Journal of Technology Research (JTR)
Volume 3, Issue 2, (2025), pp 103-112, ISSN 3005-639X

# Clustering-Based Classification of Student Dropout Patterns: A Case Study in College of Industrial Technology – Misrata, Libya

Hoda B. Abugharsa[1]*

[1]Electronic Engineering Department, Collage of Industrial Technology, Misrata, Libya.
*Corresponding author email: hudabader82@cit.edu.ly.

Received: 15-10-2025 | Accepted: 14-10-2025 | Available online: 15-12-2025 | DOI:10.26629/jtr.2025.12

**ABSTRACT**

Student dropout remains a critical challenge for educational systems, adversely affecting graduation rates and institutional resource allocation. This study proposes an analytical framework utilizing unsupervised clustering algorithms to classify students at the College of Industrial Technology, Misurata, Libya into two distinct categories: one group comprising students who dropped out early with minimal academic engagement, and another group consisting of students who withdrew after achieving a relative degree of academic progress. This classification aims to provide a deeper understanding of the diverse pathways of student attrition. The research integrates demographic variables (e.g., gender, admission age, and enrollment year) with academic performance indicators to construct a comprehensive predictive model. The performance of K-Means and Agglomerative Clustering algorithms was evaluated using validation metrics: Silhouette Score, Davies-Bouldin Index. The findings reveal statistically significant patterns that enable the identification of high-risk student cohorts, providing actionable insights for targeted academic interventions. These results may contribute to the enhancement of student retention policies and support data-driven decision-making.

**Keywords:** Student dropout, clustering algorithms, KMeans, , educational data mining.

# تصنيف أنماط تسرب الطلاب باستخدام خوارزميات التجميع

# دراسة حالة في كلية التقنية الصناعية – مصراتة، ليبيا

هدى أبوغرسة

قسم الهندسة الإلكترونية، كلية التقنية الصناعية، مصراتة، ليبيا.

## ملخـــــص البحـــــث

لا يزال تسرب الطلاب يمثل تحديًا حاسمًا أمام الأنظمة التعليمية، حيث يؤثر سلبًا في معدلات التخرج وتخصيص الموارد المؤسسية. تقترح هذه الدراسة إطارًا تحليليًا يستخدم خوارزميات التجميع غير المراقب (Unsupervised Clustering) لتصنيف طلاب كلية التقنية الصناعية – مصراتة، ليبيا إلى فئتين متميزتين: الفئة الأولى تضم الطلاب الذين انسحبوا في وقت مبكر مع حد أدنى من المشاركة الأكاديمية، والفئة الثانية تضم الطلاب الذين انسحبوا بعد تحقيق قدر نسبي من التقدم الأكاديمي. يهدف هذا التصنيف إلى توفير فهم أعمق للمسارات المتنوعة لظاهرة تسرب الطلاب. يُدمج البحث بين المتغيرات الديموغرافية (مثل الجنس، وعمر القبول، وسنة

الالتحاق) ومؤشـــرات الأداء الأكاديمي لبناء نموذج تنبؤي شـــامل. تم تقييم أداء خوارزميتي K-Means و **Agglomerative Clustering** باسـتخدام مقاييس التحقق Silhouette Score وDavies-Bouldin Index. تكشـف النتائج عن أنماط ذات دلالة إحصـــائية تتيح تحديد مجموعات الطلاب الأكثر عرضـــة لخطر التســـرب، مما يوفر رؤى قابلة للتطبيق لتوجيه التدخلات الأكاديمية المستهدفة. وقد تسهم هذه النتائج في تعزيز سياسات الاحتفاظ بالطلاب ودعم اتخاذ القرارات القائمة على البيانات.

**الكلمات الدالة:** تسرب الطلاب، خوارزميات التجميع، خوارزمية K-Means، تقييم البيانات التنبؤية.

## 1. INTRODUCTION

Student dropout is a significant issue faced by educational institutions, negatively affecting graduation rates and incurring financial and social costs [1]. Early identification and classification of dropout patterns enable institutions to design effective intervention strategies that improve student retention and academic success. With the increasing volume of educational data, machine learning techniques, especially unsupervised clustering, have gained traction as effective tools for detecting latent patterns in student behaviors without requiring labeled data [2].

