A logo of a person holding a circle

Description automatically generated

Journal of Computers for Society

Journal homepage: https://ejournal.upi.edu/index.php/JCS

*Journal of Computers for Society x*(x) (2024) xxx-xxx

Uncovering Success Patterns in Online Education: A Data-Driven Analysis Using Principal Component Analysis and K-means Clustering

*Leena Ardini Abdul Rahim1,\*, Salal Ahmad2, Nor Syazwani Mohd Pakhrudin3 , Azka Naufal Nurrahman4*

1College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Merlimau, Malacca, Malaysia

2Department of Computer Science, University of Swabi, Swabi, Pakistan

3 School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia

4 Computer Science, Universitas Pendidikan Indonesia, Indonesia

\*Correspondence: E-mail: leenaardini120100@gmail.com

|  |  |  |
| --- | --- | --- |
| **A B S T R A C T** |  | **A R T I C L E I N F O** |
| The rapid expansion of online education necessitates a deeper understanding of what makes courses successful, as many online courses struggle with low completion rates and engagement. This research aims to uncover patterns in online course performance by analysing a dataset using Principal Component Analysis (PCA) and K-means clustering. Data was sourced from Kaggle, including attributes like course duration, enrolled students, completion rates, platform, price, and ratings. PCA was used to reduce dimensionality, and K-means clustering identified distinct groups of courses based on these attributes. The results indicated that clusters 1 and 3 had the highest mean ratings for 4.34 and 4.45, respectively, and completion rates of 80.23% and 85.69%, respectively, highlighting their success and high engagement levels. In contrast, Cluster 0 had the lowest mean rating at 3.66 and completion rate at 62.42%, indicating a need for improvement. The study's significance lies in its potential to guide educators and course designers in creating more effective and engaging online courses. By providing actionable insights into course characteristics associated with success, this research addresses the critical issue of improving student engagement and course completion in online education.  © 2023 Universitas Pendidikan Indonesia |  | ***Article History:***  *Submitted/Received 00 xxx 2024*  *First Revised 00 xxx 2024*  *Accepted 00 xxx 2024*  *First Available Online 00 xxx 2024*  *Publication Date 00 xxx 2024*  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  ***Keyword:***  *Online education,*  *Principal Component Analysis (PCA),*  *K-means clustering,*  *Course performance,*  *Student engagement.* |



**1. INTRODUCTION**

The increasing prevalence of online education has revolutionized how knowledge is disseminated, making learning more accessible and flexible. As online courses proliferate, understanding the factors that contribute to their success becomes critical (Maitra & Anurekha, 2021). This research delves into the intricacies of online course performance, leveraging data analysis techniques to uncover patterns that can enhance course design and student engagement (Xiaoyu & Tobias, 2023; Zhou & Bhat, 2021). Recent studies have employed various analytical methods to evaluate and improve online education (Castro & Tumibay, 2021; Maitra & Anurekha, 2021; ONAN, 2021; Ouyang, Zheng, & Jiao, 2022). For instance, Principal Component Analysis (PCA) and K-means clustering are prominent techniques used to handle high-dimensional data and identify distinct groupings within datasets (Mousavian Anaraki, Haeri, & Moslehi, 2021; Salih Hasan & Abdulazeez, 2021). PCA reduces the dimensionality of data, retaining its most significant features, while K-means clustering groups data points based on similarity (Ikotun, Ezugwu, Abualigah, Abuhaija, & Heming, 2023). These methods have been applied to diverse datasets, including educational data, to extract meaningful insights that can inform better course offerings and personalized learning experiences​​.

While substantial research has explored various factors influencing online course success, there remains a gap in understanding how a combination of course characteristics collectively impacts student engagement and completion rates (Caskurlu, Richardson, Maeda, & Kozan, 2021; Qureshi, Khaskheli, Qureshi, Raza, & Yousufi, 2023; Wang, Mirzaei, Xu, & Lin, 2022). Existing studies often focus on individual factors, such as course content or delivery methods, without sufficiently integrating multiple dimensions of course attributes (Zamecnik et al., 2023). Moreover, the application of advanced analytical techniques like PCA and K-means clustering in the context of online education is still relatively underexplored (Khan et al., 2024; Miraftabzadeh, Colombo, Longo, & Foiadelli, 2023). This research seeks to fill this gap by employing these techniques to provide a holistic analysis of online course performance, offering a more comprehensive understanding of the elements that contribute to effective and engaging online learning environments.

