Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 425−438
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10265

Comparative Study of Agglomerative Hierarchical Clustering and K-Means for Student Academic Stress Grouping

Irfan Arifin, Iwan Iskandar*, Elvia Budianita, Novi Yanti, Fitri Insani

Faculty of Science and Technology, Informatics Engineering, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
Email: [1]12250113743@students.uin-suska.ac.id, [2,*]iwan.iskandar@uin-suska.ac.id, [3]elvia.budianita@uin-suska.ac.id, [4]novi_yanti@uin-suska.ac.id, [5]fitri.insani@uin-suska.ac.id
Correspondence Author Email: iwan.iskandar@uin-suska.ac.id
Submitted: **10/06/2026**; Accepted: **29/06/2026**; Published**: 30/06/2026**

Abstract−Academic stress is a common problem experienced by college students due to high academic demands, parental expectations, and social pressures during their college years. The high levels of academic stress experienced by students underscore the need for a data-driven approach to more accurately identify and map students' stress levels. This research aims to compare the performance of the Agglomerative Hierarchical Clustering (AHC) and K-Means methods in clustering students' academic stress levels and to determine which method produces the best clustering quality. Data were obtained from the distribution of the Perception of Academic Stress Scale (PAS) questionnaire, consisting of 18 statement items, with 361 valid respondents from the Informatics Engineering Program at UIN SUSKA Riau, class of 2022–2025. The selection of the best linkage method in AHC was performed using the Cophentic Correlation Coefficient (CCC), where Ward Linkage was selected with the highest CCC value of 0.8180. Comparative evaluation was conducted using the Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index for variations in the number of clusters from K=2 to K=7. The test results showed that AHC Ward Linkage with K=2 was the best configuration with a Silhouette Coefficient of 0.4407 and a Davies-Bouldin Index of 0.8373, outperforming K-Means, which only excelled in the Calinski-Harabasz Index with a value of 419.7405 The clustering resulted in two clusters: High Stress with 244 students (67.6%) and Low Stress with 117 students (32.4%). The 2023 and 2024 cohorts had the highest proportions of high stress at 90.4% and 90.6%, respectively. This research contributes empirical evidence comparing hierarchy-based and partition-based clustering methods for academic stress data, while also demonstrating the use of the Cophenetic Correlation Coefficient as an objective basis for linkage method selection in AHC. It is hoped that the results of this study can serve as a basis for the institution in designing targeted mental health intervention programs for students.

Keywords: Academic Stress; Agglomerative Hierarchical Clustering; K-Means; Perception of Academic Stress Scale; University Students

1. INTRODUCTION

College is a stage of education that requires students to adapt to various challenges, both academic and non-academic. During this period, students are generally in the transitional age range from late adolescence to early adulthood a developmental stage that is inherently vulnerable to emotional stress, given the complexity of the social, academic, and personal issues they face simultaneously [1]. Throughout their academic journey, students often face various sources of stress, ranging from a heavy course load and expectations from their parents to pressure stemming from their social environment and peer groups [2]. Among these various sources of pressure, academic issues are believed to be the most common cause of stress among college students, with study-related activities such as completing assignments, preparing for exams, and competing for grades being the primary triggers in their daily lives [3]. If these pressures are not managed effectively, this situation has the potential to develop into academic stress a condition in which students experience psychological pressure characterized by various physical and emotional responses resulting from academic demands placed on them by professors and parents, as well as the need to complete assignments on time [4]. Furthermore, when students experience stress that exceeds their tolerance threshold, they tend to experience physiological and psychological effects such as chronic fatigue, loss of appetite, headaches, and digestive problems. In fact, academic stress has been empirically linked to a range of broader negative consequences, including health problems, anxiety, depression, and a decline in academic performance [3].

Various research findings also indicate that academic stress is quite common among college students. In the context of online learning, for example, it was found that the largest proportion of students 39.2% experienced moderate academic stress, while the remainder experienced stress ranging from high to very high [5]. In particular, students majoring in Informatics Engineering and other engineering fields tend to experience higher levels of academic stress compared to students in other majors. This is due to a heavy academic workload, a large volume of assignments, and complex course material [6]. On a global scale, data compiled by the Mental Health Foundation in the UK reveals that approximately 60% of individuals aged 18–24 report high levels of stress [7]. This situation underscores that academic stress among students articularly those in the Informatics Engineering program is not merely an individual issue, but rather a systemic phenomenon that requires serious attention, including through data-driven approaches to identify and map students' circumstances in a more precise and structured manner.

In this context, data mining technology particularly clustering techniques has emerged as a proven approach for extracting meaningful information from student data. By definition, clustering is the process of grouping a set of data objects into distinct clusters based on shared characteristics among the objects [8]. Previous research has explored the application of various clustering algorithms in mapping students' academic stress levels. In an investigation by

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 425−438
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10265

Wiranti et al., the K-Means algorithm was applied to categorize 507 students based on their academic stress levels using the Perception of Academic Stress Scale (PAS) questionnaire. The evaluation results, with a Davies-Bouldin Index (DBI) of 1.43 and a Silhouette Coefficient of 0.27, yielded two groups: 229 students with low stress levels and 278 students with high stress levels [7]. A subsequent analysis by Alfaiza et al. examined a similar topic by applying the Fuzzy C-Means (FCM) algorithm to students in the Faculty of Science and Technology, resulting in two optimal clusters with a Silhouette Coefficient of 0.3056, which divided the students into 313 individuals in the high-stress category and 274 individuals in the low-stress category [9]. Further work by Wijaya et al. combined the K-Means and K-Modes algorithms to cluster stressors among current college students, taking into account external variables such as environmental, financial, family, friendship, romantic, career, and hobby-related factors [10]. In their paper, Trifani et al. applied the C4.5 classification to determine the stress levels of senior students based on structured academic and non-academic attributes [11].