This study employs clustering techniques to stratify students within an industrial technical college into two distinct cohorts: those demonstrating consistent academic progression and those exhibiting premature discontinuation. By synthesizing demographic variables with academic performance indicators, the study seeks to elucidate the multifaceted interdependencies underlying attrition phenomena. Furthermore, the research prioritizes the rigorous quantitative assessment of clustering efficacy to substantiate the methodological robustness and practical utility of the derived classifications.

## 2. LITERATURE REVIEW

The use of clustering algorithms in educational data mining has evolved significantly over the past decade. Smith et al. (2017) used KMeans to classify students based solely on GPA and attendance rates, achieving initial insights into at-risk populations [1].

By 2020, research began incorporating additional features such as socio-economic factors and engagement data. Lee and Kim (2020) demonstrated how combining academic performance with demographic attributes like gender and age improved dropout classification in university environments [2].

In 2021, Martínez and Torres highlighted the advantages of hierarchical clustering (Agglomerative Clustering) in discovering hidden structures in educational datasets, especially when student groups are heterogeneous [3].

They argued that unlike partitioning methods like KMeans, hierarchical algorithms capture nested patterns in the data. Further advancements came with hybrid models. In 2022, Zhao et al. proposed combining hierarchical and partitioning approaches to enhance clustering accuracy and interpretability in educational settings [4]. Their study emphasized that relying on a single algorithm might lead to biased or oversimplified results. The need for model explainability led to the adoption of Explainable AI techniques in educational clustering by 2023. Nguyen et al. applied SHAP and LIME algorithms to interpret why students belonged to certain clusters, making clustering results more transparent to academic decision-makers [5].

In 2023 and 2024, several studies emphasized the importance of evaluating clustering performance using multiple validity indices. Rousseeuw's Silhouette score [6] and the Davies-Bouldin index [7] were frequently cited for objectively determining the optimal number

of clusters and validating the separability of groups. Most recent works (e.g., Wang et al., 2024) advocate for rigorous comparative analysis between clustering algorithms to avoid biased conclusions based on a single method [8].

While prior research has demonstrated the efficacy of clustering algorithms in analyzing student performance and identifying at-risk cohorts, important gaps remain. Most existing studies focus on academic or behavioral metrics in isolation, often overlooking the combined influence of demographic factors such as admission year, enrollment age, and gender. Additionally, the context of technical and vocational education—particularly within technical colleges in developing regions—has received limited attention, despite its unique academic progression patterns and dropout risk factors. To address these gaps, this study integrates both demographic and academic features, offering a comprehensive clustering framework that not only identifies distinct dropout trajectories but also provides actionable insights tailored to the specific context of an industrial technical college in Libya. This approach enhances both the **methodological rigor** and the **practical relevance** of student dropout analysis.

This study addresses these gaps through two primary contributions:

**Integrated Feature Approach:** By combining demographic and academic indicators, the study develops a more holistic clustering model that better reflects the multifaceted nature of student dropout risk.

**Comparative Algorithmic Evaluation:** The research provides a comparative analysis of KMeans and Agglomerative clustering within the specific context of an industrial technical college, highlighting how different unsupervised learning approaches reveal complementary insights into student progression and dropout behaviors.

## 3. RESEARCH OBJECTIVES

This study aims to achieve the following objectives:

### *3.1 Implement unsupervised clustering techniques:*

K-means and agglomerative hierarchical clustering to categorize students at an industrial technical college into two distinct groups: those demonstrating steady academic progression and those at risk of early dropout.

### *3.2 Enhance classification robustness:*

by incorporating demographic variables (e.g., gender, age at admission, and enrollment year) alongside traditional academic performance metrics, thereby addressing the limitations of purely grade-based approaches.

3.3 Rigorously evaluate clustering outcomes:

Using multiple validity indices (Silhouette score, Davies-Bouldin index) to objectively determine the optimal number of clusters and ensure statistically meaningful student groupings.

### *3.4 Analyze the characteristics of each cluster:*

to derive actionable recommendations for tailored academic interventions, ultimately supporting data-driven strategies to mitigate dropout rates.

