldin
K-means	7	0.41s	94 785	0.59	0.51
GMM	7	0.19s	90 314	0.58	0.52
K-means	25	0.53s	310 359	0.54	0.50
GMM	25	0.38s	294 330	0.54	0.50

Considering the lack of a clear winner between K-means and GMM using the objective approach, the GMM method was ultimately chosen for implementation in the web-based application. This choice was driven by several factors, including its probabilistic nature and its capacity to provide comprehensive groupings of VLE engagements. The probabilistic nature of GMMs allows for modeling the uncertainty associated with the number of engagement levels selected and the probability of a student belonging to each engagement level.[63] This capability addresses the complexities arising from students potentially belonging to multiple engagement levels.

Web-based Clustering Application

A web-based clustering application using the Gaussian Mixture Model (GMM) was developed through the R Shiny framework. The GMM clustering application is accessible online via the following link: https://ds-analytics.shinyapps.io/Student-Segmentation/. The application comprises three distinct tabs: (1) the GMM Data Analysis tab, (2) the IM tab, and (3) the Student Characteristics tab. Additionally, an informative modal feature has been incorporated, offering users detailed insights into each tab's functionality, as visually depicted in Figure 3.

[63] Bradley Boehmke, "Model-Based Clustering," in *Hands-On Machine Learning with R* (https://bradleyboehmke.github.io/HOML/model-clustering.html, 2022), https://bradleyboehmke.github.io/HOML/model-clustering.html.

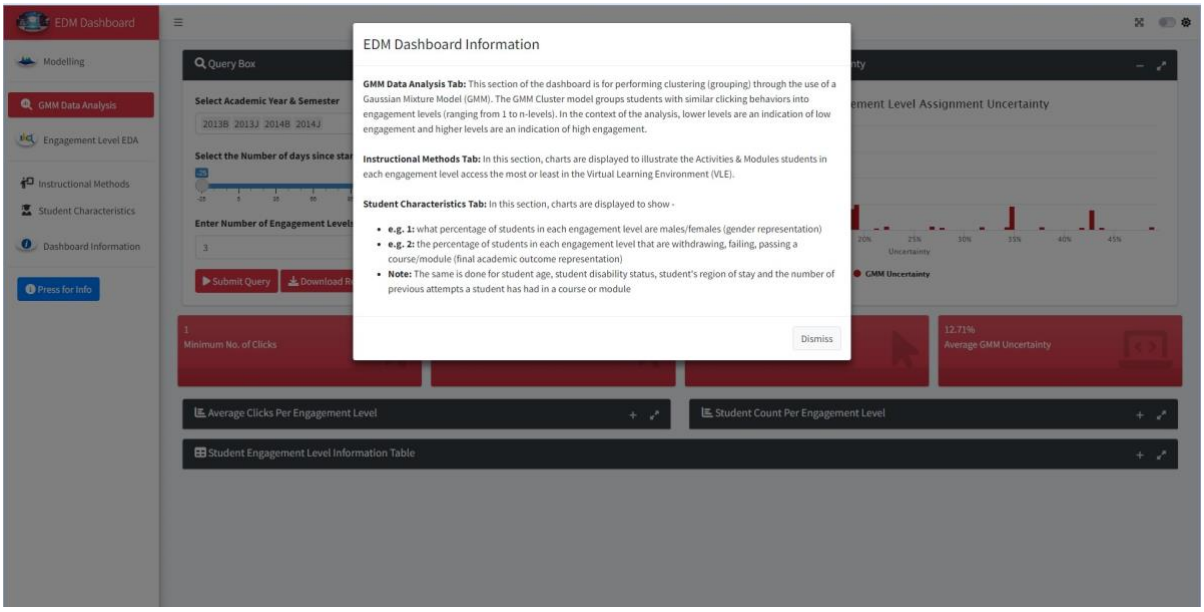

Figure 3. Web-based application with a modal

GMM Data Analysis Tab

Upon initial loading, the application defaults to the GMM Data Analysis tab, where users can engage in cluster analysis of student VLE data using the GMM method. This tab is equipped with a user guide feature designed to offer step-by-step instructions on performing GMM cluster analysis and interpreting the resultant outcomes. This instructional element has been integrated to assist users who may not possess technical expertise.

