

**AN ANALYTICAL SYSTEM FOR LIFELONG
LEARNING ACHIEVEMENTS: INTEGRATING
EDP-MEANS CLUSTERING AND EDU-ETL
PROCESSES**



GANT GAW WUTT MHON

UNIVERSITY OF COMPUTER STUDIES, YANGON

JUNE, 2024

**An Analytical System for Lifelong Learning Achievements:
Integrating EDP-Means Clustering and Edu-ETL
Processes**

Gant Gaw Wutt Mhon

University of Computer Studies, Yangon

A thesis submitted to the University of Computer Studies, Yangon in partial
fulfillment of the requirements for the degree of

Doctor of Philosophy

June, 2024

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

.....

Date

.....

Gant Gaw Wutt Mhon

ACKNOWLEDGEMENTS

First of all, I would like to thank the Union Minister, the Ministry of Science and Technology for granting me for full facilities support during the Ph. D Course at the University of Computer Studies, Yangon.

I would like to express my special appreciation and deepest gratitude to **Dr. Mie Mie Thet Thwin**, Former Rector of the University of Computer Studies, Yangon, for giving me kindness and morality supports.

Secondly, I would like to express very special thanks to **Dr. Mie Mie Khin**, the Rector, the University of Computer Studies, Yangon, for allowing me to do this research and giving me technical supports during the period of my research.

I would like to describe my special thanks to my supervisor, **Dr. Nilar Aye**, Professor, Faculty of Information Science, the University of Computer Studies, Yangon. She took care of me. Because of her blessings I got the confidence to manage this research and all other hardships encountered in my research journey. This research would not be possible without her supervision, invaluable guidance and constructive ideas. I really gratitude for her supports which lead me to this point.

I would like to thank and respect **Dr. Sabai Phyu**, Professor, and former Dean of the Ph.D 11th Batch, the University of Computer Studies, Yangon, for her clear guidance, inspiration, and encouragement.

I would like to thank **Dr. Tin Thein Thwel**, Professor, and former Dean of the Ph.D 11th Batch, the University of Computer Studies, Yangon, for her advice and encouragement during my research.

I would like to thank **Dr. Si Si Mar Win**, Professor, and Dean of the Ph.D 11th Batch, the University of Computer Studies, Yangon, for her guidance and support during my research. Her insightful feedback, unwavering support, and commitment to fostering a productive research environment have significantly contributed to the successful completion of this dissertation.

I would like to express my gratitude regard to **Daw Aye Aye Khine**, Professor and Head of the English Department, the University of Computer Studies, Yangon, for her careful assistance from the language point of view and pointed out

the correct usage in my dissertation.

Finally, I wish to express my appreciation to all the people who greatly contributed and supported me throughout the writing of this dissertation.

Last but not least, I am profoundly grateful to my beloved family for their endless love and unwavering support. I extend my warmest thanks to my entire family, especially my aunts, who patiently endured this long process with me. Their constant encouragement and nurturing have been instrumental in the completion of my dissertation. Without their support, this achievement would not have been possible.

ABSTRACT

In the field of education, analyzing academic performance is vital for understanding student learning behaviors, identifying areas needing enhancement, and developing targeted interventions to improve educational outcomes. Traditional assessment methods typically depend on simple metrics like grades or standardized test scores; which often fail to capture the complexities of student proficiency and behavior. To overcome these limitations, educational researchers have increasingly adopted advanced data mining techniques and machine learning algorithms for a more granular and comprehensive analysis of academic performance data. This research proposes an Enhanced Dirichlet Process Means (EDP-Means) clustering algorithm combined with Educational Extract, Transform, Load (Edu-ETL) processes to evaluate academic performance across various educational levels. The proposed approaches aim to offer greater assurance and clarity in evaluating and supporting student achievements throughout their educational journey. The integration of Edu-ETL processes ensures data quality and consistency, preparing educational datasets for thorough analysis. The architecture of the proposed system utilizes the EDP-Means clustering algorithm, an improvement over the original DP-Means, for enhanced clustering performance. While both algorithms assign data points to clusters based on distance and threshold, EDP-Means introduces iterative optimization steps for improved accuracy and stability. In the original DP-Means algorithm, the number of clusters K and the threshold parameter λ were typically fixed or set based on heuristic choices. In EDP-Means, these parameters are dynamically adjusted based on the data characteristics and clustering quality, leading to more accurate and reliable clustering results. This study demonstrates that EDP-Means performs better and is comparable to traditional K-Means and original DP-Means algorithms in clustering educational data. To validate and prove the performance of EDP-Means, datasets from different fields were used to further experiment EDP-Means and ensure its effectiveness. Furthermore, the analysis of the PySpark environment underscores how the utilization of PySpark enhances the scalability and efficiency of EDP-Means, particularly in processing large-scale datasets.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	i
ABSTRACT	iii
TABLE OF CONTENTS	iv
LIST OF FIGURES	vii
LIST OF TABLES	x
LIST OF EQUATIONS	xi
1. INTRODUCTION	1
1.1 Focus of Research	3
1.2 Motivation of Research	5
1.3 Objectives of Research	6
1.4 Contributions of Research	7
1.5 Organization of Research	8
2. LITERATURE REVIEW AND RELATED WORK	10
2.1 Introduction to Educational Data Analysis (EDA)	11
2.2 Exploring Lifelong Learning in Education	14
2.3 Optimizing Educational Insights: Advanced Data Analysis and ETL Processes	15
2.4 Investigating Clustering Techniques in Educational Research	16
2.5 Educational Big Data Mining using PySpark	18
2.6 Chapter Summary	21
3. THEORETICAL BACKGROUND	23
3.1 DP-Means: A Nonparametric Extension of K-Means	23
3.1.1 Role of the Threshold Parameter (λ)	24
3.1.2 Key Concepts of DP-Means Clustering Technique	25
3.1.3 Key Steps of DP-Means Clustering Algorithm	29
3.1.4 K-Means and DP-Means Clustering	31
3.1.5 Critical Evaluation of DP-Means Clustering: Limitations and Challenges	33
3.2 Enhanced DP-Means: EDP-Means Clustering Technique	35
3.3 Unsupervised Evaluation Metrics	39
3.4 ETL and Edu-ETL Processes	41

3.5 Trends and Applications of Big Data in Education	47
3.5.1 Educational Aspects of Big Data	47
3.5.2 Big Data Framework: Apache Spark	48
3.6 Chapter Summary	49
4. THE ARCHITECTURE OF THE PROPOSED SYSTEM	51
4.1 Initial Data Acquisition	54
4.2 Data Analysis with Edu-ETL Processes	55
4.3 Clustering Process: K-Means, Original DP-Means, EDP-Means	56
4.4 Clustering Process in PySpark: K-Means, Original DP-Means, EDP-Means.....	59
4.5 Cluster Validation and Analysis.....	60
4.6 Analyze Learning Outcomes and Success Factors.....	61
4.7 Chapter Summary	61
5. IMPLEMENTATION OF THE EDU-ETL PROCESSES	62
5.1 Edu-ETL Processes in Action: Experiments and Case Studies	62
5.1.1 Dataset Overview and Categorization	62
5.1.2 Experimental Analysis of Datasets Using Edu-ETL Processes ..	67
5.2 System Demonstration	80
5.3 Chapter Summary	82
6. EXPERIMENTAL RESULTS AND EVALUATIONS	84
6.1 Experimental Setup and Procedures	85
6.1.1 Datasets Selection and Preparation	85
6.1.2 Experimental Methodology	86
6.2 Result and Analysis of the Proposed Analytical System	87
6.2.1 Experimental Evaluation of the EDP-Means Clustering Algorithm	88
6.2.2 Experimental Evaluation of the EDP-Means Clustering Algorithm with Edu-ETL Processes	98
6.3 Result and Analysis for Lifelong Learning Achievements of the Proposed Analytical System	112
6.4 Chapter Summary	115

7. CONCLUSION AND FUTURE WORKS	116
7.1 Dissertation Summary	117
7.2 Advantages and Limitations	119
7.3 Future Works	121
LIST OF ACRONYMS	122
AUTHOR’S PUBLICATION	123
BIBLIOGRAPHY	124

LIST OF FIGURES

Figure 2.1	Key Phases of EDA’s Life Cycle	11
Figure 2.2	Data-Driven Approach to Lifelong Learning	14
Figure 2.3	Key Applications of Big Data in Education	19
Figure 2.4	Model of Big Data Mapping in Higher Education System.....	20
Figure 3.1	Overview of ETL Processes	42
Figure 4.1	Architecture of the Proposed System	53
Figure 4.2	EDP-Means Clustering Algorithm	57
Figure 4.3	Overview of Clustering Processes in PySpark Environment	59
Figure 5.1	Distribution and Variation of Total_SPro and Total_S11 Attributes	69
Figure 5.2	Distribution of Different Competency Scores and Total Scores in University	69
Figure 5.3	Correlation Matrix for Attributes of Academic Performance	71
Figure 5.4	Correlation Matrix for Attributes of Secondary Performance	71
Figure 5.5	Heatmap Visualization for Attributes from Different Records	72
Figure 5.6	Implementation of Categorical Attributes with Values	72
Figure 5.7	Mean (2ND_DECILE) Values by School Nature and Type with Min and Max Values	73
Figure 5.8	Mean (Total_S11) Values by School Nature and Type with Min and Max Values	74
Figure 5.9	Mean (QUARTILE) Values by School Nature and Type with Min and Max Values	74
Figure 5.10	Average (SEL, SEL_IHE) Scores by Father Educational Level	75
Figure 5.11	Average (SEL, SEL_IHE) Scores by Mother Educational Level	75
Figure 5.12	Heatmap Analysis Reveals Correlations Between Academic (Professional), Secondary Achievement and Household Socioeconomic Status Levels	77

Figure 5.13	Correlation Coefficients Between Four Attributes	79
Figure 5.14	Main View of Program Demonstration	81
Figure 5.15	“Clustering” View of Program Demonstration	81
Figure 5.16	“Clustering on Spark” View of Program Demonstration	82
Figure 6.1	Comparative Analysis of Clustering Algorithms Using Silhouette Scores Across Different Datasets.....	90
Figure 6.2	Comparative Analysis of Clustering Algorithms Using CH Scores Across Different Datasets.....	90
Figure 6.3	Comparative Analysis of Clustering Algorithms Using DB Scores Across Different Datasets.....	90
Figure 6.4	Comprehensive Processing Time (ms) for Clustering, Optimal Cluster Numbers and Threshold Parameters Finding in K-Mean, DP-Means, and EDP-Means	91
Figure 6.5	CH Scores with λ , λ^* Values for DP-Means and EDP-Means	93
Figure 6.6	Comparative Analysis of Clustering Algorithms Using Silhouette Scores Across Different Datasets in PySpark Environment	94
Figure 6.7	Comprehensive Processing Time (ms) for Clustering, Optimal Cluster Numbers and Threshold Parameters Finding in K-Mean, DP-Means, and EDP-Means (PySpark Environment)	95
Figure 6.8	Clustering Result Visualization Using EDP-Means Algorithm for the “Diabetes” Dataset	97
Figure 6.9	Clustering Result Visualization Using EDP-Means Algorithm for the “Spotify Popular Music” Dataset	98
Figure 6.10	Processing Time (ms) Comparison for Clustering in K-Means, DP-Means, and EDP-Means	100
Figure 6.11	Comparative Analysis of Clustering Algorithms Using Silhouette Scores Based on:” Academic Evaluation” Dataset	102
Figure 6.12	Comparative Analysis of Clustering Algorithms Using CH Scores Based on:”Academic Evaluation” Dataset	102
Figure 6.13	Comparative Analysis of Clustering Algorithms Using DB Scores Based on:” Academic Evaluation” Dataset	102
Figure 6.14	Clustering Result Visualization Using DP-Means Algorithm for Key Attributes -Mat_S11, QR_PRO	106
Figure 6.15	Clustering Result Visualization Using EDP-Means Algorithm for Key Attributes -Mat_S11, QR_PRO	106

Figure 6.16	Clustering Result Visualization Using K-Means Algorithm for Key Attributes -Mat_S11, QR_PRO	107
Figure 6.17	Clustering Result Visualization Using DP-Means Algorithm for Key Attributes -ENG_S11, ENG_PRO	107
Figure 6.18	Clustering Result Visualization Using EDP-Means Algorithm for Key Attributes -ENG_S11, ENG_PRO	108
Figure 6.19	Clustering Result Visualization Using K-Means Algorithm for Key Attributes -ENG_S11, ENG_PRO	108
Figure 6.20	Clustering Result Visualization Using DP-Means Algorithm for Key Attributes -CC_S11, CC_PRO	109
Figure 6.21	Clustering Result Visualization Using EDP-Means Algorithm for Key Attributes -CC_S11, CC_PRO	109
Figure 6.22	Clustering Result Visualization Using K-Means Algorithm for Key Attributes -CC_S11, CC_PRO	110
Figure 6.23	Clustering Result Visualization Using DP-Means Algorithm for Key Attributes -CR_S11, CR_PRO	110
Figure 6.24	Clustering Result Visualization Using EDP-Means Algorithm for Key Attributes -CR_S11, CR_PRO	111
Figure 6.25	Clustering Result Visualization Using K-Means Algorithm for Key Attributes -CR_S11, CR_PRO	111

LIST OF TABLES

Table 3.1	Summarizing the Role and Characteristics of λ in DP-Means Clustering Algorithm	24
Table 3.2	Notations for the PDF of a Multivariate Gaussian Distribution	26
Table 3.3	Notations for the PDF of a GMM	26
Table 3.4	Notations Used in the CRP Analogy	28
Table 3.5	Applications of DP-Means Clustering with GMM and CRP	28
Table 3.6	Key Steps in the DP-Means Clustering Algorithm	30
Table 3.7	Notations Used in Key Steps of DP-Means Clustering Algorithm	31
Table 3.8	Comparison of K-Means and DP-Means Clustering Algorithm	32
Table 3.9	Steps of EDP-Means Clustering Algorithm	36
Table 3.10	Comparison of Original DP-Means and EDP-Means Clustering Algorithms ...	37
Table 3.11	Edu-ETL Processes for In-Depth Data Analysis	43
Table 3.12	Differences between ETL and Edu-ETL Processes	45
Table 4.1	Notations for the EDP-Means Clustering Algorithm	58
Table 5.1	Description of Numerical Attributes	63
Table 5.2	Description of Categorical Attributes	64
Table 5.3	Each Level Description of Categorical Attributes	65
Table 5.4	Summary of Metrics by QUARTILE	78
Table 6.1	Datasets in Different Sizes and Domains	86
Table 6.2	Cluster Quality Scores for K-Means, DP-Means, and EDP-Means	88
Table 6.3	Silhouette Scores Comparison for K-Means, DP-Means, and EDP-Means in PySpark Environment	94
Table 6.4	Comparison of Cluster Quality Scores and Processing Time (Clustering) for Key Attributes Using K-Means, DP-Means and EDP-Means.....	99
Table 6.5	Optimal Numbers of Clusters and Threshold Values for Key Attributes Using K-Means, DP-Means, and EDP-Means	104
Table 6.6	Comparison of the Optimal Number of Clusters Based on SSE Values	104
Table 6.7	Cluster Interpretation of Prevalent Patterns in Each Cluster Group	112

LIST OF EQUATIONS

Equation 3.1 25
Equation 3.2..... 26
Equation 3.3 27
Equation 3.4 27
Equation 3.5 40
Equation 3.6 40
Equation 3.7 40

CHAPTER 1

INTRODUCTION

In the realm of education, assessing academic performance plays a crucial role in understanding student learning patterns, identifying areas of improvement, and tailoring interventions to enhance educational outcomes. Because of developments in today's educational domain, there is now a wealth of data related to students, aimed at being more efficient and gaining a better understanding of students. The automation of student activities, facilitated by technology-enhanced learning tools, generates extensive datasets. Analyzing and processing this data yields valuable insights into students' knowledge levels and their engagement with academic tasks, paving the way for informed decision-making and targeted interventions to bolster student success [39]. However, despite these advancements, there is still a need for effective approaches to analyze student performance. Traditional methods of academic performance evaluation often rely on simplistic metrics such as grades or standardized test scores, which may overlook the nuanced aspects of student proficiency and learning behavior.

The research proposes Enhanced Dirichlet Process Means (EDP-Means) based clustering, combined with Educational Extract, Transform, Load (Edu-ETL) processes, to evaluate academic performance spanning various educational levels in continuous learning environments. Dirichlet Process Means (DP-Means) clustering, a variant of the K-Means algorithm, offers the advantage of automatically determining the optimal number of clusters while accommodating varying cluster shapes and sizes. However, this capability also introduces uncertainties and complexities, emphasizing the importance of refining and adapting the algorithm for specific applications. Addressing these challenges becomes pivotal in fully unlocking the algorithm's potential in research. Efforts to enhance its effectiveness and scalability are imperative for effectively tackling real-world clustering dilemmas [2][18]. Additionally, improvements in the algorithm are necessary to effectively utilize prior knowledge for the identification of the threshold parameter (λ) value in DP-Means clustering.

The ETL process stands as a versatile technique widely utilized in data management and analytics, serving to maintain data integrity within the warehouse by

implementing standardization and eliminating redundant entries [4]. Gathering and preparing student data from various educational outlets poses considerable challenges owing to its intricate and expansive nature. The incorporation of elements such as prior academic records and familial context adds layers of complexity, requiring meticulous data reconciliation to ensure precision. ETL integration ensures the consolidation, cleansing, and organization of data derived from various sources, rendering it conducive to thorough analysis.

The proposed approach begins with Edu-ETL processes, specifically designed for in-depth data analysis to investigate student learning achievement. These processes are tailored to handle and analyze educational datasets, addressing their unique requirements. An effective ETL process ensures data quality and consistency in cluster analysis, laying the groundwork for robust analysis and interpretation. Once the data is prepared through Edu-ETL processes, EDP-Means clustering is employed to categorize students into distinct groups based on their educational outcomes, considering various factors such as secondary and higher education accomplishments, familial information, and socioeconomic criteria for each student. Unlike traditional clustering algorithms in Machine Learning (ML) that require specifying the number of clusters in advance, DP-Means automatically determines the optimal number of clusters based on the data distribution, thereby mitigating the need for subjective decisions and enhancing the robustness of the clustering results. However, the DP-Means clustering algorithm is sensitive to initial parameters such as the initial number of clusters or cluster centers, which may impact the final clustering results. EDP-Means improves upon traditional DP-Means clustering by incorporating optimization techniques to achieve better clustering performance, particularly in complex datasets.

Forecasting student performance is paramount in educational settings like schools and universities, as it enables the development of efficient mechanisms to enhance academic outcomes and prevent dropout rates, among other benefits. The academic community encounters difficulties in thoroughly examining and assessing student academic performance. The classification of student performance poses a complex challenge, with recent research employing cluster analysis and statistical techniques, which are deemed inefficient [53]. Besides, examining lifelong learning through clustering poses several challenges, including the heterogeneous nature of educational data, high dimensionality, temporal dynamics, sparsity, and imbalance, which complicate the identification and interpretation of meaningful clusters.

Moreover, ensuring the interpretability and validation of clustering results in the absence of ground truth labels and addressing ethical and privacy concerns surrounding data usage are essential considerations. Overcoming these challenges necessitates interdisciplinary collaboration, methodological innovation, and ethical frameworks to effectively analyze educational data, derive actionable insights, and inform evidence-based educational practices.

DP-Means clustering proves to be well-suited for the examination of lifelong learning, owing to its capability to automatically determine optimal cluster numbers based on data. This adaptability aligns well with the intricate nature of lifelong learning. When applied to longitudinal learning achievement data, DP-Means clustering discerns distinct clusters of lifelong learners exhibiting similar skill development patterns, educational accomplishments, and learning engagement. This provides nuanced insights into factors such as learning preferences, motivation, prior knowledge, and access to learning resources. However, the intricate task of determining the initial cluster centers and the optimal value of λ in DP-Means clustering poses significant challenges. To overcome these limitations and enhance the effectiveness of clustering analysis, the utilization of enhanced DP-Means clustering becomes crucial in this research endeavor. Therefore, through the proposed integration of EDP-Means clustering with Edu-ETL processes, this research aims to provide a comprehensive solution for exploring patterns in lifelong learning achievements and advancing educational data analytics.

1.1 Focus of Research

Understanding lifelong learning achievements is vital for educational data analysis and students themselves. In educational data analysis, comprehending student performance is crucial for enhancing educational outcomes. However, traditional assessment methods often lack granularity, overlooking the multifaceted aspects of student proficiency and learning behavior. To address this limitation, a comprehensive analysis of various features of student datasets is essential. These datasets encompass diverse dimensions such as academic achievements, familial context, and socioeconomic background, which collectively contribute to student learning trajectories. These factors significantly shape student learning trajectories, providing valuable insights into their long-term academic progress. Such insights enable

educators to tailor interventions and support mechanisms effectively to meet students' evolving needs. However, in contemporary educational data analysis research, the exploration of lifelong learning achievements remains a pertinent area of inquiry. Moreover, in an increasingly dynamic and interconnected world, the ability to adapt and continuously learn is becoming increasingly critical for success in both academic and professional spheres.

Rethinking the original ETL processes is necessary to effectively manage the intricacies of educational data. This involves ensuring data quality, consistency, and integration across diverse sources. By consolidating and preparing data for analysis, ETL processes enable researchers to derive meaningful insights. Similarly, there is a growing recognition of the necessity to modernize traditional clustering techniques by adopting DP-Means clustering. Unlike conventional methods, DP-Means clustering offers the advantage of automatically determining optimal cluster numbers while accommodating varying data distributions. However, it remains imperative to address its limitations, including sensitivity to initial parameters, potential challenges in handling large datasets, or difficulties in dealing with data of high dimensionality or complex structures.

The significant problem definitions related to why Edu-ETL and EDP-Means are needed:

Problem 1. Traditional ETL processes may not adequately address the intricate nature of educational datasets, which often encompass diverse dimensions such as academic achievements, familial context, and socioeconomic background.

Problem 2. Identifying the λ value from prior knowledge in DP-Means clustering can be challenging and may not always guarantee optimal cluster quality. It is essential to recognize the limitations of this approach and supplement it with data-driven methods or validation techniques to ensure robust and reliable clustering outcomes, especially in educational data analysis contexts.

These problem definitions converge towards the overarching goal of advancing educational data analysis methodologies to better support comprehensive examinations of students' educational trajectories.

1.2 Motivation of Research

The motivation behind this research lies in the recognition of the critical role that lifelong learning plays in shaping educational outcomes and fostering success in academic and professional spheres.

The key motivations and facts driving this research are:

1. **Complexity of Lifelong Learning Achievements:** Lifelong learning encompasses a broad spectrum of educational experiences and achievements, ranging from formal schooling to informal learning activities throughout one's life. Understanding and analyzing these achievements requires a comprehensive approach that considers various dimensions such as academic performance, socio-economic background, and learning trajectories over time.
2. **Limitations of Traditional Assessment Methods:** Traditional assessment methods, such as grades and standardized test scores, often fail to capture the full range of student abilities and learning behaviors, providing only a partial view of their capabilities. These methods overlook critical thinking, creativity, and problem-solving skills, and do not account for individual learning styles and paces.
3. **Need for In-Depth Analysis:** In-depth data analysis enables researchers and educators to identify trends, patterns, and correlations within complex datasets, paving the way for evidence-based decision-making and targeted interventions. data needed to use in building the acoustic models.
4. **Importance of ETL Integration:** By refining and optimizing ETL processes, researchers can enhance the effectiveness and efficiency of data analysis workflows.
5. **Sensitivity to Initial Parameters in DP-Means:** Sensitive to initial parameters such as the initial number of clusters or cluster centers, which may impact the final clustering results. Retaining the adaptive nature of determining the optimal number of clusters further enhances performance and flexibility through optimization techniques can enhance the accuracy and improve cluster initialization and scalability.

6. **Harnessing the PySpark Environment:** Discuss the increasing adoption of PySpark in educational data analysis due to its distributed computing capabilities and scalability and highlight the significance of comparing clustering algorithms in PySpark to identify the most suitable approach for analyzing lifelong learning achievements.

In this system, the EDP-Means clustering algorithm with Edu-ETL integration is proposed to categorize students into distinct groups based on their educational outcomes and lead to more meaningful clusters. EDP-Means builds upon DP-Means by incorporating optimization techniques to enhance clustering performance and flexibility.

1.3 Objectives of Research

The primary goal of this research is to investigate the patterns in lifelong learning achievements by proposing the fusion of enhanced original DP-Means clustering and Edu-ETL methodologies. The next aim of this research is to demonstrate that the proposed EDP-Means method exhibits enhanced performance, thereby affirming its efficacy in cluster analysis compared to the traditional K-Means and original DP-Means clustering algorithms. One of the objectives includes conducting a comparative analysis of these clustering algorithms specifically within the PySpark environment.

The specific objectives of the research are outlined as follows:

1. To realize the effectiveness of Machine Learning (ML) techniques in today's educational development
2. To develop EDP-Means based clustering combined with Edu-ETL processes
3. To evaluate academic performance across various educational levels using the proposed clustering and data integration techniques
4. To enhance the understanding of student proficiency and learning behavior through comprehensive data analysis
5. To provide insights for educators to tailor interventions and support mechanisms to meet students' evolving needs effectively
6. To address the limitations of traditional clustering techniques by integrating EDP-Means clustering with Edu-ETL processes.

7. To advance educational data analysis methodologies to better support comprehensive examinations of students' educational trajectories
8. To emphasize the goal of evaluating the performance and efficacy of each clustering algorithm in handling educational datasets within PySpark focusing on factors such as scalability, efficiency, and accuracy

1.4 Contributions of Research

The contribution of this research lies in several key areas:

1. **Advancement of Educational Data Analysis:** This research contributes to the field of educational data analysis by proposing a novel approach that combines enhanced DP-Means clustering with Edu-ETL processes. By integrating educational data, particularly in the context of lifelong learning achievements.
2. **Improved Cluster Analysis Techniques:** The research demonstrates the enhanced performance of EDP-Means clustering over traditional K-Means and original DP-Means clustering algorithms. Extends the original DP-Means by incorporating steps to find an optimal λ value based on the silhouette score, and ensures convergence by iteratively updating the λ value and cluster centers until convergence is achieved. EDP-Means clustering enhances the accuracy and effectiveness of cluster analysis, providing valuable insights into student learning trajectories.
3. **Enhanced Understanding of Student Learning Behavior:** Through in-depth data analysis facilitated by the proposed methodologies, this research contributes to a deeper understanding of student proficiency and learning behavior. By exploring patterns in lifelong learning achievements, educators gain insights that can inform tailored interventions and support mechanisms to meet the evolving needs of students effectively.
4. **Methodological Innovation:** The integration of Edu-ETL processes with EDP-Means clustering represents a methodological innovation in educational data analysis. By addressing the limitations of traditional

clustering techniques and refining the clustering process with data-driven approaches, the research advances leading methodologies for analyzing educational datasets.

5. **Comparison in PySpark Environment:** Furthermore, this research extends its contribution by conducting a comparison of clustering algorithms, including EDP-Means, traditional K-Means, and original DP-Means, within the PySpark environment. This comparative analysis explores the performance of these algorithms in distributed computing settings, providing valuable insight into their scalability and efficiency in handling large-scale datasets.

1.5 Organization of Research

The dissertation comprises seven chapters, each serving a distinct purpose. Chapter 1 sets the stage by delineating the study areas, motivations, research issues, and objectives. It provides an overview of the methodology employed and outlines the contributions made by the research work.

Chapter 2 delves into the varied approaches and viewpoints concerning the modernization of educational technology and systems, contextualizing them within existing literature. It underscores the significance of lifelong learning achievements and the intricate nature of educational data, while also exploring different clustering techniques and their applications in educational research. Additionally, the chapter discusses the potential of PySpark in enhancing teaching and learning outcomes, fostering evidence-based decision-making, and ultimately improving student success.

In Chapter 3, the theoretical underpinnings of the proposed methods, including the EDP-Means clustering technique and Edu-ETL processes, are examined. The chapter offers a thorough discussion of essential theoretical principles and methodologies relevant to these approaches, setting the stage for deeper investigation and study in the respective field.

Chapter 4 introduces the "Architecture of the Proposed System," offering a structured framework to delve into lifelong learning achievements comprehensively. It underscores the importance of these insights in educational data analysis and aims to overcome the constraints of traditional assessment methods. The system acknowledges the impact of diverse datasets, such as academic records and

socioeconomic backgrounds, on students' educational paths.

Chapter 5 explores the implementation of Edu-ETL processes, which are crucial for preprocessing and reshaping data from student profiles and datasets. The Edu-ETL processes' essential role in preparing data for clustering analysis is emphasized, offering users practical insights into the system's capabilities. Using a graphical user interface (GUI), these processes are demonstrated visually, emphasizing the importance of integrating analytical techniques and data preprocessing methods.

Chapter 6 explains the experimental results and evaluations, structured into four main sections. The first section assesses the performance of the EDP-Means clustering algorithm, comparing it with K-Means and DP-Means. Following this, experiments are extended to PySpark to demonstrate scalability and efficiency. Subsequently, comprehensive validation of clustered results is conducted, evaluating cluster quality and processing times. Finally, learning outcomes are analyzed, and key success factors are identified. These experiments rigorously test the system under various conditions, utilizing diverse datasets and validation metrics.

Finally, Chapter 7 makes the culmination of the research, highlighting both the strengths and weaknesses of the study and outlining avenues for future research endeavors.

CHAPTER 2

LITERATURE REVIEW AND RELATED WORK

This chapter provides a comprehensive review of research about lifelong learning achievements, delving into existing literature and frameworks within Educational Data Analysis (EDA) and clustering methodologies. The modernization of education technology and systems has become imperative to meet the evolving needs of students and the demands of a rapidly changing world. One of the key reasons for the need to modernize education technology is the increasing complexity of educational data. Educational institutions generate vast amounts of data from various sources such as student assessments, attendance records, learning management systems, and online platforms. Analyzing this data provides valuable insights into student learning patterns, performance trends, and areas for improvement. However, traditional methods of data analysis often fall short in handling the volume, variety, and velocity of educational data. Therefore, there is a growing need to leverage advanced data analytics techniques, such as machine learning and data mining, to extract actionable insights from educational data.

Lifelong learning achievements holds immense importance in the context of modern education. Previous researchers have primarily focused on analyzing students' academic performance using various methodologies, such as traditional grading systems and standardized test scores. However, there is a growing recognition of the need to delve deeper into the concept of lifelong learning achievements. While existing research provides valuable insights into short-term academic progress, there is a gap in understanding the long-term trajectory of student learning and its implications for lifelong success.

Additionally, various clustering algorithms, including traditional K-Means and DP-Means clustering, have been investigated for their applicability in EDA. However, limitations such as the need for predetermined cluster numbers and sensitivity to initial parameters have prompted the exploration of enhanced clustering methodologies like the proposed EDP-Means algorithm. Moreover, integrating Edu-ETL processes has emerged as a crucial step in preparing and analyzing educational datasets, ensuring data quality, consistency, and integrity. By synthesizing findings from previous studies and identifying gaps in the literature, this review lays the

groundwork for the novel contributions and advancements introduced in this research endeavor.

2.1 Introduction to Educational Data Analysis (EDA)

EDA plays a crucial role in understanding student learning patterns and improving educational outcomes. By analyzing student performance, engagement, and behavior data, educators and policymakers can gain valuable insights into how students learn and where they may need additional support. This data-driven approach allows educational institutions to identify areas for improvement, tailor instruction to meet individual student needs, and ultimately enhance the overall quality of education. Figure 2.1 illustrates the key phases of EDA's life cycle [5].

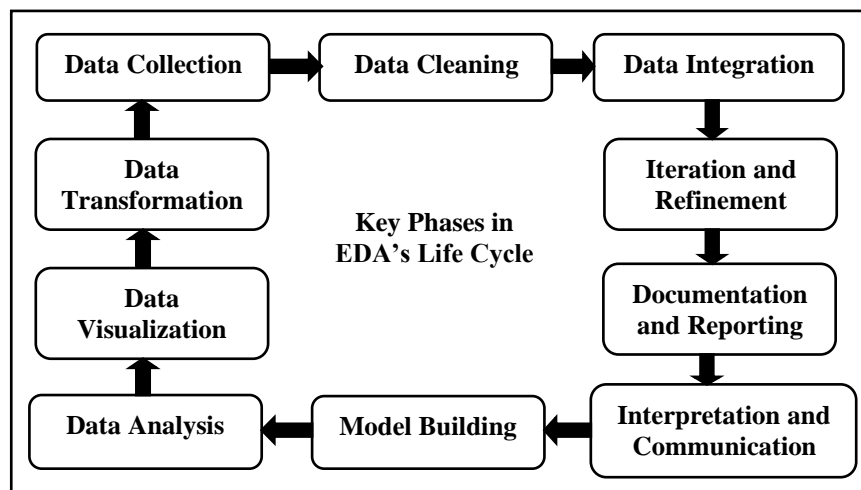


Figure 2.1 Key Phases of EDA's Life Cycle

Before delving into the specifics of related work, it is pertinent to highlight the predictive models or techniques utilized in previous studies. These models inform the proposed approach to analyzing lifelong learning achievements and provide valuable insights into the efficacy of different methodologies. By examining similarities and differences between the proposed research methodology and the predictive models employed in previous studies, it can be elucidated how this work builds upon or extends existing research in the field.

The researcher explored various ML strategies to forecast students' course grades within the context of private universities. By considering diverse factors influencing student outcomes, the study trained seven distinct classifiers to categorize final grades into four quality tiers: Excellent, Good, Poor, and Fail. Subsequently, a

weighted voting mechanism was employed to combine the outputs of these classifiers, resulting in enhanced predictive accuracy. Notably, the study achieved an accuracy rate of 81.73%, with the weighted voting classifier surpassing the performance of individual base classifiers. This research provides valuable insights into predictive modeling approaches within educational contexts, informing subsequent analyses of lifelong learning achievements [25].

The study presented in this article proposed an automated approach for observing and forecasting students' academic performance, aiming to achieve higher classification accuracy and lower root mean square error. Additionally, the research involved grouping students with similar educational backgrounds, such as those who had taken the same subjects in the same academic session. This approach generated a wealth of data requiring thorough analysis to extract actionable insights for planning and future educational development. The findings demonstrated the efficiency and relevance of machine learning technology in predicting students' performance, underscoring its potential for educational application [7].

The framework proposed in this research introduced a comprehensive Educational Data Mining (EDM) framework, presenting it as a rule-based recommender system designed not only to analyze and predict student achievement but also to elucidate the underlying reasons for such outcomes. This framework extensively examined various factors including students' demographic details, study-related attributes, and psychological factors, leveraging input from peers, educators, and parents to gather comprehensive insights. In comparison to existing frameworks, this proposed system proved effective in identifying students' weaknesses and providing pertinent recommendations, surpassing them in a practical case study involving 200 individuals [37].