## 4. THEORETICAL BACKGROUND

### *4.1 Determining the Optimal Number of* Clusters

In this study, two well-established methods were employed to determine the optimal number of clusters: the Elbow Method and the Silhouette Analysis.

**A. Elbow Method**

The Elbow Method is based on evaluating the within-cluster sum of squares (WCSS), commonly referred to as the inertia. The inertia is calculated as follows:

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \quad \text{.....(1)}$$

Where:

k is the number of clusters,

Ci represents the set of points assigned to cluster ii,

μi is the centroid of cluster ii,

||x−μi||2 is the squared Euclidean distance between a data point x and the cluster centroid.

By plotting the WCSS values against various numbers of clusters, the "elbow point" can be identified as the point where additional clusters result in a diminished decrease in inertia. This point suggests an optimal balance between reducing intra-cluster distance and avoiding overfitting.

**B. Silhouette Analysis**

The Silhouette Score provides a quantitative measure of how well each object lies within its cluster. The score for a single point ii is computed as:

Where:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad \text{.....(2)}$$

a(i) is the mean intra-cluster distance of point i.

b(i) is the smallest mean distance of i to any other cluster to which it does not belong.

The overall silhouette score is the mean of all individual scores and ranges from -1 to 1. A higher score indicates well-separated and dense clusters.

### *4.2 . Clustering Algorithms*

**A. KMeans Clustering**

KMeans clustering is a widely used partitioning method that aims to partition n observations into kk clusters in which each observation belongs to the cluster with the nearest mean. The algorithm iteratively optimizes the following objective function:

$$\min \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \quad \text{.....(3)}$$

The key steps in KMeans include:

- Randomly initializing k centroids.
- Assigning each data point to the nearest centroid.
- Recomputing the centroids as the mean of the assigned data points.
- Repeating the process until convergence (no change in cluster assignments or centroids).

**B. Agglomerative Hierarchical Clustering**

Agglomerative clustering is a hierarchical method that builds clusters incrementally. Initially, each data point is treated as an individual cluster. Pairs of clusters are then merged iteratively based on a linkage criterion until the desired number of clusters is reached. The most common linkage methods include:

- Single Linkage: Minimum distance between points in two clusters.
- Complete Linkage: Maximum distance between points in two clusters.
- Average Linkage: Average distance between all pairs of points in two clusters.

The algorithm does not rely on centroids but instead on pairwise distances, and its hierarchical nature can be visualized as a dendrogram.

The mathematical formulation of the average linkage criterion is:

$$D(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} ||a - b|| \quad \text{.....(4)}$$

Where:

A and B are clusters,

||a−b|| is the distance between point a in cluster A and point b in cluster B.

## 5. METHODOGY

### *5.1 Dataset*

This study utilized academic and demographic data collected from the student information system of the College of Industrial Technology – Misrata. The dataset includes 1,642 records, representing all students who dropped out of the college between the academic years 1989–1990 and 2023–2024. During the same period, the total number of students admitted to the college was 3,180, indicating a dropout rate of approximately 51.6%. This comprehensive dataset provides an opportunity to analyze the dropout phenomenon over an extended historical period.

Before applying clustering algorithms, the dataset underwent several preprocessing steps to ensure the quality and consistency of the input features. First, the dataset was checked for missing, inconsistent, or erroneous values. Any incomplete records were handled through removal or appropriate imputation strategies to maintain data integrity. Then preprocessing steps included:

- Handling missing and inconsistent data via removal or imputation.
- Standardizing numerical features using z-score normalization.
- Encoding binary course outcomes (pass/fail) numerically.

Next, all numeric features were standardized using z-score normalization, which transforms each feature to have a mean of 0 and a standard deviation of 1. This step is essential to prevent features with larger scales from dominating the distance calculations used by clustering algorithms. Binary features, such as those representing the passing status of specific courses, were encoded as 0 for fail and 1 for pass to reflect their categorical nature in a numerical format.

The dataset comprises the following fields, each representing key academic or demographic factors related to student progression:

Gender: A binary variable where 1 indicates male students and 0 indicates female students.

Age at Admission: The age of the student at the time of enrollment, represented as a numeric value in years.

