INVESTIGATING LIFELONG LEARNING ACHIEVEMENTS WITH IN-DEPTH DATA ANALYSIS USING ENHANCED DP-MEANS CLUSTERING AND ETL INTEGRATION

[1]Gant Gaw Wutt Mhon

Faculty of Information Technology Supporting and Maintenance,
University of Computer Studies, Yangon
Myanmar
gantgawwuttmhon@ucsy.edu.mm

[2]Nilar Aye

Faculty of Information Science,
University of Computer Studies, Yangon
Myanmar
nilaraye@ucsy.edu.mm

Abstract

In the realm of educational data analytics, the synergy of clustering algorithms and ETL (Extract, Transform, Load) integration is harnessed to explore patterns in lifelong learning achievements. This paper proposes integrating enhanced DP-Means clustering (Dirichlet Process Means Clustering) and ETL processes to examine educational outcomes, considering evaluations of secondary and higher education accomplishments, familial information, and socioeconomic criteria for each student. The enhanced DP-Means clustering utilizes two primary methods to enhance clustering performance. The choice of variables was carefully examined through comprehensive data analysis conducted via the ETL process. The study evaluates the enhanced DP-Means algorithm's performance using a benchmark dataset and compares the results with K-Means and the original DP-Means algorithm outperforms existing techniques in cluster analysis, achieving significantly higher scores across three validation indexes.

***Keywords*: Lifelong learning achievements, Enhanced DP-Means clustering, ETL processes.**

1. Introduction

Recognizing and exploring continuous learning accomplishments holds paramount significance in the context of educational institutions, and each student's educational journey. The research aims to contribute insights by employing DP-Means clustering and ETL integration techniques. DP-Means clustering identifies distinct subgroups, facilitating an exploration of diverse learning trajectories, while ETL integration ensures the aggregation, cleaning, and structure of data from diverse sources, making it suitable for in-depth analysis.

Previous studies primarily focused on categorical comparisons of clustering algorithms. This study diverges by concentrating on the non-parametric generalization of the clustering method. DP-Means employs a non-parametric approach, leading to the automatic determination of cluster numbers based on the data. While this is advantageous in some scenarios, it introduces uncertainties and complexities, making it crucial to refine and adapt the algorithm for specific applications. The effectiveness of DP-Means clustering is impacted by the distribution of data, particularly in diverse datasets. Addressing these challenges is crucial to harness the algorithms' potential in research fully. Continuous efforts to improve its robustness and scalability are essential for effectively addressing real-world clustering problems [Basu et al., (2023)]. The K-Means algorithm faces challenges in its clustering performance. These include the need for predefining cluster numbers (K) and random initialization of cluster centers, especially problematic for large datasets. Its reliance on Euclidean distance limits its ability to detect diverse cluster shapes and overlapping clusters [Ikotun et al., (2023)]. While K-Means is efficient and straightforward, DP-Means offers enhanced adaptability and robustness, albeit with increased computational complexity due to density estimation. Both algorithms cater to different clustering requirements, offering trade-offs between simplicity and flexibility. Nevertheless, it is crucial to acknowledge that the selection between DP-Means and K-Means hinges upon the distinctive variables of the data as well as the objectives of the clustering endeavor.

DP-Means stands out for its ability to automatically determine the number of clusters without requiring prior knowledge of the dataset. However, DP-Means face challenges with parallelization, limiting its scalability in large-scale applications. Building upon the insights into DP-Means' struggle with parallelization, recent advancements have aimed to overcome these limitations. DP-Means, a nonparametric extension of K-Means, to situations with an unknown number of clusters, and faces challenges in parallelization, limiting its applicability to large-scale tasks. The researcher introduced a parallel algorithm termed PDC-DP Means (Parallel Delayed Cluster DP-Means), designed to address the limitations of previous attempts to parallelize DP-Means and to provide a faster and more scalable solution [Dinari et al., (2022)]. While DP-Means presents notable potential, ongoing research and development are crucial for refining its capabilities and expanding its applicability in diverse domains beyond traditional clustering tasks.

Anticipating students' performance is one of the most important topics for learning contexts such as schools and universities since it helps to design effective mechanisms that improve academic results and avoid dropout, among other things. The automation of various student activities, facilitated by the use of technology-enhanced learning tools, generates vast amounts of data. Analyzing and processing this data offers valuable insights into students' knowledge levels and their interactions with academic tasks, providing opportunities for informed decision-making and targeted interventions to support student success [Rastrollo-Guerrero et al., (2020)]. The academic community faces challenges in scrutinizing and evaluating student academic performance. Classifying student performance is complex with recent studies utilizing cluster analysis and statistical techniques, which are inefficient [Omar et al., (2020)].

Clustering algorithms' performance can vary substantially across different applications and data types, influenced by various parameters, operational aspects in high-dimensional spaces, and challenges with noisy, incomplete, and sample data [Rodriguez et al., (2019)]. DP-Means clustering is well-suited for examining lifelong learning due to its ability to automatically determine optimal cluster numbers based on data. This flexibility accommodates the complex nature of lifelong learning. When applied to longitudinal learning achievement data, DP-Means clustering identifies distinct clusters of lifelong learners with similar skill development patterns, educational achievements, and learning engagement, offering nuanced insights into factors such as learning preferences, motivation, prior knowledge, and access to learning resources. The proposed enhanced DP-Means clustering with ETL processes integration is to facilitate parameter selection and initial cluster identification, particularly beneficial for student datasets across various educational levels in continuous learning contexts.

The subsequent sections of this paper are structured as follows: Section 2 discusses the related work concerning a comprehensive analysis of student performance, which involves examining a broad range of student information, such as personal and academic data with the overarching goal of fostering ongoing educational improvement and enhancing academic efficiency. Section 3 provides an overview of the dataset and its preparation. Section 4 discusses the methodology used to carry out the experimental work, offering a step-by-step explanation of the proposed system, process, and algorithm. Section 5 presents the analysis of the results, and the final section 6 concludes the paper by providing insights into the future direction of the system.