Although this prior research has made significant contributions, direct comparisons between hierarchy-based methods such as AHC and partition-based methods such as K-Means in the context of clustering students' academic stress have not yet been widely conducted. K-Means is known to be efficient and easy to implement, but it has the drawbacks of being sensitive to outliers and requiring the number of clusters (k) to be specified at the start of the analysis [12]. On the other hand, AHC operates in a bottom-up manner without requiring the specification of k at the outset and produces dendrogram visualizations that facilitate the interpretation of the data structure [13], [14]. The work by Sujjada et al. combined AHC with K-Means to cluster data on people with disabilities across 7 provinces in Indonesia, where AHC served as the initial centroid determiner for K-Means and produced three main clusters with the best Davies-Bouldin Index (DBI) of 0.662 for two of the clusters [15]. An evaluation by Abdulpatah et al., which compared K-Means and AHC for clustering rice-producing regions in Indonesia, showed that AHC using the Average Linkage method produced more optimal performance compared to K -Means, with a Silhouette Coefficient of 0.723 and a Davies-Bouldin Index of 0.229, while K-Means achieved a Silhouette Coefficient of 0.696 and a Davies-Bouldin Index of 0.404 [16]. Research by Husna et al., comparing AHC with K-Medoids on data regarding the types of diseases among inpatients, also demonstrated the superiority of AHC, where AHC achieved an average Silhouette Coefficient of 0.5837, far superior to K-Medoids, which only achieved a value of -0.3558 [17]. Similar findings were obtained in a publication by Tjipta et al., which compared K-Means++ and AHC for clustering healthcare workers across 38 provinces in Indonesia. AHC outperformed K-Means++ with a Silhouette Score of 0.550 and a Davies-Bouldin Index of 0.457, although K-Means++ scored higher on the Calinski-Harabasz Index [18].

This research utilized questionnaire data collected from students in the Informatics Engineering Program at Sultan Syarif Kasim Riau State Islamic University, class of 2022–2025, using the Perception of Academic Stress Scale (PAS), which consists of 18 items. This instrument was developed by Bedewy and Gabriel and includes three main subscales: academic expectations (4 items), academic load (8 items), and academic self-perception (6 items), with a reliability coefficient of 0.7 [19]. The main objective of this research was to compare the performance of the Agglomerative Hierarchical Clustering (AHC) and K-Means methods in clustering students based on their level of academic stress, as well as to determine which method produced the best cluster quality based on the evaluation metrics of the Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index. The contribution of this research lies in providing empirical evidence on the comparative performance of a hierarchy-based method (AHC) and a partition-based method (K-Means) in the context of academic stress clustering, an aspect that has rarely been explored in previous research. In addition, this research introduces the use of the Cophenetic Correlation Coefficient (CCC) as an objective basis for selecting the optimal linkage method prior to comparison with K-Means, thereby offering a more rigorous and reproducible methodological framework for future clustering research involving academic stress data.

2. RESEARCH METHODOLOGY

This section outlines the systematic procedures followed in conducting the research to address the research questions. The overall research process is presented in Figure 1.

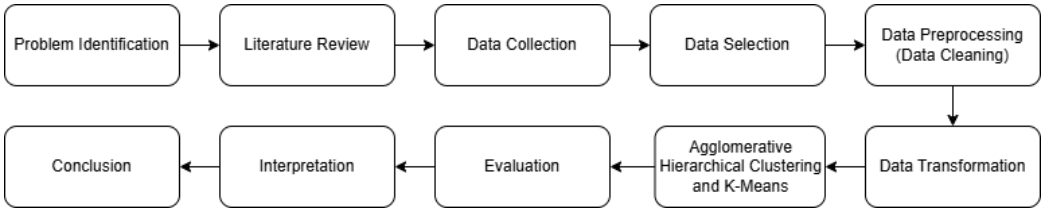

Figure 1. Research flowchart

2.1 Problem Identification

The first step in this research was to identify and define the focus of the issue to be examined. Academic stress was chosen as the primary focus because it is directly related to the learning environment and can be systematically

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 425−438
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10265

identified, making it more relevant as a basis for academic policy-making compared to non-academic stress, which is personal in nature and difficult to standardize. The Informatics Engineering Program was chosen because it is an engineering discipline known for its heavy academic workload. According to Jensen et al., engineering students consistently identify engineering workload as the primary stressor distinguishing their major from other disciplines, which includes a high volume of assignments, complex material, and intense time pressure [6]. These factors inherently influence students' levels of academic stress. Based on these conditions, this research focuses on comparing the performance of the Agglomerative Hierarchical Clustering (AHC) and K-Means methods in clustering the academic stress levels of students in the Informatics Engineering Program at UIN SUSKA Riau.

2.2 Literature Review

A literature review is a crucial stage in research that involves an in-depth examination of the issues that form the focus of the research. At this stage, the researcher systematically reviews various relevant sources of literature, including scientific journals, books, and previous studies. This process aims to establish a strong theoretical foundation to understand and address the issues raised in the research [20]. The literature review in this research encompasses four main aspects. First, theories regarding academic stress, including definitions and basic concepts, its triggering factors both internal and external as well as its impact on students' mental health and academic achievement. Second, a review of the Agglomerative Hierarchical Clustering (AHC) method, which covers the basic principles of bottom-up clustering, the mechanism for calculating distances using Euclidean Distance, the cluster merging method (linkage method), and how to read and interpret the results of dendrogram visualizations. Third, a study of the K-Means method, covering the basic principles of partition-based clustering, the centroid initialization mechanism, and the iterative process of cluster assignment and updating. Fourth, a comparative study of the two methods, covering fundamental differences in how they work, initial parameter requirements, sensitivity to outliers, and the evaluation metrics used to objectively measure and compare the quality of clustering results.

2.3 Data Collection

In this stage, students of the Informatics Engineering Program at Sultan Syarif Kasim Riau State Islamic University were designated as the research subjects. This research successfully collected data from 374 respondents (256 males and 118 females) across four cohorts: 2022 (83 students), 2023 (87 students), 2024 (91 students), and 2025 (113 students). Data collection was administered online via Google Forms. The measurement instrument used was an adaptation of the Perception of Academic Stress Scale (PAS) developed by Bedewy and Gabriel [19]. This instrument comprises three main subscales: academic expectations (4 items), academic load (8 items), and academic self-perception (6 items), totaling 18 statements, as detailed in Table 1. The instrument uses a five-point Likert scale response format, with the following options: Strongly Disagree (SD), Disagree (D), Neutral (N), Agree (A), and Strongly Agree (SA). The items in this instrument are divided into two types: Favorable (F) statements, which align with the construct being measured, and Unfavorable (UF) statements, which are reverse-worded [21].