As illustrated in Figure 4, the query box prompts users to input specific parameters, including the academic year and semester, a range for the number of days from the start to the end of a course, and the desired number of engagement levels. Various combinations of these inputs empower the GMM method to unveil diverse engagement patterns, allowing for the dynamic nature of student VLE data to be captured effectively. This is particularly valuable as it enables continuous querying of new VLE data, facilitating the extraction of fresh insights as they become available.

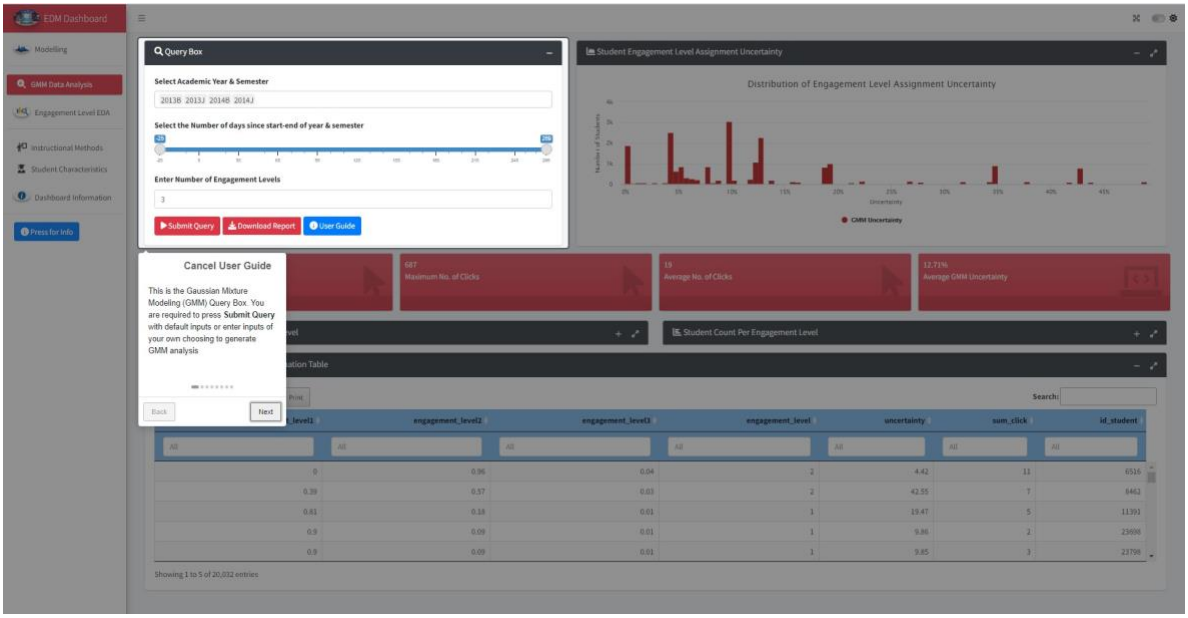

Figure 4. Web-based application with a user guide pop-up

Upon completion of the query box, Figure 5 provides users with five critical pieces of information. It includes a depiction of the distribution of engagement level assignment uncertainty, offering insights into the level of uncertainty associated with the GMM assignment of students to engagement levels. Additionally, a

statistical summary of student engagement on the VLE is presented, featuring the overall average GMM uncertainty.