This research aims to explore the landscape of online courses by analysing a dataset from Kaggle, focusing on attributes such as course duration, enrolled students, completion rates, platform, price, and ratings. The primary objectives are to apply PCA for dimensionality reduction, simplifying the dataset while retaining its essential variance, and to use K-means clustering to segment online courses into distinct groups based on their features. Additionally, the study evaluates the quality of the clusters formed, using metrics such as silhouette scores to ensure meaningful and actionable insights. By identifying characteristics of successful courses, the research seeks to provide recommendations for course design and improvement, thereby contributing valuable knowledge to the field of online education​.

**2. RESEARCH METHODS**

The objective of this research is to uncover patterns in online education by analysing course data using PCA and K-means clustering. This study aims to identify distinct groups of courses based on various attributes, providing insights that can enhance course design and improve student engagement. This research follows the flowchart illustrated in **Figure 1**.

A diagram of a process flow

Description automatically generated

**Figure 1.** Flowchart of Research Method

**2.1. Data Collection**

In this step, the data will be collected from secondary sources. Secondary source refers to existing sources of information that offer second-hand data, such as books, monographs, hospital records, and online databases (Faryadi, 2019). Specifically, it is sourced from Kaggle, which is a well-known platform for data science competitions. It also provides public datasets, ML notebooks, and tutorials. Data scientists and machine learning engineers can use Kaggle to compete in creating models, practice skills, build portfolios, and connect with industry peers (Elkabir, 2023). The dataset is called Online Courses Usage and History (Das, M., 2024). It consists of ‘Course ID’, ‘Course Name’, ‘Category, ‘Duration’ in hours’, ‘Enrolled students’, ‘Completion Rate’, ‘Platform’, ‘Price ($)’ and ‘Rating (out of 5)’.

**2.2. Data Preprocessing**

Data preprocessing is a crucial step in preparing the dataset for analysis. This process involves cleaning the data, normalizing it, and selecting relevant features. Effective pre-processing ensures that the data is in a suitable format for subsequent analysis techniques.

**2.2.1. Data Cleaning**

During the data cleaning step, the dataset will be cleaned to handle missing values. Missing values will be managed using appropriate imputation techniques or by removing incomplete records if necessary. Data cleaning helps ensure that the data is accurate, complete, and reliable. This is because, high-quality data is crucial for producing valid and trustworthy results in data analysis.

**2.2.2. Normalization**

In this step, the data will be standardized to ensure a uniform scale across all features, which is essential for PCA. This will be done using the ‘StandardScaler’ from the ‘sklearn. preprocessing’ module. Standardization transforms the data to have a mean of 0 and a standard deviation of 1, ensuring that each feature contributes equally to the analysis.

**2.2.3. Feature Selection**

In feature selection, relevant features from the dataset will be chosen. The selected features include all numeric attributes such as ‘Duration (hours)’, ‘Enrolled\_Students’, ‘Completion\_Rate (%)’, ‘Price ($)’, and ‘Rating (out of 5)’. Non-numeric and irrelevant columns such as ‘Course\_ID’, ‘Course\_Name’, ‘Category’, and ‘Platform’ will be excluded. This is because, irrelevant features can lead to overfitting, where the model learns noise in the training data instead of the actual patterns. By focusing on important features, the model generalizes better to new data.

**2.3 Principal Component Analysis (PCA)**

PCA is a dimensionality reduction technique that transforms original features into a set of uncorrelated variables called principal components that retains most of the variance in the dataset. In this study, PCA is used to handle the high dimensionality of the online course dataset. By reducing dimensions, PCA simplifies the dataset, making it easier to visualize and analyse, and enhances the performance of subsequent clustering algorithms like k-means by focusing on the most significant features.