Table 1. Kuesioner PAS

Instruments	SD	D	N	A	SA
Academic Expectations					
1. Competition with classmates to get good grades is quite intense (F)					
2. The Lecturer are critical of my academic performance (F)					
3. The Lecturer have unrealistic expectations of me (F)					
4. My parents' unrealistic expectations cause me stress (F)					
Academic Load					
5. The time allocated for classes and academic assignments is sufficient (UF)					
6. The curriculum load (course material) is too heavy (F)					
7. I feel the number of assignments given is too high (F)					
8. I struggle to catch up if I'm late on assignments (F)					
9. I have enough time to rest after studying (UF)					
10. Exam questions are usually difficult (F)					
11. Exam time feels too short to complete all the answers (F)					
12. Exam periods are very stressful for me (F)					
Academic Self-Perception					
13. I am confident that I will be a successful student (UF)					
14. I am confident that I will succeed in my future career (UF)					
15. I can make academic decisions easily (UF)					
16. I am afraid of failing a course this semester (F)					
17. I consider my anxiety about exams to be a personal weakness (F)					
18. Even if I pass the exams, I am still worried about my job prospects after graduation (F)					

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 425−438
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10265

2.4 Data Selection

The data selection process involves determining the parameters or attributes to be used from the collected dataset for subsequent steps. Only the eighteen columns of the statement were used as analysis variables. Respondent demographic data were used separately for the purpose of describing respondent profiles. After the selection stage, the data were ready for use in the preprocessing and clustering stages using the Agglomerative Hierarchical Clustering and K-Means methods. Table 2 shows the data used as the basis for the study. In the table, the labels S1 through S18 in the column section represent the 18 statements from the Perception of Academic Stress Scale (PAS) instrument that were answered by the respondents.

Table 2. Data Selection

No	S1(F)	S2(F)	S3(F)	S4(F)	S5(UF)	S6(F)	….	S18(F)
1	Strongly Agree	Neutral	Agree	Disagree	Neutral	Agree	….	Disagree
2	Strongly Agree	Agree	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree	….	Strongly Agree
3	Disagree	Neutral	Agree	Neutral	Neutral	Neutral	….	Agree
….	….	….	….	….	….	….	….	….
372	Disagree	Disagree	Disagree	Neutral	Strongly Agree	Disagree	….	Neutral
373	Neutral	Neutral	Disagree	Agree	Disagree	Strongly Agree	….	Strongly Agree
374	Disagree	Disagree	Strongly Disagree	Disagree	Agree	Disagree	….	Strongly Disagree

2.5 Data Preprocessing (Data Cleaning)

Data preprocessing is a crucial step prior to further analysis. In this research, preprocessing was limited to data cleaning, which involves examining the data to identify and remove duplicate entries in the Student ID (NIM) column. Of the total 374 respondents who completed the questionnaire, 361 clean data records were obtained following the cleaning process. This step was performed to ensure the data used is clean and does not affect the accuracy and validity of the clustering results [22].

2.6 Data Transformation

Data transformation is the process of converting questionnaire data into a numerical format and standardizing the data if necessary. This is important to ensure that each variable has an equivalent scale when calculating the distance between data points using Euclidean Distance [23]. The conversion process was carried out by converting text-based response categories into numerical values. The scoring scheme was differentiated based on statement type, with Favorable (F) statements presented in Table 3 and Unfavorable (UF) statements presented in Table 4. Implementing this reverse-scoring method ensures that the dataset maintains consistent directionality, where higher numerical values universally indicate elevated levels of academic stress. This uniformity is critically important to prevent mathematical bias during the subsequent clustering iterations of both the AHC and K-Means algorithms

Table 3. Data Transformation for Favorable (F)

Answer	Abbreviation	F
Strongly Disagree	SD	1
Disagree	D	2
Neutral	N	3
Agree	A	4
Strongly Agree	SA	5

Table 4. Data Transformation for Unfavorable (UF)

Answer	Abbreviation	UF
Strongly Disagree	SD	5
Disagree	D	4
Neutral	N	3
Agree	A	2
Strongly Agree	SA	1

Table 5 presents the transformed data, which is ready to be used as input for the Agglomerative Hierarchical Clustering and K-Means methods. Using this standardized matrix, the subsequent phase will evaluate and compare how effectively each algorithm groups the students based on their stress characteristics.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 425−438
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10265

Table 5. Data Transformation

No	S1(F)	S2(F)	S3(F)	S4(F)	S5(UF)	S6(F)	….	S18(F)
1	5	3	4	2	3	4	….	2
2	5	4	5	5	1	5	….	5
3	2	3	4	3	3	3	….	4
….	….	….	….	….	….	….	….	….
359	2	2	2	3	1	2	….	3
360	3	3	2	4	4	5	….	5
361	2	2	1	2	2	2	….	1

2.7 Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering (AHC) is a bottom-up hierarchical clustering method that treats each data point as an independent cluster in the initial stage, then gradually merges the nearest clusters based on Euclidean distance and the linkage method until a single large cluster is formed. The clustering results are visualized in the form of a dendrogram [24], [25]. The overall workflow of this method is presented in Figure 2.

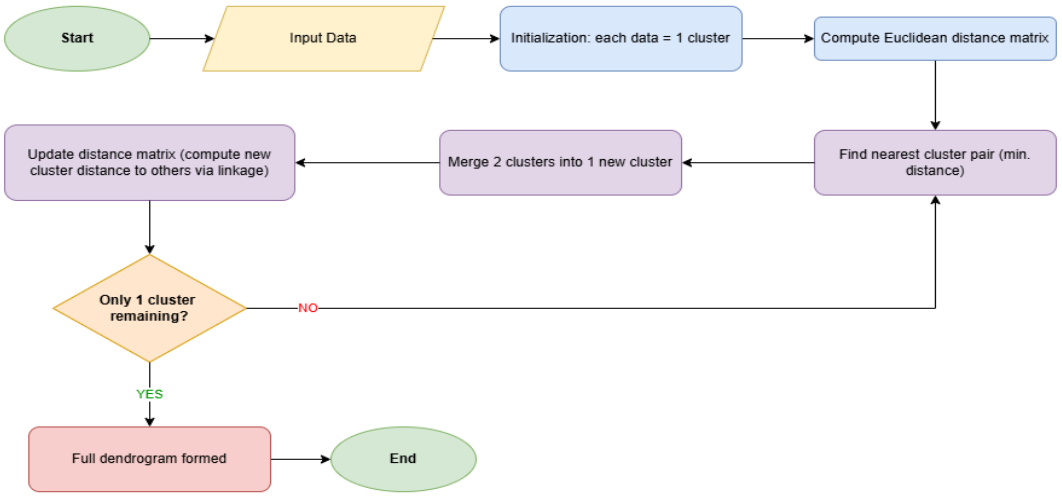

Figure 2. Flowchart of the Agglomerative Hierarchical Clustering Method

Operationally, the main steps in the Agglomerative Hierarchical Clustering (AHC) algorithm are as follows:

a. Cluster Initialization
Set each data object as a single cluster, so that if there are N data objects, there will be N clusters in the initial stage.