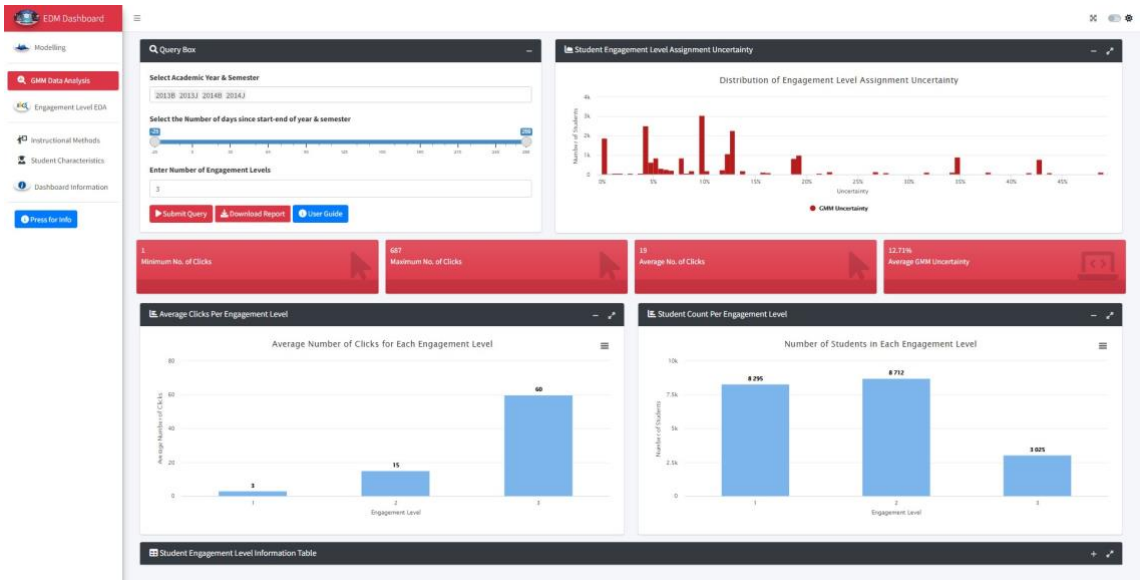

Figure 5. GMM data analysis tab

To visually represent student engagement data, two bar charts are included, illustrating the average interaction frequency of students within each engagement level with the VLE (left bar chart) and the distribution of students across the different engagement levels (right bar chart). Lastly, a table is presented, featuring columns that display: (1) the probability of a student's affiliation with each engagement level, (2) the specific engagement level to which a student belongs, (3) the percentage of uncertainty regarding a student's engagement level assignment, and (5) the count of VLE interactions undertaken by a student during the selected academic period.

Instructional Methods Tab

The charts in Figure 6 are presented to visually represent the activities and modules that students within each engagement level predominantly engage with, as well as those they engage with less frequently on the VLE. This analysis aids in the identification of the most frequently accessed modules and activities within the VLE.