The research was conducted by addressing the issue and by first preprocessing the data, particularly focusing on eliminating missing data, and subsequently applying the mentioned models to forecast student performance. The study found that among the models examined, random forest emerged as the most effective for predicting student grades in the dataset. Moreover, the research concluded that predicting students' future progress based solely on their past scores was both acceptable and reasonable. Furthermore, the study provided insights into potential reasons for the variations in outcomes observed across different algorithms [48].

A novel model employing ML algorithms was introduced to forecast the final exam grades of undergraduate students, utilizing their midterm exam grades as the primary dataset. Various ML algorithms were assessed and compared for their predictive capabilities regarding the final exam grades. This research was primarily focused on two aspects: firstly, the anticipation of academic performance based on prior achievement grades, and secondly, the evaluation of performance metrics across different ML algorithms [49].

The researcher introduced a predictive model for students' academic performance, utilizing a supervised learning techniques with an Artificial Neural Network (ANN). The model's efficacy was evaluated using a provided dataset for modelling purposes. The study identified two key attributes, namely students' IQ levels and class attendance regularity, which significantly influenced the prediction of academic performance within the model. This suggests that in practical settings, these attributes play crucial roles in determining students' success. Consequently, it is recommended that students focus on enhancing their IQ, maintaining consistent attendance in classes cultivating regular study habits, and fostering positive personality traits. Additionally, descriptive statistics were employed to identify potential attributes influencing students' academic performance [1].

The dataset utilized in this study encompassed comprehensive student records, including demographic information, academic history, personal attributes, and socio-economic factors [2]. The outcomes of this study yielded a dependable system for predicting student performance, offering valuable support to educational institutions and policymakers in making well-informed decisions and implementing early intervention strategies. The proposed framework was applied to forecast student academic outcomes utilizing both balanced and imbalanced datasets, employing the Synthetic Minority Oversampling Technique (SMOTE) [24]. The article explores the key factors to identify for predicting student performance include academic institution, sessional marks, semester progress, family occupation, as well as various methods and algorithms [14].

While there were variances in the performance of different methods, all demonstrated capability in capturing the intricate and implicit educational patterns and behaviors. Notably, ML techniques that comprehensively accounted for diverse factors exhibited superior predictive and generalization powers. Thus, achieving an accurate characterization of educational patterns and precise evaluation of academic

performance necessitates the incorporation of myriad influencing factors within the ML framework [41].

2.2 Exploring Lifelong Learning in Education

As mentioned, educational institutions play a crucial role in fostering a culture of lifelong learning by providing opportunities for skill development, professional growth, and personal enrichment. Lifelong learning achievement is not limited to academic performance but encompasses a broader spectrum of competencies, including critical thinking, problem-solving, communication, collaboration, and digital literacy. Figure 2.2 illustrates the process of leveraging data to improve lifelong learning outcomes.

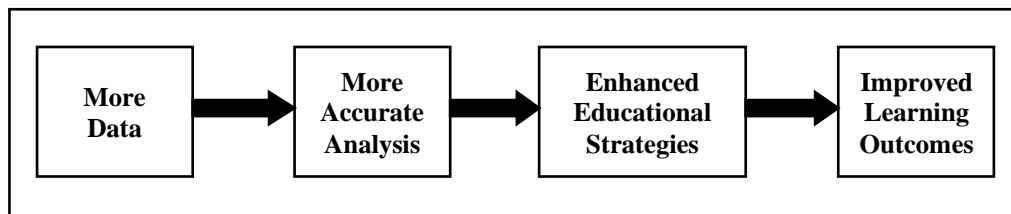


Figure 2.2 Data-Driven Approach to Lifelong Learning

The study documented by utilizing supervised ML algorithms to investigate factors negatively impacting academic performance among college students on probation, specifically underperforming students [19]. The findings highlighted study duration in the university and previous performance in secondary school as the primary factors influencing student academic achievement. A model was constructed to illustrate the significant prediction of lifelong learning competencies by computational thinking, thereby highlighting the interconnectedness between these two factors [9].

The related study investigated lifelong learning tendencies among university students using quantitative research methods [12]. It employed the general survey model and a probability-based random sampling method, with a sample size of 1312 students. The findings revealed significant differences in mean scores based on gender, school type, education type, economic status, place of residence, and type of high school attended. Notably, achievement orientation was found to impact lifelong learning tendencies.

Analyzing lifelong learning achievements allows educators to assess the

effectiveness of educational interventions, identify areas of strength and weakness, and tailor instruction to meet individual learning needs. Moreover, understanding lifelong learning trajectories enables educational institutions to design curricula, programs, and support services that facilitate continuous learning and skill development beyond formal education. The referenced studies highlight the importance of delving deeper into lifelong learning achievements and exploring advanced clustering methodologies like EDP-Means in EDA. These methodologies can help address existing limitations in EDA providing educators with more accurate insights into student learning trajectories and facilitating personalized interventions and support strategies.

2.3 Optimizing Educational Insights: Advanced Data Analysis and ETL Processes

Advanced data analysis techniques offer a more nuanced understanding by uncovering patterns and trends in educational data that may not be apparent through traditional means. These techniques also enable the exploration of lifelong learning achievements by analyzing longitudinal data and identifying patterns of skill development over time.

The research streamlined the problem through data preprocessing, which involved eliminating missing data before applying predictive models. The study found the best model for predicting student grades and then also highlights the importance of preprocessing in ensuring accurate and reliable predictions [3]. Similarly, the findings demonstrated that their proposed algorithms exhibited high performance in classification, underscoring the significance of preprocessing in optimizing algorithm performance and predictive accuracy [20]. These findings emphasize the need for preprocessing techniques, such as data cleaning and transformation, as part of the ETL process in educational data analysis.

Proper preprocessing ensures that the data is clean, standardized, and suitable for analysis, ultimately enhancing the accuracy and effectiveness of predictive models in academic performance prediction tasks. Through preprocessing and advanced data analysis, researchers can gain valuable insights into student performance and learning trajectories, enabling educators to tailor instruction and support services to meet individual learning needs effectively.

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 [21]. 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 [51]. Different tests served specific functions, revealing a research gap in advanced analysis techniques for distributed data sources [22]. By leveraging advanced ETL techniques tailored to educational contexts, institutions can unlock valuable insights to enhance teaching effectiveness, student learning outcomes, and overall educational performance. The proposed Edu-ETL processes address the complexities associated with multiple data sources, empowering educational institutions to effectively leverage their diverse data assets, derive actionable insights, and make informed decisions to enhance teaching, learning, and administrative processes.

2.4 Investigating Clustering Techniques in Educational Research

The investigation into clustering techniques within educational research encompasses diverse approaches and perspectives researchers adopt in this field. Traditional techniques like K-Means clustering have been foundational in identifying patterns and grouping similar data points. However, these methods face limitations, notably sensitivity to initial parameters and the need to specify the number of clusters beforehand, which can hinder their efficacy in educational contexts.

The study underscores the importance of developing more robust and adaptable clustering techniques tailored to educational datasets. Researchers can gain deeper insights into student learning behaviors and academic performance by overcoming these challenges. The exploration encompasses a range of clustering algorithms, including hierarchical clustering, K-Means clustering, spectral clustering, and other centroid-based techniques such as DP-Means, as well as density-based methods, showcasing their versatility in different educational research scenarios. Furthermore, the study delves into the diverse applications of clustering techniques in educational research. It highlights their role in analyzing student performance, identifying at-risk students, and uncovering hidden patterns in educational data.

To explore clustering techniques in educational research, the study aimed to enhance students' academic performance prediction while avoiding unreasonable evaluation results [20]. Employing clustering, discrimination, and convolutional neural network theories, the research proposed novel methodologies. The resulting model demonstrated promising potential for predicting prospective performance. To

validate these predictions, the model's effectiveness was evaluated using two metrics across two cross-validation methods. This comprehensive approach illustrates the potential of clustering techniques in advancing educational research and predictive analytics.

This study applied the DP-means clustering algorithm to pretest data from 264 students interacting with the tutoring system [36]. Results revealed three distinct learning clusters, with DP-means outperforming other methods. The study also noted DP-means' decreased quality with categorical data, proposing k-modes clustering as a solution. Furthermore, it highlighted the benefits of clustering in educational research for balancing adaptivity and authoring costs by grouping students with similar mental models.

In practical terms, categorizing student performance poses a significant scientific challenge. Recent studies have applied cluster analysis to evaluate student results, utilizing statistical techniques to segment their scores in relation to performance. However, this method lacks efficiency. This research combines two techniques, namely, K-means and Elbow clustering algorithms, to assess student performance [6]. They implement a methodology to construct a diverse and intriguing model based on student test scores. Its primary aim is to introduce a new clustering model that integrates the K-means algorithm with four functionalities: the Elbow method, scaling, and normalization/standardization. Following the clustering of students into groups, an improvement plan is tailored for each group, highlighting areas of weakness and recommending specific actions for enhancement, such as reviewing chapters, redoing homework, and focusing on certain topics.

Among the clustering techniques, centroid-based clustering techniques, such as K-Means clustering, are commonly used in educational research due to their simplicity and efficiency. One reason centroid-based clustering techniques are often preferred is their ease of implementation and interpretation. Additionally, this technique is well-suited for scenarios where the number of clusters is known or can be easily determined. However, it is essential to acknowledge that centroid-based clustering techniques have limitations. For instance, they are sensitive to the initial placement of cluster centroids and may converge to suboptimal solutions, especially in high-dimensional or noisy datasets. Moreover, centroid-based methods assume that clusters are spherical and of similar size, which may not always reflect the true underlying structure of the data.

DP-Means clustering can also be considered a centroid-based clustering technique, albeit with some unique characteristics compared to traditional methods like K-Means. Like other centroid-based approaches, DP-Means aims to partition the dataset into clusters based on the similarity of data points to cluster centroids. However, DP-Means introduces a probabilistic framework that allows for more flexibility in determining the number of clusters, making it particularly useful in scenarios where the optimal number of clusters is unknown or attribute. DP-Means offers a promising approach to centroid-based clustering in educational research, providing flexibility in determining the number of clusters and accommodating non-spherical cluster shapes. While DP-Means offers advantages such as automatic determination of cluster number and flexibility in cluster shapes, it also has drawbacks like computational intensity and sensitivity to parameter settings.

The system proposed EDP-Means clustering techniques to address some of the limitations and challenges associated with traditional DP-Means clustering, particularly in the context of educational research. EDP-Means techniques aim to overcome these challenges by introducing improvements or modifications to the original algorithm. Overall, the goal of proposing EDP-Means clustering techniques is to provide educational research and other domains with more effective and reliable tools for analyzing complex datasets and extracting meaningful insights. By addressing the limitations of traditional DP-Means clustering, these enhancements advance clustering methodology and facilitate accurate and insightful data analysis in educational research.

2.5 Educational Big Data Mining using PySpark

Educational Big Data Mining involves utilizing data mining methods on extensive educational datasets to derive valuable insights and knowledge, shaping educational strategies and policies. The rise of digital learning platforms, online assessments, and educational technologies has led to the generation of substantial data, capturing diverse facets of teaching and learning processes. Educational big data comprises a wide range of data types, such as student demographics, academic performance records, learning behaviors, engagement metrics, instructor feedback, and administrative data. These datasets are often intricate, varied, and multi-dimensional, presenting obstacles in analysis and comprehension.

The utilization of educational big data mining holds promises in transforming

education through evidence-based decision-making, personalized learning experiences, early intervention strategies, adaptive learning systems, and data-driven policy development. By leveraging this approach, educators and administrators gain deeper insights into student needs, preferences, and learning paths, ultimately enhancing educational outcomes and fostering student success. Nevertheless, it's imperative to meticulously address ethical considerations, privacy concerns, and data security issues during the collection, storage, and analysis phases of educational big data. This is crucial to uphold responsible data usage and safeguard student privacy rights. Furthermore, fostering interdisciplinary collaboration among education researchers, data scientists, computer scientists, and domain experts is paramount for advancing the field of educational big data mining and unlocking its full potential to enhance education globally. Figure 2.3 explores that big data has a wide range of applications in education, offering numerous opportunities to enhance teaching, learning, administration, and overall educational outcomes [46].

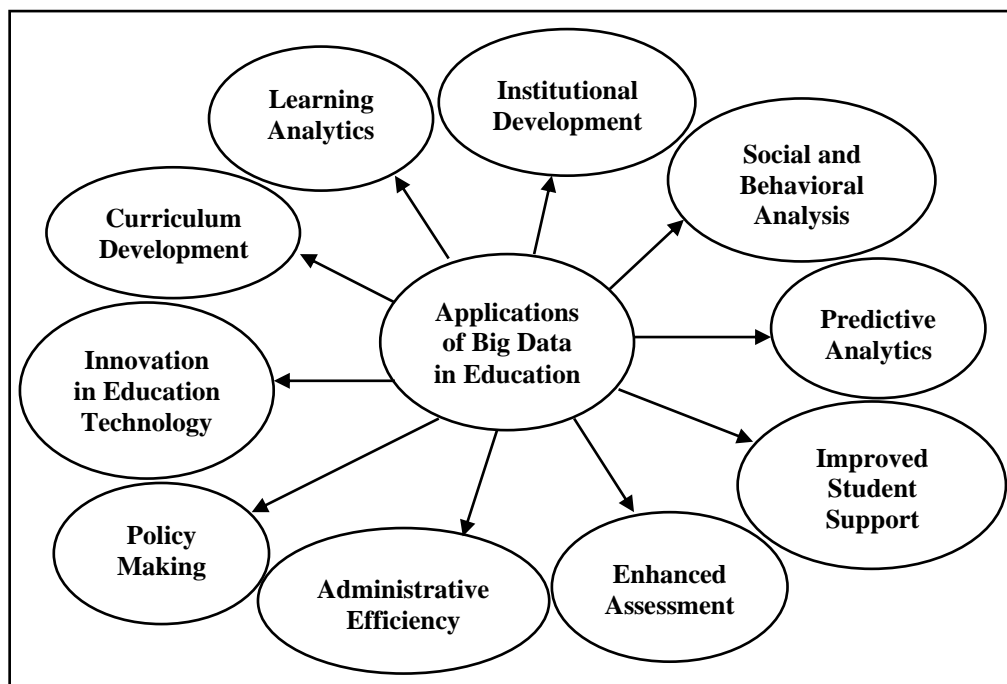


Figure 2.3 Key Applications of Big Data in Education

Figure 2.4 exemplifies the interplay among big data, learning analytics, and academic analytics within the higher education system [40]. Mapping learning and academic analytics in the context of big data enables educational stakeholders to harness the power of data-driven insights to support student success, enhance teaching practices, and drive continuous improvement in educational outcomes.

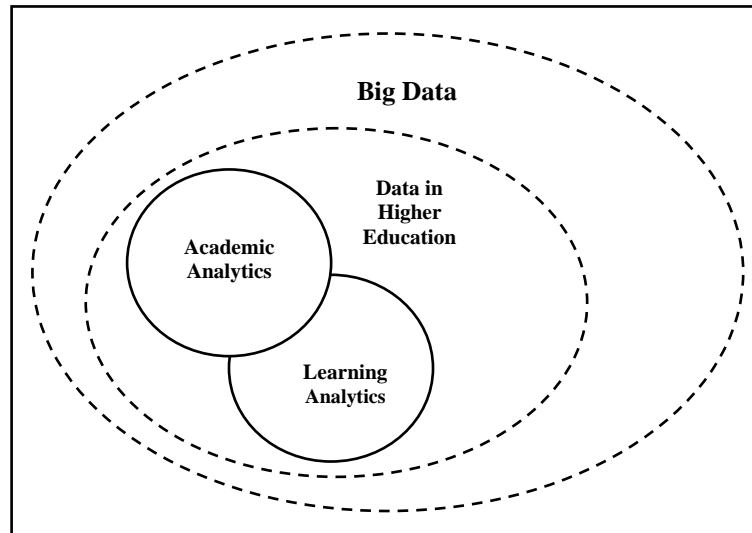


Figure 2.4 Model of Big Data Mapping in Higher Education System

Educational big data mining involves the extraction of valuable insights and patterns from large-scale educational datasets to improve teaching and learning outcomes. PySpark, a Python API for Apache Spark, plays a significant role in this process by providing powerful tools for processing and analyzing big data efficiently. PySpark provides a powerful platform for educational big data mining, offering scalability, advanced analytics capabilities, real-time processing, and seamless integration with the big data ecosystem. By leveraging PySpark, educational institutions can unlock the full potential of their data to improve teaching and learning outcomes, drive evidence-based decision-making, and enhance student success.

The researchers delved into student performance analysis utilizing Spark and harnessed machine learning algorithms to glean insights from student databases, to comprehend student behavior and pinpoint reasons for academic setbacks [10]. The findings yielded valuable insights for educational institutions seeking to improve teaching efficacy. Similarly, in another study, the use of K-Means clustering was explored for predicting student academic performance [23]. Although Spark was not explicitly mentioned in the study, it shed light on the potential application of clustering techniques for this objective.

This research endeavor entailed analyzing student academic data to discern the attributes or features contributing to student progress through clustering and classification [11]. Spark was employed for clustering due to its efficiency in handling large datasets, utilizing K-Means clustering for this purpose. Classification

tasks were carried out using logistic regression, decision tree, and random forest algorithms within the Spark environment. Similarly, in another study, various Machine Learning techniques, including Pearson correlations, as well as models like Multilayer Perceptron, Decision Tree, and Random Forest, were utilized to identify factors influencing student performance [15]. The aim was to uncover patterns or insights assisting educators in prioritizing factors for better student outcomes in educational settings.

The efforts of these research underscore the transformative potential of PySpark and advanced analytics in educational big data mining, driving continuous improvement in teaching and learning practices. The proposed system expands its contribution by comparing clustering algorithms, such as EDP-Means, traditional K-Means, and original DP-Means, within the PySpark environment. This comparative assessment investigates how these algorithms perform in distributed computing settings, offering valuable insights into their scalability and effectiveness in managing extensive educational datasets.

2.6 Chapter Summary

This chapter offered a thorough exploration of lifelong learning achievements, examining existing literature and frameworks within EDA, clustering methodologies, and Educational Big Data Mining using PySpark.

Firstly, the chapter delves into the modernization of education technology and systems to meet evolving needs. It highlights the importance of lifelong learning achievement and the challenges posed by the complexity of educational data. The discussion emphasizes the need for advanced data analytics techniques such as machine learning and data mining to extract actionable insights. This chapter also reviews and discusses various approaches and perspectives from existing research papers on EDA, Lifelong Learning, and ETL processes, comparing them with previous studies.

"Investigating Clustering Techniques in Educational Research" examines different clustering algorithms and their applicability in analysis of educational data. It discusses traditional methods like K-Means and DP-Means clustering, along with their limitations and the exploration of enhanced techniques like EDP-Means. From existing research papers, these facts are known: the effectiveness of methodologies varies based on the quality and quantity of available data. Enhanced clustering

approaches empower educators to better comprehend student performance and adapt interventions effectively. Moreover, potential limitations include issues related to the scalability or interpretability of clustering results. The section underscores the importance of clustering methodologies in understanding student learning patterns and performance trends.

"Educational Big Data Mining using PySpark" explores the role of PySpark in educational research for handling large-scale datasets efficiently. It discusses the significance of leveraging PySpark's scalability and advanced analytics capabilities in Educational Big Data Mining. The section emphasizes the potential of PySpark in improving teaching and learning outcomes, driving evidence-based decision-making, and enhancing student success.

Overall, this chapter provided a comprehensive overview of key concepts and methodologies in educational research, laying the groundwork for further exploration and advancements in the field.

CHAPTER 3

THEORETICAL BACKGROUND

Evaluating lifelong learning achievements has become crucial in modern education, requiring robust analytical frameworks. This research integrates EDP-Means and Edu-ETL methodologies to enhance this evaluation. EDP-Means, a clustering algorithm, systematically processes learning outcomes, while Edu-ETL manages and integrates diverse datasets, transforming them into actionable insights. This synergy improves the accuracy and depth of achievement evaluations, supporting data-driven decision-making in education. This section explores these approaches' theoretical foundations and relevance in evaluating lifelong learning achievements.

3.1 DP-Means: A Nonparametric Extension of K-Means

DP-Means is a clustering algorithm that extends the traditional K-Means method by incorporating principles from nonparametric Bayesian statistics, specifically the Dirichlet Process. This extension allows DP-Means to automatically determine the number of clusters, addressing one of the significant limitations of K-Means [33].

Key features of DP-Means:

1. **Dynamic Cluster Formation:** Unlike K-Means, where the number of clusters K is fixed, DP-Means allows the number of clusters to grow as needed.
2. **Threshold Parameter (λ):** DP-Means introduces a threshold parameter λ . If a data point is farther from all existing centroids than λ , a new cluster is created with that point as its centroid. λ controls the balance between forming new clusters and assigning points to existing clusters.
3. **Nonparametric Approach:** By using a nonparametric approach, DP-Means can handle datasets with unknown or varying cluster numbers more effectively.

3.1.1 Role of the Threshold Parameter (λ)

DP-Means is a clustering algorithm that extends the traditional K-Means method by incorporating principles from nonparametric Bayesian statistics, specifically the Dirichlet Process. This extension allows DP-Means to automatically determine the number of clusters, addressing one of the significant limitations of K-Means.

In the context of DP-Means clustering, λ plays a crucial role in determining the formation and assignment of clusters. Unlike traditional K-Means, where the number of clusters is fixed in advance, DP-Means uses λ to dynamically decide whether a data point should be assigned to an existing cluster or form a new cluster. This parameter is fundamental in balancing the trade-off between cluster granularity and computational efficiency. Table 3.1 provides a detailed summary of both the role and specific characteristics of the λ within the context of the DP-Means clustering algorithm [18][33].

Table 3.1 Summarizing the Role and Characteristics of λ in DP-Means Clustering Algorithm

Aspect	Description
Definition	<ul style="list-style-type: none"> • Critical distance measure determining cluster assignments and formation.
Function	<ul style="list-style-type: none"> • Decide whether a data point joins an existing cluster or forms a new one based on its distance to the nearest centroid.
Impact on Cluster Formation	<ul style="list-style-type: none"> • High λ: Fewer, larger clusters; data points are more likely to join existing clusters. • Low λ: More, smaller clusters; data points are more likely to form new clusters
Selection Methods	<ul style="list-style-type: none"> • Empirical Testing: Experiment with different values on validation datasets. • Theoretical Guidance: Use domain knowledge and expected cluster distances to set λ.
Optimal λ	<ul style="list-style-type: none"> • Balances cluster quality and computational efficiency; avoids overfitting and underfitting.

Aspect	Description
Influence on Performance	<ul style="list-style-type: none"> Cluster Quality: Affects interpretability and accuracy of clusters. Computational Efficiency: Lower λ increases computational complexity due to more clusters.
Cluster Assignment Rule	<ul style="list-style-type: none"> Assigns data points to the nearest centroid if distance $< \lambda$; otherwise, forms a new cluster.
Convergence Criteria	<ul style="list-style-type: none"> Iterates until centroids stabilize or a maximum number of iterations is reached.
Key Considerations	<ul style="list-style-type: none"> Data Distribution: Consider natural distances in data. Computational Resources: Balance λ to manage resource usage.
Common Values	<ul style="list-style-type: none"> Often determined experimentally; no universal value fits all datasets.
Effects on Clustering Results	<ul style="list-style-type: none"> High λ: May merge distinct clusters. Low λ: May split cohesive clusters into smaller ones.

3.1.2 Key Concepts of DP-Means Clustering Technique

DP-Means is a sophisticated extension of the traditional K-Means algorithm, designed to address its limitations by leveraging nonparametric methods. Two pivotal concepts that underpin DP-Means are Gaussian Mixture Models (GMMs) and the Chinese Restaurant Process (CRP).

GMMs represent a probability distribution of data points as a mixture of several Gaussian distributions. In the context of clustering, GMM is used to model the distribution of data points within each cluster. For a multivariate Gaussian distribution with mean vector μ and covariance matrix Σ , the probability density function (PDF) is given by [31]:

$$f(x|\mu, \Sigma) = \frac{1}{2\pi^{d/2} \Sigma^{1/2}} \exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)\right) \quad (3.1)$$

The notations required to describe the concept of PDF of a multivariate Gaussian distribution are shown in Table 3.2.

Table 3.2 Notations for the PDF of a Multivariate Gaussian Distribution

Symbol	Description
d	The dimensionality of the random vector x
x	A d -dimensional vector
μ	The d -dimensional mean vector
Σ	The covariance matrix of the Gaussian distribution
Σ^{-1}	The inverse of the covariance matrix Σ
$(x - \mu)^T$	The transpose of the vector $(x - \mu)$
π	The determinant of the covariance matrix Σ

In a GMM, the overall probability density function is a weighted sum of the PDFs of individual multivariate Gaussian distributions (mixture components). Each multivariate Gaussian distribution represents one component of the mixture model. The PDF of a GMM is expressed as [33]:

$$f(x|\pi, \mu, \Sigma) = \sum_{k=1}^K \pi_k \cdot f(x|\mu_k, \Sigma_k) \quad (3.2)$$

The notations used in the above equation for the PDF of a GMM are shown in Table 3.3.

Table 3.3 Notations for the PDF of a GMM

Symbol	Description
K	Number of mixture components in the GMM
π	Mixing coefficients representing the probabilities of each component
μ_k	Mean vector of the k -th Gaussian component
Σ_k	The covariance matrix of the k -th Gaussian component
$f(x \mu_k, \Sigma_k)$	PDF of the k -th Gaussian component

Each term in the summation represents the contribution of one Gaussian component to the overall PDF. The mixing coefficients π_k determine the weight of

each component in the mixture, while the mean vectors μ_k and covariance matrices Σ_k define the shape and orientation of each component. GMMs serve as a fundamental building block in the DP-Means clustering algorithm. GMMs allow DP-Means to identify clusters with different shapes and densities within the data [36].

CRP is a metaphorical model, that illustrates the probabilistic assignment of data points to clusters based on their similarity. CRP is often used in the context of Bayesian nonparametric models, including DP-Means clustering.

The CRP analogy is as follows:

- Imagine a Chinese restaurant with an infinite number of tables.
- The first customer sits at the first table.
- Subsequent customers choose a table based on the following rules:
 - If a table already has customers, the new customer sits at that table with probability proportional to the number of customers already seated.
 - If all tables are occupied, a new table is opened, and the new customer sits there.

The probability of a new customer sitting at an existing table j with m_j customers are:

$$P(\text{customer sits at table } j) = m_j / \alpha + N - 1 \quad (3.3)$$

If a new table is opened, the probability is:

$$P(\text{customer opens new table}) = \alpha / \alpha + N - 1 \quad (3.4)$$

This process continues for each customer, determining the assignment of data points to clusters based on the similarity of their features [34]. Table 3.4 provides the notations that are used in the probability equations for a new customer as well as for a new table.

Table 3.4 Notations Used in the CRP Analogy

Symbol	Description
α	The concentration parameter
N	The total number of customers
m_j	The number of customers already seated at table j

Table 3.5 outlines various applications of DP-Means, highlighting how the integration of GMM and CRP enhances the algorithm's capabilities in diverse domains.

Table 3.5 Applications of DP-Means Clustering with GMM and CRP

Application	Description
Educational Data Clustering	<ul style="list-style-type: none"> Applies DP-Means to cluster educational data, such as student performance metrics or learning behavior, identifying meaningful patterns and subgroups for personalized interventions and strategies.
Automatic Cluster Determination	<ul style="list-style-type: none"> Utilizes CRP to dynamically determine the number of clusters based on data, eliminating the need for a predefined number of clusters.
Adaptive Clustering	<ul style="list-style-type: none"> Combines GMM's flexibility in modeling clusters of varying shapes and densities with CRP's probabilistic cluster assignment for more accurate and adaptive clustering.
Data Mining	<ul style="list-style-type: none"> Applies DP-Means in data mining to discover patterns and groupings in large datasets, benefiting from GMM's ability to represent complex data distributions.
Image Segmentation	<ul style="list-style-type: none"> Leverages the adaptive nature of DP-Means for segmenting images into distinct regions, using GMM to model pixel intensity distributions and CRP for adaptive region determination.
Customer Segmentation in Marketing	<ul style="list-style-type: none"> Employs DP-Means to segment customers into distinct groups based on purchasing behavior, using GMM to capture diverse customer profiles and CRP for dynamic group sizing.

Application	Description
Anomaly Detection	<ul style="list-style-type: none"> • Uses DP-Means for detecting anomalies by identifying clusters that represent normal behavior patterns and isolating outliers, benefiting from GMM's detailed data modeling and CRP's flexible cluster assignment.
Bioinformatics	<ul style="list-style-type: none"> • Applies DP-Means in analyzing biological data, such as gene expression profiles, where GMM models the complex data distributions and CRP adapts to varying cluster sizes.
Natural Language Processing (NLP)	<ul style="list-style-type: none"> • Utilizes DP-Means for clustering words or documents in NLP tasks, leveraging GMM to model semantic similarities and CRP to dynamically determine the number of clusters based on context.

In summary, the elucidation of GMM and the CRP serves as a crucial foundation for comprehending the intricacies of DP-Means clustering. GMM allows for the flexible representation of data clusters, accommodating varying shapes and densities, while CRP provides a probabilistic framework for dynamically determining the number of clusters based on the data itself. Together, these concepts empower DP-Means to adaptively cluster data points, offering a robust and versatile solution for exploring complex datasets in diverse domains. By integrating these fundamental concepts into the DP-Means framework, researchers gain valuable insights into the mechanisms driving adaptive clustering and the potential applications in real-world scenarios.

3.1.3 Key Steps of DP-Means Clustering Algorithm

The DP-Means clustering algorithm involves a sequence of steps that allows for the dynamic determination of the number of clusters in a dataset. Table 3.6 provides a detailed explanation of each key step in the DP-Means clustering process, illustrating how the algorithm initializes, assigns, updates, and converges clusters based on the data.

Table 3.6 Key Steps in the DP-Means Clustering Algorithm

Step	Explanation
1. Initialization	<ul style="list-style-type: none"> • Single Centroid Initialization: Start with the mean of the entire dataset as the first centroid. • Random Initialization: Randomly select a subset of data points as initial centroids. • k-means++ Initialization: Use a method that spreads out initial centroids to improve initial clustering.
2. Assignment	<ul style="list-style-type: none"> • For each data point x_i, calculate the distance to each existing centroid μ_j. • If the distance $d(x_i, \mu_j)$ to the nearest centroid μ_j is less than λ, assign x_i to cluster j. • If $d(x_i, \mu_j) \geq \lambda$ for all j, create a new cluster with x_i as the centroid.
3. Update	<ul style="list-style-type: none"> • Recalculate the centroids of the clusters. • For each cluster j, compute the new centroid μ_j as the mean of all points x_i assigned to that cluster. • Update the position of μ_j to this new mean value. Mathematically, for cluster j with n_j points, the new centroid μ_j is: $\mu_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i$
4. Convergence	<ul style="list-style-type: none"> • Ensure the algorithm stops when clusters are stable: • After each iteration, compare the new centroid positions with the previous positions. • If changes in all centroids are below a predefined threshold, or a maximum number of iterations is reached, the algorithm converges: $\max_j \ \mu_j(t+1) - \mu_j(t)\ < \epsilon$

The steps from Table 3.6 are critical for understanding the operational mechanics and practical implementation of DP-Means in clustering applications [18][36][34]. The following Table 3.7 explores notations used in key steps of the DP-Means clustering algorithm.

Table 3.7 Notations Used in Key Steps of DP-Means Clustering Algorithm

Symbol	Description
x_i	A data point in the dataset.
μ_j	The centroid of cluster j .
$d(x_i, \mu_j)$	The distance between data point x_i and centroid μ_j .
λ	A predefined threshold distance used to determine whether a new cluster should be created.
n_j	The number of data points assigned to cluster j .
t	The iteration number in the algorithm.
ϵ	A small positive value is used as a threshold for convergence.
$\ \cdot \ $	The norm used to measure the distance between points or centroids.

3.1.4 K-Means and DP-Means Clustering

Both K-Means and DP-Means are centroid-based clustering algorithms, meaning they use central points or centroids, to represent each cluster and group similar data points around these centroids. K-Means is one of the most widely used clustering algorithms, known for its simplicity and efficiency in partitioning datasets into a predefined number of clusters. However, it requires the user to specify the number of clusters in advance, which can be a significant limitation. In contrast, the DP-Means algorithm, an extension of K-Means, addresses this limitation by dynamically determining the number of clusters based on the data. This flexibility makes DP-Means particularly useful for complex datasets where the number of clusters is unknown beforehand. The following table provides a detailed comparison of K-Means and DP-Means, highlighting their key differences and similarities to illustrate how each algorithm operates and their respective advantages in various applications [38][16][50][32].

Table 3.8 Comparison of K-Means and DP-Means Clustering Algorithm

Feature	K-Means	DP-Means
Cluster Number	<ul style="list-style-type: none"> • Fixed number of clusters K 	<ul style="list-style-type: none"> • Dynamically determined the number of clusters
Initialization	<ul style="list-style-type: none"> • Requires initial selection of K centroids 	<ul style="list-style-type: none"> • Starts with a single centroid; new centroids added as needed
Cluster Assignment	<ul style="list-style-type: none"> • Assigns data points to the nearest centroid based on Euclidean distance 	<ul style="list-style-type: none"> • Assigns data points to the nearest centroid if within threshold λ ; otherwise, creates a new cluster
Update Step	<ul style="list-style-type: none"> • Centroids updated as the mean of assigned data points 	<ul style="list-style-type: none"> • Similar to K-Means; centroids updated as the mean of assigned data points
Convergence Criteria	<ul style="list-style-type: none"> • Stops when centroids no longer change significantly or after a set number of iterations 	<ul style="list-style-type: none"> • Similar to K-Means; stops when centroids stabilize or a maximum number of iterations is reached
Distance Metric	<ul style="list-style-type: none"> • Typically uses Euclidean distance 	<ul style="list-style-type: none"> • Typically uses Euclidean distance
Flexibility	<ul style="list-style-type: none"> • Less flexible; requires a predefined number of clusters 	<ul style="list-style-type: none"> • More flexible; automatically determines the number of clusters
Computational Complexity	<ul style="list-style-type: none"> • $O(n.K.d)$ per iteration (n: number of points, K: clusters, d: dimensions) 	<ul style="list-style-type: none"> • $O(n.C.d)$ per iteration (C: current number of clusters, dynamically determined)
Handling of Cluster Shapes	<ul style="list-style-type: none"> • Assumes clusters are spherical and equally sized 	<ul style="list-style-type: none"> • More adaptable to varying cluster shapes and sizes due to dynamic cluster formation
Threshold Parameter (λ)	<ul style="list-style-type: none"> • Not applicable 	<ul style="list-style-type: none"> • Essential for controlling cluster formation; determines the threshold distance for creating new clusters

Feature	K-Means	DP-Means
Use Case Suitability	<ul style="list-style-type: none"> • Suitable for well-separated, spherical clusters with a known number of clusters 	<ul style="list-style-type: none"> • Suitable for complex data with varying cluster sizes and an unknown number of clusters
Algorithm Type	<ul style="list-style-type: none"> • Parametric (fixed number of clusters) 	<ul style="list-style-type: none"> • Nonparametric (flexible number of clusters)
Underlying Model	<ul style="list-style-type: none"> • Based on minimizing within-cluster variance 	<ul style="list-style-type: none"> • Based on the Dirichlet Process, combining aspects of K-Means with nonparametric clustering

3.1.5 Critical Evaluation of DP-Means Clustering: Limitations and Challenges

Despite its flexibility and ability to dynamically determine the number of clusters, the DP-Means clustering algorithm has several disadvantages. One of the primary challenges is the selection of the threshold parameter λ , which significantly influences the clustering results. Choosing an inappropriate value for λ can lead to either over-clustering or under-clustering, making it difficult to achieve optimal performance without extensive parameter tuning. Additionally, while DP-Means can handle varying cluster shapes and sizes better than K-Means, it still assumes clusters are relatively spherical due to its reliance on Euclidean distance, which can limit its effectiveness for more complex data distributions.