b. Calculating Distances Between Objects (Distance Matrix)
Calculating the distances between data points to determine the proximity of each object using the Euclidean Distance formula as follows [1]:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

The Euclidean Distance formula is used to calculate the distance between two data objects based on their attribute values. In this formula, $d(x,y)$represents the distance between object x and object y, $x_i$and y_i represent the i-th attribute values of objects x and y respectively, while n denotes the number of data variables (attributes) involved in the calculation.

c. Cluster Merging
Identify the two clusters with the closest distance or highest similarity in the distance matrix, then merge these two clusters into a single new cluster [26]. Recalculate the distances between the newly formed cluster and the remaining clusters using the linkage method [14]. The new distances are determined based on the selected method, including:

1. Single Linkage: The shortest distance between objects.

$$d_{(UV)W} = \min\{d_{UW}, d_{VW}\} \quad (2)$$

2. Average Linkage: The average distance of all members.

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W} \quad (3)$$

3. Complete Linkage: The greatest distance between objects.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 425−438
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10265

$$d_{(UV)W} = max\{d_{UW}, d_{VW}\} \quad (4)$$

4. Ward's Method: Minimize the sum of squared errors (SSE).

$$SSE = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (5)$$

d. Iteration
Repeat steps 3 and 4, reducing the number of clusters by one at each iteration, until all data is merged into a single cluster.

e. Dendrogram Visualization
The entire process of hierarchical data merging is recorded and visualized in the form of a dendrogram. This dendrogram illustrates the sequence of cluster merging along with the distances between clusters at each stage. The optimal number of clusters is determined by cutting the dendrogram at the level where there is a sufficiently large vertical gap before the next cluster merger, resulting in more homogeneous and clearly distinct groups [27].

2.8 K-Means

K-Means is a partition-based clustering method. This method works by grouping data objects based on the mean value of each cluster, where the number of clusters (k) must be specified in advance by the user as an initial parameter. The clustering process proceeds iteratively by assigning each object to the nearest centroid, followed by updating the centroid's position based on the cluster members' mean, until convergence is achieved. However, this method has several limitations: it is prone to converging on local minima, sensitive to outliers, and can only optimally detect spherical clusters due to the use of the Euclidean metric. Despite these limitations, K-Means is still recognized for its flexibility, efficiency, and ease of implementation, making it one of the top ten clustering methods in data mining[28]. The systematic workflow of the method is illustrated as a flowchart in Figure 3.

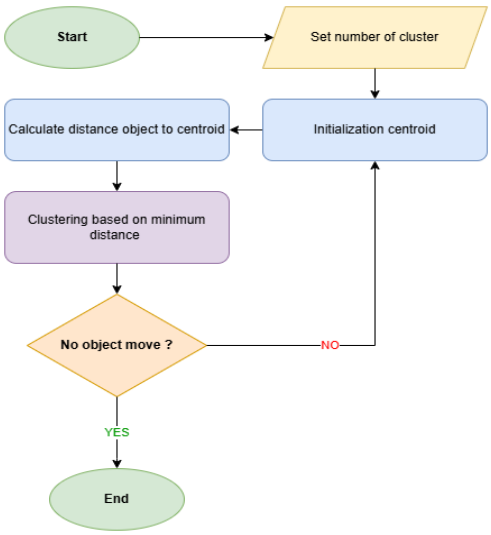

Figure 3. *Flowchart* Metode *K-Means*

Operationally, the main steps in the K-Means algorithm are as follows [29]:

a. Determining the number k and Initializing the Centroids
Determine the number k and randomly select initial centroids from the dataset, where each centroid represents the center of a cluster[30].

b. Calculating the distance between objects
Calculating the distance of each data object from all existing centroids using the Euclidean Distance equation as shown in formula (1)

c. Assigning Data to Clusters
Assigning each object to the cluster with the nearest centroid.

d. Centroid Update
After all data points have been clustered, the position of each cluster's centroid is recalculated based on the average value of all members of that cluster.

e. Iteration
Repeating steps (b) and (c) iteratively until the centroid positions no longer change, indicating that the algorithm has reached convergence and the clustering is stable.

2.9 Evaluation

a. Cophenetic Correlation Coefficient (CCC)

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 425−438
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10265

The Cophenetic Correlation Coefficient (CCC) is a coefficient calculated to evaluate the agreement between the original data distances and the distance measures generated by the clustering process. Specifically, the CCC measures the extent to which a dendrogram is able to represent the pairwise distances of the original data distance matrix. This coefficient is calculated by comparing the original distance matrix between data objects with the cophenetic distance matrix, which is a matrix that replaces the original distances between objects with distances calculated based on the cluster merger positions on the dendrogram. A high CCC value indicates that the linkage method is the most accurate clustering and distance measurement criterion for the analyzed dataset [31], [32]. In this research, CCC was used as the basis for selecting the best linkage method from the four tested candidates : Single, Complete, Average, and Ward before the selected method was compared with K-Means in the comparative evaluation stage. The CCC equation is defined as follows[31]:

$$c = \frac{\sum_{i<j}(x(i,j)-x)(t(i,j)-t)}{\sqrt{\sum_{i<j}[x(i,j)-x]^2 \sum_{i<j}[t(i,j)-t]^2}} \quad (6)$$

The Cophenetic Correlation Coefficient is used to measure how faithfully a dendrogram preserves the pairwise distances between the original data points. In this formula, c represents the cophenetic correlation coefficient, $x(i,j)$ represents the Euclidean distance between objects i and j, and $t(i,j)$ represents the dendrogram distance between the same pair of objects as represented in the hierarchical clustering tree.

b. Silhouette Coefficient

This method measures how well each data point is grouped into the appropriate cluster by comparing intra-cluster cohesion (proximity within clusters) and inter-cluster separation. The Silhouette Coefficient ranges from -1 to 1, where a value close to 1 indicates dense and well-separated [33].