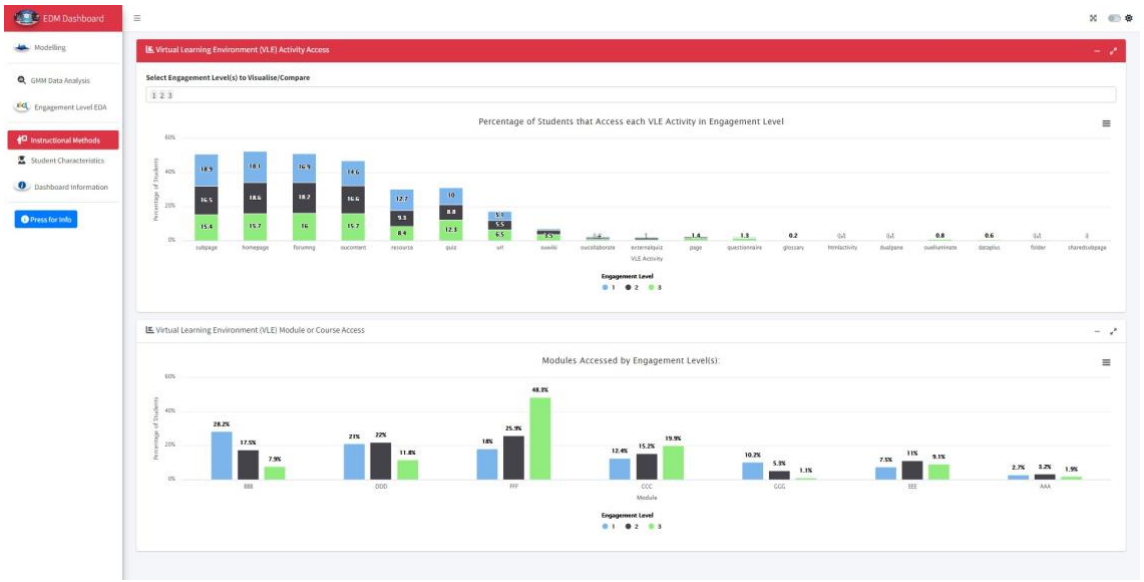

Figure 6. Instructional methods tab

Student Characteristics Tab

Additional exploration can be conducted based on the profiling of engagement levels, examining various dimensions of student data such as (1) gender, (2) age group, (3) disability status, (4) previous course attempts, (5) final academic performance, and (6) geographic region of residence. This extended analysis aims to provide deeper insights into the characteristics of students within each engagement level. Figure 7 illustrates the distribution of engagement levels across these different student attributes using bar charts.

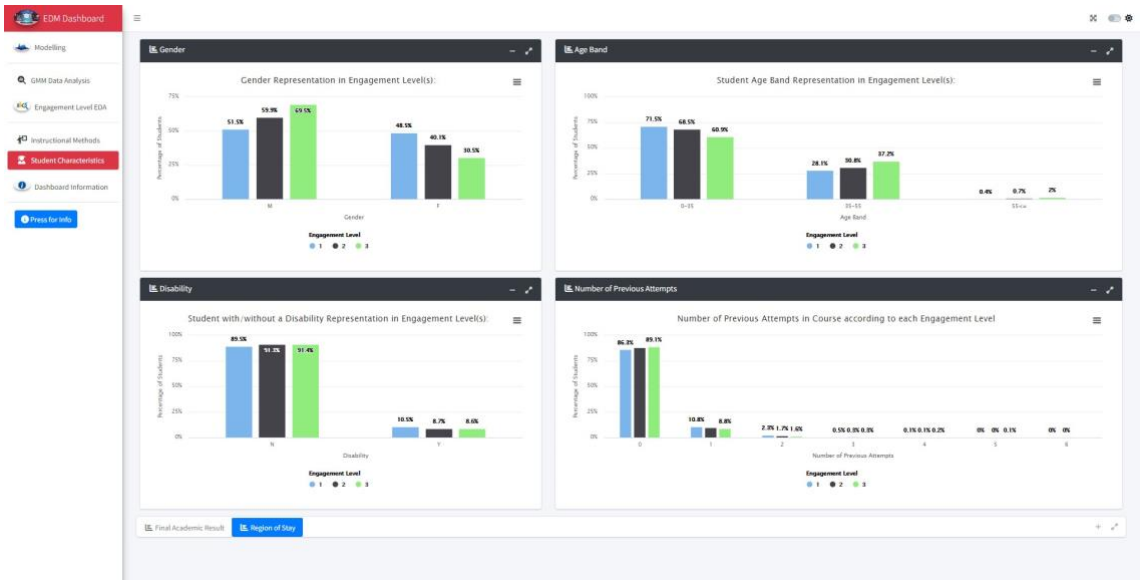

Figure 8. Student Characteristics tab

RECOMMENDATIONS

This research emphasizes the utilization of the Gaussian Mixture Model (GMM) as the preferred method for integration into web-based applications designed to analyze student engagement levels within VLEs. The choice of GMM is substantiated by its probabilistic nature, efficiency in clustering time, and its capacity to offer comprehensive characterizations of student engagement.

A key recommendation arising from this study is the determination of student engagement levels into three primary categories: low-engaged (level 1), mid-engaged (level $\frac{n}{2}$), and high-engaged (level n). This categorization simplifies the process of distinguishing between students with varying degrees of engagement in VLEs, providing a practical framework for educators and institutions.

In addition to determining engagement levels, it is advised to expand the analysis by incorporating various facets of student information. This may include considering the types of VLE activities and modules accessed, demographic factors such as gender, age group, disability status, historical course attempts, final academic performance, and geographic region of residence. Examining the distribution of engagement levels across these dimensions can yield valuable insights into instructional strategies and student characteristics.

Given the dynamic nature of VLE data, it is suggested that web-based clustering applications incorporate query boxes. This feature enables academic institutions to proactively extract fresh insights into student engagement over time. By facilitating ongoing exploration of engagement patterns, institutions can make timely adjustments to improve the learning experience.