Furthermore, DP-Means shares some of the same computational drawbacks as K-Means. It can be computationally intensive for large datasets, especially since it requires calculating distances between data points and centroids in each iteration. This can become particularly burdensome when the number of clusters dynamically increases during the clustering process. Also, like K-Means, DP-Means is sensitive to the initial placement of centroids, which can affect the final clustering outcome, potentially leading to suboptimal solutions if the initial centroids are poorly chosen.

Several studies have explored the limitations of the DP-Means algorithm and proposed modifications to address these issues. For instance, researchers have investigated various methods for optimizing the selection of the λ parameter. One approach involves adaptive techniques that adjust λ during the clustering process based on the data characteristics, aiming to improve clustering accuracy without

extensive manual tuning. λ -means, an innovative clustering algorithm designed to automatically derive the optimal λ value. Additionally, λ -means showcased remarkable speedup potential in parallel computing environments, further solidifying its position as a promising solution for overcoming the challenges of DP-Means clustering.

Another area of research focuses on enhancing the robustness of algorithm to different data distributions by incorporating alternative distance metrics or integrating DP-Means with other clustering methods. An integrated clustering approach that combines model-based and centroid-based methodologies is one of the solutions that can mitigate noise impact and eliminate the need to pre-specify the number of clusters. Statistical guarantees and rigorous evaluations demonstrate its superiority over existing algorithms.

Moreover, there have been efforts to reduce the computational complexity of DP-Means. Some proposed algorithms introduce techniques such as efficient data point assignment strategies and parallel processing to accelerate the clustering process. These advancements aim to make DP-Means more scalable and practical for large-scale datasets. Dirichlet Process Means for Clustering Extremely Large Datasets (DACE) demonstrated remarkable scalability, capable of clustering billions of sequences within a few hours. By leveraging parallel processing techniques, the Dirichlet Process Means for Clustering Extremely Large Datasets (DACE) offered a practical solution for handling extremely large-scale datasets on time, making it a promising tool for clustering high-throughput sequencing data [29]. Parallel Delayed Cluster Dirichlet Process Means (PDC-DP-Means) is addressed as a parallel algorithm that offers significant speedups and performance gains. PDC-DP-Means is recommended for datasets with moderate cluster counts, while extensions provide remedies for datasets with larger clusters, despite a slight drop in quality [18].

In summary, while DP-Means offers significant advantages over traditional K-Means by dynamically determining the number of clusters, it faces challenges related to parameter selection, computational complexity, and sensitivity to initial conditions. Ongoing research continues to address these limitations, seeking to enhance the algorithm's robustness and efficiency in various clustering applications.

3.2 Enhanced DP-Means: EDP-Means Clustering Technique

The proposed EDP-Means clustering algorithm improves upon the original DP-Means by incorporating two key optimization techniques to enhance clustering performance.

Firstly, the Elbow method is employed to determine the optimal number of clusters. This technique utilizes the Sum of Squared Errors (SSE) metric and visual analysis to identify the 'elbow point,' where additional clusters no longer significantly reduce SSE. These heuristics guide decision-making regarding the appropriate number of clusters, ensuring a balance between model complexity and clustering accuracy.

Secondly, the EDP-Means algorithm focuses on identifying the optimal λ value, which is crucial for defining cluster boundaries. The algorithm evaluates clustering quality using the Silhouette score, selecting the λ value that maximizes this metric. This ensures that the resulting clusters effectively capture the underlying data patterns, leading to more meaningful and well-defined clusters.

By iteratively assessing different cluster configurations and λ values, these enhancements enable a more fine-tuned and data-driven approach to clustering. Consequently, the resulting clusters are better at capturing the underlying patterns and structures within the dataset. This iterative process leads to improved clustering outcomes compared to the original DP-Means algorithm alone.

By employing the Elbow method and optimizing λ through the Silhouette score, the EDP-Means algorithm achieves better clustering outcomes with minimal manual intervention. Additionally, the inclusion of comprehensive performance metrics allows for a thorough evaluation of clustering quality. The improvements introduced in EDP-Means address common challenges faced by the original DP-Means, such as the need for manual parameter tuning and limitations in handling large datasets. As a result, EDP-Means stands out as a robust and scalable clustering solution. The step-by-step outline of the proposed EDP-Means clustering process is detailed in Table 3.9. Additionally, Table 3.10 offers a comprehensive comparison between the original DP-Means and EDP-Means clustering algorithms, highlighting various aspects.

Table 3.9 Steps of EDP-Means Clustering Algorithm

Step	Description	Details
Input	<ul style="list-style-type: none"> • Provide initial data and parameters 	<ul style="list-style-type: none"> • Data points (number of attributes) • Initial threshold value λ
Output	<ul style="list-style-type: none"> • Final results 	<ul style="list-style-type: none"> • Final cluster centers • Assigned labels
Begin		
Step 1: Initialization	<ul style="list-style-type: none"> • Set initial values 	<ul style="list-style-type: none"> • Set initial λ value and cluster centers
Step 2: Find the optimal number of clusters	<ul style="list-style-type: none"> • Determine the number of clusters 	<ul style="list-style-type: none"> • Read the dataset and extract relevant attributes • Iterate over a range of cluster numbers: <ul style="list-style-type: none"> • Fit DP-Means clustering for each number of clusters • Calculate the SSE (inertia) for each clustering and store it in the 'sse' list • Plot the SSE against the number of clusters to identify the 'elbow point'
Step 3: Find the optimal threshold value	<ul style="list-style-type: none"> • Determine the threshold value 	<ul style="list-style-type: none"> • Execute the optimal threshold value based on the silhouette score • Update the λ value with the optimal threshold value
Step 4: Assignment	<ul style="list-style-type: none"> • Assign data points 	<ul style="list-style-type: none"> • Assign data points to clusters based on the updated λ value

Step	Description	Details
Step 5: Update cluster centers	<ul style="list-style-type: none"> Recalculate centers 	<ul style="list-style-type: none"> Recalculate cluster centers based on the assigned data points
Step 6: Repeat until convergence	<ul style="list-style-type: none"> Ensure convergence 	<ul style="list-style-type: none"> Repeat steps 4 and 5 until convergence is achieved
End		

Table 3.10 Comparison of Original DP-Means and EDP-Means Clustering Algorithm

Aspect	Original DP-Means	EDP-Means
Initialization	<ul style="list-style-type: none"> Requires initial threshold λ Initial cluster centers set manually or randomly 	<ul style="list-style-type: none"> Similar initialization with initial threshold λ Enhanced method may include better initialization techniques
Number of Clusters	<ul style="list-style-type: none"> Automatically determined by the algorithm based on the threshold λ Number of clusters can grow as more clusters are needed to fit the data 	<ul style="list-style-type: none"> Utilizes a systematic approach to find the optimal number of clusters through iterative fitting and SSE calculation Uses the 'elbow method' to determine the optimal number of clusters
Threshold Value	<ul style="list-style-type: none"> Single fixed threshold value λ used throughout the clustering process 	<ul style="list-style-type: none"> Determines the optimal threshold value based on the silhouette score Updates λ value dynamically for better clustering performance
Convergence	<ul style="list-style-type: none"> Repeats assignment and update steps until convergence is achieved based on threshold λ 	<ul style="list-style-type: none"> Similar convergence criteria Enhanced with repeated optimization steps for more accurate clustering

Aspect	Original DP-Means	EDP-Means
Clustering Process	<ul style="list-style-type: none"> • Assigns data points to clusters based on distance and threshold value λ • Iteratively updates cluster centers until convergence 	<ul style="list-style-type: none"> • Similar process of assignment and updating cluster centers • Enhanced with iterative optimization steps and threshold adjustment
Performance	<ul style="list-style-type: none"> • May require tuning of parameters such as λ and ϵ • Simpler, may require fewer iterations depending on data and threshold value λ 	<ul style="list-style-type: none"> • Automatic determination of parameters based on data • Potentially more complex due to additional steps for finding optimal parameters • More computationally intensive due to iterative optimization steps
Use Cases Suitability	<ul style="list-style-type: none"> • Suitable for cases where a fixed threshold can effectively determine clusters 	<ul style="list-style-type: none"> • More robust for diverse datasets with varying cluster structures • Better for applications requiring optimal cluster number determination and dynamic threshold adjustment
Advantages	<ul style="list-style-type: none"> • Simple to implement and understand • Effective for datasets with well-defined clusters and appropriate threshold λ 	<ul style="list-style-type: none"> • More flexible and adaptable to different data characteristics • Improved clustering accuracy and performance through systematic optimization
Disadvantages	<ul style="list-style-type: none"> • May require trial and error to find suitable λ • Performance highly dependent on initial threshold value 	<ul style="list-style-type: none"> • More computationally intensive • Potentially higher complexity in implementation

Overall, the proposed EDP-Means provides a more comprehensive and efficient clustering method, making it a valuable technique for applications requiring precise and adaptable cluster analysis. Its enhanced capabilities enable it to deliver improved clustering accuracy and performance, making it well-suited for a wide range of data clustering challenges. To explore these capabilities of the EDP-Means algorithm, the study evaluates its performance using datasets and compares the results with those obtained from K-Means and the original DP-Means algorithms. The comparison focuses on key performance metrics to determine the effectiveness of each clustering method. This comparison uses three validation indexes: the Silhouette Score, the Calinski-Harabasz (CH) Index, and the Davies-Bouldin (DB) Index.

3.3 Unsupervised Evaluation Metrics

The evaluation of clustering algorithms is crucial for assessing their performance and understanding their effectiveness in organizing data into meaningful groups. Among the various metrics employed for this purpose, the Silhouette score, DB index, and CH index stand out as widely used measures. These metrics provide valuable insights into different aspects of clustering quality, such as cluster cohesion, separation, and overall compactness. In exploring the significance and interpretation of these three indexes, the focus will be on their contributions to assessing clustering algorithms' unsupervised performance.

1. **Silhouette Score:** The silhouette score is calculated for each data point and provides a measure of how well that point lies within its own cluster compared to other clusters. The silhouette score ranges from -1 to 1. A score closer to 1 indicates that the point is well-clustered, with a clear separation between clusters. Conversely, a score near -1 suggests that the point may be assigned to the wrong cluster. For a given data point i the silhouette score $s(i)$ is computed as [38]:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (3.5)$$

In equation (3.5),

- $a(i)$ is the average distance from the point i to all other points within the same cluster.

- $b(i)$ is the minimum average distance from the point i to all points in any other cluster, representing how well-separated the clusters are.
2. DB Index: The DB index evaluates the clustering algorithm based on the average similarity between each cluster and its most similar cluster. A lower DB index indicates better clustering, with clusters that are more distinct and well-separated. It is computed as the average of the similarity ratios for each cluster i [42]:

$$DB = \frac{1}{n} \sum_{i=1}^n \max_{j \neq i} (sim(i, j) + sim(j, i) / d(i) + d(j)) \quad (3.6)$$

In equation (3.6),

- n is the number of clusters.
 - $sim(i, j)$ represents the similarity between clusters i and j , often measured as the distance between their centroids.
 - $d(i)$ is the measure of the compactness of the cluster i , typically represented by the average distance between its points and its centroid.
3. CH Index: The CH index quantifies the ratio between-cluster dispersion to within-cluster dispersion. A higher CH index suggests better clustering, indicating clusters that are dense and well-separated. It is calculated as [38]:

$$CH = (trace(B) / trace(W)) \times (N - k / k - 1) \quad (3.7)$$

In equation (3.7),

- B is the between-cluster dispersion matrix, representing the variability between cluster centroids.
- W is the within-cluster dispersion matrix, representing the variability within clusters.
- N is the total number of data points.
- k is the number of clusters.

By considering these three metrics together, a comprehensive understanding of the clustering performance is achieved, taking into account both the compactness of clusters and their separation from each other.

3.4 ETL and Edu-ETL Processes

ETL is a standard process in data warehousing and analytics, designed to extract data from various sources, transform it into a suitable format, and load it into a target system. In the context of educational data, a specialized variant addressed as Edu-ETL is employed. Edu-ETL tailors the traditional ETL process to address the unique challenges and requirements of educational data, including handling diverse data types, ensuring data quality, and enabling meaningful educational insights. This section delves into the distinctions and nuances between traditional ETL processes and Edu-ETL processes, highlighting their respective purposes, methodologies, and applications within the realm of data management and analytics.

Figure 3.1 provides an overview of the ETL processes, illustrating the key stages involved in extracting, transforming, and loading data for analysis and decision-making purposes [35][43][13].

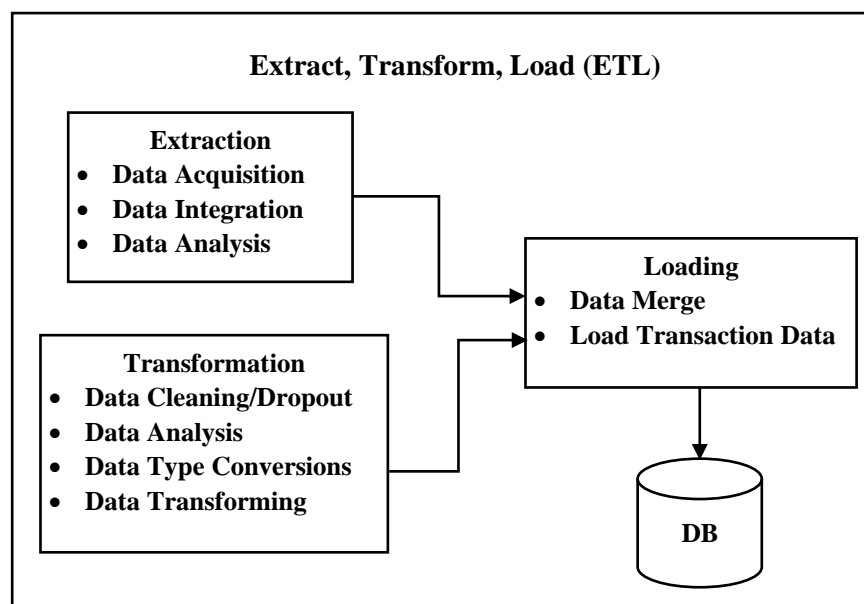


Figure 3.1 Overview of ETL Processes

ETL processes are critical for several reasons:

- **Data Quality:** Through cleaning and transformation, ETL processes improve the quality and reliability of data.

- **Data Integration:** ETL allows data from multiple sources to be combined into a single, coherent dataset, making it easier to analyze and draw insights.
- **Efficiency:** Automated ETL processes can handle large volumes of data quickly and efficiently, enabling timely access to up-to-date information.
- **Scalability:** ETL systems can scale to handle increasing data volumes and complexities as organizational data needs grow.

While ETL processes are powerful, they also come with challenges:

- **Complexity:** Designing and maintaining ETL processes can be complex, particularly when dealing with diverse and large-scale data sources.
- **Data Quality Issues:** Ensuring data quality throughout the ETL process requires robust validation and error-handling mechanisms.
- **Performance:** Processing large volumes of data efficiently can be challenging, requiring careful optimization of ETL workflows.
- **Change Management:** ETL processes must adapt to changes in source data structures and business requirements, necessitating ongoing maintenance and updates.

Given these challenges, the need for a proposed approach, Edu-ETL arises. Edu-ETL is tailored specifically for the education sector, addressing the unique data integration and transformation requirements of educational institutions. Table 3.11 presents a step-by-step outline of the proposed Edu-ETL processes designed for the educational domain to enhance the efficiency, accuracy, and relevance of the data preparation process.

Table 3.11 Edu-ETL Processes for In-depth Data Analysis

Step	Process
Step 1: Extraction	<ul style="list-style-type: none"> • Data source identification • Harvesting data • Ensuring data quality • Student data consolidation (integrating data from multiple sources based on student ID)
Step 2: Transformation	<ul style="list-style-type: none"> • Cleansing and preparation • Data Normalization

Step	Process
Step 2.1: Iteration and Optimization	<ul style="list-style-type: none"> • Data Enrichment • Data Privacy and Security • Analyzing data distribution • Aggregating competencies/marks • Checking variation • Grouping for contribution assessment (significantly or not) • Assessing categorical attribute (encode or not) • Attribute renaming/pruning for loading simplification • Grading structure implementation • Detecting correlations/prioritizing analysis with visualization
Step 3: Loading	<ul style="list-style-type: none"> • Connecting and loading datasets • Staging data • Data Validation • Indexing and partitioning • Loading schema design

Edu-ETL is built upon foundational principles of data integrity, quality, and usability, ensuring that educational data is not only accurate and reliable but also accessible and actionable for stakeholders across the educational spectrum.

A detailed discussion of each step of the proposed Edu-ETL processes:

1. **Extraction:** The initial phase of the Edu-ETL process, extraction involves gathering essential data from diverse educational sources, laying the groundwork for comprehensive analysis and insights.
 - **Data Source Identification:** Identifying relevant sources of educational data.
 - **Harvesting Data:** Collecting data from these identified sources.
 - **Ensuring Data Quality:** Verifying the accuracy and completeness of the collected data.
 - **Student Data Consolidation:** Integrating data from multiple sources based on student ID to create a unified dataset.

2. **Transformation:** Once the data is extracted, it undergoes transformation to ensure it is clean, consistent, and ready for analysis. This step includes a series of processes tailored specifically for educational data:
 - **Cleansing and Preparation:** Removing inaccuracies and inconsistencies from the data.
 - **Data Normalization:** Standardizing data to ensure uniformity.
 - **Data Enrichment:** Enhancing data with additional relevant information.
 - **Data Privacy and Security:** Ensuring compliance with data privacy regulations.
 - **Analyzing Data Distribution:** Understanding the spread and characteristics of the data.
 - **Aggregating Competencies/Marks:** Summarizing educational metrics.
 - **Checking Variation:** Identifying variations and outliers in the data.
 - **Grouping for Contribution Assessment:** Grouping data to assess contributions to educational outcomes.
 - **Assessing Categorical Attributes:** Deciding on encoding strategies for categorical data.
 - **Attribute Renaming/Pruning:** Simplifying attribute names and removing irrelevant ones.
 - **Grading Structure Implementation:** Applying grading structures to the data.
 - **Detecting Correlations:** Identifying relationships between attributes and prioritizing analyses with visualizations.
3. **Loading:** The final step involves loading the transformed data into a target system for analysis and reporting. This ensures that the data is accessible and usable for stakeholders:
 - **Connecting and Loading Datasets:** Establishing connections to the target system and loading the datasets.
 - **Staging Data:** Temporarily storing data for further processing.
 - **Data Validation:** Ensuring the loaded data is accurate and consistent.

- **Indexing and Partitioning:** Organizing data to improve query performance.
- **Loading Schema Design:** Structuring the data schema to facilitate efficient data retrieval and analysis.

By following these steps, Edu-ETL processes ensure that educational data is effectively managed, transformed, and utilized to drive insights and decision-making, ultimately enhancing educational outcomes and institutional performance. Table 3.12 provides differences between ETL and Edu-ETL processes from various aspects.

Table 3.12 Difference between ETL and Edu-ETL Processes

Aspect	ETL	Edu-ETL
Purpose	<ul style="list-style-type: none"> • General data processing for business intelligence. 	<ul style="list-style-type: none"> • Specialized for educational data analysis, focusing on student performance, learning outcomes, and institutional effectiveness within educational institutions.
Data Sources	<ul style="list-style-type: none"> • Various, often business-related (e.g., sales, finance). 	<ul style="list-style-type: none"> • Educational sources such as student records, assessments, attendance, course data, and other educational-specific datasets within educational institutions.
Transformation Focus	<ul style="list-style-type: none"> • General data cleaning and transformation. 	<ul style="list-style-type: none"> • Focused on educational metrics, grading structures, competencies, and domain-specific transformations tailored to educational analytics.
Iteration and Optimization	<ul style="list-style-type: none"> • Limited, primarily focused on initial transformation. 	<ul style="list-style-type: none"> • Continuous iteration and optimization for improving educational insights, supporting ongoing educational improvements and decision-making.

Aspect	ETL	Edu-ETL
Aggregation	<ul style="list-style-type: none"> • Business metrics (e.g., revenue, profit). 	<ul style="list-style-type: none"> • Educational metrics (e.g., total scores, competency aggregation, learning outcomes), enabling a deeper understanding of student performance and academic trends.
Handling Categorical Data	<ul style="list-style-type: none"> • Encoding based on business needs. 	<ul style="list-style-type: none"> • Specific strategies for handling categorical data in educational contexts, such as grading categories, student proficiency levels, and learning progressions.
Visualization	<ul style="list-style-type: none"> • Standard business dashboards and reports. 	<ul style="list-style-type: none"> • Specialized educational visualizations (e.g., heatmaps, performance trends, student progress trackers) tailored to educational stakeholders' needs for better data interpretation and decision-making.
Final Loading	<ul style="list-style-type: none"> • Into business intelligence platforms. 	<ul style="list-style-type: none"> • Into educational analysis platforms and tools designed for educational stakeholders, facilitating deeper insights into student learning and institutional performance.

In conclusion, while traditional ETL processes provide a robust framework for general data integration and business intelligence, Edu-ETL processes are specifically tailored to address the unique challenges and requirements of educational data analysis, offering customized transformations, continuous optimization, and specialized visualizations that empower educational institutions to make data-driven decisions and enhance student outcomes.

3.5 Trends and Applications of Big Data in Education

Big data technologies have driven innovation in various sectors, including healthcare, technology, and education. This innovation is increasingly essential in the education sector at all levels. Referred to as "Education 4.0," this evolving framework addresses the diverse needs of the educational field. This article specifically outlines the architecture and properties of big data that are well-suited for the education sector [28]. With the number of global internet users now at an impressive 5.16 billion, there has been a significant increase in the volume of continuously collected data. The challenge, however, is to effectively utilize this vast data. The COVID-19 pandemic has expedited the digital transformation in higher education, bringing it in line with other industries [5].

Since 2016, researchers have increasingly focused on the field of education, as reflected in the Scopus database. Over the past decade, 352 publications have been documented, with 98 of them (more than 27%) published in 2021 alone. These findings were obtained by searching the Scopus database for research publications containing at least the keywords "Big Data" and "Higher Education." This surge in interest has inspired the authors of this research to explore the potential of Big Data in this vital area [8].

3.5.1 Educational Aspects of Big Data

Big data technologies are leveraged to extract valuable information from vast, diverse, and continually growing datasets [47]. These massive datasets are utilized to develop various applications for mining educational data and gaining insights, thereby enhancing the intelligence of educational institutions such as schools and universities. The data within the educational system is classified as big data due to the large volume and variety of information generated regularly. This includes data on students' attitudes and interactions with learning platforms, learning activities, course information that varies significantly, and additional information that improves the quality of educational processes [30].

Big data holds the significant promise for transforming education. Today's generation of students has grown up with technology, and their daily activities generate a multitude of digital footprints. These include movements detected by motion sensors, keystrokes and mouse interactions on computers, and taps and

gestures on mobile phones and tablets. Although the trend of big data is still in its early stages, it has already shown considerable potential in the field of education [48].

Education-related data is gathered from a variety of sources across different educational settings, including traditional classrooms and alternative learning management systems (LMS). These sources encompass student records, behavior logs, examination results, social media posts, administrative data, demographic data, and IoT data [45]. The education sector has the potential to generate an immense volume of data. Similar to other extensive data mining efforts, mining large amounts of educational data involves several key steps: collecting data from various sources such as learning management systems and online assessments; cleaning and preprocessing the data to ensure accuracy and consistency; storing and managing the data in centralized systems; performing exploratory data analysis and visualization to uncover patterns; applying data mining techniques like clustering and regression to build predictive models; interpreting and evaluating the results to gain actionable insights; deploying the models in real-world settings; and ensuring ethical considerations and data privacy are maintained throughout the process.

3.5.2 Big Data Framework: Apache Spark

Big data refers to large, varied, and complex data collections that pose challenges in terms of storage, processing, and presentation for future use or outcomes. In 2021, a leading market research firm predicted the production of 74 zettabytes of data. Furthermore, the International Data Corporation (IDC) projects that the global data volume will surpass 175 zettabytes by 2025, reflecting a compound annual growth rate of 61 percent. Big data analytics involves analyzing these massive datasets to uncover hidden patterns and relationships. Research in big data is at the forefront of contemporary research and industry. Online transactions, emails, videos, audio files, pictures, click streams, logs, posts, web searches, medical records, social networking activities, scientific data, sensors, and mobile phone applications all contribute to the generation of big data. These data are stored in rapidly growing and increasingly complex databases, which makes them challenging to collect, construct, store, manage, distribute, analyze, and present using traditional database software tools.

Researchers, businesses, and individuals have defined big data in various ways, but the most common definitions focus on the three Vs: Velocity, Volume, and

Variety [40]. Apache Spark is a highly parallel in-memory processing solution designed to support both batch and stream data processing. Its primary objective is to accelerate batch data processing using in-memory computing. In the realm of in-memory analytics, Spark exhibits the potential to achieve speeds up to 100 times faster than the Hadoop MapReduce framework. The core engine of Apache Spark offers essential cluster computing capabilities coupled with in-memory functionality, including fault recovery, memory management, job scheduling, and communication with databases [27].

Apache Spark is a powerful tool for educational data mining (EDM), offering parallel processing and in-memory computing capabilities. It efficiently handles large-scale educational datasets, including student records and assessment results. Spark's flexibility allows for both batch and real-time processing, enabling timely analysis of student behavior and learning trends. Integration with machine learning libraries enables advanced analytics, personalized learning experiences, and identification of at-risk students. In conclusion, Apache Spark empowers educators to extract insights from complex datasets, enhancing student outcomes and driving innovation in education.

3.6 Chapter Summary

This chapter delves into various theoretical backgrounds relevant to the proposed approaches EDP-Means clustering technique and Edu-ETL processes.

Commencing with the DP-Means clustering technique, it serves as a nonparametric extension of K-Means clustering. The discussion encompasses the role of the threshold parameter (λ), key concepts, algorithmic steps, and a comparison between K-Means and DP-Means clustering techniques. Despite its merits, DP-Means clustering also poses limitations and challenges, which are critically evaluated. Subsequently, the EDP-Means clustering technique is examined, which improves upon DP-Means to address its shortcomings and enhance its performance.

Shifting to unsupervised evaluation metrics, techniques for assessing clustering results without predefined class labels are explored. The section further discusses the comparison between traditional ETL processes and educational-specific Edu-ETL processes, emphasizing the tailored nature of Edu-ETL for handling educational data.

Additionally, the trends and applications of big data in education are

scrutinized, highlighting its potential to revolutionize teaching and learning practices. The educational aspects of big data, the opportunities it presents, the mining process of educational data, and the role of Apache Spark as a big data framework are also examined.

Overall, this chapter provides a comprehensive overview of key theoretical concepts and methodologies in the related proposed approaches, laying the groundwork for further exploration and research in the field.

CHAPTER 4

THE ARCHITECTURE OF THE PROPOSED SYSTEM

Comprehending lifelong learning achievements is essential for educational data analysis, as it offers insights into students' academic advancement. Conventional assessment approaches frequently fail to capture the diverse facets of student proficiency and learning behavior. Various datasets, such as academic accomplishments and socioeconomic background, influence the trajectories of students' learning journeys. Educators can tailor interventions effectively based on these insights. However, exploration of lifelong learning achievements remains underexplored, despite its significance in today's dynamic world. Analyzing lifelong learning achievements is particularly critical for each student within the educational domain, as it offers a comprehensive understanding of their academic journey over time, enabling tailored interventions and support strategies to optimize their learning outcomes and overall success.

The architecture of the proposed system delineates a structured framework designed to elucidate lifelong learning achievements through a systematic process. Informed by the outlined step-by-step implementation of the proposed approaches, the system aims to analyze educational datasets, explore clustering methodologies comprehensively, and integrate EDP-Means clustering techniques. The proposed system architecture outlines six main stages, each designed to enable a thorough analysis of educational data and learning outcomes. Furthermore, the contributions of enhancing the DP-Means clustering algorithm and developing the Edu-ETL processes extend beyond the educational realm, providing valuable insights for improving clustering techniques in broader applications.

This introduction provides an overview of the main stages within the proposed system architecture, outlining its core components and objectives.

1. **Initial Dataset Acquisition Stage:** The primary focus is on acquiring the foundational data necessary for subsequent analysis. This involves gathering relevant datasets from various sources. The collected data may include student profiles, academic records, assessment results, and any other pertinent information needed for the analysis. This stage sets the groundwork for the subsequent stages of data processing and analysis within the proposed system architecture.

2. **Data Analysis Stage:** The system demonstrates adaptability by differentiating between merged and non-merged data preprocessing. For merged data (handled by Edu-ETL processes), it effectively integrates and cleans multiple datasets to form a unified dataset. In contrast, for non-merged data, it uses specific preprocessing techniques to maintain data quality for clustering analysis. Through rigorous analysis, these datasets are prepared for subsequent clustering processes.
3. **Clustering Processes:** Following the data analysis stage, the proposed system delves into three distinct clustering processes: K-Means clustering, original DP-Means clustering, and EDP-Means clustering. Each process utilizes the processed/transformed data to categorize students into distinct groups based on their educational outcomes. These clustering methodologies leverage different algorithms to identify patterns and relationships within the dataset, facilitating a nuanced understanding of student learning trajectories.
4. **Clustering Processes in PySpark:** In parallel with the traditional clustering processes, the proposed system incorporates PySpark to perform clustering algorithms in a distributed computing environment. This stage involves executing K-Means clustering, original DP-Means clustering, and EDP-Means clustering algorithms within the PySpark framework. Leveraging the scalability and efficiency of PySpark, the system aims to enhance the performance of clustering algorithms and expedite computational processes.
5. **Cluster Validation and Analysis:** Upon completion of the clustering processes, the proposed system enters the cluster validation and analysis stage. Here, clustered results from all processes are subjected to comprehensive validation metrics, including the Silhouette Score, CH Index, and DB Index. Through rigorous analysis, the system evaluates the effectiveness and performance of each clustering method, providing insights into their scalability, efficiency, and clustering quality.

6. **Analyze Learning Outcomes and Success Factors:** After the clustering processes, the proposed system proceeds to the stage of analyzing learning outcomes and identifying success factors. This stage involves examining the clustered results to uncover patterns and relationships between student groups and their academic achievements. Through thorough analysis, the system aims to identify key success factors that influence lifelong learning achievements, enabling tailored interventions and support strategies to enhance educational outcomes.

By delineating the main stages of the system, this framework facilitates a systematic exploration of educational datasets and clustering techniques, ultimately contributing to a deeper understanding of student learning outcomes and trajectories. The methodology outlined in this study is derived from the overview depicted in Figure 4.1.

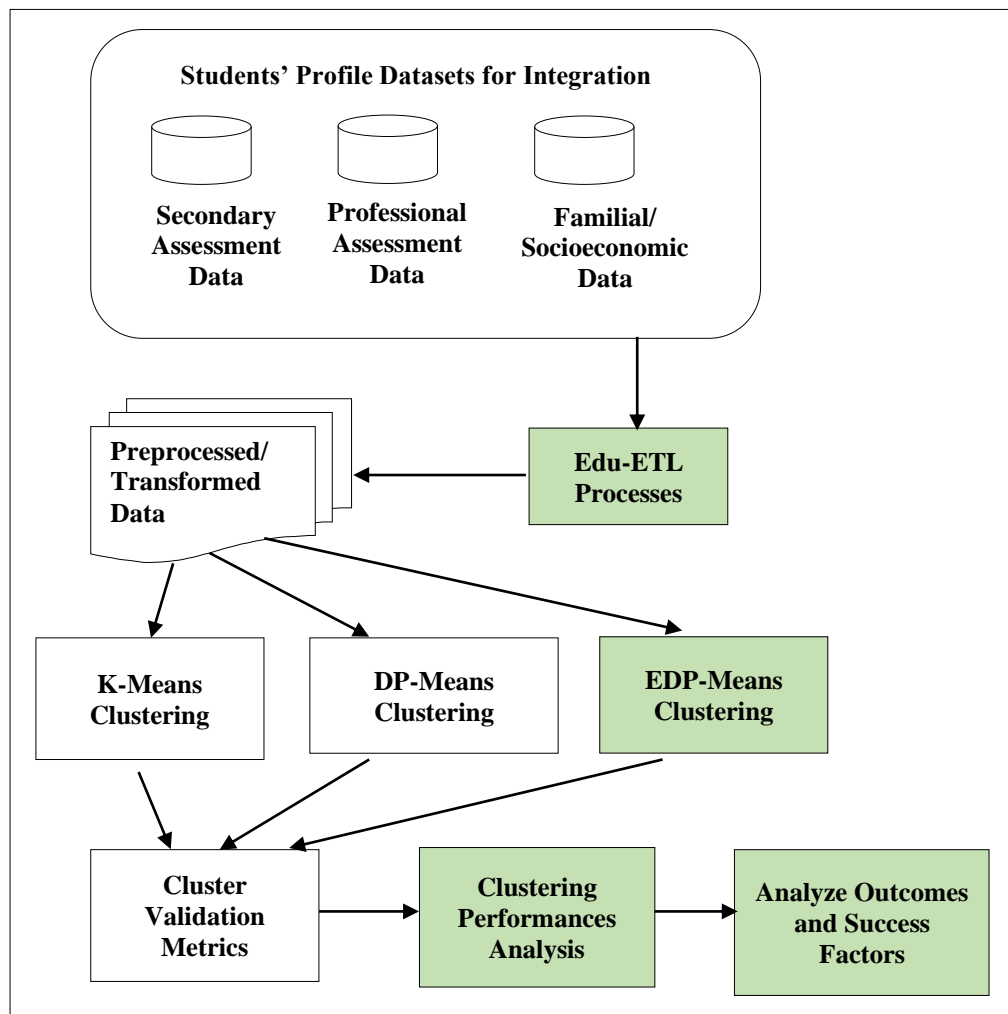


Figure 4.1 Architecture of the Proposed System

4.1 Initial Data Acquisition

The primary dataset “Academic Evaluation” used in this research was sourced from the “Mendeley Data Repository”. It offers a thorough look into the academic, social, and economic backgrounds of students. This dataset facilitates the analysis of academic performance among engineering students at two distinct junctures. The initial phase includes the outcomes of secondary evaluations, while the subsequent phase pertains to the results of professional assessments. Additionally, attributes related to the student’s social contexts have been incorporated.