Department: A categorical variable indicating the student's academic department, encoded as numerical values (e.g., 0, 1, 2) for analysis purposes.

Semesters Attended: The total number of semesters the student remained enrolled before dropping out.

Passed Courses: The total number of courses successfully completed by the student during their enrollment period.

Math1, Physics1, Computer1: Binary variables (1 for pass, 0 for fail) representing whether the student passed each of the three key foundational courses in Mathematics, Physics, and Computer Science, respectively.

These variables were selected for their relevance in characterizing student academic engagement and performance, and for their potential role in differentiating patterns of academic persistence and dropout.

Determining the Optimal Number of Clusters

We employed two complementary validation approaches - the Elbow Method and Silhouette Analysis - to establish the optimal cluster count (k) for student segmentation. The Elbow Method evaluates within-cluster variance (inertia), identifying the inflection point where additional clusters provide minimal reduction in dispersion. Silhouette Analysis quantifies cluster separation quality, with values approaching 1 indicating well-defined groupings.

**Table1.** Comparative Metrics for k=2 through k=8.

| Clusters | Inertia | Silhouette Score |
|---|---|---|
| 2 | 7922.72 | 0.529 |
| 3 | 7292.76 | 0.512 |
| 4 | 6614.50 | 0.426 |
| 5 | 5607.90 | 0.431 |
| 6 | 4634.30 | 0.492 |
| 7 | 4182.39 | 0.498 |
| 8 | 3329.40 | 0.525 |

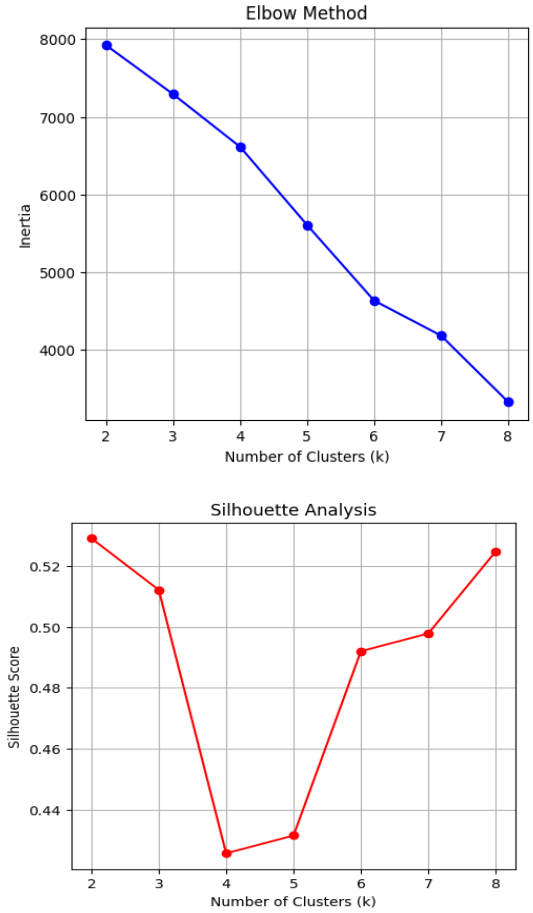


**Fig1.** Comparative Metrics for k=2 through k=8.

Table 1 and Fig 1 present comparative metrics for k=2 through k=8. While inertia monotonically decreases with increasing k (expected with finer partitioning), the silhouette coefficient peaks at k=2 (0.529). Although k=8 achieves marginally better inertia, its comparable silhouette score (0.525) doesn't justify the added complexity.

The k=2 solution optimally satisfies our primary research need to distinguish between:

**Early Dropouts:**

Students who discontinued their studies during the initial phase of their academic journey

Showed limited engagement with the curriculum before withdrawing

Typically left before establishing significant academic momentum

**Late Dropouts:**

Students who demonstrated substantial academic progress before discontinuing

Successfully completed multiple semesters of study

Represent cases where non-academic factors led to discontinuation despite academic success

This parsimonious grouping maintains interpretability while demonstrating robust statistical separation, as evidenced by:

Clear elbow point at k=2 in variance plot

Maximum silhouette cohesion-separation balance

Pedagogical relevance for intervention targeting

### 5.2 Clustering Algorithms

In this study, two clustering algorithms were applied to categorize the dropout student data into distinct groups based on their academic and demographic attributes. Both algorithms are unsupervised learning techniques designed to uncover inherent patterns in unlabeled data. A general workflow of the clustering process is illustrated in Fig 2, outlining the main steps from data preparation to the interpretation of cluster results.