**2.4 K-Means Clustering**

K-means is a clustering algorithm used to partition a dataset into distinct groups based on feature similarity. It aims to minimize the within-cluster variance by assigning each data point to the nearest cluster centre. K-means clustering is used in this study to identify distinct groups of online courses based on their attributes. This helps in uncovering hidden patterns and similarities among the courses, which can provide valuable insights for course designers and educators. By grouping similar courses together, k-means can highlight trends in course characteristics, student engagement, and performance metrics, enabling more targeted improvements and recommendations.

**2.5 Evaluation**

Evaluating the quality of clusters is a critical step in ensuring that the clustering algorithm has effectively grouped the online courses based on their features. This involves using various validation metrics to measure the cohesiveness and separation of the clusters. This evaluation is essential to confirm that the clustering results are meaningful and provide actionable insights into the characteristics and performance of the online courses.

**2.5.1 Validation**

The quality of the clusters will be validated using metrics such as silhouette scores. These metrics will assess how well the online courses are grouped, ensuring that courses within the same cluster are more similar to each other than to those in different clusters. This validation step is crucial to confirm that the clustering results provide meaningful and actionable insights into the characteristics and performance of the online courses.

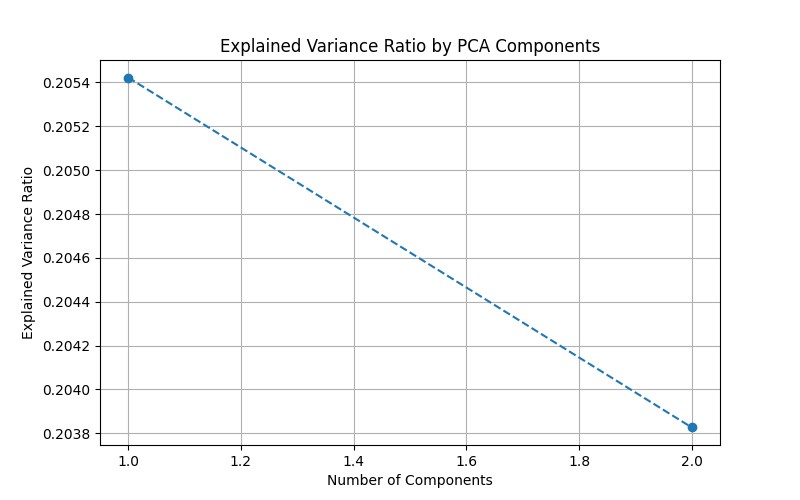
**3. RESULTS AND DISCUSSION**

This section presents the results of the analysis and discusses the findings. The study aimed to uncover patterns in online courses by using PCA for dimensionality reduction and K-means clustering for segmentation. The performance of different numbers of PCA components was evaluated using explained variance ratios and silhouette scores. The clusters were then analysed to understand their characteristics and effectiveness in segmenting the courses based on key features such as duration, enrolled students, completion rate, price, and rating.

**3.1. Explained Variance Ratio**

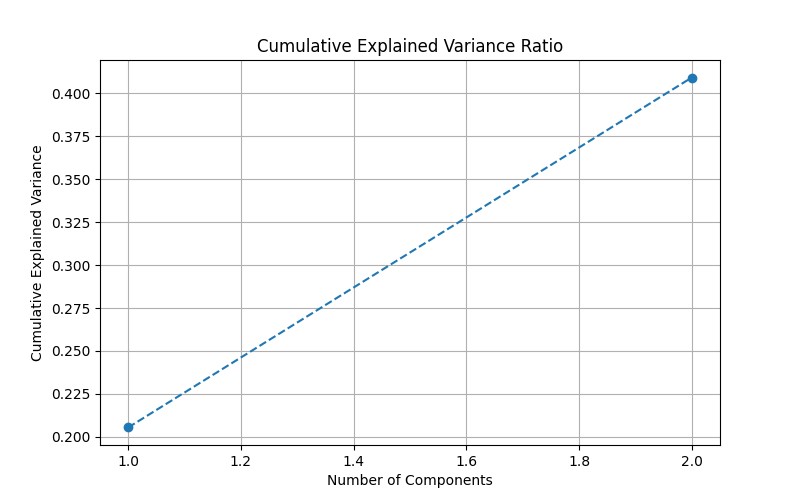
The explained variance ratio indicates how much of the total variance in the dataset is captured by each principal component. This metric helps in understanding the importance of each principal component in describing the variability of the data. Principal components with the highest explained variance ratios are usually selected to ensure that most of the dataset's variance is retained while reducing its dimensionality. However, in this study, by selecting a higher number of components, more variances can be retained from the original dataset. However, this does not always lead to better clustering performance, as seen with the silhouette score.