This dataset was obtained through the systematic integration of databases from the Colombian Institute for the Evaluation of Education [17]. 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. Specifically, attributes heavily dependent on the unique socio-economic and educational context of Colombia were adjusted to ensure broader applicability and relevance. Unnecessary or unrelated attributes can detrimentally impact both the performance and processing time of the clustering process. The selection of attributes underwent scrutiny through in-depth data analysis facilitated by the Edu-ETL processes. Before employing the proposed clustering algorithms, this dataset was preprocessed via the Edu-ETL processes to ensure it was ready for analysis.

The reason for primarily using this dataset is that it provides an excellent foundation to explore and demonstrate the effectiveness of the proposed Edu-ETL processes. Additionally, it allows for a comprehensive examination of lifelong learning achievements, revealing correlations between past and present educational outcomes, and illustrating the impact of various factors on student success. Due to privacy concerns and data protection regulations, access to student information was not granted. Consequently, it was not feasible to apply real local datasets for this research. This limitation necessitated the use of the dataset from the Mendeley Data Repository. While this dataset serves the research objectives well, it is acknowledged that local datasets might offer additional insights specific to the context of the region of interest. However, the adaptations made to the Colombian dataset aimed to ensure its relevance and applicability to the broader research goals.

In addition to this primary dataset, other datasets were also experimented with to evaluate the robustness and generalizability of the proposed system. To ensure

comprehensive data handling, the preprocessing of data was approached in two distinct forms, each catering to different data scenarios:

1. Merged Data (Edu-ETL)

Efficient Integration and Cleaning of Multiple Datasets: The Edu-ETL process involves the extraction, transformation, and loading (ETL) of educational data from various sources into a unified dataset. This method focuses on the efficient integration and cleaning of multiple datasets to ensure a cohesive and consistent data foundation. The preprocessing steps include handling missing values, normalizing data, and transforming data types to maintain data integrity and consistency.

2. Non-merged Data

Tailored Preprocessing Techniques for Data Quality: In scenarios where datasets are not merged, tailored preprocessing techniques are applied to individual datasets to enhance their quality. This involves custom preprocessing steps such as filtering, aggregating, and feature engineering specific to each dataset. The goal is to ensure that each dataset is optimized for subsequent analysis and clustering processes, maintaining its unique characteristics and improving overall data quality.

However, the focus of this research and the majority of the experiments were conducted using the dataset from the Mendeley Data Repository, as it best supports the objectives of demonstrating Edu-ETL's capabilities and exploring lifelong learning achievements.

4.2 Data Analysis with Edu-ETL Processes

This stage involves transforming raw educational data into a format suitable for detailed analysis, ensuring that the data is accurate, consistent, and enriched for meaningful insights. The Edu-ETL processes involve three key steps: Extraction, Transformation, and Loading, tailored specifically for educational data analysis. A detailed discussion of each step is mentioned in Chapter 2.

The provided Edu-ETL facilitates various data processing tasks for educational datasets aligning with the stages of the Edu-ETL process. In the extraction stage, the process involves collecting and integrating the students' profile datasets from multiple sources based on student ID to create a unified dataset. The transformation stage ensures proper column alignment, removes duplicates, addresses missing values, and drops specific columns based on in-depth data analysis. This stage also includes renaming columns, transforming categorical data into numerical

representations for specific columns, and categorizing scores into grades for analysis. Additionally, some visualizations are used for exploratory data analysis. The loading stage of the Edu-ETL process, where the cleaned and transformed data is loaded into the target system for further analysis and reporting. In this implementation, SQLite3 is used as the target database system. This stage ensures that the data is stored in a structured format, enabling efficient retrieval and further analysis. The use of SQLite3 provides a lightweight and easily manageable database solution, making the Edu-ETL process more robust and scalable. After applying the Edu-ETL processes, the processed and transformed data is ready for cluster analysis using the proposed clustering algorithms.

4.3 Clustering Process: K-Means, Original DP-Means, EDP-Means

The clustering process within the architecture of the proposed system is designed to compare execution time, cluster result quality, and accuracy across different algorithms. The proposed system's architecture integrates the EDP-Means algorithm, an improvement over the Original DP-Means, for enhanced clustering performance. The comparison between the two algorithms encompasses various aspects such as initialization, number of clusters, threshold value determination, clustering process, performance, convergence, use case suitability, and advantages/disadvantages. To validate the effectiveness of the EDP-Means algorithm, the system evaluates its performance using an educational dataset and compares the results against K-Means and the original DP-Means clustering algorithms. The architecture facilitates this comparison by leveraging processed and transformed data to categorize students into distinct groups based on their educational outcomes. This structured comparison is intended to prove that EDP-Means achieves higher accuracy and efficiency in educational data clustering, thereby validating the improvements within the proposed system's architecture.

By analyzing the clustering results across three validation indexes, the proposed system aims to demonstrate that the enhancements in the DP-Means algorithm enable EDP-Means to perform better than K-Means and original DP-Means. The system used various datasets from different fields to evaluate the algorithm's performance. Figure 4.2 presents the proposed EDP-Means clustering algorithm.

EDP-Means Clustering Algorithm

Input: $X = (x_i)_{i=1}^N \subset R^d$, λ , λ_{min} , λ_{max}

Output: $(\mu_k)_{k=1}^{K^*}$, $(z_i)_{i=1}^N$, λ^*

//Initialization

1. $K \leftarrow 1$
2. $\mu_1 \leftarrow \frac{\sum_{i=1}^N x_i}{N}$
3. $(z_i)_{i=1}^N \leftarrow 1$

//Cluster Assignment

4. **While Not Converged do**
5. **for** $i \in \{1, \dots, N\}$ **do**
6. $z_i \leftarrow \arg \min_{k \in \{1, \dots, K\}} \|x_i - \mu_k\|_2^2$

//Find optimal number of clusters

7. $SSE_k \leftarrow \sum_{i=1}^N \|x_i - \mu_{z_i}\|_2^2$
8. $sse_k \leftarrow (SSE_k)_{k=1}^K$
9. $K^* \leftarrow \arg \min_k (sse_k)$

//Find optimal λ

10. **for** $\lambda_j \in \lambda_j \in [\lambda_{min}, \lambda_{max}]$ **do**
11. $Silhouette(\lambda_j) = \frac{1}{N} \sum_{i=1}^N \frac{b_i - a_i}{\max(a_i, b_i)}$
12. $\lambda^* \leftarrow \arg \max_{\lambda_j} Silhouette(\lambda_j)$

//Cluster Adjustment

13. **for** $i \in \{1, \dots, N\}$ **do**
14. **if** $\|x_i - \mu_k\|_2^2 > \lambda^*$ **then**
15. $K^* \leftarrow K^* + 1$
16. $\mu_{K^*} \leftarrow x_i$
17. $z_i \leftarrow K^*$

//Update the cluster centers

18. **for** $k = 1 \in \{1, \dots, K^*\}$ **do**
 19. $n_k \leftarrow |\{i: z_i = k\}|$
 20. $\mu_k \leftarrow \frac{\sum_{i: z_i = k} x_i}{n_k}$
-

Figure 4.2 EDP-Means Clustering Algorithm

Table 4.1 Notations for the EDP-Means Clustering Algorithm

Symbol	Description
$X = (x_i)_{i=1}^N \subset R^d$	Set of data points in R^d
λ	Initial value for the threshold parameter
λ_{min}	Minimum value for λ
λ_{max}	Maximum value for λ
$(\mu_k)_{k=1}^{K^*}$	Set of cluster centers after convergence
$(z_i)_{i=1}^N$	Cluster assignments for each data point
λ^*	Optimized value of λ
K	Current number of clusters
μ_1	Initial cluster center, the mean of all data points
z_i	Cluster assignment for data point x_i
$\ x_i - \mu_{z_i}\ _2^2$	Squared Euclidean distance between x_i and μ_k
SSE_k	Sum of squared errors for cluster k
K^*	Optimal number of clusters
$Silhouette(\lambda_j)$	Silhouette score for a given λ_j
λ_j	A specific value of λ within the range $\{\lambda_{min}, \lambda_{max}\}$
n_k	Number of points in cluster k
μ_k	Updated cluster center for cluster k

The algorithm continues iterating until the cluster assignments (z_i) no longer change significantly or the centroids (μ_k) stabilize. The proposed EDP-Means clustering algorithm employs a multi-step process to effectively partition a dataset into clusters.

Figure 4.2 indicates the modified part with line numbers from 7 to 12. Initially, it initializes with one cluster center set to the mean of all data points. Subsequently, each data point is assigned to the nearest cluster center, forming an initial clustering configuration. Then, it computes the $SSE_k \leftarrow \sum_{i=1}^N \|x_i - \mu_{z_i}\|_2^2$ for

each cluster to assess the clustering quality and identifies the optimal number of clusters $K^* \leftarrow \arg \min_k (sse_k)$ by minimizing the SSE. Next, it iterates over a range of (λ) values: $\lambda_j \in [\lambda_{min}, \lambda_{max}]$ to maximize the silhouette score $Silhouette(\lambda_j) = \frac{1}{N} \sum_{i=1}^N \frac{b_i - a_i}{\max(a_i, b_i)}$, which measures the similarity of data points within clusters relative to neighboring clusters, aiding in selecting the optimal $\lambda^* \leftarrow \arg \max_{\lambda_j} Silhouette(\lambda_j)$. If a data point deviates too far from its assigned cluster center (beyond λ^*), a new cluster is created with the outlier as its center. Finally, the algorithm updates the cluster centers based on the current clustering assignments. This iterative process ensures the clustering structure is compact and well-separated, guided by the silhouette score criterion.

The incorporation of techniques to find the optimal K^* and λ^* are the key enhancements in the EDP-Means clustering algorithm compared to the original DP-Means algorithm. These techniques distinguish EDP-Means from its predecessor by offering improved adaptability and performance. In the original DP-Means algorithm, the number of clusters K and the threshold parameter λ were typically fixed or set based on heuristic choices. However, in EDP-Means, these parameters are dynamically adjusted based on the data characteristics and clustering quality.

4.4 Clustering Process in PySpark: K-Means, Original DP-Means, EDP-Means

The proposed system integrates advanced clustering methodologies to categorize students based on their educational outcomes at different level. Initially, the system undertakes traditional clustering processes involving K-Means clustering, original DP-Means clustering, and EDP-Means clustering algorithms. Each of these processes employs processed and transformed data to identify patterns and relationships within the dataset, facilitating a nuanced understanding of student learning trajectories. Figure 4.3 describes the flow of processes with the proposed clustering algorithms in PySpark.

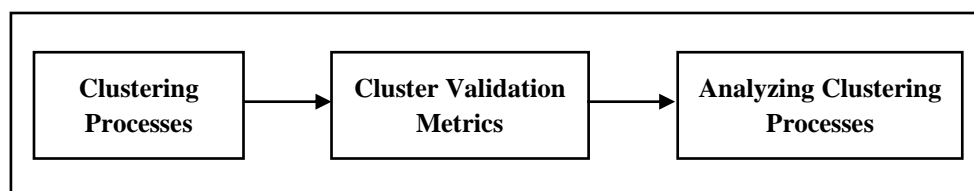


Figure 4.3 Overview of Clustering Processes in PySpark Environment

In parallel with these traditional approaches, the proposed system also incorporates PySpark in the PyCharm environment to execute these clustering algorithms in a distributed computing environment. By leveraging PySpark, the system can perform K-Means clustering, original DP-Means clustering, and EDP-Means clustering algorithms on a larger scale and with greater efficiency. PySpark's scalability and distributed computing capabilities significantly enhance the performance of these clustering algorithms, expediting the computational processes involved. This parallel implementation ensures that the system can handle large datasets more effectively, providing timely and accurate clustering results. The fundamental clustering logic remains consistent across both traditional and PySpark implementations; the key difference lies in the use of PySpark to optimize and accelerate the processing of large-scale data.

4.5 Clustering Process in PySpark: K-Means, Original DP-Means, EDP-Means

In the cluster validation and analysis stage, the system evaluates the quality and effectiveness of the clustering results obtained from all processes. Key performance metrics, including the Silhouette Score, the CH Index, and the DB Index, serve as datasets for assessing the effectiveness of each clustering method. By systematically evaluating each algorithm using these three validation indexes, and make a comprehensive comparison of their clustering performance:

- Compare the Silhouette Scores of EDP-Means, K-Means, and original DP-Means to determine which algorithm yields better cluster cohesion and separation.
- Compare the CH Index values across the three algorithms to assess their clustering performance in terms of intra-cluster similarity and inter-cluster dissimilarity.
- Compare the DB Index values to evaluate the compactness and separation of clusters generated by EDP-Means, K-Means, and original DP-Means.

Through this comparative analysis, the study aims to provide valuable insights into the strengths and limitations of EDP-Means in comparison to existing clustering techniques, thereby highlighting its potential as a preferred choice for diverse clustering applications.

4.6 Analyze Learning Outcomes and Success Factors

When the results come out from the three clustering algorithms, it is important to discuss, explore, and show the findings for each algorithm to provide a comprehensive analysis. The objective of this stage is to explore the relationship between previous student performance results, family background, socioeconomic factors, and current academic performance. By analyzing clustered results obtained from the proposed clustering algorithms and considering various attributes related to student's academic, social, and economic backgrounds, the system aims to identify key success factors that influence lifelong learning achievements.

4.7 Chapter Summary

The "Architecture of the Proposed System" chapter presents a structured framework aimed at comprehensively understanding lifelong learning achievements. Recognizing the significance of such insights in educational data analysis, the system is designed to address the limitations of conventional assessment approaches. It emphasizes the influence of various datasets, including academic accomplishments and socioeconomic backgrounds, on students' learning journeys.

The system's architecture comprises six main stages, starting with the Initial Dataset Acquisition Stage, where foundational data is collected. Subsequently, the Data Analysis Stage employs Edu-ETL processes to prepare datasets for clustering. The Clustering Processes stage utilizes K-Means, original DP-Means, and EDP-Means clustering methods to categorize students based on their educational outcomes. Leveraging PySpark, the system conducts clustering in a distributed computing environment, enhancing performance. Cluster Validation and Analysis follow, assessing clustering quality using metrics like Silhouette Score and CH Index. Finally, the system analyzes Learning Outcomes and Success Factors, uncovering patterns between student groups and academic achievements. By delineating these stages, the framework facilitates a systematic exploration of educational datasets and clustering techniques, ultimately enhancing understanding of student learning outcomes and trajectories.

CHAPTER 5

IMPLEMENTATION OF THE EDU-ETL PROCESSES

In this chapter, the implementation of the proposed Edu-ETL processes for an analytical system designed to track lifelong learning achievements is detailed. Following the system pipeline described in Chapter 4, these Edu-ETL processes preprocess and transform data from student profile datasets sourced from multiple origins. Initially, the student profile datasets are utilized to demonstrate the fundamental capabilities and procedures of the Edu-ETL processes. Then, the implementation steps for the Edu-ETL processes are demonstrated, with system functionality illustrated through user-friendly Graphical User Interface (GUI) demonstrations, offering a clear and concise visual explanation.

5.1 Edu-ETL Processes in Action: Experiments and Case Studies

The dataset utilized in this study comprises 12,411 observations, each representing a student with 44 attributes that capture personal information (categorical) and assessment results (numerical) in high school and university [17]. This educational dataset used in this experiment is instrumental in implementing the proposed system. However, the proposed system is versatile and can also be tested with other educational datasets as well as various kinds of datasets from different domains.

5.1.1 Dataset Overview and Categorization

Academic assessments are recorded at two significant points in a student's life:

1. Final Year of High School

- Mathematics (MAT_S11): Assesses students' skills in solving problems using mathematical tools.
- Critical Reading (CR_11): Evaluates skills needed to understand and interpret texts in everyday and academic contexts.
- Citizen Competencies (CC_S11): Measures knowledge and skills to understand social phenomena from the perspective of social sciences.

- Biology (BIO_S11): Tests the ability to explain natural phenomena based on scientific knowledge.
- English (ENG_S11): Assesses the competence to communicate effectively in English

2. Final year of University Studies

- Critical Reading (CR_SPRO): Measures the ability to understand and critically analyze texts.
- Quantitative Reasoning (QR_PRO): Assesses the ability to manipulate quantitative data in various representations.
- Citizen Competencies (CC_PRO): Evaluates the understanding of citizenship and inclusive coexistence.
- Written Communication (WC_PRO): Tests the ability to effectively communicate ideas in writing.
- English (ENG_PRO): Measures competence in communicating effectively in English.

Personal information collected at exam enrollment includes attributes such as socioeconomic level, participation in government aid programs (e.g., ‘sisben’), and household amenities (eg., Internet, TV, Computer, etc.). Table 5.1 and 5.2 describe attributes from datasets.

Table 5.1 Description of Numerical Attributes

Attribute	Full Name	Levels
MAT_S11	Mathematics	1-100
CR_S11	Critical Reading	1-100
CC_S11	Citizen Competencies S11	1-100
BIO_S11	Biology	1-100
ENG_S11	English	1-100
QR_PRO	Quantitative Reasoning	1-100
CR_PRO	Critical Reading	1-100
CC_PRO	Citizen Competencies of SPRO	1-100
ENG_PRO	English	1-100

Attribute	Full Name	Levels
WC_PRO	Written Communication	1-100
FEP_PRO	Proficiency of students in designing, planning, and engineering project	1-300
G_SC	Global Score (Overall achievement level of individual students)	1-300
PERCENTILE	Percentile (based on their observations)	1-100
2ND_DECILE	Second Decile (based on their position)	1-5
QUARTILE	Quartile (based on academic achievements in their position)	1-4
SEL	Socioeconomic Level	1-4
SEL_IHE	Socioeconomic Level in Higher Education	1-4

Table 5.2 Description of Categorical Attributes

Attribute	Full Name	Levels
GENDER	Gender	2
EDU_FATHER	Father's Education	12
EDU_MOTHER	Mother's Education	12
OCC_FATHER	Father's Occupation	13
OCC_MOTHER	Mother's Occupation	13
STRATUM	Stratum	7
SISBEN	Sisben	6
PEOPLE_HOUSE	People in the House	13
INTERNET	Internet	2
TV	TV	2
COMPUTER	Computer	2
WASHING_MCH	Washing machine	2
MIC_OVEN	Microwave oven	2

Attribute	Full Name	Levels
CAR	Car	2
DVD	DVD	2
FRESH	Fresh	2
PHONE	Phone	2
MOBILE	Mobile	2
REVENUE	Revenue	3
JOB	Job	8
SCHOOL_NAME	School Name	3735
SCHOOL_NAT	Nature of School	2
SCHOOL_TYPE	Type of School	4
Cod_SPRO	Code Saber Pro	12411
Cod_S11	Code Saber 11	12411
UNIVERSITY	University Name	134
ACADEMIC_PROGRAM	Academic Program	23

Table 5.3 Each Level Description of Categorical Attributes

Attribute	Levels	Description
GENDER	2	{'F', 'M'}
EDU_FATHER	12	{'0', 'Complete Secondary', 'Complete Primary', 'Complete Professional Education', 'Complete Technique or Technology', 'Incomplete Professional Education', 'Incomplete Secondary', 'Incomplete Primary', 'Incomplete Technical or Technological', 'None', 'Not Sure', 'Postgraduate Education'}
EDU_MOTHER	12	{'0', 'Complete Secondary', 'Complete Primary', 'Complete Professional Education', 'Complete Technique or Technology', 'Incomplete Professional Education', 'Incomplete Secondary', 'Incomplete Primary', 'Incomplete Technical or Technological', 'None', 'Not

Attribute	Levels	Description
		Sure', 'Postgraduate Education'}
OCC_FATHER	13	{'0', 'Auxiliary or Administrative', 'Entrepreneur', 'Executive', 'Home', 'Independent', 'Independent Professional', 'Operator', 'Other Occupation', 'Retired', 'Small Entrepreneur', 'Technical or Professional Level Employee'}
OCC_MOTHER	13	{'0', 'Auxiliary or Administrative', 'Entrepreneur', 'Executive', 'Home', 'Independent', 'Independent Professional', 'Operator', 'Other Occupation', 'Retired', 'Small Entrepreneur', 'Technical or Professional Level Employee'}
PEOPLE_HOUSE	13	{'0', 'Eight', 'Five', 'Four', 'Nine', 'Eleven', 'One', 'Seven', 'Six', 'Ten', 'Three', 'Twelve or More', 'Two'}
REVENUE	8	{'0', '10 or more LMMW', 'Between 1 and less than 2 LMMW', 'Between 2 and less than 3 LMMW', 'Between 3 and less than 5 LMMW', 'Between 5 and less than 7 LMMW', 'Between 7 and less than 10 LMMW', 'Less than 1 LMMW'}
JOB	6	{'0', 'No', 'Yes, 20 hours or more per week', 'Yes, less than 20 hours per week'}
SCHOOL_NAT	2	{'PRIVATE', 'PUBLIC'}
SCHOOL_TYPE	4	{'ACADEMIC', 'NOT APPLY', 'TECHNICAL', 'TECHNICAL/ACADEMIC'}
INTERNET	2	{'Yes', 'No'}
TV	2	{'Yes', 'No'}
COMPUTER	2	{'Yes', 'No'}
WASHING_MCH	2	{'Yes', 'No'}
MIC_OVEN	2	{'Yes', 'No'}
CAR	2	{'Yes', 'No'}
DVD	2	{'Yes', 'No'}
FRESH	2	{'Yes', 'No'}

Attribute	Levels	Description
PHONE	2	{'Yes', 'No'}
MOBILE	2	{'Yes', 'No'}

Table 5.3 comprehensively describes the categorical attributes used in the research, detailing the levels and possible values for each attribute. These attributes are integral to the research study and are sourced from the “Mendeley Data Repository”. The numerical attributes typically represent academic performance metrics, while the categorical attributes provide demographic and background information about the students.

5.1.2 Experimental Analysis of Datasets Using Edu-ETL Processes

In the proposed system, the Edu-ETL processes streamline data handling by identifying and organizing relevant attributes from datasets. This process begins by focusing on attributes such as gender, parental education, household characteristics, and academic performance at different educational levels.

1. Extraction Phase

Given the dataset’s composition, data were collected from various sources and categorized into three types of records:

1. Academic (Professional) Performance Records: 11 attributes
(RECod_Sid, Cod_SPRO, G_SC, QUARTILE, FEP_PRO, ACADEMIC_PROGRAM) and (Five Competencies - QR_PRO, CR_PRO, CC_PRO, WC_PRO, ENG_PRO)
2. Secondary Performance Records: 11 attributes
(RECod_Sid, Cod_S11, 2ND_DECILE, SCHOOL_NAT, SCHOOL_TYPE, PERCENTILE) and (Five Competencies - MAT_S11, CR_S11, CC_S11, BIO_S11, ENG_S11)
3. Household Socioeconomic Status Records: 21 attributes
(RECod_Sid, GENDER, EDU_FATHER, EDU_MOTHER, OCC_FATHER, OCC_MOTHER, PEOPLE_HOUSE, REVENUE, JOB, INTERNET, TV, COMPUTER, WASHING_MCH, MIC_OVEN, CAR, DVD, FRESH, PHONE, MOBILE, SEL, SEL_IHE)

A new attribute, RECod_Sid, was introduced as a student identifier to ensure

confidentiality while linking relevant information across different records. Datasets are integrated into the system using the RECod_Sid attribute, making the Cod_SPRO and Cod_S11 attributes unnecessary.

2. Transformation Phase

In the transformation phase, several preprocessing tasks are performed to prepare the data for analysis.

Data Cleaning:

- Handling Missing Values: Missing values are either dropped or imputed to ensure data completeness.
- Removing Duplicates: Duplicate records are identified and removed to maintain data integrity.

Attributes Selection:

- Attributes heavily influenced by socioeconomic and country-specific factors are excluded to focus on relevant features.
- Excluded Attributes: SISBN, STRATUM, SCHOOL_NAME, ACADEMIC_PROGRAM

Aggregating Scores:

- Two additional attributes, Total_S11 and Total_SPro, are created to represent cumulative scores in secondary and professional education. Introducing new attributes requires examining their variation and distribution to ensure relevance and impact on the overall analysis. Figure 5.1 displays the distribution and variation of the attributes Total_SPro and Total_S11, showing a normal (Gaussian) distribution.
- Total_S11 and Total_SPro attributes are applied to grade cumulative scores into levels 1 through 4. These aggregated scores facilitate comparisons and trend identification across different groups or periods. Figure 5.2 explores the distribution of different competency scores and the Total_SPro in the dataset. This figure shows that the distribution is normally distributed, the data variability is well spread, and there are no unusual or extreme values.
- The grading structures are adapted to align with the decisions, policies, and structural frameworks of individual educational institutions. This ensures that grading systems are fair, accurate, and reflective of specific academic environments and objectives.

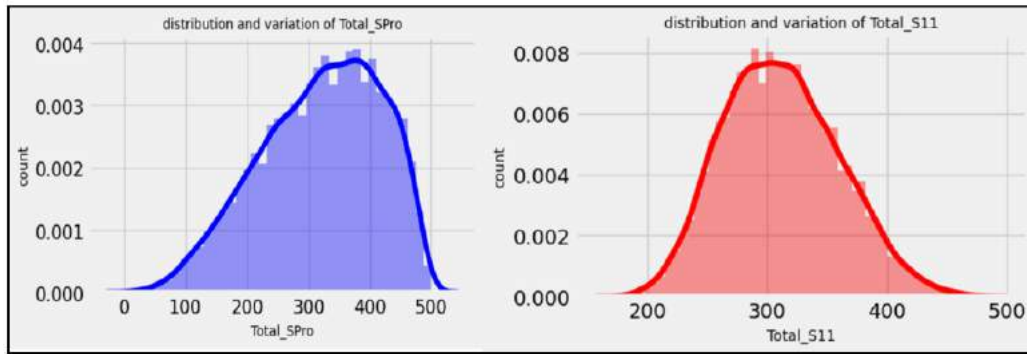


Figure 5.1 Distribution and Variation of Total_SPro and Total_S11 Attributes

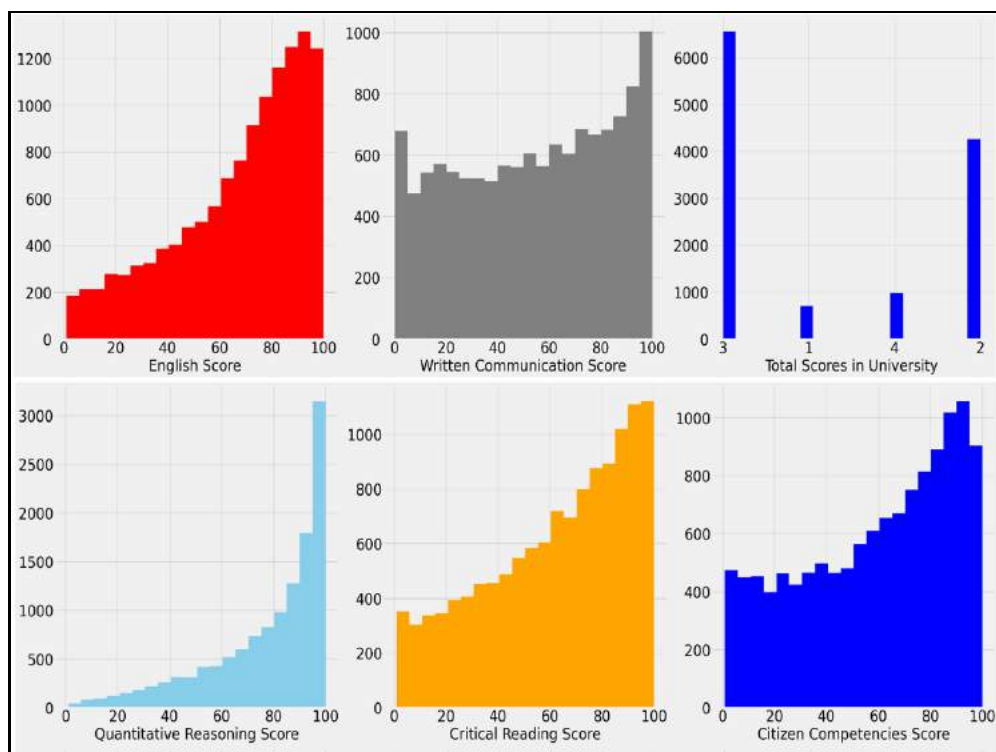


Figure 5.2 Distribution of Different Competency Scores and Total Scores in University

- The conversion of the G_SC attribute to a format similar to total scores is a strategic step to standardize metrics, leading to clearer insights and more effective decision-making. This categorizes the data based on the distribution of global scores.
- The competencies from different educational levels are also provided for these reasons: several attributes represent similar subjects at both high school and university levels. Attributes are:

- (a) MAT_S11 (Mathematics in high school) and QR_PRO (Quantitative Reasoning in university)
- (b) ENG_S11 (English in high school) and ENG_PRO (English in university)
- (c) CR_S11 (Critical Reading in high school) and CR_PRO (Critical Reading in university)
- (d) CC_S11 (Citizen Competencies in high school) and CC_PRO (Citizen Competencies in university)
- These pairs of attributes allow for tracking the development of specific competencies from high school to university, providing insights into how students' proficiency in subjects like mathematics and English evolves. This allows for a more comprehensive understanding of student development and can be crucial for clustering and further educational analysis.

Attribute Analysis:

- Professional Records: Analyzing attributes are G_SC, FEC_PRO, and QUARTILE.
- Figure 5.3 displays a heatmap showing the correlation matrix of numerical attributes from academic performance records. Key observations include:
 - (a) Total_SPro, G_SC, and QUARTILE attributes demonstrate nearly equal strong correlations with each other, indicating similar impacts on other attributes.
 - (b) The FEP_PRO attribute suggests minimal impact on other attributes due to lower test scores. Figure 5.5 also proves lower test scores between PERCENTILE, 2ND_DECILE, and FEP_PRO.
- Figure 5.4 is similar to the previous heatmap, correlation coefficients range from -1 to 1, with values closer to 1 indicating a strong positive correlation, values closer to -1 indicating a strong negative correlation, and values near 0 indicating no correlation. Key observations include:
 - (a) Students' percentile and decile ranks are strong indicators of the total score.
 - (b) PERCENTILE and 2ND_DECILE attributes demonstrate nearly equal strong correlations with each other, suggesting they provide similar information.



Figure 5.3 Correlation Matrix for Attributes of Academic Performance

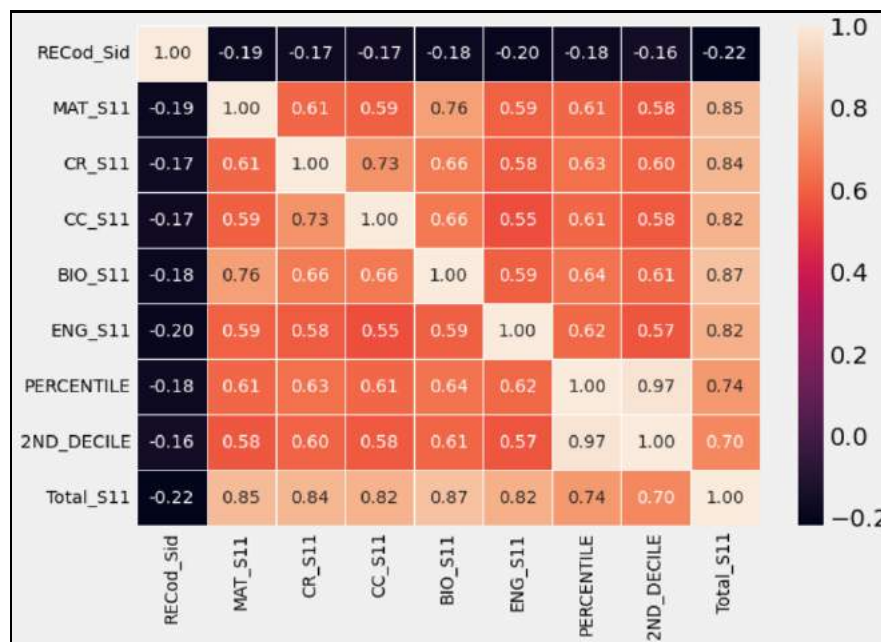


Figure 5.4 Correlation Matrix for Attributes of Secondary Performance

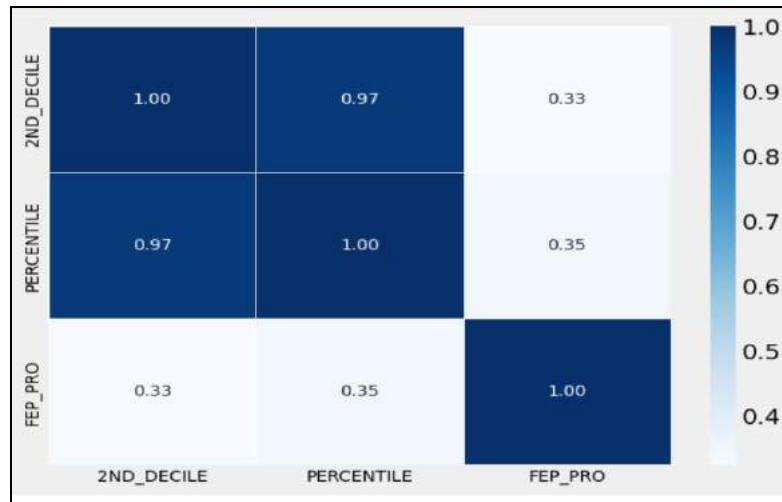


Figure 5.5 Heatmap Visualization for Attributes from Different Records

Categorical to Numerical Conversion:

- Given the need for clustering algorithms like DP-Means, categorical attributes are converted to numerical formats. This ensures that the algorithms function effectively and avoid inaccuracies.
- Secondary Performance Records: Analyzing attributes are SCHOOL_NAT and SCHOOL_TYPE (categorical); competencies, PERCENTILE, and 2nd_DECILE (numerical).
- Figure 5.6 presents the count of each categorical attribute based on their values.