2. Related Work

While integrating testing and data holds promise for enhancing educational outcomes, it also presents challenges and considerations. Privacy concerns, data security issues, and the ethical use of student data necessitate careful deliberation. Moreover, the complexity of data integration and analysis poses technical and logistical challenges for institutions. Building upon the insights from previous studies, the following literature review aims to explore the multifaceted challenges and opportunities associated with data-driven decision-making in education, with a particular focus on using cluster analysis techniques and various Machine Learning (ML) methods.

In [Khayi et al., (2019)], researchers applied various clustering algorithms on pre-test data collected from 264 high-school students who participated in a five-week experiment with an intelligent tutoring system to identify clusters of students with similar knowledge patterns. This investigation aimed to analyze student groups based on misconceptions and mastered concepts, providing insights for instructional task authoring and within-task instructional strategies. DP-Means clustering yielded excellent results with binary response data. while k-modes performed better with categorical response data. Individualized strategies for each learner could potentially be achieved through automated methods like reinforcement learning, such approaches require more experimental data, which is currently unavailable.

In [Vettori et al., (2020)], researchers put forth theoretical and practical implications discussions by employing a person-oriented approach to examine cognitive, socio-cultural, affective/motivational, and attributional/regulative components of learning. Data collection is a sample of 243 university students who completed the Learning Conceptions Questionnaire (LCQ), a validated self-report instrument. Non-hierarchical cluster analysis, employing a two-step method, was used to identify distinct profiles among the participants. This study, explored interactions between three primary profiles and variables such as gender, educational level, and academic disciplines, providing insights into how different factors may influence learning profiles.

In [Nagahi et al., (2020)], researchers explored the relationship between students' systems thinking (ST) skills and proactive personality (PP) on the academic performance of engineering students. This review is interested in considering the important contribution decided to data collection which involved the administration of established instruments for ST skills, comprising seven dimensions, and PP, comprising one dimension, through a web-based cross-sectional survey using Qualtrics. Studies in the literature developed three distinct machine-learning models using SPSS to classify students based on their Cumulative Grade-Point Average (CGPA) and current Semester Grade-Point Average (SGPA).

In [Asikainen et al., (2020)], researchers discussed the importance of considering students' learning profiles in addressing study-related burnout, suggesting that effective time and effort management along with deep-level understanding are crucial for reducing burnout and promoting well-being. Additionally, inadequacy and exhaustion were found to be negatively associated with study success, underscoring the significance of addressing study-related burnout for overall academic achievement. The experiment, involved 339 first-year students who voluntarily participated, followed by confirmatory factor analysis conducted on the complete dataset to identify consistent profiles across various disciplines.

In [Han, H. (2023)], researchers proposed the Heuristic Fuzzy C-Means Clustering Algorithm (HFCA) to analyze college students' stress levels, psychological well-being, and academic performance. Data from the Kaggle stress dataset are utilized, with a focus on identifying psychological factors affecting academic success. Fuzzy clustering algorithms are employed to uncover critical aspects of student engagement and satisfaction, aiding in developing targeted interventions to address mental health challenges. Despite its effectiveness in emotional feedback provision, limitations in small dataset performance suggest potential improvements with large datasets and varied student feedback evaluations.

In [Orsoni et al., (2023)], researchers presented the effectiveness of two clustering techniques for grouping students based on cognitive abilities. Comparing a two-level approach combining Kohonen's Self-Organizing Map (SOMs) and K-Means, 292 students aged 11-15 years participated. Results indicated the two-level approach as the optimal solution, with AdaBoost and ANN-supervised algorithms used for profile prediction. This study, conducted on Italian secondary school students, measures cognitive abilities through an online digital game. Hybrid clustering, incorporating SOMs and K-Means, enhanced clustering reliability, particularly in analyzing cognitive profiling among groups with and without specific learning difficulties (SLD). These findings implied the utility of hybrid clustering techniques in psychology, improving clustering reliability and solution accuracy.

In [Rajendran et al., (2023)], researchers predicted middle and high school academic performance using ML algorithms. The primary emphasis is on the consideration of various socio-demographic, school-related, and student-related variables. Grade Point Average (GPA) served as the model output and evaluation metrics include precision, recall, and F1-score. This study, examined health-conscious lifestyle positively impacts academic performance, while stress negatively affects it and then gender is not a significant predictor of academic performance.

In [Priyambada et al., (2023)], researchers suggested a two-layer ensemble learning technique, combining ensemble learning and ensemble-based progressive prediction, utilizing students' learning behavior data and domain knowledge. Data from academic information systems, focusing on students' course-taking behavior, facilitated predictions aligned with curriculum guidelines. Course-taking behavior served as an indicator of potential performance, considering alignment with curriculum guidelines and domain knowledge. This research indicated the results with improved accuracy for predicting students' performance is crucial for both students and academic stakeholders in higher education. The main focus of this paper is the acknowledgment that past performance also requires consideration.

In [Issash et al., (2023)], researchers reviewed based on 84 selected publications highlighting various ML methods used to analyze student characteristics and their impact on academic performance. The review also identified academic and demographic attributes as the primary factors influencing performance, with classification and decision trees being the most commonly employed methods. GPA, an indicator of academic accomplishment, is explored to strongly influence researchers' perceptions of student success. This paper addressed gaps in research on basic academic performance and intervention plans for poor performance, benchmarked datasets, and early intervention strategies are needed.

Through an extensive analysis of existing research, the review sheds light on the predictive capabilities of academic performance from various learning systems. Studies have been utilized to explore the predictive factors influencing academic success and the methodologies employed to forecast students' performance. This research emphasizes the multifaceted nature of academic achievement prediction, urging the need for comprehensive data analysis and targeted interventions to support students' educational journey effectively.

3. Dataset

The dataset sourced from the “Mendeley Data Repository” offers a comprehensive look into the academic, social, and economic backgrounds of 12,411 students and has 44 variables. The identification of criteria enabled the analysis of academic performance among engineering students at two distinct junctures. The initial phase encompasses the outcomes of secondary evaluations, while the subsequent phase pertains to the results of professional assessments. Furthermore, variables about the social context in which the students reside have been incorporated into the dataset. This dataset was obtained through a systematic integration of databases from the Colombian Institute for the Evaluation of Education (ICFES) [Delahoz-Dominguez et al., (2020)].