**Figure 2** shows the variance explained by each principal component. It helps identify how much variance each individual component captures, showing that the first few components capture more variance than the later ones. The first two components capture about 41% of the total variance.



**Figure 2.** Explained Variance Ratio by PCA Components

**Figure 3** shows the cumulative variance explained by the PCA components. It helps in determining the optimal number of components needed to capture the majority of the variance in the data. For example, two components capture around 40% of the total variance.

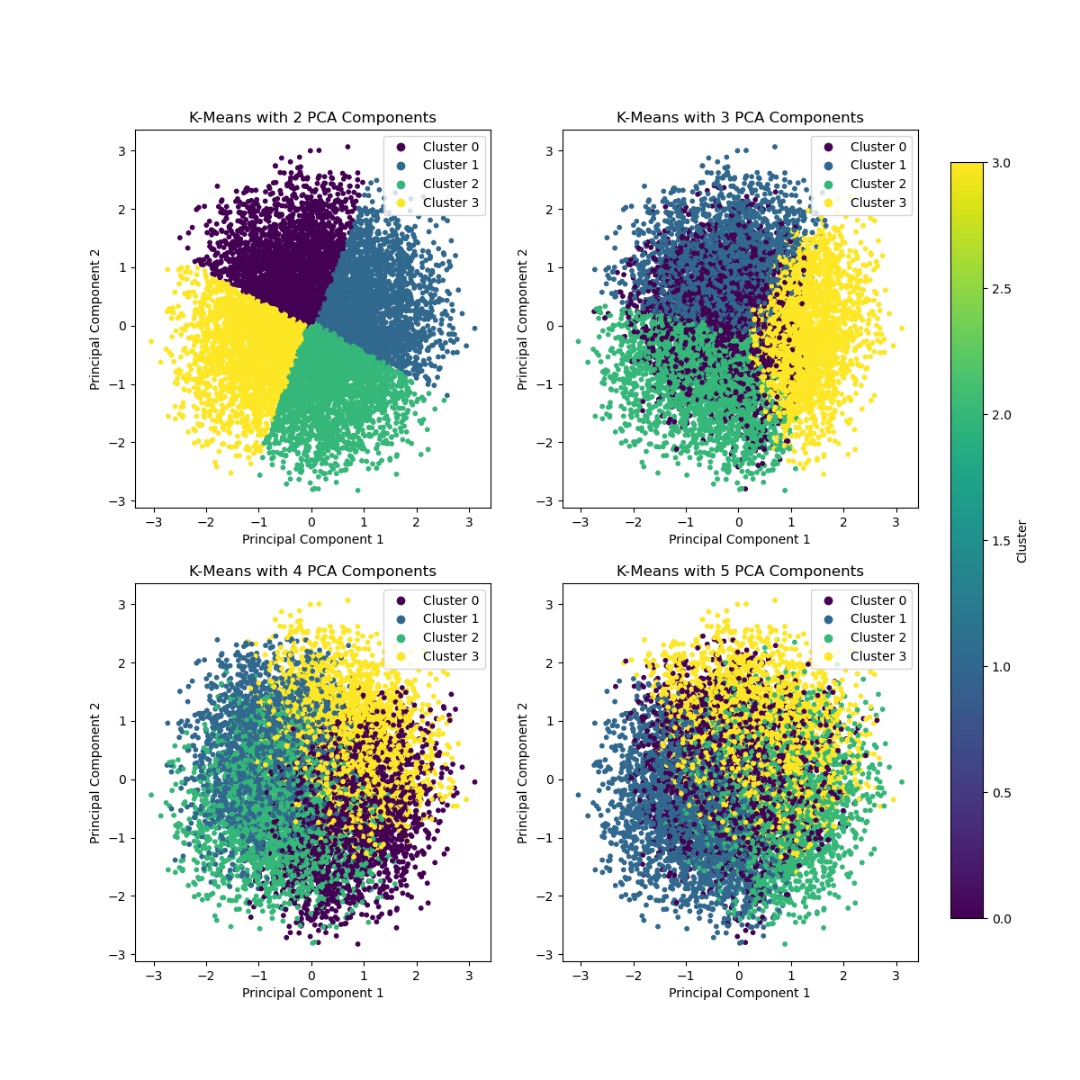


**Figure 3.** Cumulative Explained Variance Ratio by PCA Components

**3.2. Comparison of K-Means Clustering with Different PCA Components**

**Figure 4** displays the results of K-means clustering applied to a dataset after reducing its dimensions using PCA with different numbers of components. Each subplot represents clustering results with varying numbers of PCA components which are two, three, four, and five. The clustering with two PCA components shows the most distinct and well-separated clusters, indicating clear cluster boundaries. As the number of components increases to three, four, and five, the clusters become less distinct and more overlapping in the 2D space defined by the first two principal components. This suggests that additional components introduce noise rather than improving the separation of clusters.

The optimal number of PCA components for this dataset is two, as it balances variance retention and noise reduction, resulting in the clearest cluster separation with K-means clustering. It also visualizes the distribution of online courses across four clusters in the reduced PCA space. Each point represents a course, and the colour indicates the cluster to which it belongs. This visualization helps in understanding the clustering outcome and assessing the separation of clusters. The clusters show distinct groupings of courses, suggesting effective clustering based on the PCA-reduced features.