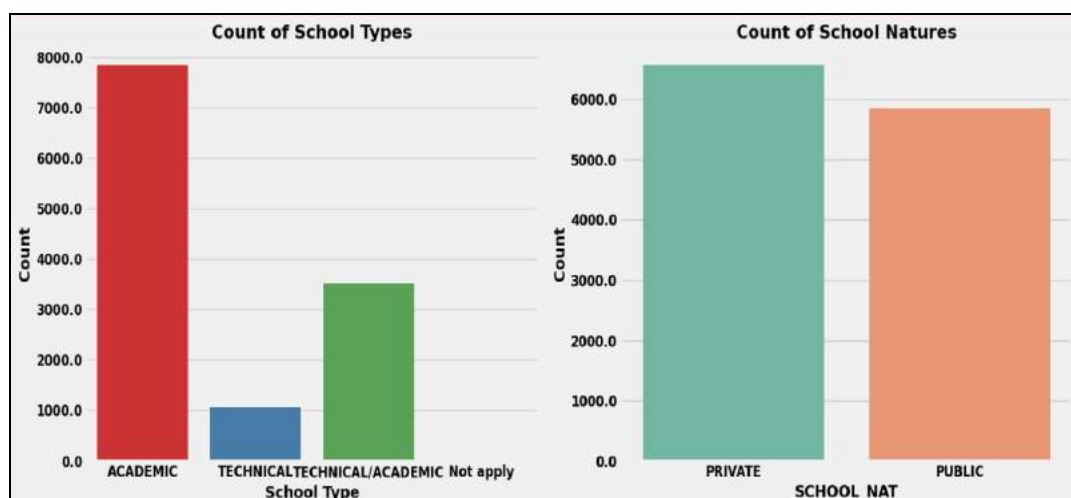


Figure 5.6 Implementation of Categorical Attributes with Values

- From the analysis results recommended from Figure 5.4, using either the PERCENTILE or the 2ND_DECILE attribute for further analysis on SCHOOL_NAT and SCHOOL_TYPE, with 2ND_DECILE potentially useful for more granular analysis. Figure 5.7 provides a comprehensive summary of performance across different school types and programs, highlighting variations in scores within and across private and public institutions. This figure also provides descriptive statistical measures for the 2ND_DECILE attribute grouped by SCHOOL_NAT and SCHOOL_TYPE, showing that these attributes are not strongly related to the representative attribute of high school records.
- Figure 5.8 provides the distribution and summary statistics of Total_S11 by the attributes SCHOOL_NAT and SCHOOL_TYPE, ensuring the analysis process for these two categorical attributes.
- Figure 5.9 generates summary statistics of academic performance QUARTILE grouped by SCHOOL_NAT and SCHOOL_TYPE, ensuring precision and comprehensiveness.

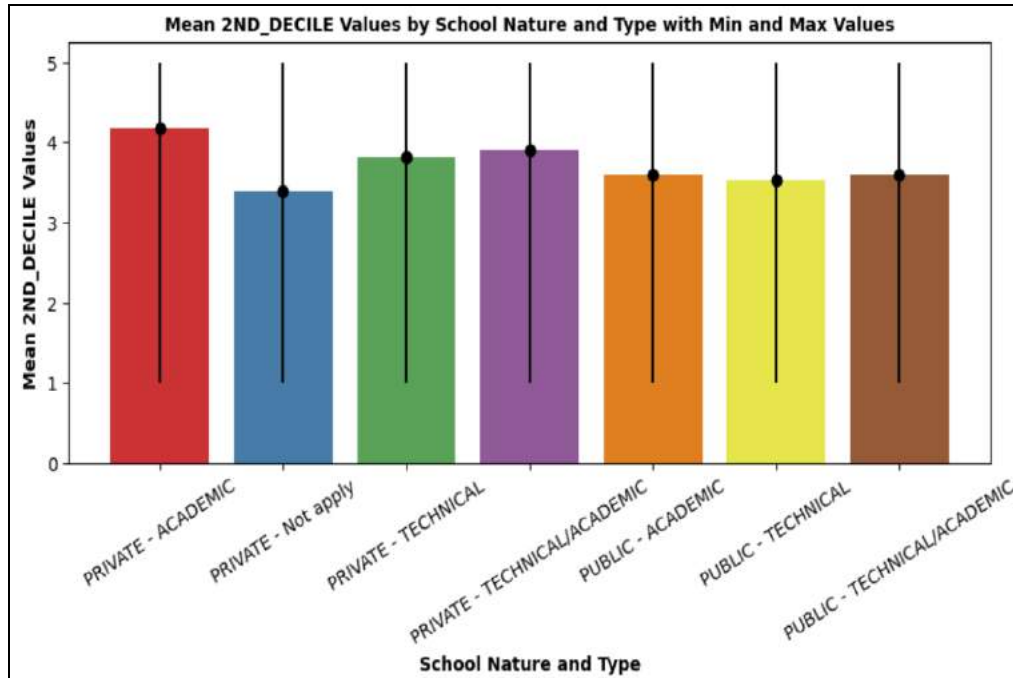


Figure 5.7 Mean (2ND_DECILE) Values by School Nature and Type with Min and Max Values

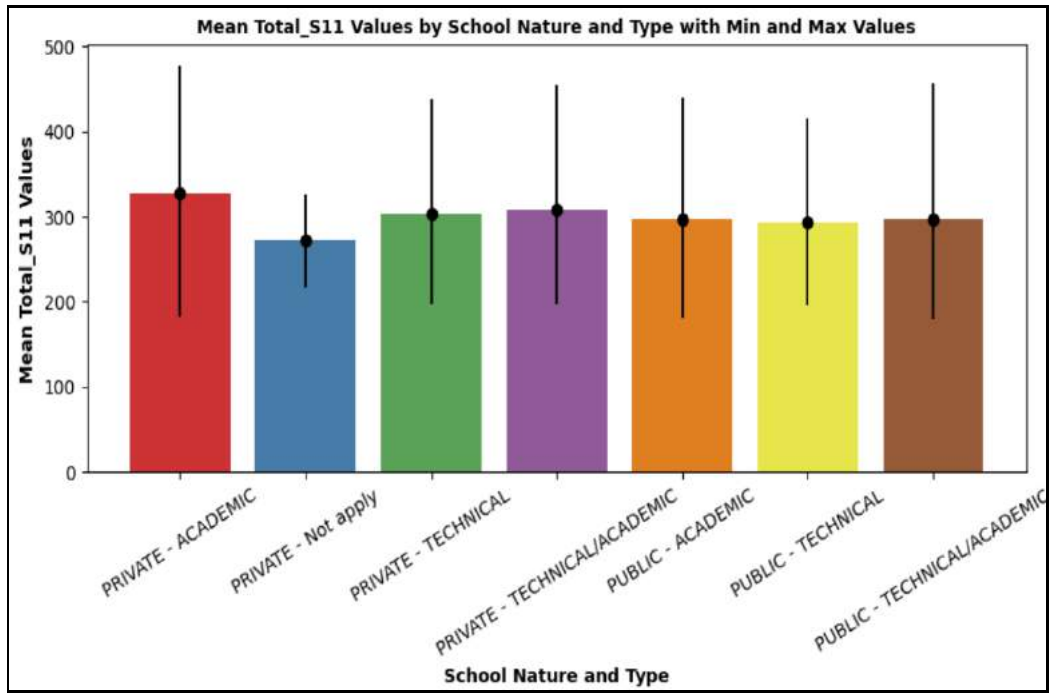


Figure 5.8 Mean (Total_S11) Values by School Nature and Type with Min and Max Values

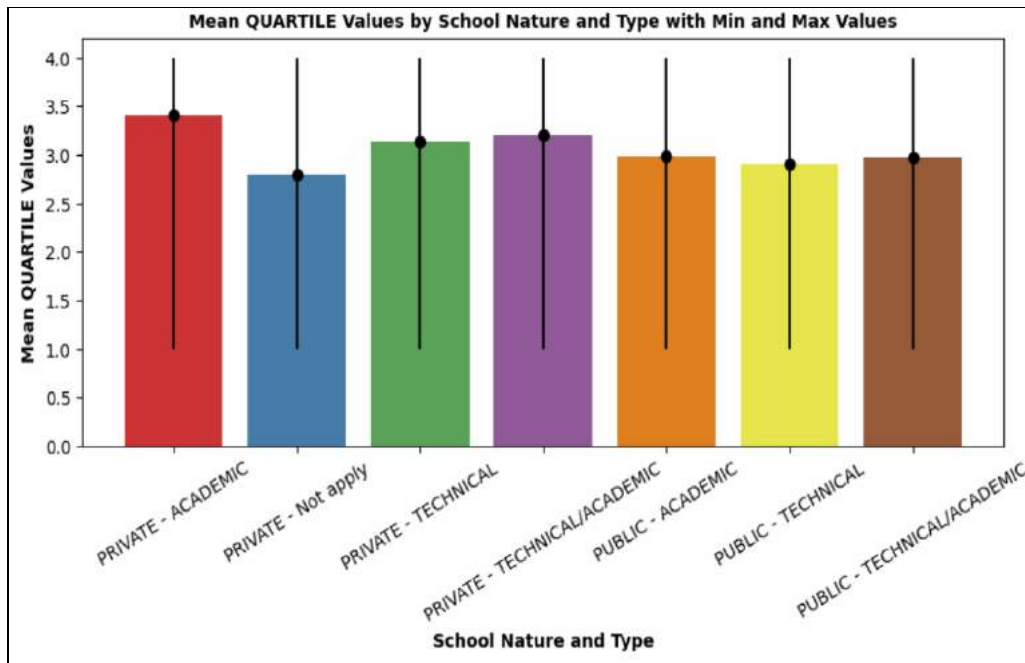


Figure 5.9 Mean (QUARTILE) Values by School Nature and Type with Min and Max Values

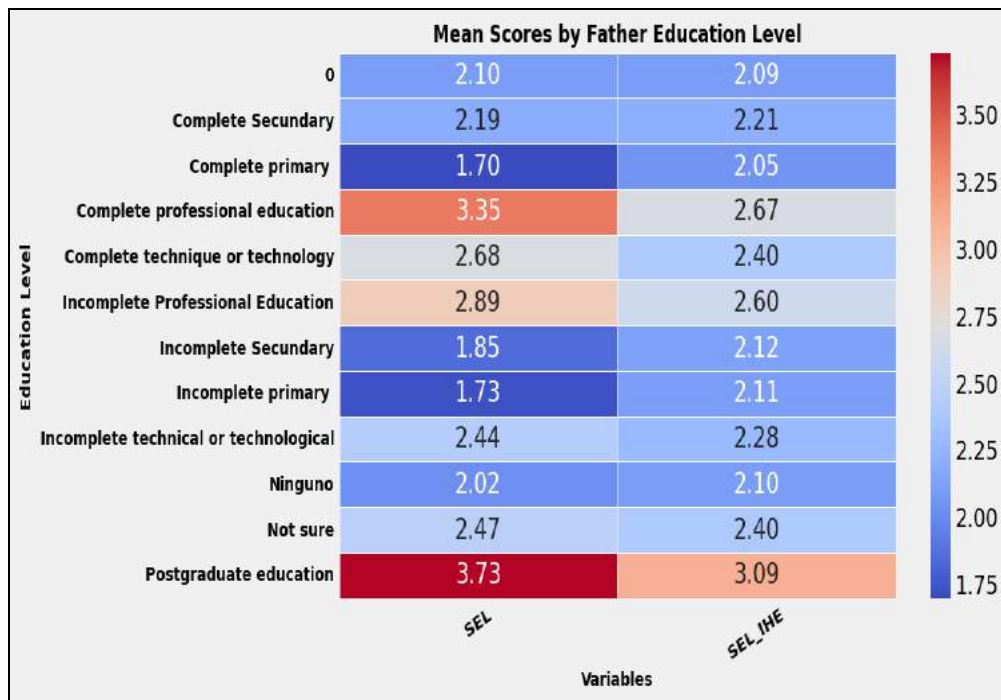


Figure 5.10 Average (SEL, SEL_IHE) Scores by Father Education Level

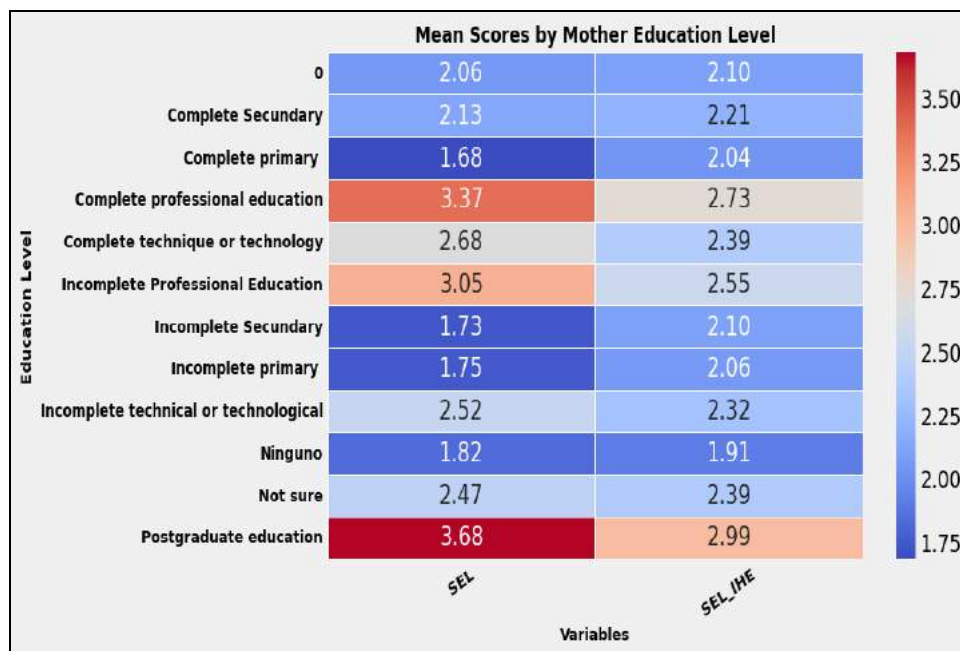


Figure 5.11 Average (SEL, SEL_IHE) Scores by Mother Education Level

- Household Socioeconomic Status: Analyzing attributes are EDU_MOTHER, EDU_FATHER, OCC_FATHER, OCC_MOTHER, INTERNET, TV, COMPUTER, WASHING_MCH, MIC_OVEN, CAR, DVD, FRESH, PHONE, and MOBILE (categorical); SEL and SEL-IHE (numerical).
- Figures 5.10 and 5.11 focus on average scores (SEL and SEL_IHE) across different levels of the father's (EDU_FATHER) and mother's (EDU_MOTHER) education. It was found that SEL and SEL_IHE exhibit nearly equal strong correlations, indicating their comparable significance within the socio-economic context.
- The mean scores of the other attributes such as INTERNET, TV, COMPUTER, WASHING_MCH, MIC_OVEN, CAR, DVD, FRESH, PHONE, and MOBILE, were also examined like EDU_FATHER and EDU_MOTHER.
- Based on the analysis results, the SEL and SEL_IHE attributes have been identified as representative attributes in socioeconomic records. Consequently, these attributes are used as key metrics in further analyses to evaluate the effectiveness of educational interventions and to design targeted strategies aimed at improving academic outcomes. For instance, the GENDER column is converted from categorical to numerical values.

After the processes of the transformation phase of the Edu-ETL applied to three types of records, attributes such as RECod_Sid, QR_PRO, CR_PRO, CC_PRO, WC_PRO, G_SC, QUANTILE, Total_SPro, MAT_S11, CR_S11, CC_S11, BIO_S11, ENG_S11, DECILE (as an attribute renaming of 2ND_DECILE), Total_S11, SEL, SEL_IHE, GENDER_F, GENDER_M were loaded into the database as the final set of attributes for subsequent clustering analysis. In Figure 5.12, it is observed that SEL and SEL_IHE fail to significantly influence or correlate with other attributes. In contrast, among the indicative level identification attributes in each record, QUANTILE, G_SC, and 2ND_DECILE exhibit significant effects and correlations with other attributes. Their impact on the attributes is nearly equivalent.

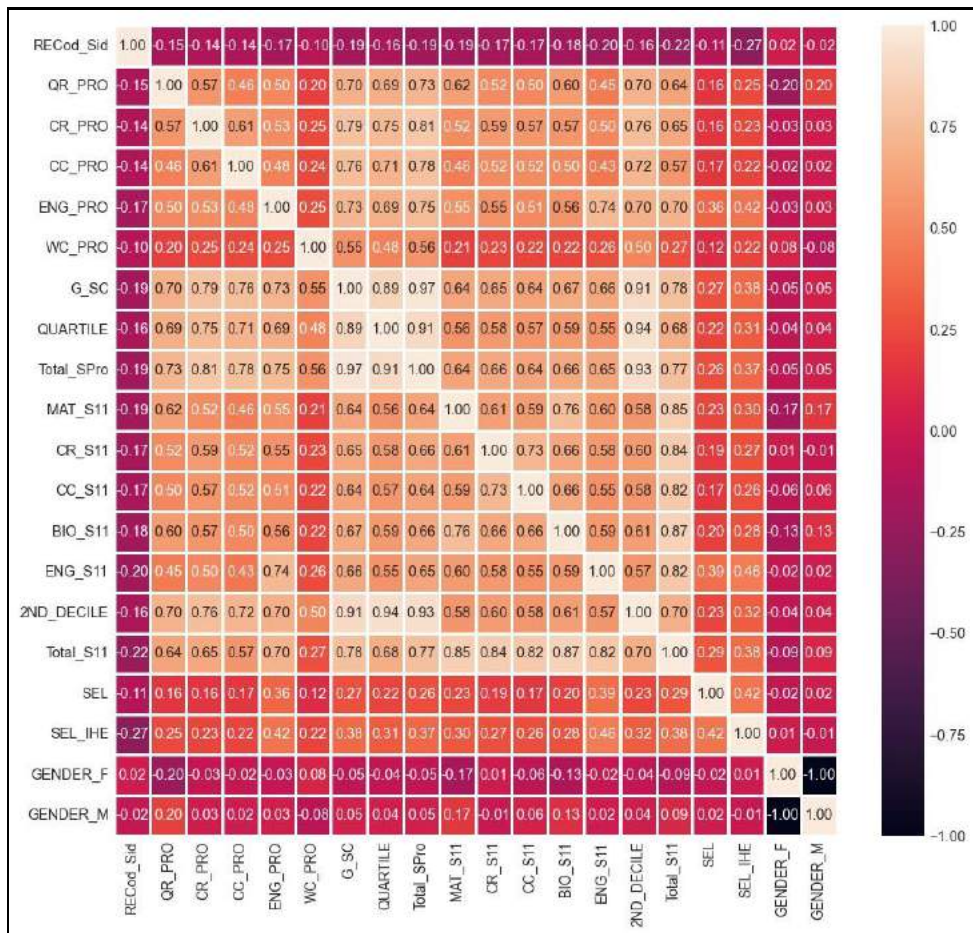


Figure 5.12 Heatmap Analysis Reveals Correlations between Academic (Professional), Secondary Achievements and Household Socioeconomic Status Levels

Table 5.4 (metrics overview by QUARTILE) presents a summary of various academic and socioeconomic metrics distributed across four quartiles. Each quartile represents a different level of achievement. The data suggests a relationship between quartile scores and academic performance metrics from both secondary and professional education. However, socioeconomic indicators show no significant correlation with quartile scores, suggesting that academic success in secondary education is closely linked to professional performance but not to socioeconomic status.

Figure 5.13 provides a visual comparison of the strength and direction of the correlations between the representative attributes in each field. Higher correlation coefficients indicate stronger relationships between attributes, while lower coefficients suggest weaker relationships.

Table 5.4 Summary of Metrics by QUARTILE

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

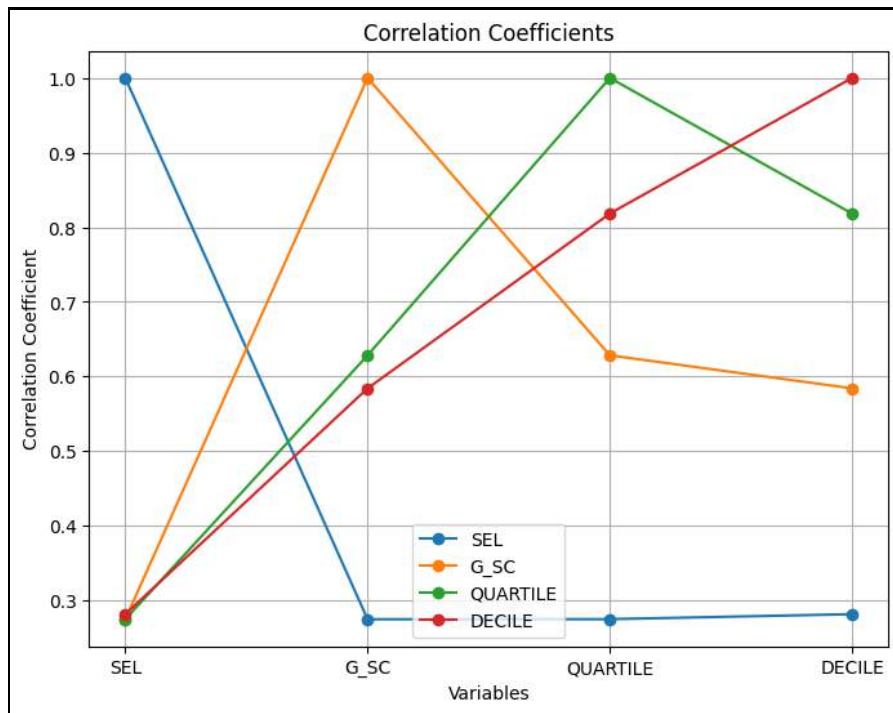


Figure 5.13 Correlation Coefficients between Four Attributes


3. Loading Phase

In the loading phase, the transformed data is loaded into the analytical system for further processing and analysis.

- The refined datasets (total number of students profile records = 12411 records and 19 attributes), consistent and cleaned, are loaded into the system for analysis. This includes integrating various records using the RECod_Sid attribute to ensure seamless information linking.
- In this proposed system, the loading phase uses SQLite, a lightweight disk-based database, to store the cleaned and transformed educational data. This structured approach ensures that the data is ready for further educational analysis, enabling researchers to draw insights and conclusions based on the data.
- Each values represents a single student profile record output after the loading phase of the Edu-ETL processes. The values correspond to specific attributes of a student's academic, secondary and socioeconomic information.

The detailed representation of each record is as follows:

A Single Student Record

- 
- RECod_Sid: 227461
 - QR_PRO: 94
 - CR_PRO: 86
 - CC_PRO: 84
 - WC_PRO: 98
 - G_SC: 2
 - QUARTILE: 4
 - Total_SPro: 4
 - MAT_S11: 68
 - CR_S11: 85
 - CC_S11: 79
 - BIO_S11: 86
 - ENG_S11: 74
 - DECILE: 4
 - Total_S11: 3
 - SEL: 4
 - SEL_IHE: 2
 - GENDER: 1

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

5.2 System Demonstration

The program demonstration experiments the system's functionality and capabilities, offering users a comprehensive view of its features through meticulously tested results. By providing visual aids, it reveals the effectiveness of system presentations, ensuring clarity and engagement.

The program demonstration of the system entails three main distinct stages, each designed to offer a detailed exploration the system's capability.

1. Selecting 'Clustering' and 'Clustering on Spark' in main view
2. Analyzing and demonstrating for 'Edu-ETL Processes', 'K-Means', 'Original DP-Means' and 'Enhanced DP-Means' in 'Clustering' view
3. Analyzing and demonstrating on PySpark environment with 'K-Means', 'Original DP-Means' and 'Enhanced DP-Means' in 'Clustering on Spark'

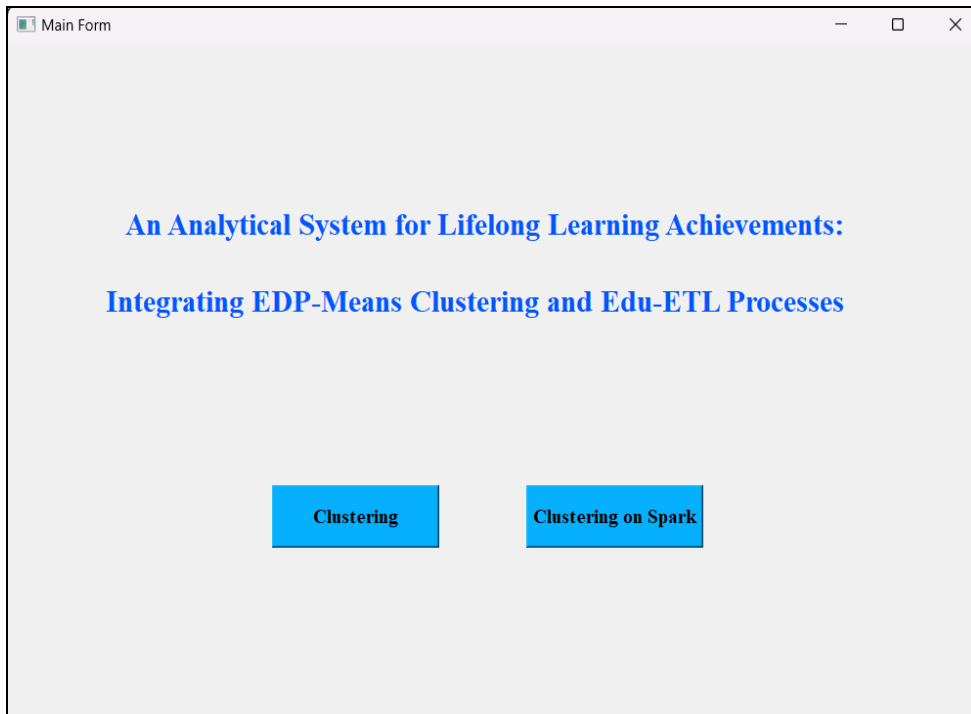


Figure 5.14 Main View of Program Demonstration

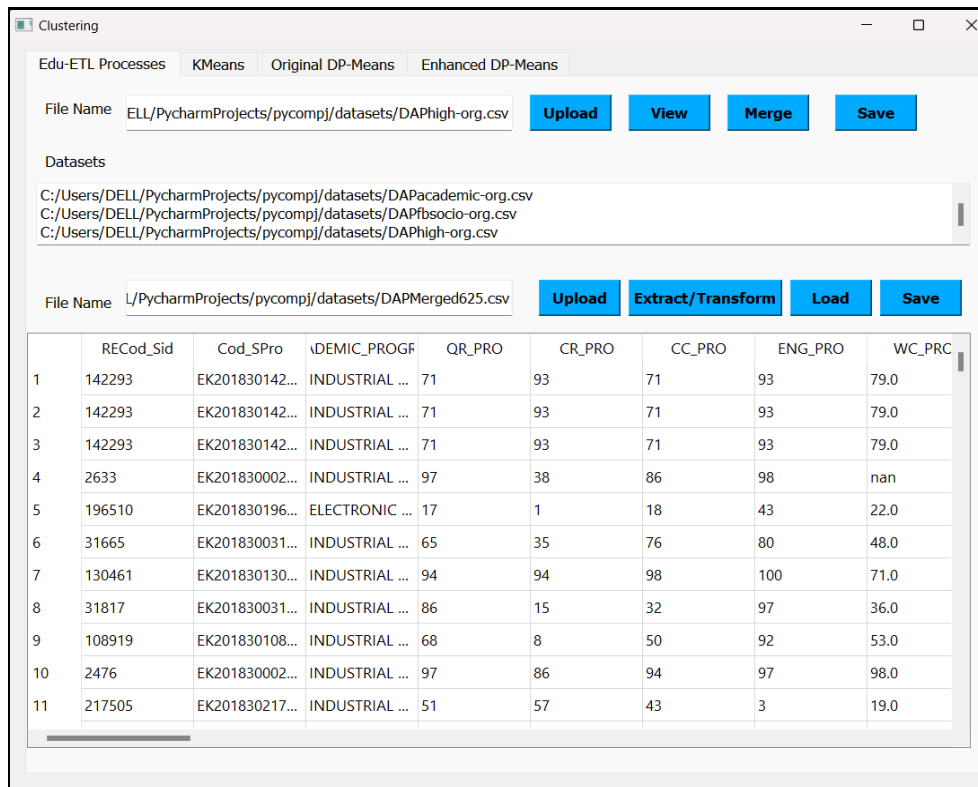


Figure 5.15 “Clustering” View of Program Demonstration

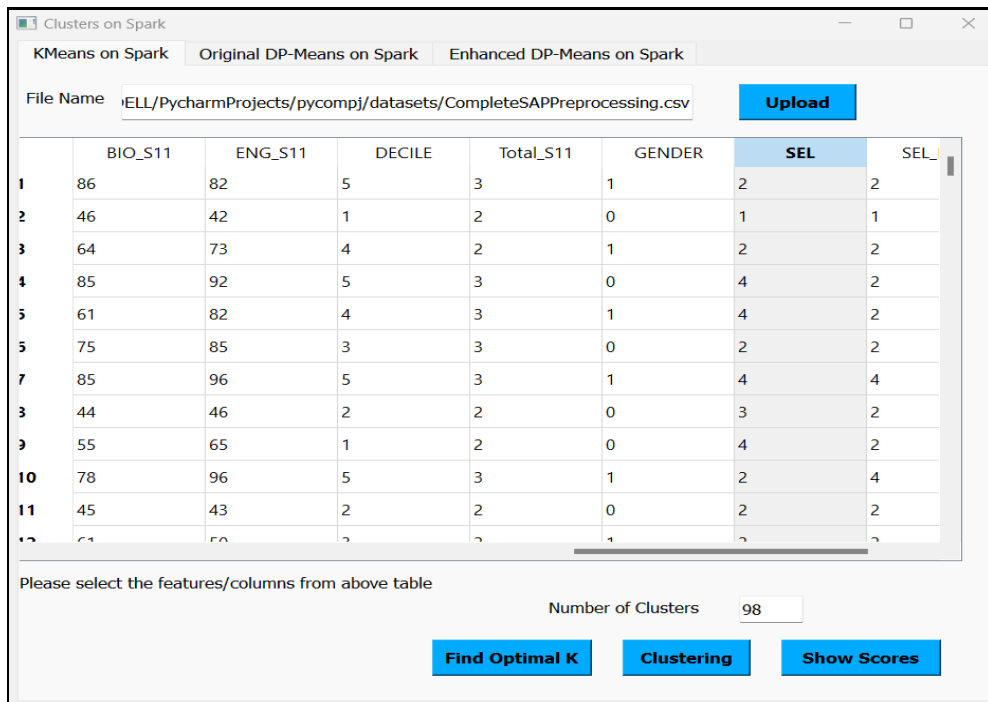


Figure 5.16 “Clustering on Spark” View of Program Demonstration

5.3 Chapter Summary

The chapter details the implementation of Edu-ETL processes to preprocess and transform data from student profiles datasets. This dataset plays a crucial role in implementing the proposed system. However, the system's adaptability allows for testing with diverse educational datasets and data from different domains. These processes are demonstrated using GUI to facilitate clear and concise visual explanations.

The first section begins by describing the dataset used, comprising 12,411 observations, with attributes capturing personal and academic information. Academic assessments at significant points in a student's life, along with personal information such as socioeconomic status and household amenities, are included. The chapter categorizes and describes both numerical and categorical attributes, providing comprehensive insights into the dataset's composition.

In the second section, the Edu-ETL processes streamline data handling by organizing attributes from datasets into three record types: academic performance records, secondary performance records, and household socioeconomic status records. The transformation phase involves preprocessing tasks like handling missing values

and transforming attributes, ensuring compatibility with clustering algorithms.

The final section concludes with a program demonstration, experimenting the system's functionality and capabilities through three distinct stages. Users can explore features such as clustering algorithms (K-Means, DP-Means, EDP-Means) and Edu-ETL processes within a user-friendly GUI. The demonstration provides visual aids to enhance understanding and engagement, ensuring clarity in system presentation and evaluation.

Overall, this chapter underscores the importance of integrating analytical techniques and data preprocessing methods to effectively analyze lifelong learning achievements. The Edu-ETL processes play a pivotal role in preparing data for clustering analysis, while the program demonstration offers users a hands-on experience with the system's capabilities.

CHAPTER 6

EXPERIMENTAL RESULTS AND EVALUATIONS

This chapter provides a comprehensive evaluation of the proposed analytical system for lifelong learning achievements, focusing on the performance of the EDP-Means clustering algorithm. Building on the detailed analysis of Edu-ETL processes presented earlier, this chapter delves into the effectiveness and efficiency of the EDP-Means clustering algorithm.

The evaluation begins with an investigation into the performance of the EDP-Means algorithm using various datasets from different fields and sizes. By comparing cluster quality and accuracy against the K-Means and original DP-Means algorithms, the analysis demonstrates that EDP-Means yields superior clustering outcomes in terms of precision and reliability. The investigation further extends to a distributed computing environment using PySpark, highlighting the scalability and efficiency of the clustering processes when applied to diverse datasets. This extension illustrates the practical applicability of the approach in real-world scenarios, particularly through the effective handling of large datasets.

A thorough validation of the clustered results is conducted for both standalone and PySpark implementations. Multiple validation metrics, including the Silhouette Score, CH index, and DB index, are employed to assess cluster quality. Additionally, processing times of the different algorithms are compared to highlight their efficiency, thereby substantiating the robustness and performance advantages of the EDP-Means algorithm.

Following this, the chapter examines the performance of the EDP-Means clustering algorithm using educational preprocessed data from the Edu-ETL processes. As with the earlier datasets, a comprehensive validation of the clustered results is performed using the same set of validation metrics. The comparison of processing times for different algorithms continues to underscore their efficiency, reinforcing the robustness and performance benefits of the EDP-Means algorithm.

The final part of the analysis focuses on learning outcomes and identifies key success factors influencing lifelong learning achievements. By examining the clustered data, the system aims to uncover patterns and insights that can inform tailored interventions and support strategies. The ultimate goal is to enhance educational outcomes by leveraging the analytical capabilities of the system to

pinpoint critical success factors.

The experiments are designed to rigorously test the proposed system under various conditions, ensuring a fair and comprehensive evaluation. Diverse datasets are utilized to reflect a practical and manageable scope for validating the improvements. The clustering processes are implemented and tested in both standalone and distributed computing environments to demonstrate the flexibility and scalability of the system. The chosen validation metrics provide a multi-faceted view of cluster quality, ensuring thorough and reliable assessments. The following sections present detailed experimental results, analysis, and insights, highlighting the efficacy of the EDP-Means clustering algorithm and the overall analytical system for lifelong learning achievements.

6.1 Experimental Setup and Procedure

This section focuses on evaluating the performance of the proposed EDP-Means clustering algorithm. The EDP-Means algorithm enhances the original DP-Means by dynamically updating the threshold parameter λ based on the silhouette score, thus improving clustering accuracy. Additionally, EDP-Means determines the optimal number of clusters through iterative fitting, enhancing precision and stability. This evaluation is crucial as it demonstrates the efficacy of EDP-Means in providing high-quality clustering results, which is fundamental for accurately analyzing lifelong achievements.