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (7)$$

The Silhouette Coefficient is used to evaluate the quality of clustering by measuring how well each data point fits within its assigned cluster compared to other clusters. In this formula, $s(i)$ represents the silhouette value for the i-th data point, $a(i)$ represents the average distance between the i-th data point and all other points in the same cluster, while $b(i)$ represents the average distance between the i-th data point and the points in the nearest neighboring cluster.

c. Davies-Bouldin index (DBI)

This index measures the average similarity between clusters by comparing intra-cluster dispersion to inter-cluster separation. The lower the DBI value, the better the clustering quality, as it indicates more compact and clearly separated clusters [34].

$$DBI = \frac{1}{k}\sum_{i=1}^{k} R_i \quad (8)$$

The Davies-Bouldin Index is used to evaluate clustering quality by measuring the average similarity ratio between each cluster and its most similar counterpart. In this formula, DBIrepresents the Davies-Bouldin Index value, $k$represents the number of clusters formed, and $R_i$represents the similarity ratio of the i-th cluster relative to its closest neighboring cluster.

d. Calinski-Harabasz index

In addition to the Silhouette Coefficient and the Davies-Bouldin Index, the quality of the results of Agglomerative Hierarchical Clustering is also evaluated using the Calinski-Harabasz Index (CHI). According to Chicco et al., although CHI is a popular metric, its performance is less consistent compared to the Silhouette Coefficient and the Davies-Bouldin Index in cases of convex clusters with two groups. However, CHI remains useful because it provides a different perspective on the ratio of between- and within-cluster variance [33].

$$CH = \frac{BCSS/(k-1)}{WCSS/(n-k)} \quad (9)$$

The Calinski-Harabasz Index is used to evaluate clustering quality by comparing the dispersion between clusters to the dispersion within clusters. In this formula, CH represents the Calinski-Harabasz Index value, $BCSS$ represents the Between-Cluster Sum of Squares which measures the separation between clusters, $WCSS$ represents the Within-Cluster Sum of Squares which measures the compactness of data points within each cluster, k represents the number of clusters formed, and n represents the total number of data points. A higher CH value indicates better clustering performance, as it reflects well-separated and compact clusters.

3. RESULT AND DISCUSSION

This chapter presents the results of the processing and analysis of data obtained from the study respondents. The cleaned and transformed data were analyzed using the Agglomerative Hierarchical Clustering (AHC) and K-Means methods to compare the quality of clustering among Informatics Engineering students based on their academic stress

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 425−438
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10265

patterns. The clustering results are then discussed in depth to derive interpretations and implications regarding students' academic stress conditions.

3.1 Characteristics of Respondents' Answers

Based on the distribution of respondent answers, the Class of 2022 exhibits a pattern dominated by the "Neutral" response for nearly all statements related to academic workload. This reflects an uncertain attitude, where pressure is beginning to be felt but has not yet solidified into a firm certainty. Indicators such as grade competition (S1), faculty expectations (S3), and assignment workload (S7) were mostly in the range of "Disagree" to "Neutral," indicating that routine academic workload has not yet been predominantly felt. On the other hand, for statements such as confidence in career success (S14) and stressful exams (S12), the percentage of responses leans strongly toward "Agree" and "Strongly Agree." This pattern indicates that exam pressure is beginning to feel tangible as the final stage of studies approaches.

Conversely, the Class of 2023 showed a more pronounced pattern toward "Agree" and "Strongly Agree" on indicators of curriculum workload (S6), fear of failing (S16), and concerns about job prospects (S18), likely driven by increasing awareness of the professional world. The Class of 2024 exhibits the most extreme pattern, with a dominance of "Strongly Agree" regarding curriculum workload, fear of failing, and stressful exams. However, this group most frequently answered "Neutral" regarding confidence in success as a student (S13), indicating that the high pressure is beginning to erode their self-confidence. This pattern stands in stark contrast to the Class of 2025, which is dominated by "Strongly Disagree" responses to stress-inducing statements, yet "Strongly Agree" on self-confidence and sufficient rest time (S5 and S9). This clearly reflects the condition of freshmen who have not yet experienced the peak of academic pressure, despite having a fairly wide range of individual variations.

3.2 Dendrogram Visualization Using the Agglomerative Hierarchical Clustering Method

The clustering process using the Agglomerative Hierarchical Clustering (AHC) method produced dendrogram visualizations for four linkage methods: Single, Complete, Average, and Ward, using the Euclidean distance metric. The results of these four dendrogram visualizations are shown in Figure 4.

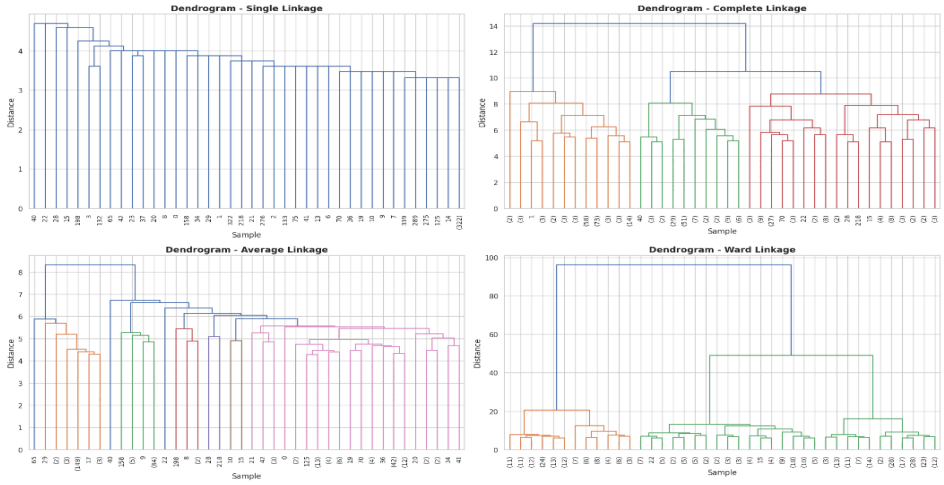

Figure 4. Visualisasi *Dendogram*

Based on Figure 4, each dendrogram illustrates the bottom-up cluster merging process of the data, starting from individual clusters and progressing to form a single large hierarchy. The vertical axis shows the Euclidean distance between the merged clusters, while the horizontal axis shows the sample index or the size of the resulting clusters.

The Single Linkage dendrogram shows the chaining effect, characterized by the sequential and gradual merging of data with very small distances (ranging from 0 to 4.5), resulting in an unbalanced hierarchical structure that is difficult to interpret. The Complete Linkage dendrogram produces a more balanced structure with the highest merging distance reaching 14, indicating a clearer separation between major groups. The Average Linkage dendrogram produces a fairly proportional hierarchical structure with the highest merging distance around 8.3, showing two main groups that form naturally. The Ward Linkage dendrogram produces the clearest and easiest-to-interpret structure, with the highest merge distance reaching nearly 100, where two main groups are visually separated by a very large distance, indicating the most distinct cluster separation compared to the other three methods.