Unnecessary or unrelated variables can detrimentally impact both the performance and processing time of the clustering process. While the data originates from Colombia, this research primarily aims to investigate lifelong learning achievements across various educational levels. Therefore, modifications were made to this dataset to align it with the research objectives. The selection of variables underwent thorough scrutiny through an in-depth data analysis facilitated by the ETL processes. Before employing the proposed clustering algorithms, this dataset went through preprocessing via the ETL processes.

4. Proposed System and Methodology

The methodology outlined in this study is derived from the overview depicted in Fig. 1, which illustrates the proposed system. This methodology can be divided into three distinct subsections. The initial subsection provides the data analysis conducted, including the ETL processes with an elucidation of the benchmark dataset utilized in the experimentation. Subsequently, the second subsection delves into the enhanced DP-Means clustering algorithm, incorporating modifications aimed at determining both the threshold parameter (λ) and identifying the initial number of clusters. Finally, the last section encompasses the evaluation metrics used to compare the performance of the enhanced DP-Means algorithm.

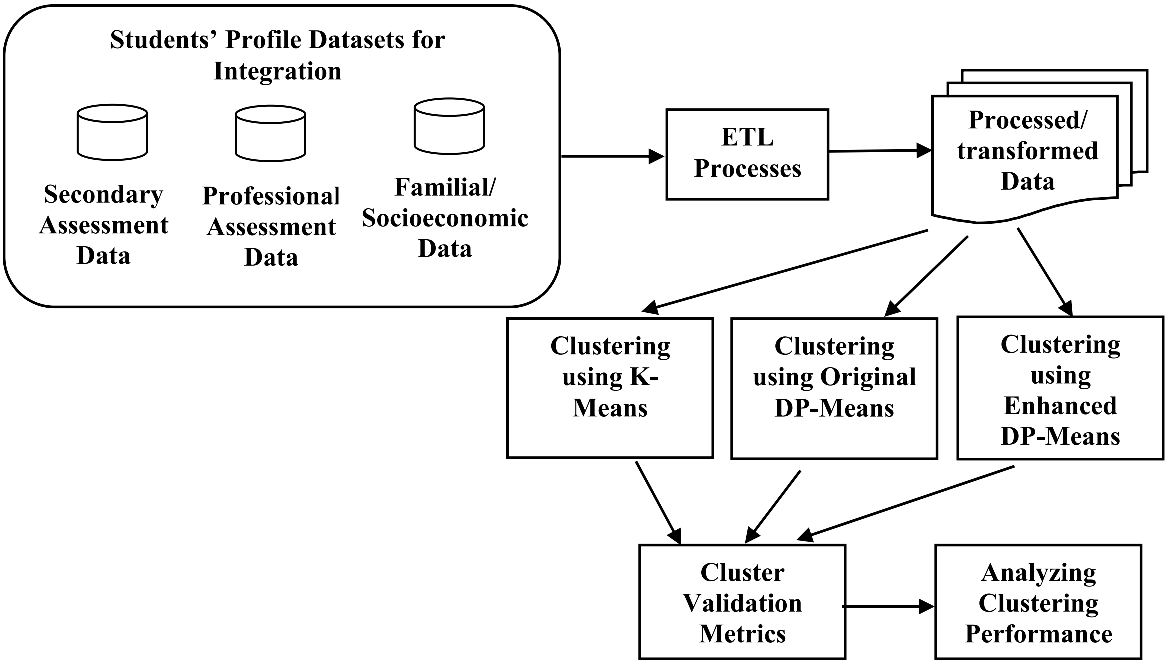
Students’ Profile Datasets for Integration
Secondary Assessment Data
Professional Assessment Data
Familial/ Socioeconomic Data
ETL Processes
Processed/ transformed Data
Clustering using K-Means
Clustering using Original DP-Means
Clustering using Enhanced DP-Means
Cluster Validation Metrics
Analyzing Clustering Performance

Fig. 1. Overview of the Proposed System.

4.1. *Data Analysis with ETL Process*

Collecting and preprocessing student-related data from diverse educational sources presents challenges due to its complexity and extensive nature. Integrating factors like previous educational results and family background further complicates the process, necessitating data reconciliation for accuracy. However, advanced data analytics techniques offer the potential to reveal insights into determinants of student academic success, facilitating evidence-based interventions to promote educational equity and excellence.

The ETL process is a versatile method commonly employed in data management and analytics and also ensures data integrity in the warehouse through standardization and the removal of redundant entries [Geetha, K. (2020)]. This process automates selecting, collecting, and conditioning data from a data warehouse, ensuring the output data is formatted optimally for further processing or business purposes [Galici et al., (2020)]. Different tests served specific functions, revealing a research gap in advanced analysis techniques for distributed data sources [Yulianto, A.A. (2019)].

Table 1 outlines the details of variables employed for experimental purposes. Based on the benchmark dataset's composition, data need to be collected from various sources. In this study, data are intended to be categorized into three types of records: academic (professional) performance records consist of nine variables, secondary performance records comprise nine variables, and household socioeconomic status records encompass twenty-three variables. Some variables are being dropped due to their limited relevance in capturing the underlying patterns or characteristics of the student cohort being evaluated for achievement. A new variable (RECod_Sid) has been introduced as a student identifier with an automatic generation method to ensure confidentiality while enabling the linkage of each student's relevant information across various types of records. Subsequently, following the application of the ETL process, two additional variables surfaced: Total_S11 and Total_SPro. These variables serve as aggregating and grading structures, respectively, categorizing total marks into levels 1 through 4 to represent the cumulative scores of five generic academic competencies in secondary and professional education.