Future research directions should focus on combining the GMM method with interpretable machine learning models. This approach goes beyond descriptive analytics and aims to provide diagnostic analytics for students within each engagement level. This would enable educators to gain deeper insights into the factors influencing student engagement and take targeted actions.

While this study primarily employed K-means and GMM, there is a call for future studies to explore a broader range of clustering techniques. These may include Latent Class Analysis, DBSCAN, K-medians, Mean-shift, K-prototyping, Fuzzy C-means, and Hierarchical clustering. Diversifying the clustering methods can offer alternative perspectives on student interactions within VLEs.

Lastly, there is an opportunity for research to delve into the application of cloud platforms within the domain of EDM. Investigating best practices for deploying EDM applications on cloud infrastructure can

enhance scalability, accessibility, and the overall effectiveness of using VLE data for educational improvement. This avenue holds the potential to transform the landscape of data-driven decision-making in education.

CONCLUSION

This research showcases the creation of a web-based clustering application tailored for the purpose of determining and understanding student engagement levels within a virtual learning environment (VLE). The findings and insights presented herein stand as a valuable asset to educational institutions, practitioners in the field of EDM and researchers. This research not only exemplifies the practical application of EDM principles but also provides essential guidance for the development and implementation of EDM applications using VLE data. Its contributions serve to enrich the landscape of data-driven decision-making in education, fostering improved learning experiences and outcomes for students.

ACKNOWLEDGEMENTS

The support of the DSI-NICIS National e-science Postgraduate Teaching and Training Platform (NEPTTP) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author(s) and are not necessarily to be attributed to the NEPTTP.

BIBLIOGRAPHY

Agnihotri, Lalitha, Ani Aghababyan, Shirin Mojarad, Mark Riedesel, and Alfred Essa. “Mining Login Data for Actionable Student Insight.” *International Educational Data Mining Society*, 2015.

Bilici, Zehra, and Durmuş Özdemir. “Data Mining Studies in Education: Literature Review for the Years 2014-2020.” *Bayburt Eğitim Fakültesi Dergisi* 17, no. 33 (2022): 342–76.

Boehmke, Bradley. “Model-Based Clustering.” In *Hands-On Machine Learning with R*. https://bradleyboehmke.github.io/HOML/model-clustering.html, 2022. https://bradleyboehmke.github.io/HOML/model-clustering.html.

Caliński, Tadeusz, and Jerzy Harabasz. “A Dendrite Method for Cluster Analysis.” *Communications in Statistics-Theory and Methods* 3, no. 1 (1974): 1–27.

Casalino, Gabriella, Giovanna Castellano, and Corrado Mencar. “Incremental and Adaptive Fuzzy Clustering for Virtual Learning Environments Data Analysis.” In *2019 23rd International Conference Information Visualisation (IV)*, 382–87. IEEE, 2019.

Dangeti, P. “Unsupervised Learning.” In *Statistics for Machine Learning*, 1st Edition., 313–14. Birmingham, United Kingdom: Packt Publishing Ltd, 2017.

Dutt, Ashish, Saeed Aghabozrgi, Maizatul Akmal Binti Ismail, and Hamidreza Mahroeian. “Clustering Algorithms Applied in Educational Data Mining.” *International Journal of Information and Electronics Engineering* 5, no. 2 (2015): 112.

Gaikar Vilas, B, M Joshi Bharat, B Jaywant, N S Mhatre, K Chitra, S Cheriyan, and G Rane Caroleena. “An Impact of Covid-19 on Virtual Learning: The Innovative Study on Undergraduate Students of Mumbai Metropolitan Region.” *Academy of Strategic Management Journal* 20 (2021): 1–19.

Hermawati, Rachma, and Imas Sukaesih Sitanggang. “Web-Based Clustering Application Using Shiny Framework and DBSCAN Algorithm for Hotspots Data in Peatland in Sumatra.” *Procedia Environmental Sciences* 33 (2016): 317–23.

Kuzilek, Jakub, Martin Hlosta, and Zdenek Zdrahal. “Open University Learning Analytics Dataset.” *Scientific Data* 4, no. 1 (2017): 1–8.

