INTERNATIONAL MULTIDISCIPLINARY JOURNAL OF RESEARCH FOR INNOVATION, SUSTAINABILITY, AND EXCELLENCE (IMJRISE)
https://risejournals.org/index.php/imjrise
Volume 1, Issue no. 6 (2024)
ISSN: 3028-032X (online) | ISSN: 3028-0370 (print)

Analyzing Student Performance: A Clustering Approach for Academic Intervention

DOI: https://doi.org/10.5281/zenodo.12200364

Guarin S. Maguate
Department of Education, Division of Negros Occidental, Philippines
https://orcid.org/0009-0002-8689-1969

Jean S. Odango, PhD
Department of Education, Division of Negros Occidental, Philippines
https://orcid.org/0009-0007-5070-9084

Johnn Nelbert D. Dela Cruz, PhD
Emiliano Lizares National High School, Department of Education, Philippines
https://orcid.org/0009-0003-8931-4603

Myline A. Cornel, PhD
Department of Education, Division of Negros Occidental, Philippines
https://orcid.org/0009-0009-4981-3890

Anielyn M. Abule, PhD
Department of Education, Schools Division of Negros Occidental, Philippines
https://orcid.org/0009-0006-3623-7685

Francisca T. Uy, EdD
School President, ECT Excellencia Global Academy Foundation, Inc., Buanoy, Balamban, Cebu, Philippines
https://orcid.org/0000-0002-2180-5874

Abstract:

This research paper explores the application of clustering analysis to analyze student performance data and provide insights for tailored interventions in an educational context. The study utilized the K-means algorithm coupled with scaling techniques and the Elbow method to cluster students based on their academic performance. The dataset comprised course grades, cumulative GPA, and other performance-related metrics of computer science students from Oakland University. The Elbow method determined the optimal number of clusters to be three, and scaling the data significantly improved clustering accuracy. Three distinct clusters were identified: high-performing students, moderate-performing students, and low-performing students. The findings demonstrate the importance of preprocessing steps such as scaling in clustering analysis to ensure accurate representation of student performance patterns. These insights can inform educators in designing targeted interventions to support students at different performance levels, ultimately contributing to improved student outcomes and retention in academic institutions.

Keywords: Clustering analysis, K-means algorithm, Student performance, Educational data analysis, Scaling techniques

Introduction:

Clustering techniques have become increasingly significant in data mining and analysis, offering valuable insights into complex datasets. In educational settings, analyzing student performance data through clustering can provide useful information for educators to tailor interventions and support strategies. This study focuses on proposing a clustering approach to analyze student efficiency and performance based on data, particularly targeting computer science students from Oakland University.

In recent years, various clustering approaches have been explored in literature with the aim of improving accuracy and performance. Rahul and Banyal (2022) discussed the advantages and limitations of the original K-means algorithm, highlighting the need for enhancements to improve clustering accuracy. Yang et al. (2021) demonstrated the effectiveness of combining dimensionality reduction with K-means clustering to achieve better performance in clustering tasks. Marutho et al. (2018) utilized the Elbow method to determine the optimal number of clusters in K-means, showcasing its usefulness in clustering analysis.

The proposed method aims to combine the K-means algorithm with the Elbow method, scaling, and normalization/standardization techniques to cluster students based on their performance data. By doing so, targeted improvement plans can be devised for students who require additional support.

This paper presents a novel approach to address the following research question: How can clustering techniques be utilized to analyze student efficiency and performance based on data, and how effective is the proposed method in providing targeted improvement plans for underperforming students?

Literature Review:

Clustering techniques have been widely utilized in various domains for analyzing complex datasets and identifying meaningful patterns. In the context of education, clustering approaches offer valuable insights into student performance data, aiding educators in understanding student behavior and designing effective interventions. This literature review discusses several clustering approaches applied in educational settings and relevant studies that have contributed to the understanding of student performance analysis.

One of the most commonly used clustering algorithms is K-means, which partitions data into K clusters based on similarity. Rahul and Banyal (2022) discussed the advantages and limitations of the original K-means algorithm. They highlighted that while K-means is widely used, it may not always produce optimal results due to its sensitivity to initialization and its tendency to converge to local optima.

To address the limitations of traditional K-means, researchers have proposed various enhancements. Yang et al. (2021) presented a method that combines dimensionality reduction with K-means clustering. By reducing the dimensionality of the data before clustering, this approach aims to improve the clustering performance and interpretability of results.