**Figure 4.** K-Means Clustering Results with Varying Numbers of PCA Components

**3.3. Silhouette Scores for Different PCA Components**

Silhouette scores are a metric used to evaluate the quality of clusters in a clustering algorithm. A high silhouette score, close to 1, indicates well-defined and distinct clusters, where data points are well-clustered within their own groups and far from other clusters. A low silhouette score, around 0, suggests overlapping clusters or points near cluster boundaries, indicating that clusters are not well-separated. A negative silhouette score, close to -1, reveals poor clustering performance, suggesting that data points may have been assigned to incorrect clusters. By analysing silhouette scores, the number of clusters can be assessed and optimized to achieve the best clustering performance.

**Table 1** shows the explained variance ratio and silhouette scores for different numbers of PCA components. It is used to evaluate the quality and consistency of the clusters formed with varying PCA components. The highest silhouette score was observed with two PCA components, indicating well-defined clusters.

**Table 1.** Explained Variance Ratio and Silhouette Scores for Different Numbers of PCA Components.

|  |  |  |
| --- | --- | --- |
| **No. of Components** | **Explained Variance Ratio** | **Silhouette Score** |
| 2 | 0.3172475151989326 | 0.34227249240236507 |
| 3 | 0.4305547567695816 | 0.23923108217613182 |
| 4 | 0.5428306558463938 | 0.18124855754879893 |
| 5 | 0.6541322634412092 | 0.13861516011206404 |

**3.4. Descriptive Statistics of Clusters**

**Table 2** provides detailed descriptive statistics for each cluster, specifically mean values for key numeric features such as Duration, Enrolled Students, Completion Rate, Price and Rating. This key features values will determine the success rates of the courses for each cluster based on its criteria.