6.1.1 Datasets Selection and Preparation

For the experiments, various datasets from different fields were selected, ensuring a mix of small, medium, and large datasets from different fields to comprehensively evaluate the algorithm's performance. The datasets include both numerical and categorical data to reflect real-world complexity. Each dataset underwent preprocessing steps such as normalization, encoding of categorical attributes, and handling of missing values to ensure consistency and comparability. The preprocessing steps in the system are divided into two distinct approaches based on whether the data is merged or non-merged. For merged data (Edu-ETL process for educational datasets), preprocessing is more elaborate and involves merging different datasets. The non-merged preprocessing pipeline is robust for most clustering

algorithms. Table 6.1 provides an overview of the three datasets used in the analysis summarizing their domain, size, and data types. A dataset size is generally considered small if typically, fewer than 1,000 records. If ranges from 1,000 to 100,000 records, this dataset size is medium. For large dataset sizes, usually more than 100,000 records, potentially into the millions or billions.

Table 6.1 Datasets in Different Sizes and Domains

Dataset	Domain	Records	Dimensions	Data Size	Data Types
Diabetes	Healthcare	769	9	Small	Numerical
Universal Bank	Finance	5000	14	Small	Numerical
Spotify Popular Music	Music	114000	20	Large	Numerical, Categorical

6.1.2 Experimental Methodology

Execution Environment: The algorithms were implemented using Python, leveraging libraries such as scikit-learn for K-Means and custom implementations for DP-Means and EDP-Means. PySpark was used to handle large-scale datasets and perform distributed computations efficiently. PySpark, an interface for Apache Spark in Python, was integrated into the development environment using PyCharm, providing a robust setup for scalable data processing. The experiments were executed on a system with a multi-core processor and ample memory to handle large-scale computations.

Proposed EDP-Means Clustering Processes: The primary focus of this experiment is to assess the performance of the EDP-Means clustering algorithm compared to the original DP-Means and K-Means algorithms. The evaluation is based on cluster quality, measured by metrics such as SSE and the silhouette score, as well as computational efficiency.

1. Optimal Number of Clusters (K^* Determination)

The EDP-Means algorithm computes the SSE for each cluster: $SSE_k \leftarrow$

$\sum_{i=1}^N \|x_i - \mu_{z_i}\|_2^2$. This metric evaluates the clustering quality by measuring the compactness of the clusters. By minimizing SSE, the algorithm identifies the optimal number of clusters K^* : $K^* \leftarrow \arg \min_k (sse_k)$. Minimizing SSE ensures well-defined clusters with minimal intra-cluster variance, enhancing clustering accuracy.

2. Dynamic Adjustment of Threshold Parameter (λ)

EDP-Means explores a range of λ values between λ_{min} and λ_{max} . This dynamic adjustment allows the algorithm to adapt the threshold parameter based on data characteristics. The optimal λ value is determined by maximizing the silhouette score: $Silhouette(\lambda_j) = \frac{1}{N} \sum_{i=1}^N \frac{b_i - a_i}{\max(a_i, b_i)}$ where a_i is the average distance between a data point and other points in the same cluster, b_i and is the average distance to points in the nearest neighboring cluster. The optimal λ value λ^* is selected to maximize the silhouette score: $\lambda^* \leftarrow \arg \max_{\lambda_j} Silhouette(\lambda_j)$. This optimization ensures compact and well-separated clusters, improving clustering quality.

3. Iterative Optimization for Enhanced Accuracy and Stability

During the clustering process, if a data point deviates significantly from its assigned cluster center (beyond the threshold λ^*), a new cluster is created with the outlier as its center. This step ensures adaptive clusters and effective handling of outliers. The algorithm continuously updates the cluster centers based on current assignments, recalculating the mean for each cluster. This iterative process continues until cluster assignments stabilize or change insignificantly, ensuring convergence to an optimal solution.

6.2 Result and Analysis of the Proposed Analytical System

This section presents a thorough comparison of the clustering performance among three prominent algorithms: K-Means, DP-Means, and EDP-Means. Evaluation criteria encompass three cluster validation indices: Silhouette Score, DB index, and CH index. Furthermore, the impact of threshold values on DP-Means and EDP-Means algorithms' performance is scrutinized.

The experimental analyses are conducted across diverse datasets to ensure comprehensive assessment. Following this, the performance of EDP-Means is meticulously analyzed. Subsequently, an analytical system is proposed for lifelong learning achievements, leveraging the integration of EDP-Means and Edu-ETL processes. Educational datasets are employed to experiment with the proposed

system, and the resulting outcomes are meticulously analyzed to elucidate the success factors for lifelong learning achievements.

6.2.1 Experimental Evaluation of the EDP-Means Clustering Algorithm

Before applying the clustering algorithm, all datasets used in the analysis undergo a comprehensive preprocessing procedure to ensure data quality and consistency. The preprocessing steps for non-merged data is used. These steps include:

1. Checking for initial missing values and handling them using the “SimpleImputer” with mean strategy, followed by the removal of duplicate entries
2. Encoding categorical data using the “OneHotEncoder”, then dropping the original categorical columns and appending the newly encoded columns to the dataset
3. Scaling the features to standardize the data
4. Verifying the presence of any NaN values post-scaling.

This systematic approach to data preprocessing is crucial for enhancing the accuracy and reliability of the clustering results. Table 6.2 presents the clustering performance metrics for K-Means, DP-Means, and EDP-Means algorithms across the datasets in Table 6.1. The performance is evaluated using three key metrics.

Table 6.2 Cluster Quality Scores for K-Means, DP-Means, EDP-Means

Dataset	Algorithm	Silhouette Score	CH Score	DB Score
Diabetes	K-Means	0.36	636.07	0.75
Diabetes	DP-Means	0.37	568.60	0.70
Diabetes	EDP-Means	0.37	617.57	0.74
Universal Bank	K-Means	0.51	4303.97	0.69
Universal Bank	DP-Means	0.51	3514.91	0.72

Dataset	Algorithm	Silhouette Score	CH Score	DB Score
Universal Bank	EDP-Means	0.54	5075.18	0.66
Spotify Popular Music	K-Means	0.35	96654.36	0.78
Spotify Popular Music	DP-Means	0.35	96604.82	0.78
Spotify Popular Music	EDP-Means	0.35	96374.03	0.79

Performance Metrics Analysis for Experimental Results: The EDP-Means Algorithm Outperforms and is Comparable to Traditional Approaches Across Diverse Domains

The performance metrics in Table 6.2 suggest that the EDP-Means algorithm generally outperforms or matches the traditional K-Means and DP-Means algorithms across different datasets and domains. EDP-Means consistently achieves better or comparable Silhouette and DB Scores, indicating higher cluster quality and robustness. The significant improvement in CH Scores for certain datasets highlights EDP-Means’s ability to form more compact and well-separated clusters. These findings validate the effectiveness of the proposed EDP-Means algorithm in handling diverse and complex datasets, making it a superior choice for clustering tasks in varied domains, including education. Applying the Edu-ETL process followed by EDP-Means clustering to the educational dataset can uncover meaningful insights into learning patterns and success factors.

Figures 6.1, 6.2, and 6.3 illustrate the comparative performance of the K-Means, DP-Means, and EDP-Means algorithms across different datasets, evaluated using Silhouette Score, CH Score, and DB Score, respectively. These figures highlight the clustering quality, compactness, and separation capabilities of each algorithm, providing a comprehensive overview of their effectiveness in handling diverse datasets.

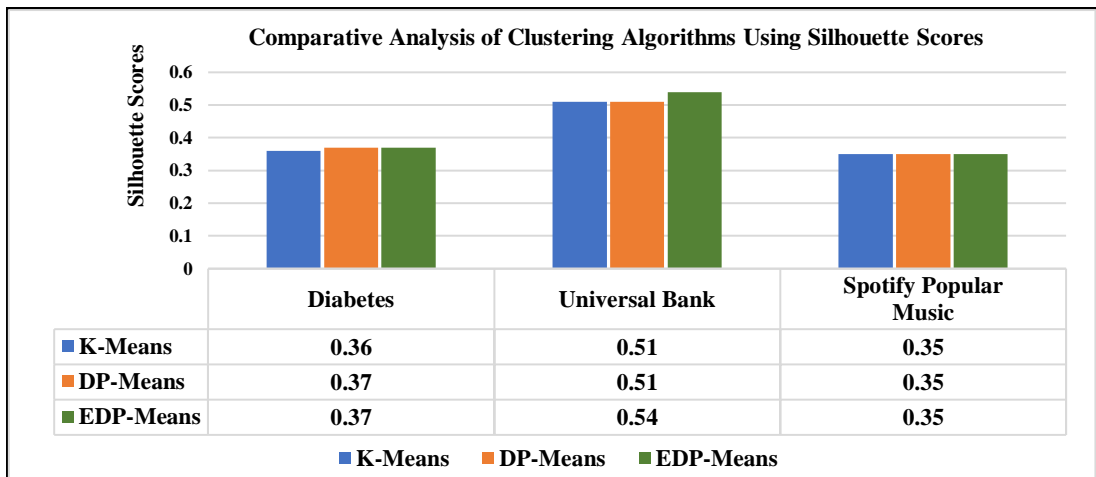


Figure 6.1 Comparative Analysis of Clustering Algorithms Using Silhouette Scores Across Different Datasets

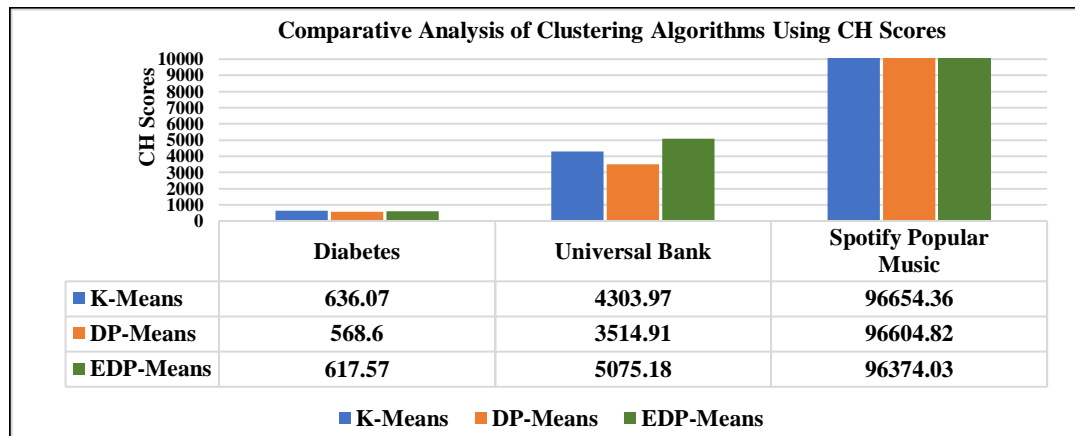


Figure 6.2 Comparative Analysis of Clustering Algorithms Using CH Scores Across Different Datasets

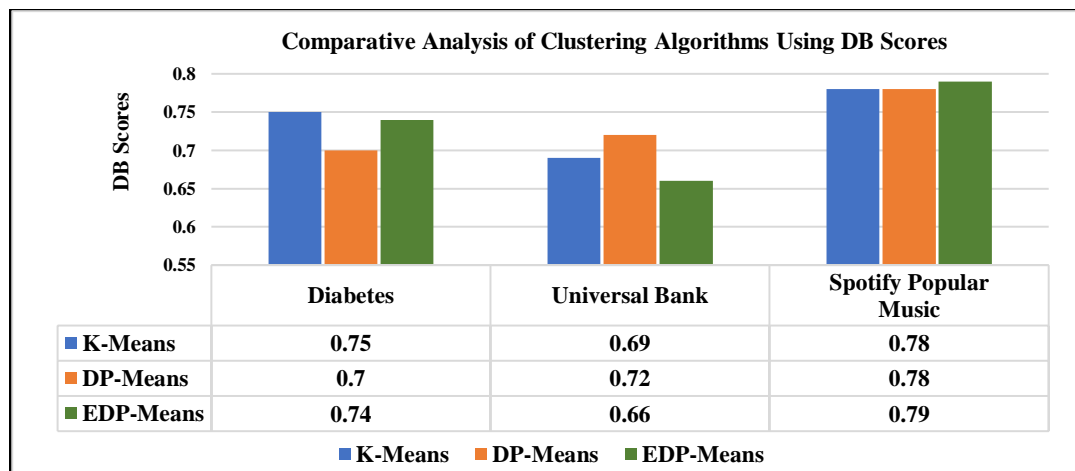


Figure 6.3 Comparative Analysis of Clustering Algorithms Using DB Scores Across Different Datasets

Processing Time Analysis for Experimental Results in Three Clustering Algorithms Across Diverse Domains

Figure 6.4 illustrates a processing time comparison of the K-Means, DP-Means, and EDP-Means clustering algorithms and the y-axis represents the processing time in milliseconds (ms). At the same time, the x-axis lists the different datasets.

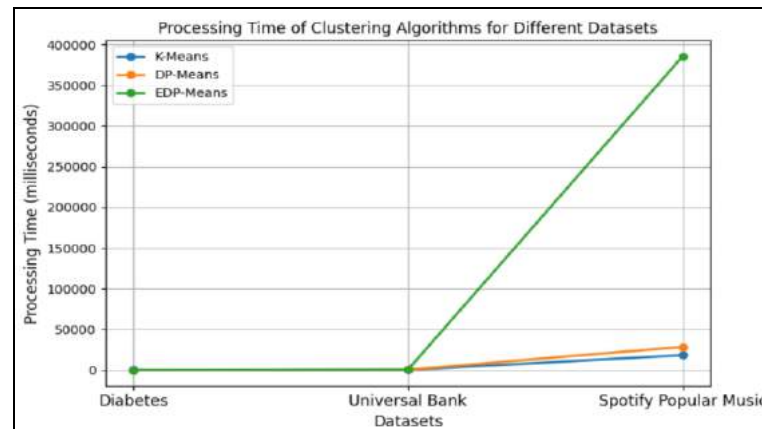


Figure 6.4 Comprehensive Processing Time (ms) for Clustering, Optimal Cluster Numbers and Threshold Parameters Finding in K-Means, DP-Means, and EDP-Means

Analysis Results in “Diabetes and Universal back” Datasets: for smaller and less complex datasets, all three algorithms exhibit efficient processing times with no significant differences, indicating that these datasets do not impose a heavy computational load on any of the algorithms.

Analysis Results in “Spotify Popular Music” Dataset: this dataset is much larger and more complex compared to the other dataset. EDP-Means, while providing enhanced clustering capabilities, incurs a substantial computational cost for this dataset. This indicates that the dynamic threshold adjustment and iterative fitting processes in EDP-Means significantly increase processing time when handling very large datasets.

Analysis Results in Algorithms Suitability: K-Means and DP-Means offer consistent and low processing times across all datasets, making them more suitable for scenarios where processing time is critical. EDP-Means should be reserved for cases where clustering quality is paramount and computational resources are ample.

This figure emphasizes the importance of balancing clustering accuracy with processing efficiency, particularly when dealing with large-scale datasets.

Analysis of Experimental Results with Threshold Parameters between DP-Means and EDP-Means Clustering Algorithms

As a result, as mentioned in Table 6.2, EDP-Means show significantly higher CH Scores in the Universal Bank, suggesting more compact and well-separated clusters. Therefore, Figure 6.5 plot shows the CH Scores for DP-Means and EDP-Means algorithms along with their respective λ and λ^* values across four datasets. Based on the Figure 6.5:

- **λ value Sensitivity:** The impact of λ value on clustering quality across datasets. In some cases, tuning λ significantly improves clustering, whereas in others, the impact is minimal.
- **Algorithm Performance:** EDP-Means tends to slightly outperform or match DP-Means across datasets, which could be attributed to its ability to adaptively find the optimal λ^* .
- **Dataset Characteristics:** The structure and nature of the dataset greatly influence how sensitive the clustering quality is to λ . For large and complex datasets like Spotify Popular Music, both algorithms seem to stabilize at a similar λ value, indicating a possible inherent structure that both algorithms coverage upon.

In Figure 6.5, the plot demonstrates the importance of parameter tuning in clustering algorithms and highlights how adaptive methods like EDP-Means can sometimes offer marginal improvements over fixed-parameter methods like DP-Means.

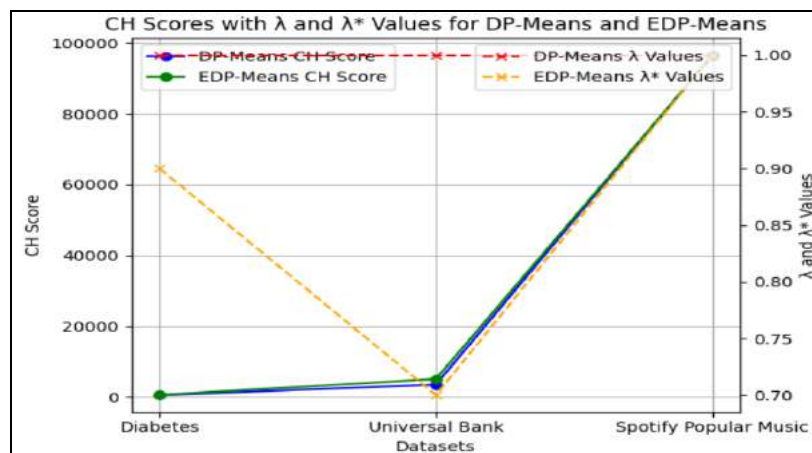


Figure 6.5 CH Scores with λ and λ^* Values for DP-Means and EDP-Means

Experimental Results and analysis of K-Means, DP-Means and EDP-Means in PySpark Environment

Table 6.3 presents the Silhouette Scores of three clustering algorithms applied to three different datasets in a PySpark environment.

The Silhouette score analysis highlights that K-Means tends to form more distinct and well-defined clusters compared to DP-Means and EDP-Means in the PySpark environment. While DP-Means and EDP-Means show similar performance, they generally produce lower silhouette scores compared to K-Means, indicating less distinct clusters. The performance differences observed in the Silhouette Scores can indeed be attributed, at least in part, to the limitations of DP-Means in a parallel computing environment like PySpark.

- **Strengths in Parallel Computing:** PySpark is particularly well-suited for running K-Means, as it can leverage the distributed computing capabilities effectively, resulting in better performance and higher Silhouette Scores.
- **Challenges in Parallel Computing:** DP-Means adds new clusters dynamically based on a λ , making it harder to parallelize. The process depends on the current cluster configuration, requiring more synchronization and communication between processors. These dependencies can cause inefficiencies in distributed environments like PySpark, resulting in lower Silhouette Scores compared to K-Means.
- **Potential Improvements:** The performance and Silhouette Scores of EDP-Means in PySpark can be slightly better than DP-Means but may still lag behind K-Means due to inherent complexities in the clustering mechanism. However, for the “Universal Bank” dataset, which consists of 5000 records and is relatively small in size, the Silhouette Scores of EDP-Means in PySpark (0.68) surpass those of DP-Means (0.60) and slightly exceed those of K-Means (0.66). This indicates that EDP-Means is highly effective for smaller datasets, even with its inherent complexities in the clustering mechanism.

This underscores the importance of considering the computational environment when selecting clustering algorithms for large-scale data processing.

Table 6.3 Silhouette Scores Comparison for K-Means, DP-Means and EDP-Means in PySpark Environment

Dataset	Algorithm	Silhouette Score
Diabetes	K-Means	0.57
Diabetes	DP-Means	0.50
Diabetes	EDP-Means	0.52
Universal Bank	K-Means	0.66
Universal Bank	DP-Means	0.60
Universal Bank	EDP-Means	0.68
Spotify Popular Music	K-Means	0.50
Spotify Popular Music	DP-Means	0.35
Spotify Popular Music	EDP-Means	0.35

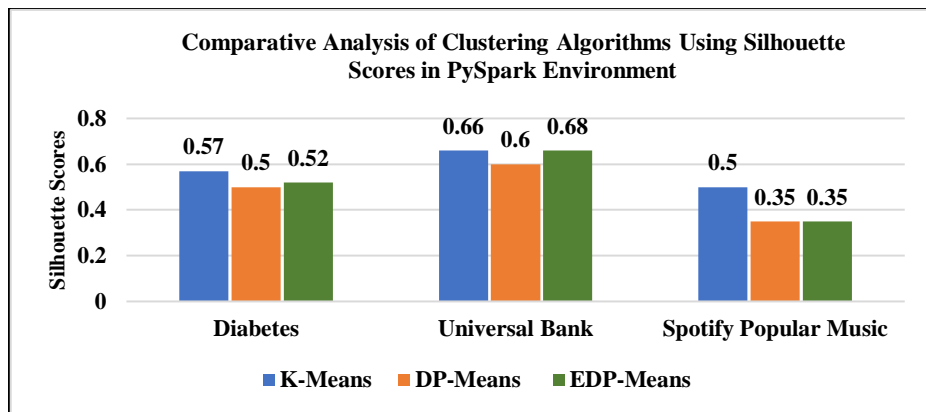


Figure 6.6 Comparative Analysis of Clustering Algorithms Using Silhouette Scores Across Different Datasets in PySpark Environment

Processing Time Analysis for Experimental Results Using Three Clustering Algorithms Across Diverse Domains in PySpark Environment

Figure 6.7 compares processing time between K-Means, DP-Means, and EDP-Means in the PySpark environment across different datasets. Based on the processing time comparison of Figures 6.4 and 6.7, PySpark excels with large-scale datasets and complex distributed computations. However, for smaller datasets or improperly configured clusters, the overhead associated with distributed computing can result in

longer processing times.

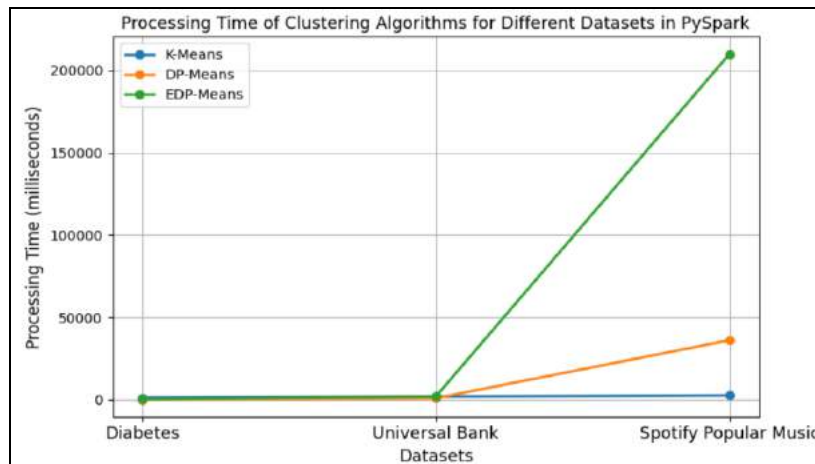


Figure 6.7 Comprehensive Processing Time (ms) for Clustering, Optimal Cluster Numbers and Threshold Parameters Finding in K-Means, DP-Means, and EDP-Means (PySpark Environment)

The difference in processing time between using PySpark and not using PySpark can be attributed to several key factors related to how data processing is handled. PySpark is optimized for large-scale data processing. For smaller datasets, the overhead of distributing tasks across a cluster, managing executors, and maintaining fault tolerance might outweigh the benefits of parallel processing. The significant education in processing time when using PySpark for the Spotify Popular Music dataset highlights the advantages of parallel computing for large-scale data processing. PySpark’s ability to distribute and parallelize tasks results in more efficient data processing, leading to faster clustering times than a non-distributed environment.

Cluster Visualization Results Analysis of EDP-Means Clustering Algorithm

Cluster visualization is a crucial aspect of understanding the performance and effectiveness of clustering algorithms. By examining how data points are grouped, the compactness and separation of clusters, and the positioning of centroids, significant insights can be drawn into the efficiency of the clustering methods. This section focuses on comparing the cluster visualization results of the EDP-Means algorithm. Key aspects of cluster visualization are: Cluster Distribution, Cluster Separation, Centroid Positioning and Data Density and Outliers.

Figure 6.8 and 6.9 represent a scatter plot of the “Diabetes” and “Spotify Popular Music” datasets, clustered using the EDP-Means. Each data point is color-coded according to its assigned cluster, with different colors representing different

clusters. The centroids of the clusters are marked with larger, black “X” symbols.

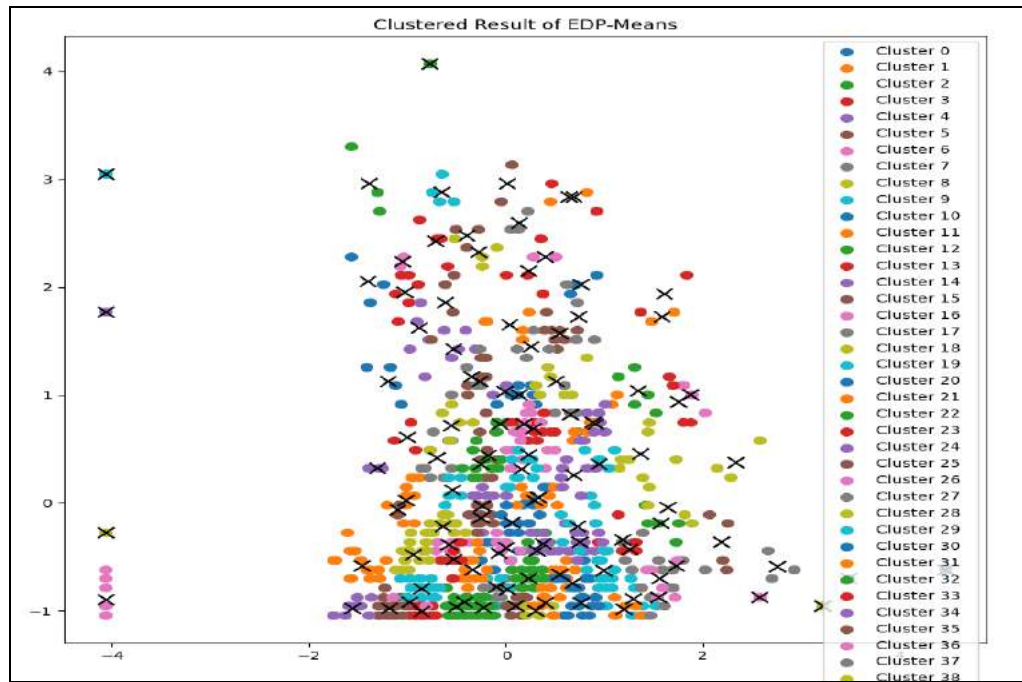


Figure 6.8 Clustering Result Visualization Using EDP-Means Algorithm for the “Diabetes” Dataset (number of attributes = 3, $\lambda^* = 0.9$)

Cluster Distribution: Figure 6.8 shows a wide distribution of clusters and λ^* value is 0.9 as optimal, with each cluster containing a varying number of data points. Some clusters are densely populated, while others are more sparse. In Figure 6.9, the plot shows a dense and uniform distribution of clusters, with each cluster containing a considerable number of data points at λ^* value is 1.0. Clusters are well-defined and cover the entire space, indicating effective partitioning of the dataset by the EDP-Means algorithm.

Cluster Separation: In Figures 6.8 and 6.9, most clusters are well-separated, indicating that the EDP-Means algorithm successfully identifies distinct groups within the dataset. However, in Figure 6.9, the separation between clusters is more pronounced compared to the “Diabetes”, demonstrating the algorithm’s ability to handle more complex and varied data structures.

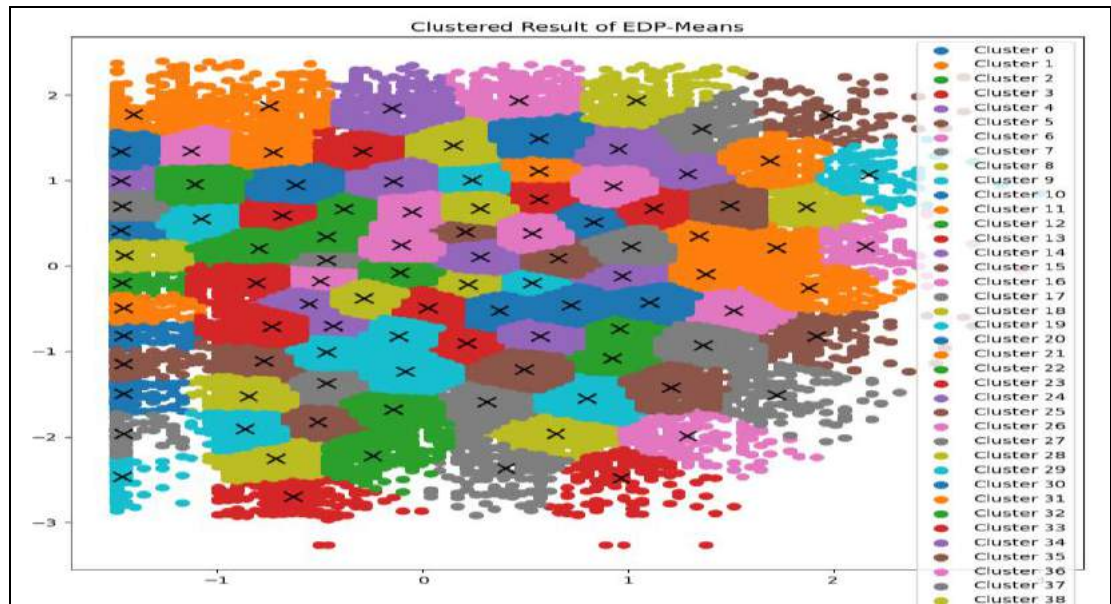


Figure 6.9 Clustering Result Visualization Using EDP-Means Algorithm for the “Spotify Popular Music” Dataset (number of attributes = 2, $\lambda^* = 1.0$)

Centroid Locations: In Figures 6.8 and 6.9, the centroids are strategically located to represent the central points of their respective clusters. The centroids for densely populated clusters are centrally located within the cluster, while those for sparser clusters are placed to best represent the spread of the data points.

Data Density and Outliers: In Figure 6.8, there are a few points (around the coordinates (-4,3) and (3,4)) that are significantly distant from the other clusters. These points are likely outliers or anomalies within the dataset. The EDP-Means algorithm has assigned these points to clusters that are nearest in terms of the distance metric used, but their distant locations suggest they might not fit well into any cluster. However, in Figure 6.9, the plot indicates high data density, especially in the central regions of the clusters. This reflects the large volume and complexity of the “Spotify Popular Music” dataset. There are fewer outliers compared to the “Diabetes” dataset, implying that the data points in this dataset are more uniformly distributed across clusters.

Overall, the visual comparison underscores the advantages of EDP-Means, particularly in terms of cluster quality, separation, and handling of outliers. These improvements make EDP-Means a more effective choice for clustering educational data and analyzing lifelong learning achievements, providing deeper insights into learner behaviors and success factors.

Insights from EDP-Means Clustering Performance Analysis

After a thorough performance analysis using datasets of varying sizes, EDP-Means demonstrates its suitability for educational data clustering. The algorithm excels in handling outliers and noise, differentiating true anomalies from significant data points, which is crucial for identifying unique learning patterns. The dynamic threshold adjustment results in clearer boundaries between learner groups, enhancing the ability to distinguish between different learning trajectories and success factors. These strengths make EDP-Means a valuable algorithm for clustering educational data, offering comprehensive insights into learner behaviors and achievements.

6.2.2 Experimental Evaluation of the EDP-Means Clustering Algorithm with Edu-ETL Processes

This section focuses on evaluating the performance of EDP-Means clustering in conjunction with the Edu-ETL processes described in Chapter 5. The dataset used comprises 12,411 observations, each representing a student with 44 attributes capturing personal information (categorical) and academic assessment results (numerical). Following the Edu-ETL processes, as detailed in Table 3.11, the dataset of the educational domain has been meticulously prepared to ensure it is ready for clustering analysis.

Attributes selected for clustering: QR_PRO, CR_PRO, CC_PRO, ENG_PRO, WC_PRO, Total_SPro, G_SC, MAT_S11, CR_S11, CC_S11, BIO_S11, ENG_S11, Total_S11, DECILE, SEL, SEL_IHE, QUARTILE and GENDER. These attributes were chosen through the analysis processes of Edu-ETL, which are based on their relevance to academic performance and their representation of different dimensions of student profiles. The goal is to analyze the clustered results to uncover patterns and identify key success factors influencing lifelong learning achievements. The key attributes that are focused such as WC_PRO, DECILE and SEL, BIO_PRO, QUARTIL and SEL, MAT_S11 and QR_PRO, ENG_S11 and ENG_PRO, CC_S11 and CC_PRO, CR_S11 and CR_PRO as they provide a comprehensive overview of a lifelong learning achievements analysis and specific strengths with other attributes. These attributes were critical in defining the clusters and distinguishing between different student groups.

Although the five competencies form the foundation for clustering, it is important to consider the impact of other subjects on continuous educational life.

Written communication (WC_PRO) and biology (BIO_S11) are key indicators of a student's capability to articulate ideas and understand scientific concepts, respectively. When these subjects are clustered with students' performance levels (QUARTILE, DECILE) and socioeconomic backgrounds (SEL), it allows for a deeper analysis of how these factors influence learning outcomes. This approach helps in identifying specific needs and tailoring educational strategies to support diverse learner groups effectively. The performance metrics in Table 6.4 provides a comprehensive comparison of clustering algorithms based on various performance metrics for specific key attributes.

Table 6.4 Comparison of Cluster Quality Scores and Processing Time (Clustering) for Key Attributes Using K-Means, DP-Means and EDP-Means

Key Attributes	Algorithm	Silhouette Score	CH Score	DB Score	Processing Time (ms)
WC_PRO, DECILE, SEL	K-Means	0.33	81517.86	0.99	496.81
	DP-Means	0.44	48966.21	0.97	5526.34
	EDP-Means	0.33	81307.91	0.99	905.71
BIO_PRO, QUARTILE, SEL	K-Means	0.42	26671.84	0.96	469.85
	DP-Means	0.51	22543.33	0.82	2712.17
	EDP-Means	0.41	26228.71	0.98	734.14
MAT_S11, QR_PRO	K-Means	0.36	17389.46	0.80	1262.81
	DP-Means	0.38	15400.57	0.69	24678.67
	EDP-Means	0.38	17465.09	0.80	1691.79
ENG_S11, ENG_PRO	K-Means	0.40	24109.64	0.79	1223.69
	DP-Means	0.44	23918.59	0.66	25043.19
	EDP-Means	0.41	24918.92	0.76	1609.40
CC_S11, CC_PRO	K-Means	0.36	19581.23	0.78	1265.48
	DP-Means	0.39	19217.10	0.68	24944.36
	EDP-Means	0.37	20095.39	0.79	1767.05

Key Attributes	Algorithm	Silhouette Score	CH Score	DB Score	Processing Time (ms)
CR_S11, CR_PRO	K-Means	0.35	18330.94	0.79	1306.86
	DP-Means	0.38	17338.63	0.69	24669.23
	EDP-Means	0.36	18674.67	0.80	1700.65

Performance Metrics Analysis for Experimental Results: The EDP-Means Algorithm Outperforms and is Comparable to Traditional Approaches when clustering is based on Preprocessed/Transformed data from Edu-ETL Processes

Table 6.4 indicates that the key attributes analyzed are combinations of performance indicators (secondary, academic) and socioeconomic factors, which are critical for understanding lifelong learning achievements. Figure 6.10 provides a clear visual representation of how each algorithm scales with the complexity of the data, offering valuable insights into their practical applicability and efficiency.