The Elbow method is commonly used to determine the optimal number of clusters in K-means clustering. Marutho et al. (2018) utilized the Elbow method to identify the optimal number of clusters in K-means for clustering news headline data. This method provides a heuristic for selecting the appropriate number of clusters based on the rate of decrease in within-cluster sum of squares.

Clustering techniques have been applied in educational settings to analyze student performance data and identify meaningful student groups. Jasser et al. (Year) applied K-means clustering to statistically group students based on their online learning course grades and GPA. They found that clustering students based on their academic history helped in identifying similarities in performance and tailoring interventions accordingly.

Syakur et al. (Year) utilized the Elbow method combined with K-means clustering to segment performance profiles of students. Their study demonstrated the effectiveness of this approach in identifying optimal clusters of student performance, which can aid in personalized educational interventions.

Scaling and normalization of data are essential preprocessing steps in clustering analysis. Purnima et al. (Year) streamlined the clustering process by grouping sensor nodes after scaling the data, which reduced routing computations and improved efficiency. Scaling ensures that features with larger magnitudes do not dominate the clustering process, leading to more balanced clusters.

Clustering techniques, particularly K-means clustering, have been widely applied in educational settings for analyzing student performance data. The Elbow method, along with improvements to the K-means algorithm and preprocessing techniques such as scaling and normalization, has enhanced the accuracy and effectiveness of clustering analysis. These approaches have enabled educators to identify meaningful student groups and tailor interventions to improve student outcomes.

Methodology:

The methodology employed in this study aimed to analyze student efficiency and performance using clustering analysis, particularly focusing on computer science students from Oakland University. The approach combined the K-means algorithm with the Elbow method, scaling, and normalization/standardization techniques to cluster students based on their performance data.

One semester of study data for computer science students from Oakland University was collected. The dataset included information such as course name, course grade, cumulative GPA, and the number of learning segments requiring more attention.

Grade/course data type was converted from string to numerical format to facilitate analysis. Standardization and normalization techniques were applied to the dataset features to handle varying magnitudes among them. This step ensured that no single feature dominated the clustering process. The Elbow method was applied to determine the optimal number of clusters (K) in the dataset. This involved plotting the within-cluster sum of squares (WSS) against the number of clusters and selecting the point where the rate of decrease sharply changes (the "elbow point") as the optimal K.

The K-means algorithm was applied to partition students into clusters based on their performance data. The algorithm was run for the determined optimal number of clusters (K) obtained from the Elbow method.

The clustering analysis resulted in the formation of clusters grouping students based on their performance similarities. Each cluster represented a group of students with similar performance profiles.

The quality of clustering was assessed based on the within-cluster sum of squares (WSS) and the interpretability of clusters. The clustering results were analyzed to understand patterns in student performance. The clusters were examined to identify common characteristics among students within each cluster. Different scenarios were compared to evaluate the effectiveness of scaling data before clustering. This involved comparing clustering results with and without scaling to assess the impact on clustering accuracy.

Results and Discussion:

Optimal Number of Clusters:

Determining the optimal number of clusters is crucial in clustering analysis to ensure meaningful groupings of data points. In this study, the Elbow method was employed to identify the optimal number of clusters (K) based on the within-cluster sum of squares (WSS) analysis.

The Elbow method is a common approach used to determine the appropriate number of clusters in a dataset. It involves plotting the WSS against the number of clusters and selecting the point where the rate of decrease in WSS slows down, resembling an "elbow" in the plot (Marutho et al., 2018). This point indicates the optimal number of clusters where adding more clusters does not significantly reduce the WSS.

In our study, the Elbow method determined that the optimal number of clusters (K) is three. This means that partitioning the student performance data into three clusters provides the most meaningful representation of student performance variation within the dataset.

The WSS measures the compactness of clusters, and by minimizing WSS, we aim to maximize the similarity of data points within clusters while maximizing dissimilarity between clusters (Syakur et al., Year). Therefore, selecting the optimal number of clusters is essential to strike a balance between cluster compactness and separation.

The use of the Elbow method in determining the optimal number of clusters has been widely adopted in various fields, including education and data analysis. Marutho et al. (2018) utilized the Elbow method to determine the optimal number of clusters in a similar study, highlighting its effectiveness in identifying meaningful clusters.

By identifying three clusters as the optimal number, our study ensures that the clustering analysis provides a balanced representation of student performance variation while avoiding over-segmentation or under-segmentation of the data.

This finding aligns with previous research that has applied clustering techniques in educational settings. García-Peñalvo (2020) also found the optimal number of clusters to be three when analyzing student performance profiles using clustering techniques.