**Table 2.** Detailed Descriptive Statistics for each Cluster.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Cluster** | **Count** | **Mean Duration (hours)** | **Mean Enrolled Students** | **Mean Completion Rate (%)** | **Mean Price ($)** | **Mean Rating**  **(out of 5)** |
| 0 | 2387 | 61.15 | 3032.47 | 62.42 | 128.79 | 3.66 |
| 1 | 2340 | 49.48 | 3041.12 | 80.23 | 83.77 | 4.34 |
| 2 | 2612 | 37.04 | 3019.14 | 75.36 | 116.14 | 4.30 |
| 3 | 2661 | 74.24 | 2896.19 | 85.69 | 137.52 | 4.45 |

Based on **Table 2**, the clusters show distinct groupings of courses based on their features. Cluster 1 has the highest mean rating at 4.34 and a relatively high completion rate for 80.23%, suggesting high-quality courses that are well-received by students. Similarly, Cluster 3 has the highest mean completion rate for 85.69% and a high rating at 4.45, indicating highly engaging and successful courses. In contrast, Cluster 0 has the lowest mean rating for 3.66 and the lowest completion rate at 62.42%, suggesting that these courses might need improvement.

The success criteria, defined as a completion rate of 80% or higher and a rating of 4.0 or higher, were met primarily by clusters 1 and 3. This indicates that these clusters consist of successful courses. The clustering effectively segments courses based on success and user engagement metrics.

Courses in Clusters 1 and 3 are more likely to be successful and well-rated by students. This insight can help in designing and promoting future courses to match these clusters' characteristics. Conversely, clusters with lower ratings and completion rates, such as Cluster 0, can be analysed further to identify areas for improvement.By using different PCA components, it showed vary clustering patterns and silhouette scores, with 2 components giving the best silhouette score. This suggests that a lower-dimensional representation can effectively capture the essential variance for clustering in this dataset, thereby providing a clearer segmentation of course quality and engagement.

**4. CONCLUSION**

This study demonstrates the effectiveness of using PCA and K-means clustering to analyse and enhance online course performance. By examining a comprehensive dataset from Kaggle, the distinct clusters of courses that vary significantly in terms of duration, enrolment, completion rates, price, and ratings are identified. Notably, clusters 1 and 3 emerged as the most successful, featuring the highest mean ratings and completion rates, which underscores their high engagement and quality. Conversely, Cluster 0, with its lower ratings and completion rates, highlights areas needing improvement. The results of the success and user engagement metrics show that cluster 1 has the highest average value for 4.34 and a relatively high completion rate of 80.23%. Cluster 3 had the highest average completion rate of 85.69% and high ranking at 4.45. Cluster 0 had the lowest mean score for 3.66 and the lowest completion rate at 62.42%. The success criteria (Completion Rate >= 80%, Rating >= 4.0) shows that cluster 1 and cluster 3 can be grouped into quality and successful courses and can help in designing and promoting courses in the future. Meanwhile, cluster 0 with a low level of completion rate is a cluster that requires improvement.

The findings underscore the importance of key course characteristics in determining student engagement and course success. Educators and course designers can leverage these insights to refine their offerings, focusing on the attributes that correlate with higher student satisfaction and completion. This research not only contributes to the academic understanding of online education dynamics but also provides practical recommendations for enhancing the effectiveness and appeal of online courses. By addressing the critical issues of student engagement and course completion, this study paves the way for more effective online learning experiences.

**6. AUTHORS’ NOTE**

The authors declare that there is no conflict of interest regarding the publication of this article. The authors confirmed that the paper was free of plagiarism.