3.3 Evaluation

3.3.1 Evaluation of Linkage Methods Using the Cophenetic Correlation Coefficient (CCC)

Before conducting a comparative evaluation between AHC and K-Means, the best linkage method was first determined from the four candidates tested using the Cophenetic Correlation Coefficient (CCC). The results of the CCC test for the four linkage methods are presented in Table 6.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 425−438
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10265

Table 6. Results of the Cophenetic Correlation Coefficient Evaluation

Linkage Method	Cophenetic Correlation Coefficient (CCC)
Single Linkage	0,4613
Average Linkage	0,7827
Complete Linkage	0,7848
Ward's Method	0,8180

Based on Table 6, the Ward Linkage method achieved the highest CCC value of 0.8180, followed by Complete Linkage (0.7848), Average Linkage (0.7827), and Single Linkage (0.4613). The highest CCC value for Ward Linkage indicates that the dendrogram generated by this method is the most accurate in representing the distance structure of the original data compared to the other three methods. Therefore, Ward Linkage was selected as the best AHC configuration and was subsequently used to represent the AHC method in the comparative evaluation with K-Means.

3.3.2 Comparative Evaluation of AHC (Ward Linkage) and K-Means

A comparative evaluation was conducted to determine which method AHC, Ward Linkage, or K-Means yields the best clustering quality. The test was conducted on a range of cluster numbers from K=2 to K=7 using three evaluation metrics: the Silhouette Coefficient, the Davies-Bouldin Index (DBI), and the Calinski-Harabasz Index (CHI). The complete results are presented in Table 7.

Table 7. Test Result

Method	K	Silhouette Coefficient	DBI	CHI
AHC (Ward)	2	0.4407	0.8373	386.2701
AHC (Ward)	3	0.3556	1.2373	336.8808
AHC (Ward)	4	0.2969	1.7922	249.2395
AHC (Ward)	5	0.1657	2.1135	199.2958
AHC (Ward)	6	0.1589	2.4315	166.3327
AHC (Ward)	7	0.1555	2.2866	144.1472
K-Means	2	0.4396	0.8888	419.7405
K-Means	3	0.3712	1.1748	353.0966
K-Means	4	0.2367	1.7534	254.7183
K-Means	5	0.1936	2.1586	204.5796
K-Means	6	0.1728	2.2421	172.0275
K-Means	7	0.1373	2.3414	148.5524

The best model was selected based on three evaluation metrics: the highest Silhouette Coefficient (close to 1), the lowest DBI (close to 0), and the highest CHI. The results of the comparison of the best models based on each metric are presented in Table 8.

Table 8. Comparison of the Best Models Based on Each Metric

Evaluation Metric	Linkage Method	K	Value
Silhouette Coefficient	AHC (Ward Linkage)	2	0.4407
Davies-Bouldin Index	AHC (Ward Linkage)	2	0.8373
Calinski-Harabasz Index	K-Means	2	419.7405

Based on Table 8, there are differences in performance between the two methods across each metric. AHC with Ward Linkage at K=2 outperforms the other method on two of the three metrics, namely the highest Silhouette Coefficient of 0.4407 and the lowest Davies-Bouldin Index of 0.8373, indicating that the clusters produced by AHC are more compact and exhibit better separation between clusters. Meanwhile, K-Means with K=2 achieved the highest Calinski-Harabasz Index value of 419.7405, indicating a higher ratio of inter-cluster variance to intra-cluster variance.

Considering all three evaluation metrics as a whole, the AHC (Ward Linkage) method with K=2 was selected as the best and used for interpreting the clustering results. This method yields the highest Silhouette Coefficient value of 0.4407, followed by a fairly good Davies-Bouldin Index (DBI) value of 0.8373, and a Calinski-Harabasz Index (CHI) value of 386.2701. A Silhouette Coefficient value of 0.4407 falls into the "weak cluster" category [35]. This low value is most likely influenced by the large number of respondents who selected "neutral" answers on the questionnaire, resulting in overlap between clusters and less distinct separation between groups.

3.4 Interpretation

Based on the results of clustering using the AHC (Ward Linkage) method with K=2, all 361 student respondents from the Informatics Engineering Program at UIN SUSKA Riau, class of 2022–2025, were successfully grouped into two clusters. A visualization of the data point distribution using Principal Component Analysis (PCA) is shown in Figure 5.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 425−438
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10265

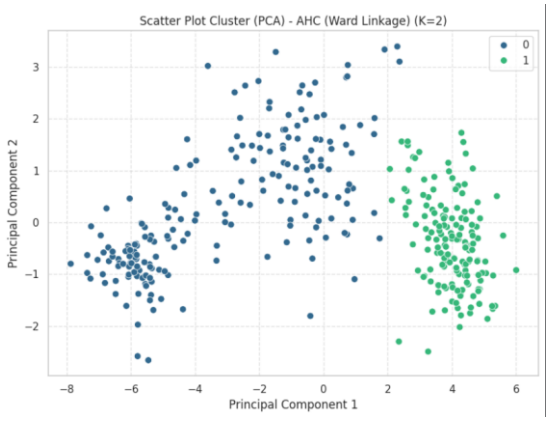

Figure 5. Cluster Distribution via PCA

Figure 5 shows a visualization of the distribution of data points resulting from clustering using AHC (Ward Linkage) with K=2, projected into two principal dimensions via Principal Component Analysis (PCA). It can be seen that the two clusters, namely cluster 0 (blue) and cluster 1 (green), form two fairly well-defined groups in the PCA projection space. Although there are some data points located in the overlapping region between the two clusters, in general, the separation between the groups is quite evident.

3.4.1 Average Scores per Questionnaire Item

A comparison of the average scores for each item of the PAS questionnaire for the two clusters resulting from clustering using the AHC (Ward Linkage) method is presented in Table 9.

Table 9. Average Score for Each Statement Item by Cluster Using the AHC (Ward Linkage) Method

Statement Item	Cluster 0	Cluster 1	Score Difference
S1(F)	3.04	1.58	1.46
S2(F)	3.38	1.85	1.53
S3(F)	3.25	1.73	1.52
S4(F)	3.52	2.32	1.20
S5(UF)	3.30	1.38	1.92
S6(F)	4.17	1.57	2.60
S7(F)	3.93	1.94	1.99
S8(F)	3.99	2.04	1.95
S9(UF)	3.08	1.44	1.64
S10(F)	4.07	1.66	2.41
S11(F)	3.97	1.98	1.99
S12(F)	4.43	2.30	2.13
S13(UF)	2.54	1.58	0.96
S14(UF)	2.20	1.48	0.72
S15(UF)	2.61	1.48	1.13
S16(F)	4.26	2.11	2.15
S17(F)	3.97	2.15	1.82
S18(F)	4.27	2.19	2.08

The “Score Difference” column in Table 9 shows that the items with the largest differences between the two clusters are S6(F) (difference of 2.60), S10(F) (difference of 2.41), S16 (F) (difference of 2.15), S12(F) (difference of 2.13), and S18(F) (difference of 2.08).