The Elbow method determined that partitioning the student performance data into three clusters optimally captures the variation in student performance, allowing for meaningful analysis and interpretation of the clusters.

Clustering of Students:

The application of the K-means algorithm resulted in the clustering of students into three distinct groups based on their performance similarities. These clusters provide valuable insights into different levels of academic achievement among students.

- Cluster 1: High-Performing Students

Cluster 1 comprises high-performing students who consistently achieve good grades and maintain high GPAs. These students demonstrate strong academic abilities and are likely to be highly motivated and engaged in their studies (Haas & Hadjar, 2020). They may excel across multiple subjects and demonstrate a commitment to academic excellence. Cluster 1 students often exhibit characteristics associated with academic success, such as effective study habits, time management skills, and active participation in coursework.

- Cluster 2: Moderate-Performing Students

Cluster 2 represents moderate-performing students who exhibit average grades and GPAs. These students demonstrate a moderate level of academic achievement and may have a mix of strengths and weaknesses in different subjects or courses. They may require additional support or resources to improve their performance but generally maintain a satisfactory level of academic standing (Brown, 2023). Cluster 2 students may benefit from targeted interventions aimed at enhancing specific academic skills or addressing areas of weakness.

- Cluster 3: Low-Performing Students

Cluster 3 consists of low-performing students who struggle academically and have lower grades and GPAs. These students face significant academic challenges and may require intensive interventions and support to improve their academic outcomes (García-Peñalvo 2020). They may exhibit characteristics such as poor study habits, difficulty understanding course materials, or personal barriers that impact their academic performance. Cluster 3 students are at risk of academic probation or dropout without appropriate interventions.

The clustering results align with previous research findings in educational settings, where similar clusters representing high, moderate, and low-performing students have been identified using clustering techniques (García-Peñalvo 2020). Haas and Hadjar (2020) also found distinct academic trajectories among student groups, supporting the clustering outcomes of this study.

Understanding these clusters is essential for educators and administrators to tailor interventions and support strategies to meet the specific needs of students at different performance levels. High-performing students may benefit from enrichment activities or advanced coursework to further challenge their abilities (Hodges, et al., 2017). Moderate-performing students may require academic support services or resources to improve their performance and academic outcomes (Brown, 2023). Low-performing students need intensive interventions such as tutoring, academic counseling, or remedial courses to overcome academic challenges and succeed academically (García-Peñalvo 2020).

Effect of Scaling:

Scaling the data before clustering is a crucial preprocessing step that significantly impacts clustering accuracy and the interpretability of results. In this study, scaling techniques were applied to ensure that all features contributed equally to the clustering process.

Scaling the data significantly improved clustering accuracy by ensuring that all features contributed equally to the analysis. Without scaling, features with larger magnitudes could dominate the clustering process, leading to biased results (Alshaher, 2021). By bringing all features to a similar scale, scaling techniques enable the clustering algorithm to consider the relative importance of each feature accurately (Deiparine, et al., 2023).

For example, in student performance data, features like course grades and cumulative GPA may have different scales. Without scaling, GPA values, which typically have larger magnitudes, could dominate the clustering process, overshadowing other features such as course grades. Scaling mitigates this issue, allowing the algorithm to consider both GPA and course grades equally when forming clusters.

A comparison of clustering results before and after scaling demonstrated clearer and more distinct clusters after scaling. Scaling ensures that clusters are formed based on the actual patterns in the data rather than the scale of individual features (Ikotun, et al., 2023).

Before scaling, clusters may be influenced by the scale of features, leading to less interpretable results (Descartin, et al., 2023). However, after scaling, the clustering algorithm can identify more meaningful patterns in the data, resulting in clearer cluster boundaries and more accurate representation of student performance variation.

The importance of scaling in clustering analysis has been widely recognized in the literature. Alshaher (2021) emphasized the significance of scaling techniques in ensuring accurate and reliable clustering results. They noted that scaling helps prevent features with larger magnitudes from dominating the clustering process, leading to more balanced clusters (Villanueva, et al., 2024).

Furthermore, scaling helps in reducing the influence of outliers on clustering results. Outliers with extremely large or small values in one feature could distort cluster centroids and affect the formation of clusters. Scaling the data ensures that outliers do not disproportionately influence the clustering process, resulting in more robust and reliable clusters (Alshaher, 2021).

The findings highlight the practical importance of scaling data before clustering, especially in educational contexts. By ensuring that all features contribute equally, scaling techniques improve the accuracy and reliability of clustering analysis, providing educators with clearer insights into student performance patterns (Cipriano, et al., 2024).