Variable	Description
RECod_Sid	Student's ID
Categorical Variables	
GENDER	Gender
EDU_ FATHER	Father's education
EDU_ MOTHER	Mother's education
OCC_FATHER	Father's occupation
PEOPLE_ HOUSE	People in the house
INTERNET	Internet
TV	TV
COMPUTER	Computer
WAHING_ MCH	Washing machine
CAR	Car
DVD	DVD
FRESH	Fresh
PHONE	Phone
MOBILE	Mobile
REVENUE	Revenue
JOB	Job
SCHOOL_NAT	Nature of school
SCHOOL_TYPE	Type of school
ACADEMIC_PROGRAM	23 Academic programs
Numerical Variables	
MAT_S11	Math problem-solving skills (1-100)
CR_S11	Understanding text locally or globally (1-100)
CC_S11	Citizenship and inclusive coexistence concept (1-100)
BIO_S11	Explaining natural phenomena based on observations and scientific knowledge (1-100)
ENG_S11	Effective communication in English (1-100)
Total_S11	Level of total marks in five generic competencies (1-4)
QR_PRO	Understanding and manipulating quantitative data (1-100)
CR_PRO	Understanding text locally or globally (1-100)
CC_PRO	Citizenship and inclusive coexistence concept (1-100)
ENG_PRO	Effective communication in English (1-100)
WC_PRO	Written Communication (1-100)
Total_SPro	Level of total marks in five generic academic competencies (1-4)
FEP_PRO	Average engineering project test score (1-300)
G_SC	Overall professional evaluation score (1-300)
PERCENTILE	Secondary level identification (1-100)
2ND_DECILE	Secondary level identification (1-5)
QUARTILE	Academic level identification (1-4)
SEL	Socioeconomic Level (1-4)
SEL_IHE	Institution's Socioeconomic Level (1-4)

Table 1. Description of variables used.

Table 2 presents a step-by-step outline of the proposed ETL processes designed to analyze the benchmark dataset. These processes play a foundational role in data management, involving the extraction of data from various sources, its transformation into a suitable format, and loading it into a target system. In the context of cluster analysis, an effective ETL process ensures that the data is properly cleansed, integrated, and structured, laying the groundwork for accurate and meaningful clustering results.

Process	Description
Step 1: Extraction	- Harvesting data - Ensuring data
Step 2: Transformation Step 2.1: Iteration and Optimization	- Analyzing data distribution - Aggregating five competencies into a total score - Checking variation - Grouping for contribution assessment (significantly or not) - Assessing categorical variable (encode or not) - Variable renaming/ pruning for loading simplification - Grading Structure Implementation - Visualization with heatmap (detect correlations/prioritize analysis)
Step 3: Loading	- Connecting and loading datasets - Data staging - Loading schema design

Table 2. ETL processes for in-depth data analysis to investigate student learning achievement.

The ETL process was applied to three types of records from benchmark dataset, as outlined in Table 2. Variables such as RECod_Sid, QR_PRO, CR_PRO, CC_PRO, WC_PRO, G_SC, QUARTILE, Total_SPro, MAT_S11, CR_S11, CC_S11, BIO_S11, ENG_S11, DECILE (as a variable renaming of 2ND_DECILE), Total_S11, SEL, SEL_IHE, GENDER_F, GENDER_M were loaded into the database as the final set of variables for subsequent clustering analysis. In Fig. 2, it is observed that SEL and SEL_IHE fail to significantly influence or correlate with other variables.

In contrast, among the indicative level identification variables in each record, QUANTILE, G_SC, and

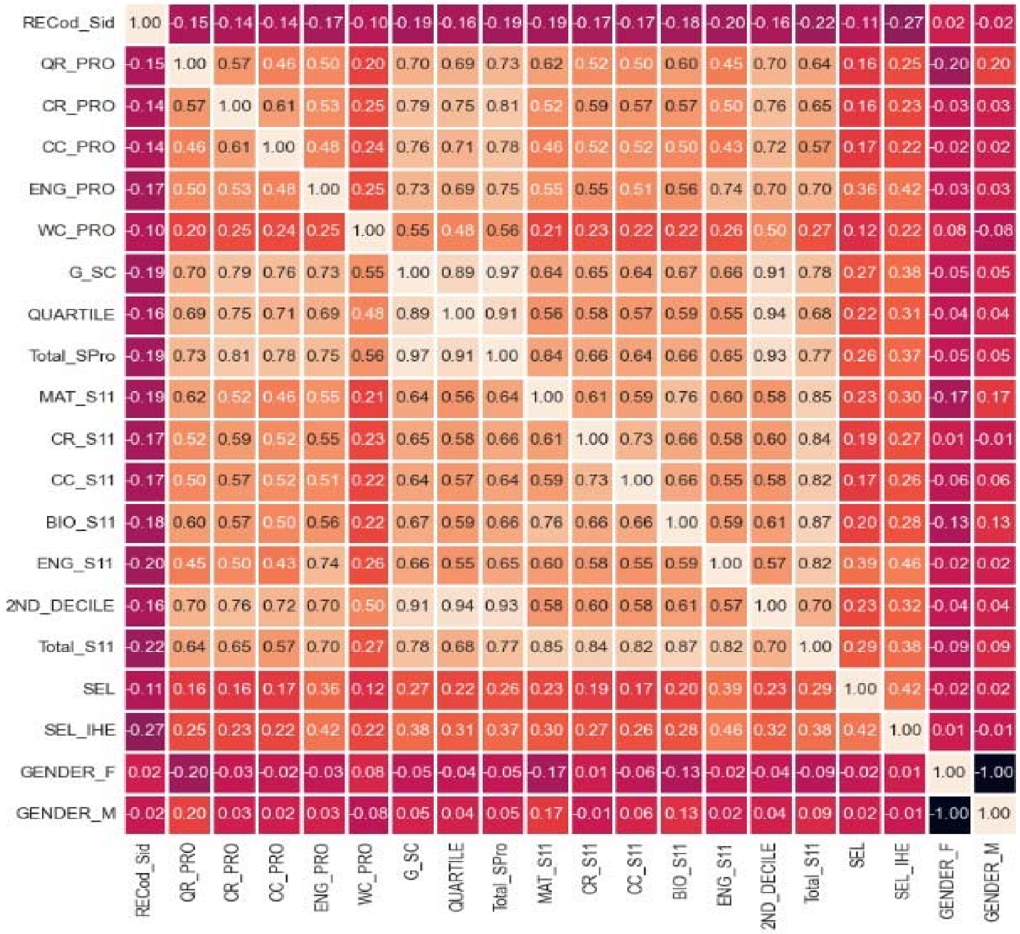

Fig. 2. Heatmap analysis reveals correlations between academic (professional), secondary achievements and household socioeconomic status levels.