3.4.2 Analysis of Cluster Characteristics

Based on the mapping results in Table 9, there are significant differences in respondent answer patterns between Cluster 0 and Cluster 1.

a. Characteristics of Cluster 0
   In Cluster 0, the overall average scores for all items tend to be very high. The three items with the highest scores in this cluster are S12(F) (average 4.43), S18(F) (average 4.27), and S16(F) (average 4.26). The high scores on these items indicate the dominance of stress stemming from academic workload, exams, and expectations of academic performance. Other indicators such as S6(F) (4.17) and S10(F) (4.07) also show values approaching the scale’s maximum, reflecting the widespread perception of academic stress. In the unfavorable item group, the scores were mostly still above the midpoint of the scale, indicating a high influence of environmental expectations

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 425−438
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10265

and negative self-perception. Demographically, Cluster 0 was dominated by 166 male students (68.0%) and 78 female students (32.0%).

b. Characteristics of Cluster 1
Conversely, in Cluster 1, the average scores for all items were well below 2.50 (below the midpoint of the Likert scale). Although S4(F) (2.32) and S12(F) (2.30) were the highest-scoring items in this group, their scores remained significantly lower compared to the corresponding indicators in Cluster 0. Unfavorable items in this cluster also showed very low scores, such as S5(UF) (1.38) and S9(UF) (1.44). This response pattern indicates that respondents in this cluster do not feel excessively burdened by academic demands and possess a relatively positive self-perception. The gender composition of respondents in Cluster 1 consists of 79 male students (67.5%) and 38 female students (32.5%).

c. Conclusion on Cluster Naming
Based on the contrasting characteristics of the two groups, it is clear that Cluster 0 consistently represents a group of students with high levels of stress and mental burden, while Cluster 1 represents a minimal perceived academic burden. Therefore, it can be concluded that Cluster 0 is categorized as the High-Stress Cluster, while Cluster 1 is categorized as the Low-Stress Cluster.

3.4.3 Distribution of Clustering Results

A visualization of the distribution of clustering results is shown in Figure 6.

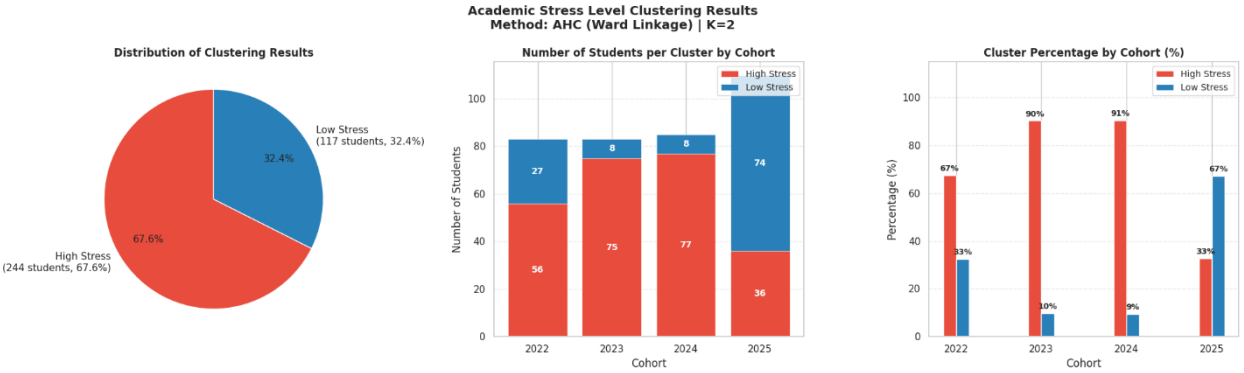

Figure 5. Distribution of Clustering Results

Based on Figure 5, the pie chart displays the overall cluster distribution of the 361 student respondents. The High Stress cluster dominates with a proportion of 67.6% (244 students), while the Low Stress cluster accounts for only 32.4% (117 students). This composition indicates that more than two-thirds of students in the Informatics Engineering Program at UIN SUSKA Riau experience high levels of academic stress.

The cluster distribution by cohort is shown via a bar chart in the center of Figure 5. The 2022 cohort consists of 27 students in the Low Stress cluster and 56 students in the High Stress cluster. The 2023 cohort shows a wider disparity, with only 8 students in the Low Stress cluster compared to 75 in the High Stress cluster, and a similar pattern is observed in the 2024 cohort, with 8 students in the Low Stress cluster and 77 in the High Stress cluster. In contrast, the 2025 cohort shows a significantly different pattern, where the Low Stress cluster dominates with 74 students compared to 36 High Stress students.

The differences in proportions across cohorts are more clearly illustrated by the percentage chart on the right side of Figure 5. The 2023 and 2024 cohorts have the highest proportions of High Stress, at 90% and 91% respectively, while the 2022 cohort stands at 67%. The 2025 cohort is the only one with a more dominant Low Stress proportion, at 67%, in contrast to the other cohorts. This pattern indicates an accumulation of academic pressure that increases with each year of study, peaking in the middle cohorts (2023 and 2024), before declining again among the incoming freshmen of the 2025 cohort, who have not yet fully faced the full academic workload.