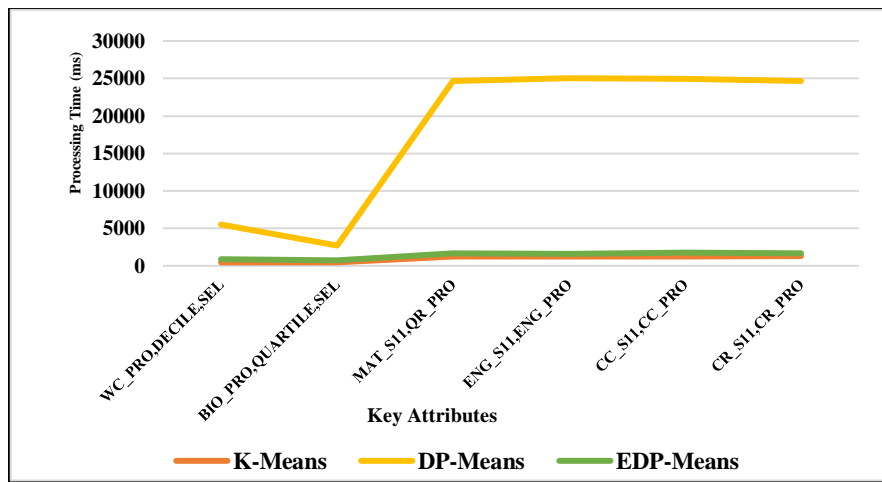


Figure 6.10 Processing Time (ms) Comparison for Clustering in K-Means, DP-Means, and EDP-Means

Analysis Results in Processing Time: This analysis focuses solely on the clustering process, excluding the time required to find the optimal number of clusters and the optimal threshold parameters. The comprehensive processing time, which includes finding the optimal number of clusters and the optimal threshold parameters, is analyzed and proved in Figure 6.7, which shows the EDP-Means has the longest processing time despite its clustering quality advantages. However, in Table 6.4, while K-Means is the most efficient in terms of clustering time. EDP-Means offers a

balanced trade-off between clustering quality and processing time. DP-Means, though effective in certain clustering quality metrics, is the least efficient in terms of processing time.

Analysis of Cluster Quality Scores for K-Means, DP-Means, and EDP-Means: The comparison is based on the Silhouette Score, CH Score, and DB Score, which comprehensively evaluate the clustering performance in Table 6.4.

- DP-Means performs particularly well with level-based attributes (such as QUARTILE, DECILE, and SEL), creating distinct clusters with high Silhouette Scores. On the other hand, EDP-Means shows its strength with attributes represented by continuous marks (such as QR_PRO, MAT_S11, ENG_S11), offering a balanced performance between clustering quality and processing time.
- K-Means and EDP-Means generally achieved higher CH Scores, indicating better-defined clusters, while DP-Means consistently had the lowest CH Scores.
- EDP-Means offers a balanced performance, providing good cluster quality as indicated by the Silhouette and CH Scores, though it often has higher DB Scores suggesting less distinct clusters.

Overall, K-Means is favored for its speed and simplicity, making it a go-to algorithm for many clustering tasks. However, EDP-Means can perform well and provide good overall clustering quality, catching up to K-Means and offering improvements in certain scenarios, particularly where data complexity and outliers are significant factors. DP-Means excels with level-based attributes, achieving high Silhouette scores, but consistently has the lowest CH Scores, indicating less well-defined clusters compared to K-Means and EDP-Means. For educational data analysis, where both cluster quality and efficiency are important, EDP-Means stands out as a robust and practical choice.

Figure 6.11, 6.12 and 6.13 illustrates the comparative performance of the K-Means, DP-Means and EDP-Means algorithms based on “Academic Evaluation” dataset using Silhouette Score, CH Score, and DB Score, respectively.

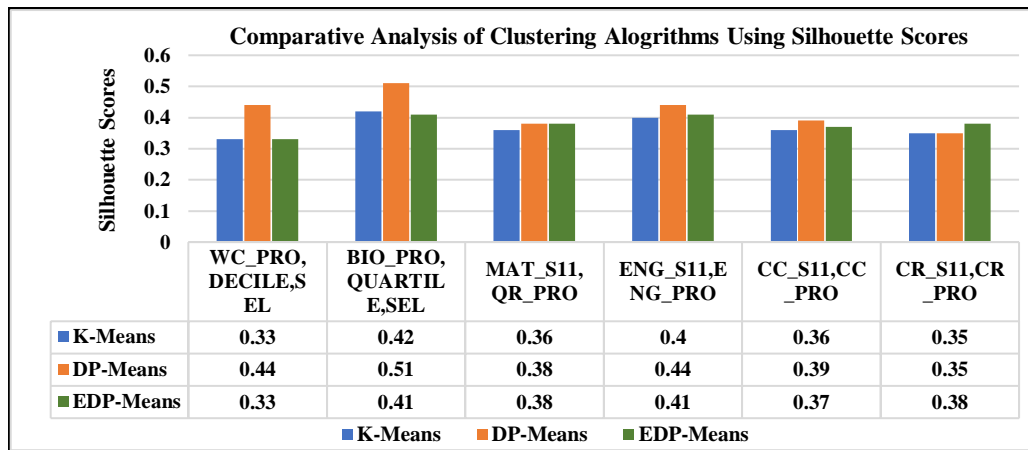


Figure 6.11 Comparative Analysis of Clustering Algorithms Using Silhouette Scores Based on “Academic Evaluation” Dataset

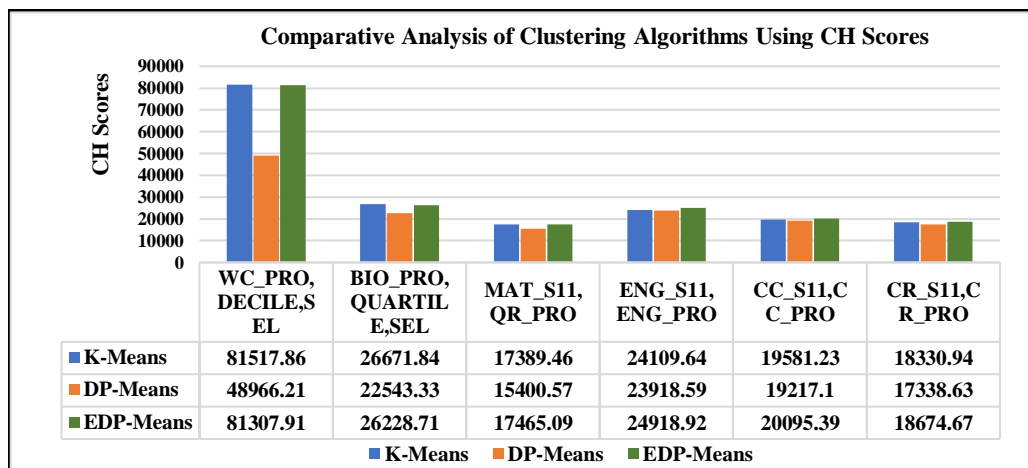


Figure 6.12 Comparative Analysis of Clustering Algorithms Using CH Scores Based on “Academic Evaluation” Dataset

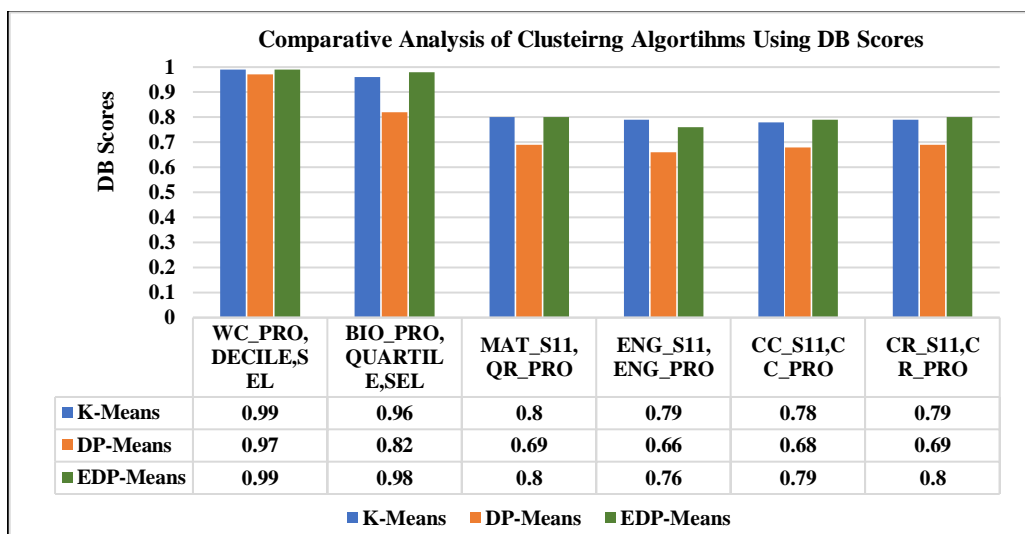


Figure 6.13 Comparative Analysis of Clustering Algorithms Using DB Scores Based on “Academic Evaluation” Dataset

Experimental Results Analysis of determining optimal cluster numbers and threshold values for key attributes Using K-Means, DP-Means, and EDP-Means

Table 6.5 provides a clear overview of the purpose of the table, which is to compare the optimal number of clusters and threshold values for key educational attributes across three different clustering algorithms. The methods used to find the optimal number of clusters and threshold parameters are critical in enhancing the clustering quality. The selected key attributes for clustering analysis were chosen due to their relevance in evaluating student performance and socioeconomic levels. Each algorithm was evaluated based on its ability to cluster the selected attributes effectively.

Approaches for the Optimal Number of Clusters and Threshold Values: The proposed EDP-Means determines the optimal number of clusters K^* by using the elbow method with SSE values. EDP-Means explored a range of λ values to dynamically adjust the threshold and find the optimal cluster number. Therefore, the algorithm's ability is highlighted to adapt threshold values dynamically, providing a balance between cluster quality and processing time.

Analysis Results of Three Clustering Algorithms from Table 6.5:

- ***K-Means:*** Best for a broad and consistent overview, especially with level-based attributes.
- ***DP-Means:*** Provides highly detailed clustering, particularly effective for level-based attributes but can result in overfitting.
- ***EDP-Means:*** Strikes a balance between detailed and manageable clustering, adapting well to both level-based and continuous attributes, making it a versatile choice for various data types.

These insights highlight the importance of selecting the appropriate clustering algorithm and parameters based on the nature of the data and the specific requirements of the analysis.

Analysis for Experimental Results of DP-Means (λ) and EDP-Means (λ^*) from Table 6.5:

- EDP-Means employs variable λ^* values, ranging from 0.3 to 0.8, tailored to the specific characteristics of each dataset. This variability allows EDP-Means to adapt to the data's nature, optimizing cluster formation by setting appropriate thresholds for different types of attributes. Lower λ^* values suggest a more conservative approach, forming fewer but more

meaningful clusters.

- In the DP-Means, the value of λ is typically selected based on prior knowledge or empirical observations from simpler datasets. However, when applying DP-Means to a more complex dataset- “Academic Evaluation” (educational datasets with continuous attributes like student marks), it becomes challenging to determine an appropriate lambda value. Therefore, in this experimental analysis, the default value of λ is set to 1.0. This default value did not capture the nuances and variations within the data.

Table 6.5 Optimal Numbers of Clusters and Threshold Values for Key Attributes Using K-Means, DP-Means and EDP-Means

Key Attributes	Optimal K (K-Means)	Optimal K, λ (DP-Means)	Optimal K^*, λ^* (EDP-Means)
WC_PRO, DECILE, SEL	100	307, 1.0	98, 0.8
BIO_PRO, QUARTILE, SEL	100	256, 1.0	99, 0.5
MAT_S11, QR_PRO	248	308, 1.0	280, 0.3
ENG_S11, ENG_PRO	248	307, 1.0	280, 0.3
CC_S11, CC_PRO	248	308, 1.0	280, 0.3
CR_S11, CR_PRO	248	308, 1.0	280, 0.3

Table 6.6 compares the optimal number of clusters K and the SSE values for K-Means and EDP-Means clustering algorithms across various key attributes.

Table 6.6 Comparison of the Optimal Number of Clusters Based on SSE Values

Key Attributes	Algorithm	Optimal K, K^*	SSE Value
WC_PRO, DECILE, SEL	K- Means	100	17064.58
	EDP-Means	98	6945.53
BIO_PRO,	K-Means	100	7288.46

Key Attributes	Algorithm	Optimal K, K^*	SSE Value
QUARTILE, SEL	EDP-Means	99	4634.01
MAT_S11, QR_PRO	K-Means	248	22971.32
	EDP-Means	280	21475.96
ENG_S11, ENG_PRO	K-Means	248	21607.83
	EDP-Means	307	21894.85
CC_S11, CC_PRO	K-Means	248	29350.17
	EDP-Means	308	22155.01
CR_S11, CR_PRO	K-Means	248	28789.05
	EDP-Means	308	21894.85

Analysis for Experimental Results of SSE Values from Table 6.6:

- ***WC_PRO, DECILE, SEL:*** EDP-Means shows a significant reduction in SSE value (6945.53) compared to K-Means (17064.58), indicating more compact and well-defined clusters with EDP-Means.
- ***BIO_PRO, QUARTILE, SEL:*** EDP-Means outperforms K-Means with an SSE of 4634.01 versus 7288.46, again suggesting tighter clusters with EDP-Means.
- ***Multiple Variables:*** For multi-variable combinations, EDP-Means generally shows lower SSE values, implying better performance despite having a higher number of clusters. For instance, for CC_S11, CC_PRO, EDP-Means achieved an SSE of 22155.01, which is significantly lower than K-Means' 29350.17.
- ***Optimal Clusters:*** The difference in the optimal number of clusters between K-Means and EDP-Means indicates that EDP-Means can capture more complexity in the data, though this comes at the cost of identifying more clusters.
- ***SSE Values:*** The lower SSE values for EDP-Means across many attribute combinations demonstrate its effectiveness in creating tighter, more defined clusters.

Figures 6.14, 6.15, 6.16 represent the scatter plots of the key attribute MAT_S11, QR_PRO, clustered using the DP-Means, EDP-Means and K-Means clustering algorithms.

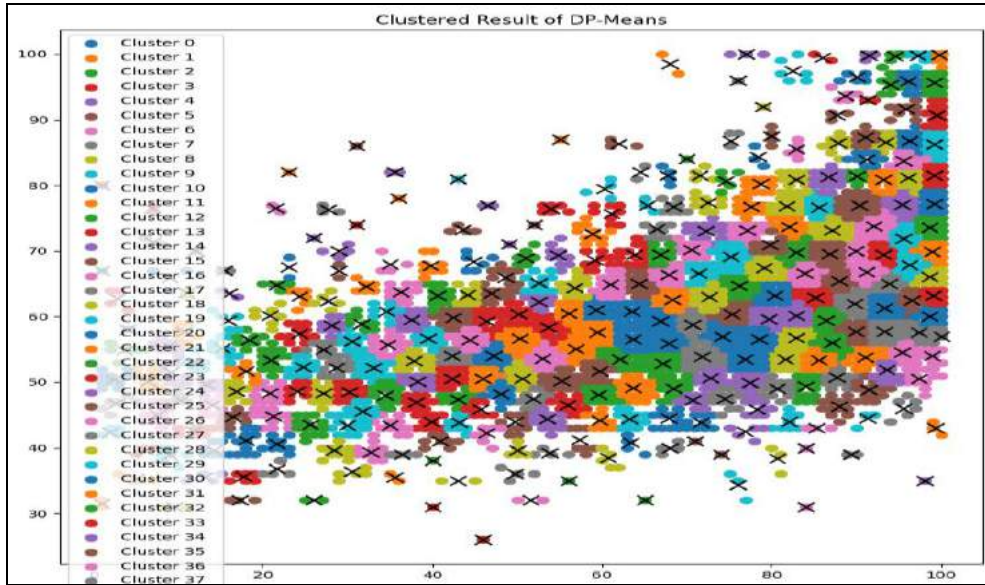


Figure 6.14 Clustering Result Visualization Using DP-Means Algorithm for Key Attribute - MAT_S11, QR_PRO

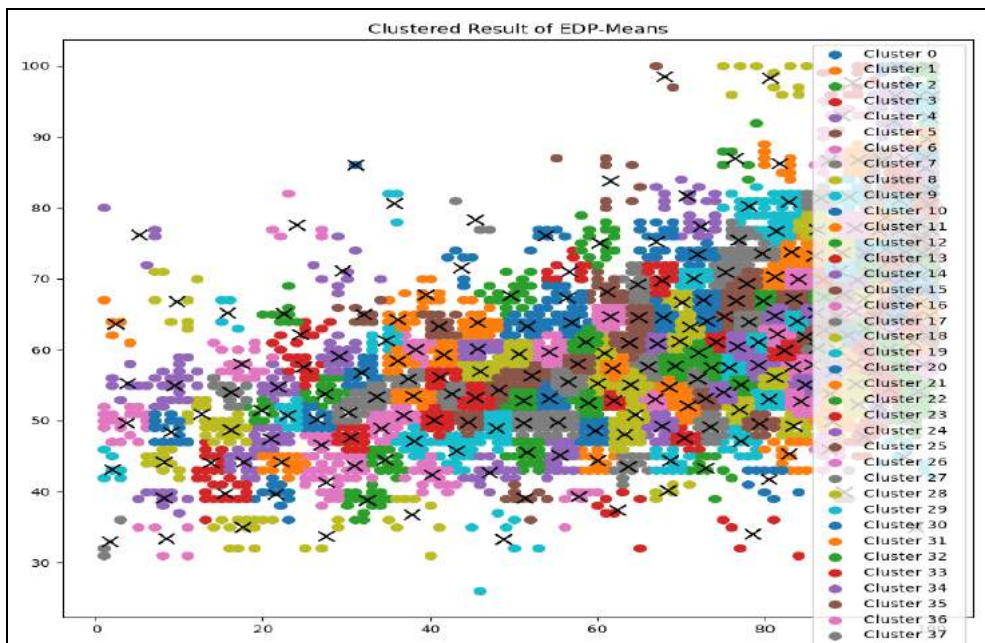


Figure 6.15 Clustering Result Visualization Using EDP-Means Algorithm for Key Attribute - MAT_S11, QR_PRO

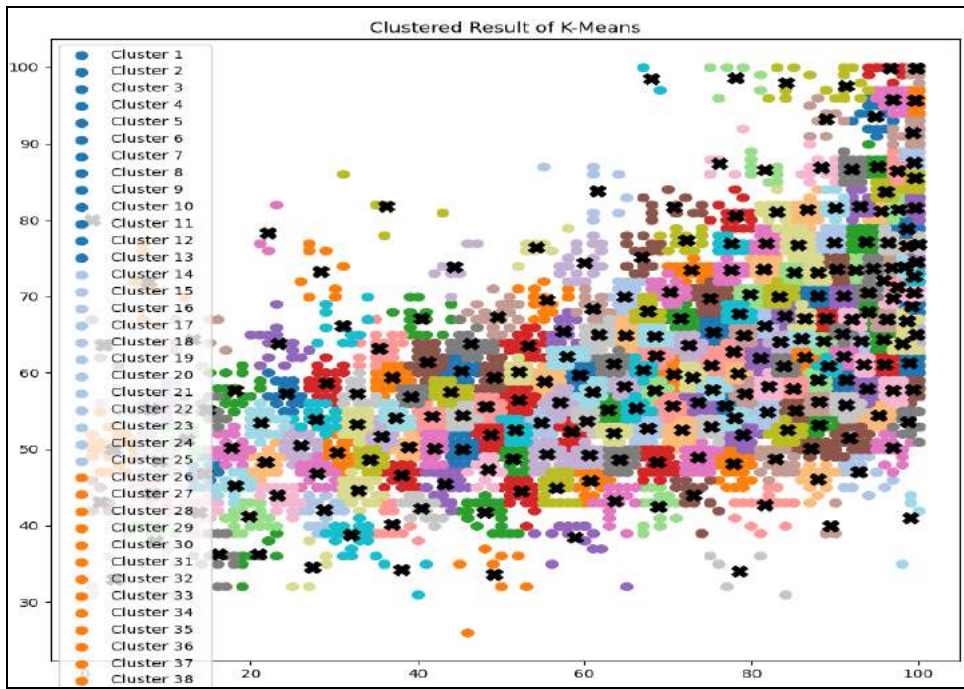


Figure 6.16 Clustering Result Visualization Using K-Means Algorithm for Key Attribute - MAT_S11, QR_PRO

Students are grouped based on their ENG_S11 and ENG_PRO scores, and the resulting clusters are visualized in the scatter plots shown in Figures 6.17, 6.18, and 6.19.

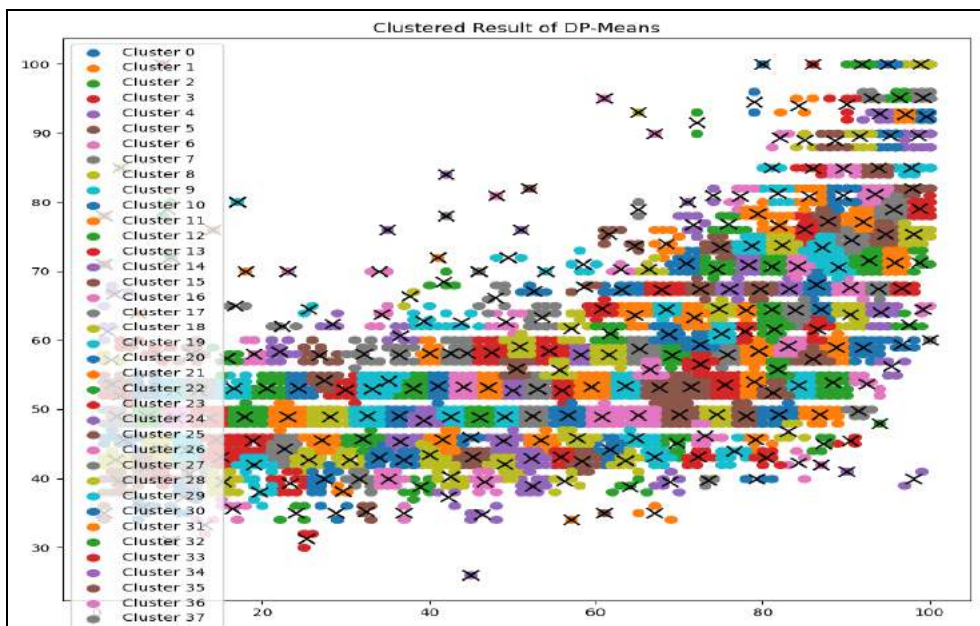


Figure 6.17 Clustering Result Visualization Using DP-Means Algorithm for Key Attribute - ENG_S11, ENG_PRO

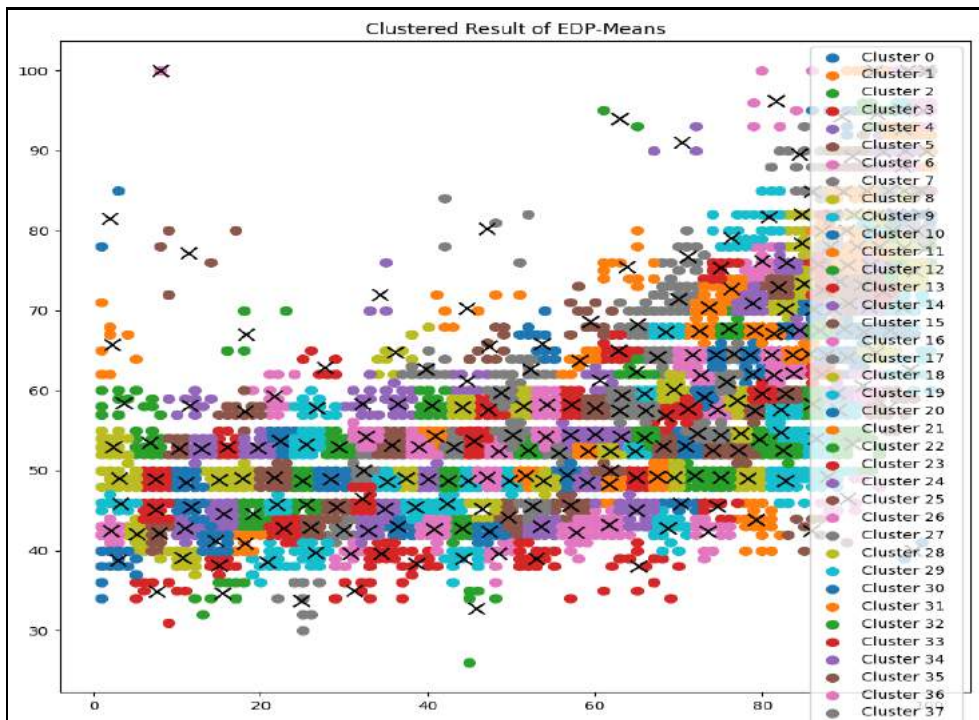


Figure 6.18 Clustering Result Visualization Using EDP-Means Algorithm for Key Attribute - ENG_S11, ENG_PRO

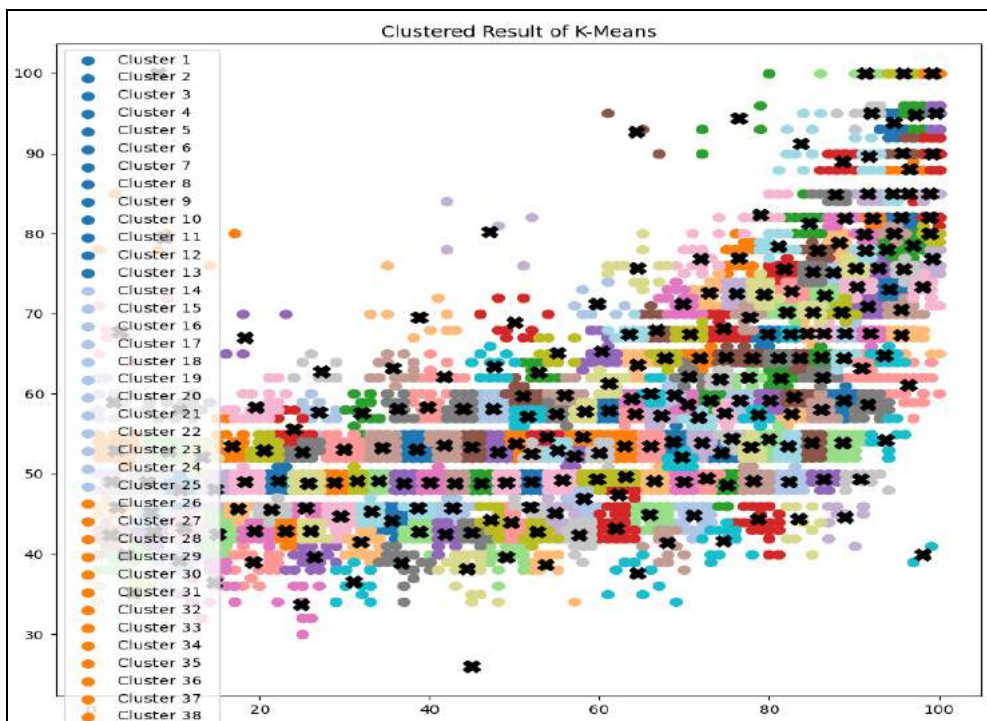


Figure 6.19 Clustering Result Visualization Using K-Means Algorithm for Key Attribute - ENG_S11, ENG_PRO

Students are grouped based on their CC_S11 and CC_PRO scores, and the resulting clusters are visualized in the scatter plots shown in Figures 6.20, 6.21, and 6.22.

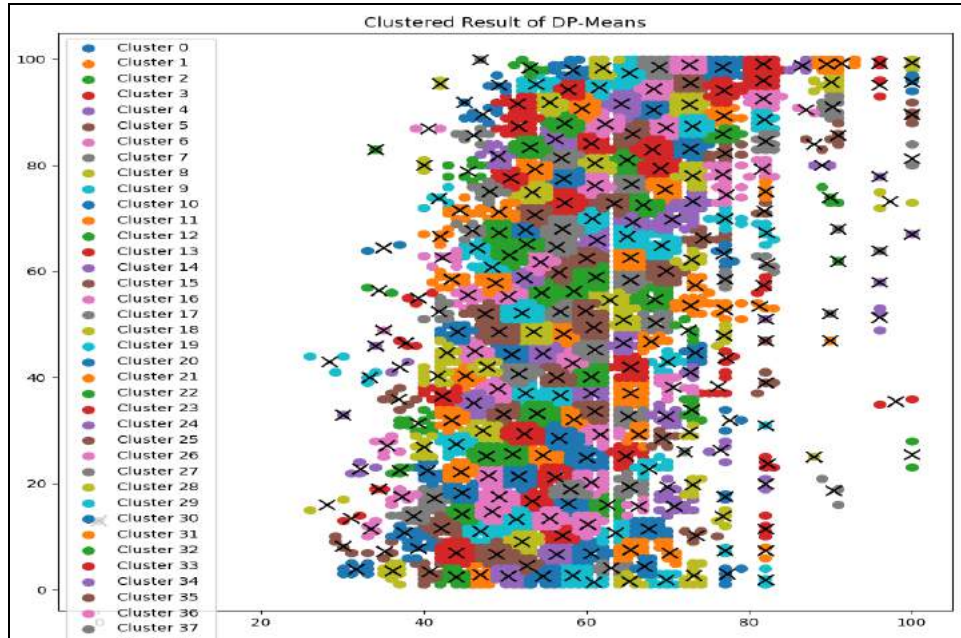


Figure 6.20 Clustering Result Visualization Using DP-Means Algorithm for Key Attribute - CC_S11, CC_PRO

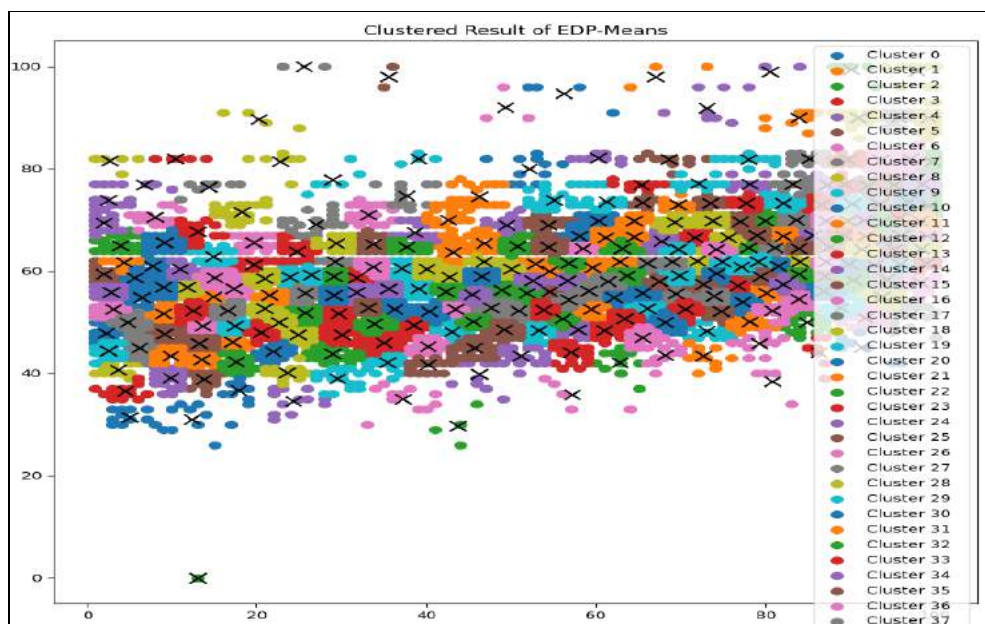


Figure 6.21 Clustering Result Visualization Using EDP-Means Algorithm for Key Attribute - CC_S11, CC_PRO

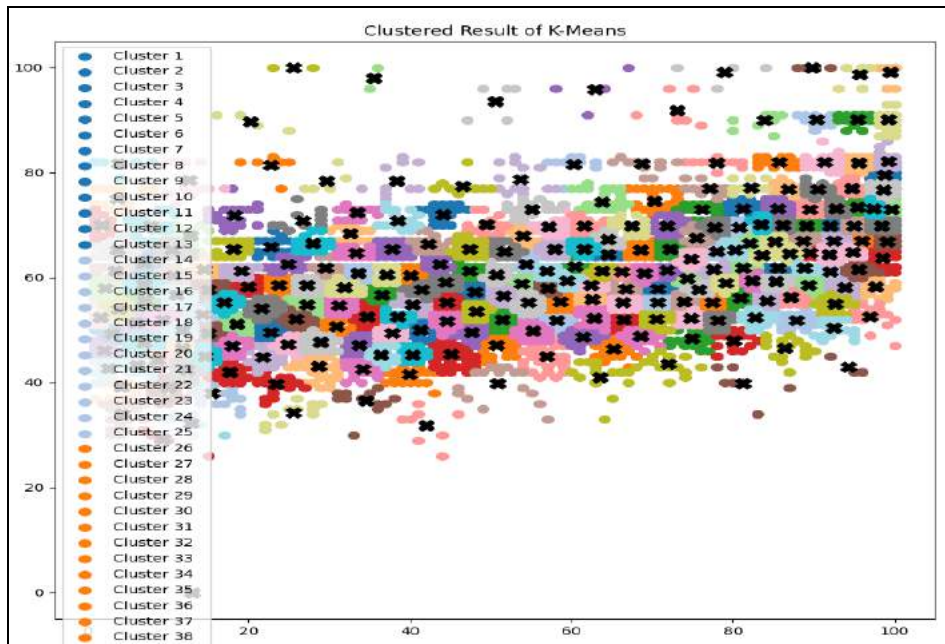


Figure 6.22 Clustering Result Visualization Using K-Means Algorithm for Key Attribute - CC_S11, CC_PRO

Students are grouped based on their CR_S11 and CR_PRO scores, and the resulting clusters are visualized in the scatter plots shown in Figures 6.23, 6.24, and 6.25.