DECILE exhibit significant effects and correlations with other variables. Their impact on the variables is nearly equivalent.

Metric	1	2	3	4
QR_PRO	40.49	59.68	74.76	90.19
CR_PRO	21.36	35.58	54.81	80.55
CC_PRO	19.49	33.00	50.51	77.88
ENG_PRO	31.43	46.02	60.98	83.17
WC_PRO	22.85	37.71	47.28	66.63
Total_SPro	135.61	211.99	288.34	398.43
G_SC	119.97	138.73	154.35	181.07
MAT_S11	51.83	56.14	60.77	70.58
CR_S11	49.62	53.36	58.09	66.16
CC_S11	49.50	53.62	58.08	65.96
BIO_S11	51.27	55.62	60.88	70.02
ENG_S11	48.65	51.68	56.48	69.69
Total_S11	250.87	270.42	294.29	342.41
DECILE	1.26	2.45	3.65	4.86
SEL	2.17	2.27	2.45	2.84
SEL_IHE	1.93	2.04	2.20	2.70
GENDER_F	0.44	0.42	0.43	0.38
GENDER_M	0.56	0.58	0.57	0.62

Table 3. Summary of Metrics by QUARTILE.

Table 3 illustrates the distribution of various academic and socioeconomic metrics across four QUANTILES, each representing distinct levels of achievement. It suggests a correlation between QUANTILE scores and secondary education metrics, indicating that students excelling in QUARTILE scores also perform well academically. However, there is no significant correlation between QUANTILE scores and socioeconomic indicators implying that academic success in secondary education is closely linked to professional performance but not strongly associated with socioeconomic variables. To provide a clearer representation of the five generic competencies, the Total_S11 and Total_SPro variables are utilized in Fig. 2 and Table 3 before implementing the grading structure in the proposed ETL process. Fig. 3 provides a visual comparison of the strength and direction of the correlations between the representative variables in each field. Higher correlation coefficients indicate stronger relationships between variables, while lower coefficients suggest weaker relationships.

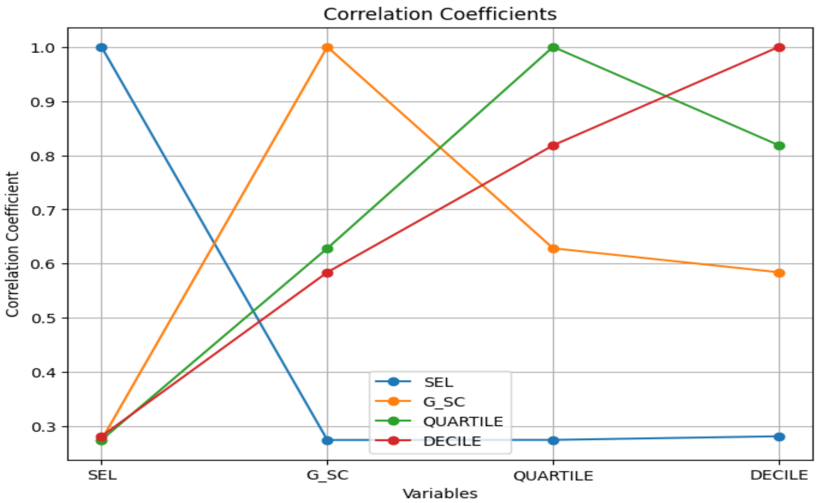

Fig. 3. Correlation coefficients between four variables.

After employing the proposed ETL process for data analysis, it was found that encoding all categorical variables as numerical is unnecessary. Prior to encoding, a process to analyze the relationship between these categorical variables and learning achievements is conducted. As a result, only the specified variables are deemed necessary for loading into the clustering analysis. This research was conducted utilizing PyCharm Professional.

4.2. *Enhanced DP-Means Clustering Algorithm*

The determination of the initial cluster and the λ value in the DP-Means clustering algorithm is critical. Initially, each data point is assigned to its cluster, with the threshold controlling the maximum cluster radius. This parameter affects the algorithm's sensitivity to new clusters, with smaller values creating more clusters. Finding the optimal λ is crucial for cluster granularity. Through iterative adjustments, the algorithm dynamically adapts to data distribution, uncovering underlying structures without predefining cluster numbers. The choice of initial cluster and λ profoundly impacts clustering accuracy and the algorithm's ability to capture data complexities.

While DP-Means uses a fixed threshold parameter to determine cluster membership, generalized DP-Means introduces a more flexible approach by allowing each cluster to have its λ. This enables the algorithm to better adapt to datasets with varying cluster densities and shapes [Kobayashi and Watanabe, (2021)]. While DP-Means directly assigns data points to clusters based on their distances to cluster centers, the Dirichlet Process Mixture Model Clustering (DPMC) employs a probabilistic framework, where each data point is assigned a probability distribution over all clusters. This probabilistic assignment allows for more flexible cluster assignments of data points to clusters [Melit Devassy and George, (2023)]. DP-Means provides a more deterministic clustering approach when the data distribution is well-defined and when clear, distinct clusters are expected. Hence, the proposed ETL process is employed before the application of the DP-Means clustering method. The step-by-step outline for the enhanced DP-Means clustering algorithm is given in Algorithm 1.

```
Input: data points(number of variables), initial threshold value $\lambda$
Output: final cluster centers, assigned labels

Begin:
Step 1: Initialization
          (1) set initial $\lambda$ value and cluster centers
Step 2: Find the optimal number of clusters
          (1) read the dataset and extracts relevant variables
          (2) iterate over a range of cluster numbers:
              (a) fits DP-Means clustering for each number of clusters
              (b) calculate the SSE(inertia) for each clustering and store it in the 'sse' list
          (3) plot the SSE against number of clusters to identify the 'elbow point"
Step 3: Find the optimal threshold value
          (1) execute the optimal threshold value based on the silhouette score
          (2) update the $\lambda$ value with the optimal threshold value
Step 4: Assignment
          (1) assign data points to clusters based on the updated $\lambda$ value
Step 5: Update cluster centers
          (1) recalculate cluster centers based on the assigned data points
Step 6: Repeat until convergence
          (1) repeat steps 4 and 5 until convergence is achieved
End
```

Algorithm 1. Enhanced DP-Means Clustering Algorithm.