### A. K-Means Clustering

The K-Means algorithm is a partitioning method that divides the dataset into k distinct, non-overlapping clusters. The process starts by

initializing k centroids, followed by iterative refinement through the following steps:

Assignment step: Each data point is assigned to the nearest centroid based on the Euclidean distance, forming preliminary clusters.

Update step: The centroids are recalculated as the mean of all data points assigned to each cluster.

These steps are repeated until convergence, either when the assignments no longer change significantly or when a maximum number of iterations is reached.

K-Means is computationally efficient and suitable for spherical, equally sized clusters, making it appropriate for high-dimensional numeric datasets such as the one used in this study. The number of clusters (k) was predetermined based on the clustering quality metrics discussed previously.

### B. Agglomerative Clustering

Agglomerative Clustering is a hierarchical clustering technique that builds nested clusters through a bottom-up approach. The algorithm begins by treating each data point as an individual cluster and successively merges the closest pair of clusters based on a distance metric, until the desired number of clusters is reached. The main steps include:

Computing the distance matrix: Initially, the Euclidean distances between all individual data points are calculated.

Merging clusters: At each iteration, the two clusters with the smallest inter-cluster distance are merged according to a specified linkage criterion (e.g., Ward's method, average linkage).

Stopping criterion: The merging process continues until the number of clusters matches the predefined k.

Agglomerative Clustering provides flexibility in capturing various cluster shapes and hierarchies, offering an alternative perspective to the flat partitions produced by K-Means.

Both algorithms were applied to the standardized dataset using the optimal number of clusters determined in the previous section. The comparison of their results enables a deeper understanding of the underlying student groups and validates the robustness of the clustering process.

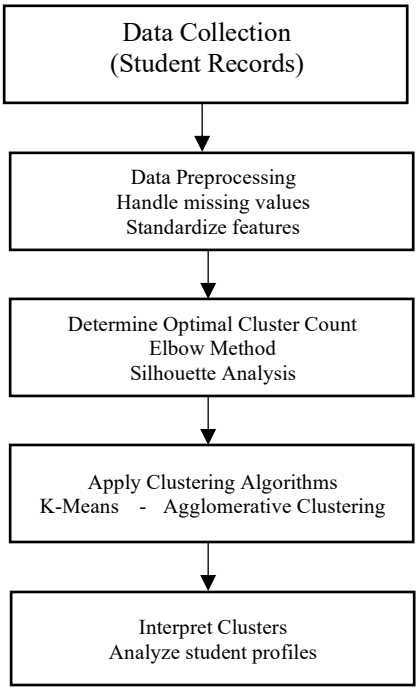


**Fig 2**. Main steps from data preparation to the interpretation of cluster results.

### Experimental Setup

The experimental implementation of this study was conducted using the Python programming language due to its flexibility and efficiency in handling data analysis and machine learning tasks. The entire workflow, including data preprocessing, clustering, and result visualization, was programmed from scratch without the use of pre-built clustering libraries.

The student dataset was first read from an Excel file and then processed to extract the relevant features for clustering. The clustering algorithms were implemented manually following their mathematical formulations. Once the algorithms completed the clustering process, the results including the assigned

clusters and performance metrics were output for further analysis.

## 6. RESULTS

This study applied two clustering algorithms KMeans and Agglomerative Clustering to categorize dropout students into two distinct groups. The goal was to identify patterns distinguishing students who dropped out early from those who persisted longer in the academic system.