**7. REFERENCES**

Das, M. (2024, July 2). Online courses usage and history dataset. Kaggle. https://www.kaggle.com/datasets/mitul1999/online-courses-usage-and-history-dataset/data

Elkabir, A. (2023, September 21). What is Kaggle? how to compete in Kaggle Competitions. Built In. https://builtin.com/data-science/what-is-kaggle

Caskurlu, S., Richardson, J. C., Maeda, Y., & Kozan, K. (2021). The qualitative evidence behind the factors impacting online learning experiences as informed by the community of inquiry framework: A thematic synthesis. *Computers & Education, 165*, 104111. doi:<https://doi.org/10.1016/j.compedu.2020.104111>

Castro, M. D. B., & Tumibay, G. M. (2021). A literature review: efficacy of online learning courses for higher education institution using meta-analysis. *Education and Information Technologies, 26*(2), 1367-1385. doi:10.1007/s10639-019-10027-z

Faryadi, Q. (2019). PhD Thesis Writing Process: A Systematic Approach—How to Write Your Methodology, Results and Conclusion. *Creative Education, 10*, 766-783. doi:10.4236/ce.2019.104057

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences, 622*, 178-210. doi:<https://doi.org/10.1016/j.ins.2022.11.139>

Khan, O., Ali, V., Parvez, M., Alhodaib, A., Yahya, Z., Yadav, A. K., & Ağbulut, Ü. (2024). Exploring the performance of biodiesel-hydrogen blends with diverse nanoparticles in diesel engine: A hybrid machine learning K-means clustering approach with weighted performance metrics. *International Journal of Hydrogen Energy, 78*, 547-563. doi:<https://doi.org/10.1016/j.ijhydene.2024.06.303>

Maitra, M. A., & Anurekha, S. (2021). A systematic review of factors influencing the effectiveness of online sessions and open educational resources on students amidst pandemic. *An Asian Perspective, 6*.

Miraftabzadeh, S. M., Colombo, C. G., Longo, M., & Foiadelli, F. (2023). K-Means and Alternative Clustering Methods in Modern Power Systems. *IEEE Access, 11*, 119596-119633. doi:10.1109/ACCESS.2023.3327640

Mousavian Anaraki, S. A., Haeri, A., & Moslehi, F. (2021). A hybrid reciprocal model of PCA and K-means with an innovative approach of considering sub-datasets for the improvement of K-means initialization and step-by-step labeling to create clusters with high interpretability. *Pattern Analysis and Applications, 24*(3), 1387-1402. doi:10.1007/s10044-021-00977-x

ONAN, A. (2021). Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach. *Computer Applications in Engineering Education, 29*(3), 572-589. doi:<https://doi.org/10.1002/cae.22253>

Ouyang, F., Zheng, L., & Jiao, P. (2022). Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. *Education and Information Technologies, 27*(6), 7893-7925. doi:10.1007/s10639-022-10925-9

Qureshi, M. A., Khaskheli, A., Qureshi, J. A., Raza, S. A., & Yousufi, S. Q. (2023). Factors affecting students’ learning performance through collaborative learning and engagement. *Interactive Learning Environments, 31*(4), 2371-2391. doi:10.1080/10494820.2021.1884886

Salih Hasan, B. M., & Abdulazeez, A. M. (2021). A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. *Journal of Soft Computing and Data Mining, 2*(1), 20-30. Retrieved from <https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/8032>

Wang, C., Mirzaei, T., Xu, T., & Lin, H. (2022). How learner engagement impacts non-formal online learning outcomes through value co-creation: an empirical analysis. *International Journal of Educational Technology in Higher Education, 19*(1), 32. doi:10.1186/s41239-022-00341-x

Xiaoyu, Z., & Tobias, T. (2023). Exploring the Efficacy of Adaptive Learning Technologies in Online Education: A Longitudinal Analysis of Student Engagement and Performance. *International Journal of Science and Engineering Applications, 12*(12), 28-31.

Zamecnik, A., Kovanovíc, V., Joksimovíc, S., Grossmann, G., Ladjal, D., Marshall, R., & Pardo, A. (2023). Using online learner trace data to understand the cohesion of teams in higher education. *Journal of Computer Assisted Learning, 39*(6), 1733-1750. doi:<https://doi.org/10.1111/jcal.12829>

Zhou, J., & Bhat, S. (2021). *Modeling Consistency Using Engagement Patterns in Online Courses*. Paper presented at the LAK21: 11th International Learning Analytics and Knowledge Conference, Irvine, CA, USA. <https://doi.org/10.1145/3448139.3448161>