The enhanced DP-Means clustering employs two key techniques to optimize clustering performance. Firstly, the Elbow method serves as a valuable heuristic for determining the optimal number of clusters. This method utilizes the Sum of Squared Errors (SSE) metric and visual analysis to identify the point at which additional clusters cease to significantly reduce SSE, thereby guiding decision-making regarding cluster quantity. Secondly, the algorithm prioritizes the identification of the λ value, crucial for defining cluster boundaries.

By evaluating clustering quality through the Silhouette score, it selects the λ value that maximizes this metric, ensuring that resulting clusters effectively capture underlying data patterns. By iteratively evaluating different cluster configurations and λ values, these additional steps allow for a more fine-tuned and data-driven approach to clustering. Consequently, the resulting clusters are better able to capture the underlying patterns and structures present in the dataset, leading to improved clustering outcomes compared to the original algorithm alone. The incorporation of these supplementary procedures contributes to more accurate and meaningful clustering results in enhanced DP-Means clustering.

To compare the performance of the enhanced DP-Means clustering with the original DP-Means and K-Means clustering algorithms are used. The K-Means algorithm is powerful but has limitations like random initialization issues and the need to predefine the K value, causing problems with convergence, cluster shapes, and outliers. It also struggles with diverse data types [Ahmed et al., (2020)]. The Elbow method assesses performance using SSE, finding the inflection point in K values with simple complexity. However, determining K can be challenging if the inflection point isn't clear due to the relationship between K and distance values [Yuan and Yang, (2019)]. The effectiveness of original DP-Means and K-Means depends on factors like data characteristics, objectives, and computational resources.

4.3. *Evaluation Metrics*

Three cluster validation measures, namely the Silhouette Coefficient (Silhouette score), Calinski-Harabasz (CH) Index, and Davies-Bouldin (DB) Index, are employed to evaluate the formed clusters and facilitate score comparisons. The Silhouette coefficient assesses how well each data point fits its assigned cluster while considering the separation between clusters.

A silhouette score close to +1 indicates correct clustering, near 0 suggests ambiguity, and close to -1 implies misclustering. The highest silhouette score determines the optimal cluster count [Shahapure and Nicholas, (2020)]. The CH index operates under the premise that clusters are highly compact and adequately separated to create a favorable distribution. A higher CH index value indicates dense, well-separated clusters, although no universally accepted cut-off value exists. Like the Silhouette and CH methods, the DB method assesses both cluster separation and compactness. Unlike the former two, a decrease in the DB index indicates improved clustering [Rachwał et al., (2023)].

5. Results and Analysis

Firstly, a series of experiments was conducted utilizing a dataset of 50 students to elucidate the experimental clustering outcomes. Assessing the proposed algorithms' clustering effectiveness involves evaluating clustering quality through diverse validation methods and comparing multiple clustering outcomes derived from varying variable counts, initial cluster numbers, and centroid settings tailored to the algorithms' clustering methodology. Table 4 offers a set of results of original DP-Means clustering using two variables.

Variables	K_0	Cluster instances in each cluster
['QUARTILE', 'Total_S11']	8	Cluster 0 → 8 Cluster 1 → 22 Cluster 2 → 6 Cluster 3 → 6 Cluster 4 → 3 Cluster 5 → 5

Table 4. Clustering results with two variables in original DP-Means at (λ=0.1).

The choice of variables used in clustering can significantly impact the resulting clusters. Type of student datasets have unique characteristics compared to other datasets, such as different distributions, densities, or patterns. These differences can influence the performance of clustering algorithms and result in divergent clustering outcomes. The number of variables included in the analysis affects the dimensionality of the dataset. Depending on the dataset's structure and characteristics, different initializations may lead to different partitioning of the data into clusters. In Table 4, the initial number of clusters K_0 is determined based on prior knowledge and the selected variables, with the resulting clusters and their instances presented.

QUANTILE	Total_S11	Number of cases
4	3	22
4	2	6
3	2	6
3	3	5
2	2	8
1	2	3

Table 5. Explanation of clustered results with two variables in original DP-Means.

Referring to Table 5, clusters resulting from the original DP-Means clustering algorithm are outlined. Specifically, twenty-two students are at academic (professional) level four, signifying secondary grading level three, and six students at academic (professional) level four correspond to secondary grading level two, and six students at academic (professional) level three align with secondary grading level two, and five students at academic (professional) level three correspond to secondary grading level three, and eight students at academic level two match secondary grading level two, and three students at academic (professional) level one coincide with secondary grading level two.

Table 6 provides the cluster validation measurement based on a different number of variables, K and K_0 values with a dataset of 12,411 students. The number of variables is six, with K equaling 39 and K_0 equaling 100, while for five variables, K is 32 and K_0 is 80. These differences are slight, with no significant variance observed in the three validation indexes. In cases where the differences in performance are subtle or unclear. The variability in clustering outcomes is notably heightened when confronted with a substantial number of variables, owing to the heightened intricacy of the data and the expanded dimensionality of the space. In contrast, a smaller set of variables tends to mitigate such discrepancies, resulting in clustering algorithms generating more uniform results.

While the default λ value of 0.1 may serve adequately in some cases, it's essential to adjust the λ value based on the dataset's characteristics, including its size, dimensionality, and the diversity of variables. Each type of student dataset offers unique insights into different aspects of students' academic, experiences, behaviors, and outcomes. The enhanced DP-Means algorithm with adaptive penalty mechanisms can provide more robust and reliable clustering results, particularly in datasets with different numbers of variables. These enhancements reduce the need for manual parameter tuning, ultimately improving the clustering process's efficiency and effectiveness.

Variables, K, K_0	**CQS**	**K-Means**	**DP-Means**

Variables = 7,
K = 300,
K_0 = 120
	Silhouette score	0.65	0.72

	CH score	486.49	256.04
	DB score	0.35	0.67

Variables = 4,
K = 36,
K_0 = 60
	Silhouette score	0.87	0.54

	CH score	456.67	432.02
	DB score	0.21	0.25

Variables = 2,
K = 6,
K_0 = 8
	Silhouette score	0.95	0.89

	CH score	564.73	734.82
	DB score	0.25	0.25

Table 6. Cluster Quality Scores (CQS) for different variable counts of K-Means and the original DP-Means.