To simplify the presentation of clustering results, the original feature names in the tables (e.g., Gender, Age at Admission, Department) are replaced with numerical labels (Feature 1, Feature 2, etc.). Below is the mapping of features to their corresponding labels:

Feature 1 (F1): Gender

Feature 2 (F2): Age at Admission

Feature 3 (F3): Department

Feature 4 (F4): Semesters Attended

Feature 5 (F5): Passed Courses

Feature 6 (F6): Math1 Success Rate

Feature 7(F7): Physics1 Success Rate

Feature 8 (F8): Computer1 Success Rate

### *6.1 KMeans Clustering Results*

The KMeans algorithm divided students into two clusters, summarized in Table2:

Cluster 0: Represents students with lower academic progress, characterized by fewer semesters attended (1.42), fewer passed courses (1.23), and very low success rates in core subjects (e.g., Math1: 0.01, Physics1: 0.06, Computer1: 0.06).

Cluster 1: Includes students who persisted longer, averaging 4.47 semesters attended, 11.57 passed courses, and significantly higher success rates in core subjects (Math1: 0.77, Physics1: 0.91, Computer1: 0.91).

Gender distribution also differed: Cluster 0 had ~89% male students, while Cluster 1 had ~72%.

**Table 2.** Summary Statistics of KMeans Clustering.

| Features | Cluster | |
|---|---|---|
| | **0** | **1** |
| F1 | 0.887 | 0.726 |
| F2 | 20.89 | 20.34 |
| F3 | 0.04 | 0.36 |
| F4 | 1.42 | 4.47 |
| F5 | 1.23 | 11.57 |
| F6 | 0.01 | 0.77 |
| F7 | 0.06 | 0.91 |
| F8 | 0.06 | 0.91 |

Principal Component Analysis PCA visualization (Figure 3) shows clear separation between the two clusters in a reduced 2D feature space. PCA is a dimensionality reduction technique that transforms high-dimensional data into fewer components while retaining maximum variance, helping visualize complex datasets.

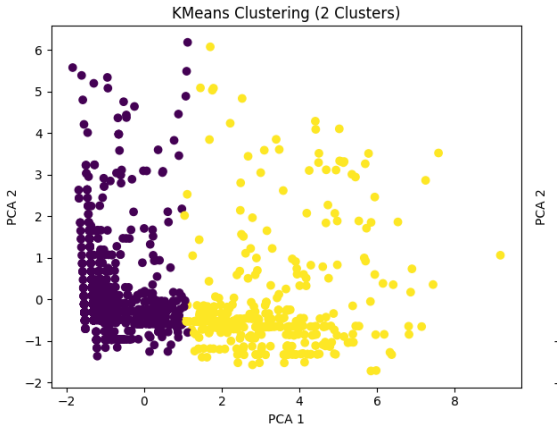


**Fig 3.** Separation between the two clusters for KMeans.

### *6.2 Agglomerative Clustering Results*

Similarly, Agglomerative Clustering produced two groups (Table 2):

Cluster 0: Students with higher academic performance (4.20 semesters, 10.96 passed courses, and strong success rates in core subjects).

Cluster 1: Students who dropped out early, with lower achievements (1.50 semesters, 1.42 passed courses, and near-zero success in Math1).

Gender distribution showed Cluster 0 had ~68% males, while Cluster 1 had ~90%.

**Table 3**. Summary Statistics of Agglomerative Clustering.

| Features | Cluster | |
|---|---|---|
| | 0 | 1 |
| F1 | 0.678 | 0.901 |
| F2 | 20.93 | 20.71 |
| F3 | 0.48 | 0.005 |
| F4 | 4.20 | 1.50 |
| F5 | 10.96 | 1.42 |
| F6 | 0.82 | 0 |
| F7 | 0.79 | 0.095 |
| F8 | 0.82 | 0.083 |

The PCA scatter plot (Fig 4) also confirmed distinct groupings, validating clustering effectiveness.

### *6.3 Comparative Analysis*

Both algorithms produced clusters with similar patterns in academic performance and gender distribution, but key differences emerged:

#### *A. Cluster Balance:*

KMeans formed more balanced clusters in size and feature distribution.

Agglomerative Clustering slightly prioritized hierarchical relationships, leading to a more uneven split (e.g., Cluster 1 in Agglomerative had near-zero Math1 success, unlike KMeans).

#### *B. Gender Differences:*

KMeans showed a moderate male majority (72%) in the persistent cluster.

Agglomerative had a lower male proportion (68%), suggesting slight variations in gender-based grouping.