Liang, Kun, Yiying Zhang, Yeshen He, Yilin Zhou, Wei Tan, and Xiaoxia Li. “Online Behavior Analysis-Based Student Profile for Intelligent E-Learning.” *Journal of Electrical and Computer Engineering* 2017 (2017).

Moubayed, Abdallah, Mohammadnoor Injadat, Abdallah Shami, and Hanan Lutfiyya. “Student Engagement Level in an E-Learning Environment: Clustering Using k-Means.” *American Journal of Distance Education* 34, no. 2 (2020): 137–56.

Murphy, Kelvin P. “Introduction.” In *Machine Learning: A Probabilistic Perspective*, 1st Edition., 1–2. London, England: MIT Press, 2012.

Nimy, Eli, Moeketsi Mosia, and Colin Chibaya. “Identifying At-Risk Students for Early Intervention—A Probabilistic Machine Learning Approach.” *Applied Sciences* 13, no. 6 (2023): 3869.

Palani, Kamalesh, Paul Stynes, and Pramod Pathak. “Clustering Techniques to Identify Low-Engagement Student Levels.” In *CSEDU (2)*, 248–57, 2021.

Park, Jihyun, Kameryn Denaro, Fernando Rodriguez, Padhraic Smyth, and Mark Warschauer. "Detecting Changes in Student Behavior from Clickstream Data." In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 21–30, 2017.

Patel, Eva, and Dharmender Singh Kushwaha. "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model." *Procedia Computer Science* 171 (2020): 158–67.

Pokhrel, Sumitra, and Roshan Chhetri. "A Literature Review on Impact of COVID-19 Pandemic on Teaching and Learning." *Higher Education for the Future* 8, no. 1 (2021): 133–41.

Rasmitadila, Rasmitadila, Rusi Rusmiati Aliyyah, Reza Rachmadtullah, Achmad Samsudin, Ernawulan Syaodih, Muhammad Nurtanto, and Anna Riana Suryanti Tambunan. "The Perceptions of Primary School Teachers of Online Learning during the COVID-19 Pandemic Period." *Journal of Ethnic and Cultural Studies* 7, no. 2 (2020): 90–109.

Regueras, Luisa M, María Jesús Verdú, Juan-Pablo De Castro, and Elena Verdu. "Clustering Analysis for Automatic Certification of LMS Strategies in a University Virtual Campus." *IEEE Access* 7 (2019): 137680–90.

Rodrigues, Marcos Wander, Seiji Isotani, and Luiz Enrique Zarate. "Educational Data Mining: A Review of Evaluation Process in the e-Learning." *Telematics and Informatics* 35, no. 6 (2018): 1701–17.

Rousseeuw, Peter J. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics* 20 (1987): 53–65.

Sahu, Pradeep. "Closure of Universities Due to Coronavirus Disease 2019 (COVID-19): Impact on Education and Mental Health of Students and Academic Staff." *Cureus* 12, no. 4 (2020).

Tetzlaff, Leonard, Florian Schmiedek, and Garvin Brod. "Developing Personalized Education: A Dynamic Framework." *Educational Psychology Review* 33 (2021): 863–82.

Zandvliet, Daryl. "Towards Effective Learning Analytics for Higher Education: Returning Meaningful Dashboards to Teachers." Vrije Universiteit, 2020.

ABOUT AUTHOR(S)

Eli Nimy is currently pursuing a Master of Science degree in eScience while concurrently holding a part-time position as a Junior Data Scientist at the Centre for Teaching, Learning, and Programme Development at Sol Plaatje University. Eli's academic and professional interest centers on the intersection of Data Science and Bayesian Statistics, with a particular passion for their applications in the fields of Education and Healthcare.

Dr. Moeketsi Mosia is currently the Director of the Centre for Teaching, Learning, and Programme Development at Sol Plaatje University. In addition to his administrative role, he also serves as a lecturer for postgraduate students in the Department of Computer Science, Information Technology, and Data Science. Dr. Mosia's research interests primarily revolve around Learning Analytics, Probabilistic Machine Learning, and Student Learning in Higher Education.