Variables, Optimal λ, Optimal K_0	**Silhouette score**	**CH score**	**DB score**
Variables = 7 Optimal λ = 0.3, Optimal K0 = 321	0.99	525.16	0.32
Variables = 4 Optimal λ = 0.1, Optimal K0 = 33	0.99	641.91	0.27
Variables = 2 Optimal λ = 0.1, Optimal K0 = 9	0.99	954.91	0.27

Table 7. Cluster Quality Scores (CQS) for different variable counts of enhanced DP-Means.

As illustrated in Table 7, the optimal values of λ and K_0 are intricately linked, as they jointly influence the clustering outcomes and execution efficiency of the enhanced DP-Means algorithm. The variations observed in the number of variables are accompanied by changes in both λ and K_0. This table also provides the cluster quality scores for enhanced DP-Means across different variable counts. The enhanced DP-Means consistently achieves high Silhouette scores across all configurations, indicating well-defined clusters. Adjusting these parameters appropriately is essential for achieving optimal clustering results tailored to specific dataset configurations. In this research study, clustering of student datasets was conducted with careful consideration given to the selection of variables, chosen based on their significant effects, correlations with other variables, and representativeness of each field. Fig. 4 compares the execution times of three clustering algorithms across different numbers of variables. As the number of variables decreases, the execution times generally decrease for all algorithms. K-Means consistently has the highest execution times, followed by original DP-Means, while enhanced DP-Means generally has the lowest execution times across all variable counts.

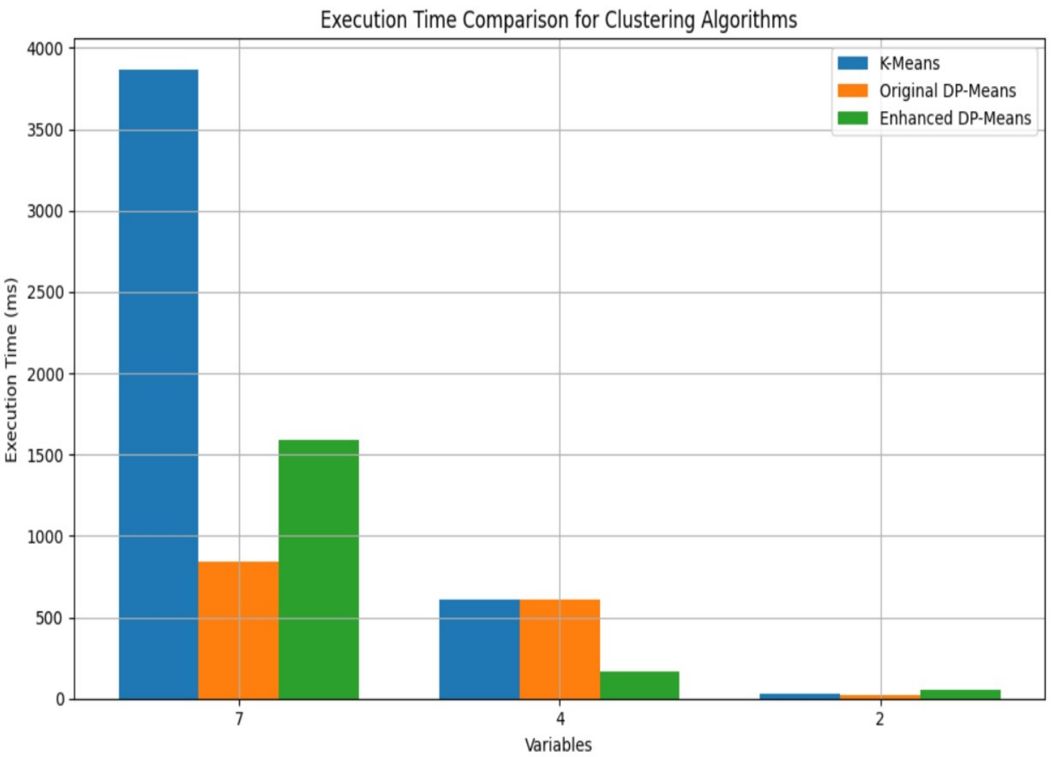

Fig. 4. Execution time comparisons for K-Means, original DP-Means and improved DP-Means.

QUANTILE	Total_S11	Number of cases	Clusters
4	4	27	8
4	3	5366	0
4	2	858	5
3	3	1228	3
3	2	1953	2
2	2	1666	1
2	3	271	6
1	2	962	4
1	3	80	7

Table 8. Explanation of clustered results with two variables in enhanced DP-Means.

Table 8 illustrates the clusters generated by the enhanced DP-Means clustering algorithm with 12,411 students. 5,366 students are categorized at academic (professional) level four, indicating secondary grading level three. Afterward, the clustering results indicate that student achievement in lifelong learning is influenced by their previous learning level, with some slight impact from socioeconomic factors.

6. Conclusion

This research aims to explore the patterns in lifelong learning achievements through the integration of clustering algorithms and ETL processes with the educational data analytics domain. By proposing the fusion of enhanced DP-Means clustering and ETL methodologies, the study delves into an in-depth examination of educational outcomes, encompassing secondary and higher educational achievements, familial information, and socioeconomic criteria for individual students. Through careful analysis facilitated by the ETL processes, the enhanced DP-Means clustering method leverages two primary techniques to bolster clustering performance. A benchmark dataset is employed to evaluate the performance of enhanced DP-Means algorithms, contrasting the results with those of the K-Means and original DP-Means clustering algorithms. By meticulously selecting variables, the clustering process prioritizes those with notable effects and correlations, ensuring the representation of each student’s performance level. The study’s findings underscore the superior performance of the proposed DP-Means algorithm, as evidenced by significantly higher scores across three validation indexes. The system’s future direction includes applying the enhanced DP-Means clustering algorithm to various datasets, aiming to uncover meaningful clusters and understand underlying data structures across different domains, facilitating informed decision-making.

Acknowledgments

I am immensely grateful to my esteemed supervisor for their unwavering guidance motivation, and steadfast support throughout this study. Their wealth of knowledge and innovative thinking have been invaluable assets to my research.