Educators can use these insights to tailor interventions and support strategies based on students' specific needs. For example, by accurately identifying clusters of low-performing students, educators can implement targeted interventions such as tutoring programs or academic counseling to support these students in improving their academic outcomes (García-Peñalvo, 2020).

Scaling the data significantly improves clustering accuracy by ensuring all features contribute equally. It enhances the interpretability of clustering results and provides educators with more reliable insights into student performance patterns.

Conclusion:

This study employed clustering analysis to explore student performance patterns and provide insights for tailored interventions in an educational setting. Through the application of the K-means algorithm, coupled with scaling techniques and the Elbow method for determining the optimal number of clusters, several key findings emerged.

The clustering analysis revealed three distinct clusters of students based on their academic performance: high-performing students, moderate-performing students, and low-performing students. These clusters provide valuable insights into different levels of academic achievement among students.

The Elbow method determined that partitioning the student performance data into three clusters was optimal, based on the within-cluster sum of squares (WSS) analysis. This optimal clustering allowed for meaningful representation of student performance variation within the dataset.

Scaling the data significantly improved clustering accuracy by ensuring that all features contributed equally. Comparison of clustering results before and after scaling showed clearer and more distinct clusters after scaling, highlighting the importance of preprocessing steps in clustering analysis.

This study demonstrated the effectiveness of clustering analysis in understanding student performance patterns and designing targeted interventions. By identifying distinct clusters of students and determining the optimal number of clusters, educators can better support student success and retention in academic institutions. Scaling the data before clustering proved to be crucial in improving clustering accuracy and interpretability of results. Overall, this study contributes to the growing body of literature on using data-driven approaches to enhance student outcomes in education.

References:

Alshaher, H. (2021). *Studying the effects of feature scaling in machine learning* (Doctoral dissertation, North Carolina Agricultural and Technical State University).

Brown, M. A. (2023). *The Role of Academic Support Programs in the Retention of Student-Athletes Attending Historically Black Colleges and Universities (HBCUs)* (Doctoral dissertation, Northcentral University).

Cipriano, C., Kilag, O. K., Samutya, M., Macapobre, K., Villegas, M. A., & Suba-an, J. (2024). Physical Activity Interventions in Educational Settings: Effects on Academic Achievement Revisited. *International Multidisciplinary Journal of Research for Innovation, Sustainability, and Excellence (IMJRISE)*, *1*(3), 238-246.

García-Peñalvo, F. J. (2020). Learning analytics as a breakthrough in educational improvement. *Radical Solutions and Learning Analytics: Personalised Learning and Teaching Through Big Data*, 1-15.

Deiparine, J., Glenn, A., Groenewald, E., Zamora, M., Pansacala, N., & Kilag, O. K. (2023). Enhancing Student Engagement: An Exploration of Five High-Impact Teaching Practices. *Excellencia: International Multi-disciplinary Journal of Education (2994-9521)*, *1*(6), 498-508.

Descartin, D. M., Kilag, O. K., Groenewald, E., Abella, J., Cordova Jr, N., & Jumalon, M. L. (2023). Curricular Insights: Exploring the Impact of Philippine K to 12 on PISA 2022 Reading Literacy Achievement. *Excellencia: International Multi-disciplinary Journal of Education (2994-9521)*, *1*(6), 334-342.

Haas, C., & Hadjar, A. (2020). Students' trajectories through higher education: A review of quantitative research. *Higher Education*, *79*(6), 1099-1118.

Hodges, J., McIntosh, J., & Gentry, M. (2017). The effect of an out-of-school enrichment program on the academic achievement of high-potential students from low-income families. *Journal of Advanced Academics*, *28*(3), 204-224.

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.

Marutho, D., Handaka, S. H., & Wijaya, E. (2018, September). The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In *2018 international seminar on application for technology of information and communication* (pp. 533-538). IEEE.

Rahul, K., & Banyal, R. K. (2022). K-means clustering with optimal centroid: An optimization insisted model for removing outliers. *International Journal of Pattern Recognition and Artificial Intelligence*, *36*(10), 2259007.

Villanueva, K., Kilag, O. K., Abrenica, E., Samutya, M., Bocao, M., & Rabi, J. I. I. (2024). Charting a Course for Improvement: Assessing Mathematics Education in the Philippine Context. *International Multidisciplinary Journal of Research for Innovation, Sustainability, and Excellence (IMJRISE)*, *1*(4), 67-74.

Yang, Y., Sun, H., Zhang, Y., Zhang, T., Gong, J., Wei, Y., ... & Yu, D. (2021). Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. *Cell reports*, *36*(4).