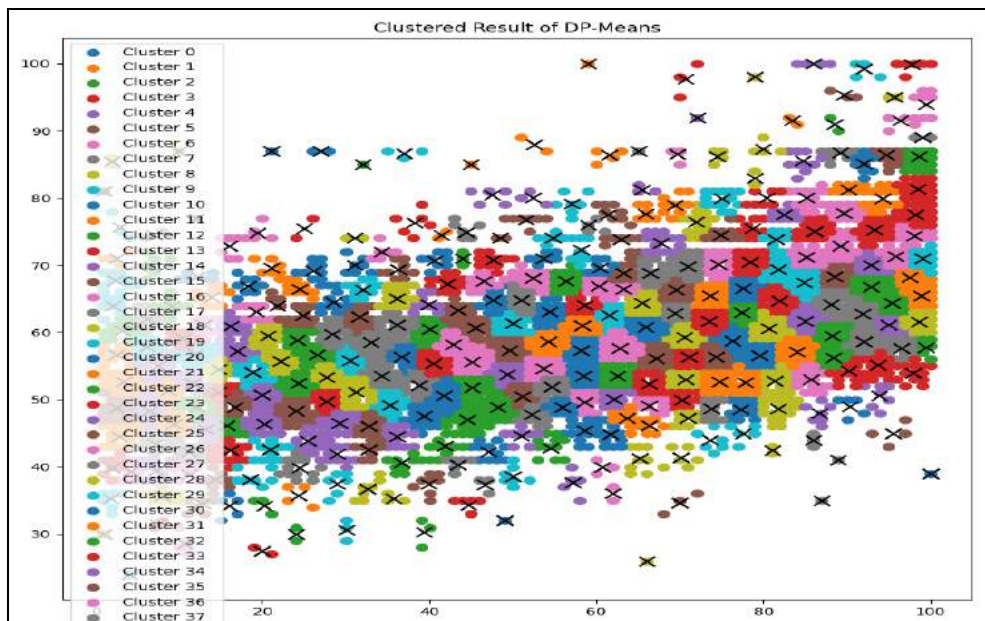


Figure 6.23 Clustering Result Visualization Using DP-Means Algorithm for Key Attribute - CR_S11, CR_PRO

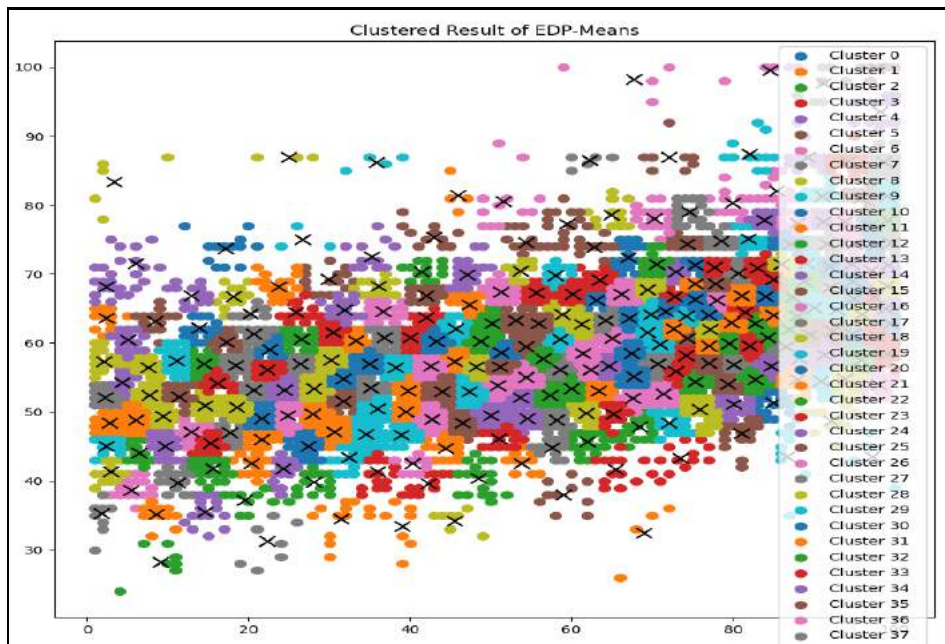


Figure 6.24 Clustering Result Visualization Using EDP-Means Algorithm for Key Attribute - CR_S11, CR_PRO

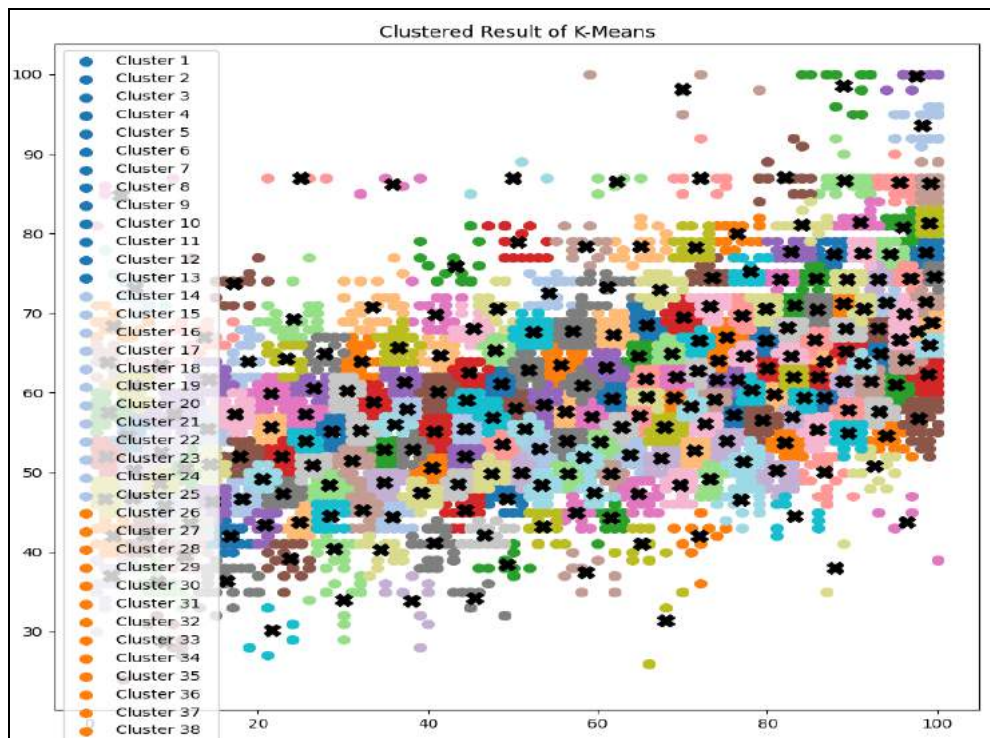


Figure 6.25 Clustering Result Visualization Using K-Means Algorithm for Key Attribute - CR_S11, CR_PRO

6.3 Result and Analysis for Lifelong Learning Achievements of the Proposed Analytical System

The analysis of prevalent patterns in each clustering group helps to identify key success factors influencing lifelong learning achievements. Table 6.7 presents an interpretation of prevalent patterns across various clusters, highlighting how different attributes interact and contribute to learning outcomes.

Table 6.7 Cluster Interpretation of Prevalent Patterns in Each Cluster Group

Key Attributes	Cluster Group	Number of Data Points/Cases	Prevalent Patterns
WC_PRO, DECILE, SEL	Cluster 36	262	[91 4 4], [92 5 4]
	Cluster 78		[93 5 4], [95 5 4], [94 5 4]
BIO_PRO, QUARTILE, SEL	Cluster 12	319	[58 3 2], [58 3 1], [58 2 2]
	Cluster 46	257	[65 4 2], [65 4 3]
MAT_S11, QR_PRO	Cluster 110	208	[100 83], [100 82], [99 82]
	Cluster 85	116	[95 61], [97 62]
ENG_S11, ENG_PRO	Cluster 11	122	[97 82], [95 82], [96 82]
	Cluster 6	94	[81 67], [80 68], [79 67], [78 67]
CC_S11, CC_PRO	Cluster 113	116	[93 68], [91 66], [92 66], [93 66]
	Cluster 27	112	[68 66], [69 64], [70 66]
CR_S11, CR_PRO	Cluster 99	113	[97 75], [98 75], [98 74]
	Cluster 171	112	[95 70], [96 69], [96 72]

Experimental Results Analysis and Interpretation to identify Key Success Factors in Lifelong Learning Achievements Based on Key Attributes:

1. Key Attributes: WC_PRO, DECILE, SEL

Cluster 36: The prevalent patterns [91, 4, 4] and [92, 5, 4] with 262 data points indicate that students with WC_PRO scores around 91-92 and DECILE scores of 4-5 tend to have a SEL score of 4.

Cluster 78: The patterns [93, 5, 4], [95, 5, 4], and [94, 5, 4] suggest a similar trend, with higher WC_PRO scores (93-95) paired with DECILE and SEL scores of 5 and 4, respectively.

Key Insight: Higher WC_PRO scores are associated with consistent SEL and DECILE scores, indicating a balanced proficiency.

2. Key Attributes: BIO_S11, QUARTILE, SEL

Cluster 12: Patterns [58, 3, 2], [58, 3, 1], and [58, 2, 2] with 319 data points show that students with a BIO_PRO score of 58 are distributed across QUARTILE scores of 2-3 and SEL scores of 1-2.

Cluster 46: The patterns [65, 4, 2] and [65, 4, 3] with 257 data points indicate students with a BIO_PRO score of 65 have higher QUARTILE and SEL scores (4 and 2-3, respectively).

Key Insight: Higher BIO_PRO scores are linked to higher QUARTILE and SEL scores, suggesting that better performance in biological subjects more correlates with higher academic rather than socio-economic quartiles.

3. Key Attributes: MAT_S11, QR_PRO

Cluster 110: Patterns [100, 83], [100, 82], and [99, 82] with 208 data points show very high MAT_S11 scores paired with high QR_PRO scores.

Cluster 85: The patterns [95, 61] and [97, 62] with 116 data points indicate high MAT_S11 scores with lower QR_PRO scores.

Key Insight: High performance in MAT_S11 is generally associated with high QR_PRO scores, indicating strong quantitative reasoning skills.

4. Key Attributes: ENG_S11, ENG_PRO

Cluster 11: Patterns such as [97, 82], [95, 82], and [96, 82] with 122 data points highlight high scores in both ENG_S11 and ENG_PRO.

Cluster 6: Patterns [81, 67], [80, 68], [79, 67], and [78, 67] with 94 data points indicate moderate scores in both ENG_S11 and ENG_PRO.

Key Insight: High ENG_S11 scores are consistently associated with high

ENG_PRO scores, emphasizing the correlation between English subject scores in secondary education and overall proficiency.

5. Key Attributes: CC_S11, CC_PRO

Cluster 113: Patterns like [93, 68], [91, 66], [92, 66], and [93, 66] with 116 data points show high CC_S11 scores paired with moderate CC_PRO scores.

Cluster 27: Patterns [68, 66], [69, 64], and [70, 66] with 112 data points show moderate CC_S11 scores with similar CC_PRO scores.

Key Insight: Higher creative composition scores (CC_S11) are linked to moderate creative composition proficiency (CC_PRO).

6. Key Attributes: CR_S11, CR_PRO

Cluster 99: Patterns [97, 75], [98, 75], and [98, 74] with 113 data points indicate high CR_S11 scores associated with high CR_PRO scores.

Cluster 171: Patterns [95, 70], [96, 69], and [96, 72] with 112 data points show slightly lower but still high CR_S11 scores with corresponding CR_PRO scores.

Key Insight: High critical reading scores (CR_S11) align with high critical reading proficiency (CR_PRO).

• ***Key success factors in lifelong learning achievements have been identified:***

1. Socioeconomic Status and Academic Performance: Students from higher socioeconomic backgrounds tend to perform better academically, indicating the importance of economic stability and resources in educational success.

2. Quartile Rank and Subject Proficiency: Quartile ranks serve as a significant indicator of overall academic proficiency and potential for higher lifelong learning achievements.

3. Subject-Specific Performance:

(a) **Mathematics and Quantitative Reasoning:** High scores in MAT_S11 are generally linked with high QR_PRO scores, demonstrating the importance of strong foundational skills in mathematics for overall academic success.

(b) **English Proficiency:** High scores in ENG_S11 correlate with high ENG_PRO scores, emphasizing the role of language skills in academic performance.

(c) **Critical Reading and Proficiency:** Consistently high scores in CR_S11 and CR_PRO indicate that students who perform well in critical reading tend to maintain high proficiency levels across other academic areas.

(d) **Creative Composition:** Higher scores in creative composition (CC_S11) are linked to moderate proficiency, highlighting the importance of creativity in academic success.

4. Balanced Academic and Extracurricular Engagement: Patterns within clusters show that students who perform well academically often have balanced engagement in extracurricular activities, suggesting that holistic development is crucial for lifelong learning success.

Effective educational interventions that address both academic and socioeconomic factors can significantly enhance lifelong learning achievements.

The key success factors identified from the analysis provide valuable insights into the elements that contribute to lifelong learning achievements. By understanding the impact of socioeconomic status, quartile ranks, subject-specific performance, and balanced academic engagement, educators and policymakers can develop strategies to foster an environment conducive to lifelong learning. These insights can help in creating targeted support systems to improve educational outcomes and support students' continuous learning journeys.

6.4 Chapter Summary

This chapter is dedicated to thoroughly examining the performance and effectiveness of the EDP-Means clustering algorithm.

The experimental results and evaluations are implemented in this chapter. The initial section delves into experimenting with the performance of the EDP-Means clustering algorithm. The cluster quality and accuracy of EDP-Means are meticulously evaluated, with comparisons made against the K-Means and original DP-Means algorithms to showcase the superior clustering outcomes of EDP-Means in terms of precision and reliability.

Subsequently, the clustering experiments extend to a distributed computing environment using PySpark, emphasizing the scalability and efficiency of the clustering processes across various benchmark datasets of differing sizes and fields. A comprehensive validation of the clustered results obtained from both standalone and PySpark implementations follows.

Finally, the chapter analyzes the learning outcomes and identifies key success factors influencing lifelong learning achievements. By examining the clustered data, the system aims to uncover patterns and insights that can inform tailored interventions and support strategies, ultimately enhancing educational outcomes.

In conclusion, the chapter presents detailed experimental results, analysis, and insights, highlighting the efficacy of the EDP-Means clustering algorithm and the overall analytical system for lifelong learning achievements.

CHAPTER 7

CONCLUSION AND FUTURE WORKS

The proposed system is an analytical system for lifelong learning achievements which presents a comprehensive framework that integrates the EDP-Means clustering algorithm with Edu-ETL processes. This innovative system aims to enhance the analysis of lifelong learning outcomes by leveraging advanced clustering techniques and rigorous data preprocessing methodologies. By dynamically updating threshold parameters and iteratively fitting data, EDP-Means improves clustering accuracy, while Edu-ETL processes ensure comprehensive data preprocessing. This chapter provides a summary of the dissertation, outlines the advantages and limitations of the proposed system, and discusses potential future work.

7.1 Dissertation Summary

The integration of EDP-Means clustering with Edu-ETL processes in the analytical system for lifelong learning achievements offers several advantages. EDP-Means consistently outperforms traditional K-Means and DP-Means algorithms in cluster quality, especially in educational datasets. While EDP-Means has higher processing times for large datasets, its ability to handle complex data and outliers makes it a superior choice for in-depth educational data analysis. The summary of the analyzing results of the system is:

1. Performance Metrics Analysis

- The EDP-Means algorithm outperforms traditional K-Means and DP-Means algorithms in cluster quality, achieving higher or comparable Silhouette and DB Scores.
- EDP-Means shows improved CH Scores in certain datasets, indicating its ability to form compact and well-separated clusters.
- In educational datasets, EDP-Means combined with Edu-ETL processes reveals significant insights into learning patterns and success factors.

2. Processing Time Analysis

- For smaller datasets, K-Means, DP-Means, and EDP-Means exhibit

efficient processing times.

- For larger datasets like "Spotify Popular Music," EDP-Means has a higher computational cost due to dynamic threshold adjustments, while K-Means and DP-Means offer lower processing times.

3. Threshold Parameter Sensitivity

- EDP-Means adapts its λ^* values to each dataset's characteristics, optimizing cluster formation and often outperforming DP-Means.
- Both algorithms stabilize at similar λ values for large datasets, indicating a converging inherent structure.

4. Performance in PySpark Environment

- K-Means forms more distinct clusters with higher Silhouette Scores in parallel computing environments.
- DP-Means faces challenges in parallelization, resulting in lower Silhouette Scores, while EDP-Means performs slightly better than DP-Means but still lags behind K-Means.
- For smaller datasets, EDP-Means excels, surpassing both DP-Means and K-Means in Silhouette Scores.

5. Insights from EDP-Means Clustering

- EDP-Means is highly effective for educational data, handling outliers and noise well.
- Dynamic threshold adjustment results in clearer boundaries between learner groups, enhancing the identification of learning trajectories and success factors.

6. Key Success Factors in Lifelong Learning

- Socioeconomic Status: Higher socioeconomic backgrounds correlate with better academic performance.
- Quartile Rank: Significant indicators of overall academic proficiency.
- Subject-Specific Performance: Strong foundational skills in mathematics and language skills are crucial.
- Balanced Engagement: Successful students balance academic performance with extracurricular activities.

7. Educational Interventions:

- Addressing both academic and socioeconomic factors can enhance lifelong learning achievements.

- Insights from clustering can guide educators and policymakers in developing strategies to support continuous learning.

The algorithm's adaptive threshold adjustment enhances cluster formation, providing clear insights into learning patterns and success factors. Effective educational interventions informed by these insights can significantly improve lifelong learning outcomes.

7.2 Advantages and Limitations

Some advantages of the proposed analytical system for lifelong learning achievements are as follows:

1. **Enhanced Clustering Accuracy:** The integration of EDP-Means clustering algorithm and Edu-ETL processes results in improved clustering accuracy. EDP-Means dynamically updates threshold parameters and iteratively fits data, leading to more precise and reliable clustering outcomes compared to traditional clustering algorithms like K-Means.
2. **Comprehensive Data Preprocessing:** The proposed Edu-ETL process enhances educational data management and analytics by integrating specialized methodologies and techniques. It consolidates student data from multiple sources based on student ID, facilitating comprehensive analysis. Through iterative transformation steps like data cleansing and normalization, it refines datasets for accurate analysis, enabling the detection of correlations and data distribution assessment. Additionally, Edu-ETL streamlines the loading process, ensuring seamless integration into target systems. Overall, it empowers educational institutions and researchers to efficiently manage and analyze data, driving informed decision-making and improving learning outcomes.
3. **Adaptability to Diverse Data Scenarios:** The system shows flexibility by distinguishing between merged and non-merged data preprocessing. For merged data (Edu-ETL processes), it efficiently integrates and cleans multiple datasets to create a unified dataset, while for non-merged data, it applies tailored preprocessing techniques to ensure data quality for clustering analysis.
4. **Multi-Faceted Validation:** The system employs multiple validation metrics, such as Silhouette Score, CH index, and DB index, to assess the quality of

clustering results comprehensively. This multi-faceted validation approach ensures robust and reliable analysis of lifelong learning achievements.

5. **Insightful Learning Outcomes Analysis:** By examining clustered data, the system uncovers patterns and identifies key success factors influencing lifelong learning achievements. These insights can inform tailored interventions and support strategies to enhance educational outcomes for students.
6. **Flexibility and Adaptability:** The proposed system is flexible and adaptable to various educational contexts and datasets. It accommodates both numerical and categorical data, making it suitable for analyzing diverse educational datasets across different fields and levels of education.

On the other hand, there are some limitations and constraints in the proposed system.

1. **Processing Time:** One significant limitation of the proposed system is the increased processing time associated with the EDP-Means clustering algorithm compared to other methods such as K-Means and DP-Means. The dynamic threshold updating mechanism and iterative fitting process of EDP-Means contribute to longer execution times, especially with larger datasets. Despite efforts to mitigate this limitation by leveraging the PySpark environment for distributed computing, the processing time remains a concern, particularly for real-time or time-sensitive applications.
2. **Scalability Issues:** While PySpark offers scalability advantages for handling large-scale datasets, it does not completely alleviate the scalability limitations of EDP-Means clustering. In practice, EDP-Means may struggle to scale effectively with very large datasets, particularly when compared to its performance on smaller or medium-sized datasets. This scalability issue can impact the practical applicability of the proposed system, especially in scenarios where analysis of extensive datasets is required.

While the proposed system offers advancements in lifelong learning achievement analysis through the integration of EDP-Means clustering and Edu-ETL processes, it is incomplete without limitations. Addressing these limitations, such as optimizing algorithm efficiency and scalability, improving resource utilization, and enhancing parameter robustness, will be crucial for enhancing the system's effectiveness and practical utility in real-world applications.

7.3 Future Works

In future research endeavors, several promising avenues emerge for further advancing the analytical system for lifelong learning achievements. One potential direction involves refining the EDP-Means clustering algorithm to improve its scalability and efficiency, particularly for large-scale datasets. These facts explore alternative optimization techniques, parallel processing strategies, or incorporating advanced machine learning methodologies to enhance algorithm performance. Additionally, efforts can be directed towards enhancing the integration of Edu-ETL processes, with a focus on automating data preprocessing tasks and extending support for a broader range of educational datasets. Furthermore, investigating the applicability of deep learning and neural network approaches for lifelong learning analysis could yield valuable insights into complex patterns and relationships within educational data. Lastly, there is a need for longitudinal studies and real-world implementation trials to evaluate the effectiveness and practical utility of the proposed system in diverse educational contexts. By pursuing these avenues, future research endeavors can contribute to the ongoing evolution and refinement of analytical frameworks for lifelong learning assessment and support.

LIST OF ACRONYMS

EDP-Means	Enhanced Dirichlet Process Means
Edu-ETL	Educational Extract, Transform, Load
DP-Means	Dirichlet Process Means
ML	Machine Learning
EDA	Educational Data Analysis
EDM	Educational Data Mining
ANN	Artificial Neural Network
SMOTE	Synthetic Minority Oversampling Technique
GMMs	Gaussian Mixture Models
CRP	Chinese Restaurant Process
PDF	Probability Density Function
DACE	Dirichlet Process Means for Clustering Extremely Large
PDC-DP-Means	Parallel Delayed Cluster Dirichlet Process Means
SSE	Sum of Squared Errors
CH	Calinski-Harabasz
DB	Davies-Bouldin
LMS	Learning Management System
IDC	International Data Corporation
ETL	Extraction, Transformation, Loading
GUI	Graphical User Interface

Author's Publications

- [P1] Gant Gaw Wutt Mhon, Nang Saing Moon Kham, "Analysis of Students' Academic Performance, Behavior and Personality in Big Data", *In 2019 17th International Conference on Computer Applications (ICCA)*, The 11th International Conference on Future Computer and Communication (ICFCC 2019), pp. 208-212, 27-28 February, 2019, Yangon, Myanmar
- [P2] Gant Gaw Wutt Mhon, Nang Saing Moon Kham, "ETL, Preprocessing with Multiple Data Sources for Academic Data Analysis", *In 2020 IEEE Conference on Computer Applications (ICCA)*, The 18th International Conference on Computer Applications (IEEE ICCA), pp. 130-135, 27-28 February, 2020. Yangon, Myanmar.
- [P3] Gant Gaw Wutt Mhon, Nilar Aye, "Investigating Lifelong Learning Achievements with In-Depth Data Analysis Using Enhanced DP-Means Clustering and ETL Integration", *Indian Journal of Computer Science and Engineering (IJCSE)*, Volume 15 Issue 2, pp. 191-203, Jan-Feb 2024.

Bibliography

- [1] Abd El-Hafeez, T., Omar, A., "Student Performance Prediction Using Machine Learning Techniques," *In Proceedings of the 2022 International Conference on Data Science and Machine Learning Applications (DSMLA)*, 2022.
- [2] Abuzinadah, N., Umer, M., Ishaq, A., Al Hejaili, A., Alsubai, S., Eshmawi, A. A., "Role of convolutional features and machine learning for predicting student academic performance from MOODLE data", *PLOS ONE*, 2023.
- [3] Adane, M. D., Deku, J. K., Asare, E. K., "Performance analysis of machine learning algorithms in prediction of student academic performance", *Journal of Advances in Mathematics and Computer Science*, vol. 38, no. 5, pp. 74-86, 2019.
- [4] Agasisti, T., Bowers, A. J., "Data analytics and decision making in education: towards the educational data scientist as a key actor in schools and higher education institutions," *Handbook of contemporary education economics*, Edward Elgar, pp. 184-210, 2017.
- [5] Ahaidous, K., Tabaa, M., Hachimi, H., "Towards IoT-Big Data architecture for future education", *Procedia Computer Science*, Elsevier, vol. 220, pp. 348-355, 2023.
- [6] Ahmed, M., Seraj, R., Islam, S. M. S., "The k-means algorithm: A comprehensive survey and performance evaluation", *Electronics*, MDPI, vol. 9, no. 8, pp. 1295, 2020.
- [7] Al-Alawi, L., Al Shaqsi, J., Tarhini, A., Al-Busaidi, A. S., "Using machine learning to predict factors affecting academic performance: the case of college students on academic probation", *Education and Information Technologies*, Springer, vol. 28, no. 10, pp. 12407-12432, 2023.
- [8] Al-Kabi, M. N., Jirjees, J. M., "Survey of Big Data applications: Health, education, business & finance, and security & privacy", *Journal of Information Studies and Technology*, HBKU Press Qatar, vol. 2, no. 12, pp. 2019.
- [9] Arslan, A., "The effect of university students' achievement orientations on lifelong learning tendencies: a structural equation model study," *Cukurova University Faculty of Education Journal*, vol. 51, no. 1, pp. 106-147, 2022.
- [10] Arun Krishna Chitturi, C. Ranichandra, N. C. Senthilkumar, "Student

- Performance Analysis in Spark”, *Intelligent Computing Paradigm and Cutting-edge Technologies (ICICCT 2020)*, 22 April 2021, pp. 337–349, 2020.
- [11] Asikainen, H., Salmela-Aro, K., Parpala, A., Katajavuori, N., "Learning profiles and their relation to study-related burnout and academic achievement among university students," *Learning and Individual differences*, Elsevier, vol. 78, pp. 101781, 2020.
- [12] Ayçiçek, B., Karafil, B., "Investigation of University Students' Lifelong Learning Tendencies in Terms of Various Variables”, *African Educational Research Journal*, ERIC, vol. 9, no. 1, pp. 121-133, 2021.
- [13] Bala, M., Boussaid, O., Alimazighi, Z., "Big-ETL: extracting-transforming-loading approach for Big Data”, *In Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*, pp. 4-Jan, 2015.
- [14] Basu, S., Choudhury, J. R., Paul, D., Das, S., “Robust and automatic data clustering: Dirichlet process meets median-of-means”, *arXiv preprint arXiv:2311.15384*, 2023.
- [15] Beckham, N. R., Akeh, L. J., Mitaart, G. N. P., Moniaga, J. V., "Determining factors that affect student performance using various machine learning methods”, *Procedia Computer Science*, Elsevier, vol. 216, pp. 597-603, 2023.
- [16] Comiter, M., Cha, M., Kung, H. T., Teerapittayanon, S., “Lambda means clustering: automatic parameter search and distributed computing implementation”, *2016 23rd International Conference on Pattern Recognition (ICPR)*, IEEE, pp. 2331-2337, 2016.
- [17] Delahoz-Dominguez, E., Zuluaga, R., Fontalvo-Herrera, T., “Dataset of academic performance evolution for engineering students”, *Data in brief*, Elsevier, vol. 30, pp. 105537, 2020.
- [18] Dinari, O., Freifeld, O., "Revisiting dp-means: fast scalable algorithms via parallelism and delayed cluster creation”, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, PMLR, pp. 579-588, 2022.
- [19] Durak, H. Y., Sarıtepeci, M., Dünya, B. A., "Examining the relationship between computational thinking, lifelong learning competencies and

personality traits using path analysis”, *Bartın University Journal of Faculty of Education*, vol. 10, no. 2, pp. 281-292, 2021.

- [20] Feng, G., Fan, M., Chen, Y., "Analysis and prediction of students' academic performance based on educational data mining,”, *IEEE Access*, IEEE, vol. 10, pp. 19558-19571, 2022.
- [21] Galici, R., Ordile, L., Marchesi, M., Pinna, A., Tonelli, R., "Applying the ETL process to blockchain data. prospect and findings," *Information*, vol. 11, no. 4, pp. 204, MDPI, 2020.
- [22] Geetha, K., "Data Analysis and ETL Tools in Business Intelligence," *International Research Journal of Computer Science (IRJCS)*, vol. 7, pp. 127-131, 2020.
- [23] Goh, Y. L., Goh, Y. H., Yip, C.-C., "Prediction of student's academic performance by k-means clustering," *Peer-review under responsibility of 4th Asia International Multidisciplinary Conference 2020 Scientific Committee*, 2020.
- [24] Gupta, V., Singhal, P., Khattri, V., "Analysis of Student Academic Performance Using Machine Learning Algorithms: A Study”, *Journal of Jilin University (Engineering and Technology Edition)*, 2024.
- [25] Hussain, S., Khan, M. Q., "Student Performulator: Predicting students' academic performance at secondary and intermediate level using machine learning”, *Annals of Data Science*, Springer, vol. 10, no. 3, pp. 637–655, 2023.
- [26] H. Benbrahim, H. Hachimi, and A. Amine, “Deep transfer learning with Apache Spark to detect COVID-19 in chest x-ray images,” *Romanian Journal of Information Science and Technology*, vol. 23, pp. S117–S129, 2020.
- [27] Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., Heming, J., K-means clustering algorithms: “A comprehensive review, variants analysis, and advances in the era of big data”, *Information Sciences*, Elsevier, vol. 622, pp. 178-210, 2023.
- [28] Java, S., Mohammed, H., Bhardwaj, A. B., Shukla, V. K., "Education 4.0 and Web 3.0 applications in enhancing learning management system: Post-lockdown analysis in covid-19 pandemic," *In Knowledge Management and Web 3.0: Next Generation Business Models*, vol. 2, p. 85, 2021.

- [29] Jiang, L., Dong, Y., Chen, N., Chen, T., "DACE: A scalable DP-means algorithm for clustering extremely large sequence data. *Bioinformatics*", *Oxford University Press*, vol. 33, no. 6, pp. 834-842, 2017.
- [30] Kena, G., Hussar, W., McFarland, J., De Brey, C., Musu-Gillette, L., Wang, X., Zhang, J., Rathbun, A., Wilkinson-Flicker, S., Diliberti, M., "The Condition of Education 2016. NCES 2016-144", *National Center for Education Statistics*, ERIC, 2016.
- [31] Khan, I. H., Javaid, M., "Role of Internet of Things (IoT) in adoption of Industry 4.0", *Journal of Industrial Integration and Management*, World Scientific, vol. 7, no. 4, pp. 515-533, 2022.
- [32] Khayi, N. A., Rus, V., "Clustering Students Based on Their Prior Knowledge", *International Educational Data Mining Society*, ERIC, 2019.
- [33] Kobayashi, M., Watanabe, K., "Generalized Dirichlet-process-means for f-separable distortion measures", *Neurocomputing*, Elsevier, vol. 458, pp. 667-689, 2021.
- [34] Kulis, B., Jordan, M. I., "Revisiting k-means: New algorithms via Bayesian nonparametrics", *arXiv preprint arXiv:1111.0352*, 2011
- [35] Leo W., "Data warehouse with big data technology for higher education", *Procedia Computer Science*, vol. 124, pp. 93-99, 2017.
- [36] Li, Y., Schofield, E., Gönen, M., "A tutorial on Dirichlet process mixture modeling", *Journal of Mathematical Psychology*, Elsevier, vol. 91, pp. 128-144, 2019.
- [37] Lyu, S., "Student's Academic Performance Prediction Based on Machine Learning Regression Models," *In Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Atlantis Press, pp. 293-299, 2023.
- [38] Melit Devassy, B., George, S., "Unsupervised Clustering of Hyperspectral Data with an Unknown Number of Clusters Using Dirichlet Process Means," *In Proceedings of the 5th International Conference on Data Science and Analytics*, 5 Jun 2023.
- [39] Nagahi, M., Jaradat, R., Davarzani, S., Nagahisarchoghaei, M., Goerger, S. R., "Academic performance of engineering students", *ASEE Virtual Annual Conference Content Access*, 2020.
- [40] Prinsloo, P., Archer, E., Barnes, G., Chetty, Y., Van Zyl, D., "Big(ger) data

as better data in open distance learning," *International Review of Research in Open and Distributed Learning*, vol. 16, no. 1, pp. 284-306, 2015.

- [41] Priya, S., Ankit, T., Divyansh, D., "Student performance prediction using machine learning," *Advances in Parallel Computing Technologies and Applications*, IOS Press, pp. 167-174, 2021.
- [42] Rachwał, A., Popławska, E., Gorgol, I., Cieplak, T., Pliszczuk, D., Skowron, Ł., Rymarczyk, T., "Determining the quality of a dataset in clustering terms", *Applied Sciences*, MDPI, vol. 13, no. 5, pp. 29-42, 2023.
- [43] Rodzi, N. A. H. M., Othman, M. S., Yusuf, L. M., "Significance of data integration and ETL in business intelligence framework for higher education," *In Proceedings of the 2015 International Conference on Science in Information Technology (ICSITech)*, IEEE, pp. 181-186, 2015.
- [44] Shahapure, K. R., Nicholas, C., "Cluster quality analysis using silhouette score", *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, pp. 747-748, 2020.
- [45] S. Sagiroglu and D. Sinanc, "Big data: a review", *In Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS)*, IEEE, San Diego, CA, USA, May 2013.
- [46] Tasmin, R., Muhammad, R. N., Aziati, A. H. Nor: "Big data analytics applicability in higher learning educational system", *IOP Conference Series: Materials Science and Engineering*, vol. 917, no. 1, 2020.
- [47] Vaitis, C., Hervatis, V., Zary, N., "Introduction to big data in education and its contribution to the quality improvement processes", *Big Data on Real-World Applications*, vol. 113, pp. 58, 2016.
- [48] Yağcı, M., "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, Springer, vol. 9, no. 1, p. 11, 2022.
- [49] Yauri, R. A., Suru, H. U., Afrifa, J., Moses, H. G., "A Machine Learning Approach in Predicting Student's Academic Performance Using Artificial Neural Network", *Journal of Computational and Cognitive Engineering*, vol. 3, no. 2, pp. 203-212, 2024.
- [50] Yuan, C., Yang, H., "Research on K-value selection method of K-means clustering algorithm", *J*, MDPI, vol. 2, no. 2, pp. 226-235, 2019.
- [51] Yulianto, A. A., "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, 2019.

- [52] Zhao, L., Ren, J., Zhang, L., Zhao, H., "Quantitative Analysis and Prediction of Academic Performance of Students Using Machine Learning", *Sustainability*, MDPI, vol. 15, no. 16, p. 12531, 2023.
- [53] Zulfiker, M. S., Kabir, N., Biswas, A. A., Chakraborty, P., Rahman, M. M., "Predicting students' performance of the private universities of Bangladesh using machine learning approaches," *International Journal of Advanced Computer Science and Applications*, Science and Information (SAI) Organization Limited, vol. 11, no. 3, pp. 672-679, 2020.