Conflicts of Interest

The authors have no conflicts of interest to declare.

References

[1] Ahmed, M., Seraj, R., Islam, S. M. S. (2020): The k-means algorithm: A comprehensive survey and performance evaluation. Electronics, vol. 9, no. 8, pp. 1295, MDPI.

[2] Asikainen, H., Salmela-Aro, K., Parpala, A., Katajavuori, N. (2020): Learning profiles and their relation to study-related burnout and academic achievement among university students. Learning and Individual differences, vol. 78, pp. 101781, Elsevier.

[3] Basu, S., Choudhury, J. R., Paul, D., Das, S. (2023): Robust and Automatic Data Clustering: Dirichlet Process meets Median-of-Means. arXiv preprint arXiv:2311.15384.

[4] Delahoz-Dominguez, E., Zuluaga, R., Fontalvo-Herrera, T. (2020): Dataset of academic performance evolution for engineering students. Data in brief, vol. 30, pp. 105537, Elsevier.

[5] Dinari, O., Freifeld, O. (2022): Revisiting dp-means: fast scalable algorithms via parallelism and delayed cluster creation. Uncertainty in Artificial Intelligence, pp. 579-588, PMLR.

[6] Galici, R., Ordile, L., Marchesi, M., Pinna, A., Tonelli, R. (2020): Applying the etl process to blockchain data. prospect and findings. Information, vol. 11, no. 4, pp. 204, MDPI.

[7] Geetha, K. (2020): Param (2020). Data Analysis and ETL Tools in Business Intelligence. International Research Journal of Computer Science (IRJCS), vol. 7, pp. 127-131.

[8] Han, H. (2023): Fuzzy clustering algorithm for university students' psychological fitness and performance detection. Heliyon, vol. 9, no. 8, Elsevier.

[9] 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, vol. 622, pp. 178-210, Elsevier.

[10] Issah, I., Appiah, O., Appiahene, P., Inusah, F. (2023): A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. Decision analytics journal, pp. 100204, Elsevier.

[11] Khayi, N. A., Rus, V. (2019): Clustering Students Based on Their Prior Knowledge. International Educational Data Mining Society, ERIC.

[12] Kobayashi, M., Watanabe, K. (2021): Generalized Dirichlet-process-means for f-separable distortion measures. Neurocomputing, vol. 458, pp. 667-689, Elsevier.

[13] Melit Devassy, B., George, S. (2023.): Unsupervised Clustering of Hyperspectral Data with an Unknown Number of Clusters Using Dirichlet Process Means. Available at SSRN 4469466.

[14] Nagahi, M., Jaradat, R., Davarzani, S., Nagahisarchoghaei, M., Goerger, S. R. (2020): Academic performance of engineering students. 2020 ASEE Virtual Annual Conference Content Access.

[15] Omar, T., Alzahrani, A., Zohdy, M. (2020): Clustering approach for analyzing the student’s efficiency and performance based on data. Journal of Data Analysis and Information Processing, vol. 8, no. 3, pp. 171-182, Scientific Research Publishing.

[16] Orsoni, M., Giovagnoli, S., Garofalo, S., Magri, S., Benvenuti, M., Mazzoni, E., Benassi, M. (2023): Preliminary evidence on machine learning approaches for clusterizing students’ cognitive profile. Heliyon, vol. 9, no. 3, Elsevier.

[17] Priyambada, S. A., Usagawa, T., Mahendrawathi, E. R. (2023): Two-layer ensemble prediction of students’ performance using learning behavior and domain knowledge. Computers and Education: Artificial Intelligence, vol. 5, pp. 100149, Elsevier.

[18] Rachwał, A., Popławska, E., Gorgol, I., Cieplak, T., Pliszczuk, D., Skowron, Ł., Rymarczyk, T. (2023): Determining the quality of a dataset in clustering terms. Applied Sciences, vol. 13, no. 5, pp. 2942, MDPI.

[19] Rajendran, S., Chamundeswari, S., Sinha, A. A. (2022): Predicting the academic performance of middle-and high-school students using machine learning algorithms. Social Sciences & Humanities Open, vol. 6, no. 1, pp. 100357, Elsevier.

[20] Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., Durán-Domínguez, A. (2020): Analyzing and predicting students' performance by means of machine learning: A review. Applied sciences, vol. 10, no. 3, pp. 1042, MDPI.

[21] Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F., Rodrigues, F. A. (2019): Clustering algorithms: A comparative approach. PloS one, vol. 14, no. 1, pp. e0210236, Public Library of Science San Francisco, CA USA.

[22] Shahapure, K. R., Nicholas, C. (2020): Cluster quality analysis using silhouette score. 2020 IEEE 7th international conference on data science and advanced analytics (DSAA), pp. 747-748, IEEE.

[23] Vettori, G., Vezzani, C., Bigozzi, L., Pinto, G. (2020): Cluster profiles of university students' conceptions of learning according to gender, educational level, and academic disciplines. Learning and Motivation, vol. 70, pp. 101628, Elsevier.

[24] Yulianto, A. A. (2019): Extract transform load (ETL) process in distributed database academic data warehouse. APTIKOM Journal on Computer Science and Information Technologies, vol. 4, no. 2, pp. 61-68.

[25] Yuan, C., Yang, H. (2019): Research on K-value selection method of K-means clustering algorithm. J, vol. 2, no. 2, pp. 226-235, MDPI.

Authors Profile

Gant Gaw Wutt Mhon received the M.C.Sc. degree from the University of Computer Studies, Yangon, Myanmar in 2010. She is currently working as a Lecturer in the University of Computer Studies, Pathein, Myanmar. Her other research interests are Data Mining, Machine Learning and Deep Learning. She can be contacted at email: gantgawwuttmhon@ucsy.edu.mm.

Dr. Nilar Aye got Ph.D(IT) in 2004. Currently, she is working as a professor of the Software Engineering and Applied Data Science Lab at the University of Computer Studies, Yangon, Myanmar. She has been supervising Master's thesis and Ph.D research on Machine Learning and Web Mining. She can be contacted at email: nilaraye@ucsy.edu.mm.