#### C. Algorithm Sensitivity:

KMeans is distance-based, making it more sensitive to outliers.

Agglomerative relies on linkage methods, potentially capturing deeper hierarchical structures (e.g., departmental influences).

These differences highlight how algorithm choice impacts cluster interpretation, with KMeans favoring balanced partitions and Agglomerative reflecting underlying data hierarchies.

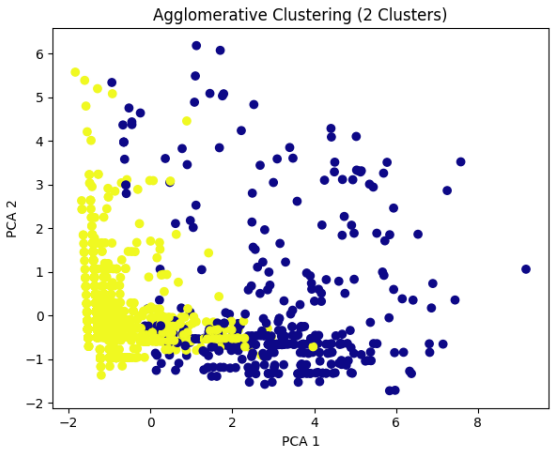


**Fig 4.** Separation between the two clusters for Agglomerative.

## 7. Results Discussion

The results reveal that student dropout from the college does not follow a uniform pattern, but rather divides into two distinct trajectories: early dropout, characterized by limited academic

engagement from the outset, and late dropout, which occurs after students have achieved some degree of academic progress. It is noteworthy that demographic factors, particularly gender and age at enrollment, played significant roles in distinguishing between these two groups. The early dropout group included a higher proportion of male students and younger enrollees. In contrast, the late dropout group showed greater gender diversity and a slightly higher average age at enrollment. This variation underscores the need to develop tailored intervention strategies that address the specific challenges of each group.

Students who drop out early demonstrate immediate difficulties in adapting to the academic environment, as evidenced by poor performance in core courses (Mathematics 1, Physics 1, and Computer 1), along with a low number of completed semesters. These indicators highlight the necessity of providing proactive support from the first semester, including academic guidance, psychological counseling, and remedial programs focused on foundational subjects. Early intervention in these issues may help prevent disengagement and enhance retention.

On the other hand, students who drop out after several semesters initially show strong academic performance, successfully transitioning to specialized departments. Their eventual withdrawal may be attributed to more complex factors, such as difficulties in keeping up with advanced coursework, declining motivation, or external socioeconomic pressures. For this group, support interventions might include intensive academic advising during the specialization phase, flexible study pathways, and career counseling services to promote persistence.

The study also emphasizes the critical role of student performance in core courses as an early warning indicator. Poor performance in Mathematics 1, Physics 1, and Computer 1 is clearly associated with early dropout, confirming the importance of monitoring these courses as part of an early warning system for predicting dropout risks.

Overall, these findings support the implementation of data-driven early warning systems that continuously monitor student performance and engagement metrics. Such systems, leveraging predictive analytics and artificial intelligence, can enable academic institutions to proactively identify at-risk students and implement targeted interventions in a timely manner before dropout occurs.

## REFERENCES

[1] Smith J, Brown A, Johnson L. Clustering student academic records for dropout prediction. *J Educ Data Min*. 2017;9(2):45-60.
[2] Lee S, Kim H. Incorporating demographic and engagement features in dropout classification. *Int J Educ Technol*. 2020;15(4):210-225.
[3] Martínez R, Torres M. Hierarchical clustering for heterogeneous student groups in education data. *Comput Educ*. 2021;158:104001.
[4] Zhao Y, Chen W, Li F. Hybrid clustering methods for enhanced educational data analysis. *IEEE Trans Learn Technol*. 2022;15(1):1-12.
[5] Nguyen T, Singh P, Patel D. Explainable AI techniques for educational clustering. *ACM Trans Interact Intell Syst*. 2023;13(2):18.
[6] Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20:53-65.
[7] Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell*. 1979;PAMI-1(2):224-227.
[8] Wang H, Zhang M, Liu S. Comparative study of clustering algorithms for educational data mining. *Expert Syst Appl*. 2024;198:116736.