3.5 Discussion

The finding of this research indicate that AHC with Ward Linkage at K=2 produced the best clustering quality, with a Silhouette Coefficient of 0.4407 and a Davies-Bouldin Index of 0.8373, while K-Means was only superior on the Calinski-Harabasz Index. This result is consistent with several previous comparative studies that also found AHC to outperform partition-based methods. Abdulpatah et al., in their comparison of K-Means and AHC for clustering rice-producing regions, reported that AHC using Average Linkage achieved a Silhouette Coefficient of 0.723 and a Davies-Bouldin Index of 0.229, both better than K-Means, which obtained 0.696 and 0.404, respectively [16]. Similarly, Husna et al. found that AHC achieved an average Silhouette Coefficient of 0.5837 when compared with K-Medoids on inpatient disease data, far exceeding K-Medoids, which only reached -0.3558 [17]. A comparable pattern was also observed by Tjipta et al., where AHC outperformed K-Means++ with a Silhouette Score of 0.550 and a Davies-Bouldin Index of 0.457, although K-Means++ scored higher on the Calinski-Harabasz Index [18], a finding that mirrors the result obtained in the present research.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 425−438
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10265

In terms of cluster composition, this research found that 67.6% of students (244 of 361) belonged to the High Stress cluster, a proportion that is notably higher than that reported by Wiranti et al., who found 278 of 507 students (54.8%) in the high-stress category using K-Means with a Davies-Bouldin Index of 1.43 and a Silhouette Coefficient of 0.27 [7]. The Davies-Bouldin Index achieved in this research (0.8373) is considerably better than that reported by Wiranti et al., suggesting that the use of AHC with Ward Linkage, supported by the Cophenetic Correlation Coefficient as a basis for linkage selection, produced more compact and well-separated clusters. The proportion of high-stress students found in this research is also higher than the 313 of 587 students (53.3%) reported by Alfaiza et al. using Fuzzy C-Means [9]. This difference may be attributed to disciplinary characteristics, as the respondents in this research were drawn specifically from the Informatics Engineering Program, a discipline previously noted by Jensen et al. to involve a particularly heavy academic workload compared with other fields [6].

The cohort-based pattern observed in this research, where the 2023 and 2024 cohorts showed the highest proportions of high stress (90% and 91%, respectively) before declining among the 2025 cohort, has not been directly examined in prior studies on academic stress clustering, since most previous research treated student populations as a single, undifferentiated group rather than disaggregating results by cohort or year of study [7], [9], [10], [11]. Wiranti et al. and Alfaiza et al., for instance, reported only the overall proportion of students in the high- and low-stress clusters without further analyzing how this distribution varied across academic years [7], [9]. By contrast, the cohort-level analysis in this research reveals that academic stress does not accumulate uniformly across the study period, but instead peaks among second- and third-year students before declining among first-year students who have not yet been exposed to the full academic workload of the program.

4. CONCLUSION

This research compares the Agglomerative Hierarchical Clustering (AHC) and K-Means methods in clustering the academic stress levels of students in the Informatics Engineering Program at UIN Suska Riau, class of 2022–2025, using the Perception of Academic Stress Scale (PAS), which consists of 18 statements. Data were obtained from 374 respondents, and after undergoing a process of removing duplicate data based on student ID numbers and transforming the questionnaire data into a numerical format based on the categories Favorable (F) and Unfavorable (UF), 361 valid data points were obtained and ready for analysis. Prior to the comparative evaluation, the selection of the best linkage method for AHC was conducted using the Cophenetic Correlation Coefficient (CCC), where Ward Linkage was selected with the highest CCC value of 0.8180. The results of the comparative evaluation using the Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index showed that AHC with Ward Linkage and K=2 outperformed the others on two primary metrics: the Silhouette Coefficient of 0.4407 and the DBI of 0.8373, while K-Means outperformed on the Calinski-Harabasz Index with a value of 419.7405; considering all three metrics, AHC with Ward Linkage was determined to be the best method for clustering academic stress data in this research. The clustering results divided 361 students into two clusters: High Stress with 244 students (67.6%) and Low Stress with 117 students (32.4%). Coursework load, exam pressure, and academic performance expectations were recorded as the most dominant sources of stress, reflected in the high mean scores on items S12(F), S18(F), and S16(F) in the High Stress cluster. In terms of distribution by cohort, the 2023 and 2024 cohorts recorded the highest proportions of high stress at 90.4% and 90.6%, respectively, while the 2025 cohort was dominated by the low stress cluster at 67.3% due to the lack of accumulated academic workload in the early semesters. These findings are expected to serve as a basis for institutions in designing targeted intervention programs, particularly psychological counseling services and academic mentoring for students in the middle of their studies. This research has limitations regarding the use of a 5-point Likert scale, which has the potential to introduce central tendency bias due to the “Neutral” option; future research is advised to use a 4-point Likert scale to produce more definitive data, and comparisons with other methods such as DBSCAN or Gaussian Mixture Models (GMM) are also recommended to provide a more comprehensive picture of the quality of clustering for students’ academic stress data.

ACKNOWLEDGMENT

The authors would like to express sincere gratitude to the academic supervisors for their invaluable guidance, patience, and constructive feedback throughout the course of this research. Appreciation is also extended to the students of the Informatics Engineering at UIN Sultan Syarif Kasim Riau who generously volunteered their time to participate as respondents in this research.

REFERENCES

[1] F. Yunita Sari, M. Sukma Kuntari, W. Ari Yati, H. Khaulasari, and M. Hafiyusholeh, “Implementasi K-Means Clustering Melalui Pemanfaatan Sampling Kombinasi Pada Pengelompokan Pola Kesehatan Mental Mahasiswa Sains dan Teknologi,” *J. Nas. Teknol. dan Sist. Inf.*, vol. 11, no. 1, pp. 9–16, Apr. 2025, doi: 10.25077/TEKNOSI.V11I01.2025.9-16.

[2] Z. Mufatihah, A. P. Zelya, R. A. Puriani, and R. M. Putri, “Fenomena Stres Akademik Pada Mahasiswa,” *EDU Res.*, vol. 6, no. 1, pp. 573–580, Mar. 2025, doi: 10.47827/JER.V6I1.564.

[3] Y. Deng *et al.*, “Family and Academic Stress and Their Impact on Students’ Depression Level and Academic Performance,”

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 425−438
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10265

Front. Psychiatry, vol. 13, p. 869337, Jun. 2022, doi: 10.3389/FPSYT.2022.869337/EPUB.
[4] R. K. Djoar *et al.*, "Faktor - Faktor Yang Mempengaruhi Stress Akademik Mahasiswa Tingkat Akhir," *Jambura Heal. Sport J.*, vol. 6, no. 1, pp. 52–59, 2024, doi: 10.37311/JHSJ.V6I1.24064.
[5] H. Lubis, A. Ramadhani, and M. Rasyid, "Stres Akademik Mahasiswa dalam Melaksanakan Kuliah Daring Selama Masa Pandemi Covid 19," *Psikostudia*, vol. 10, no. 1, pp. 31–39, 2021, doi: 10.30872/psikostudia.v10i1.5454.
[6] K. J. Jensen, J. F. Mirabelli, A. J. Kunze, T. E. Romanchek, and K. J. Cross, "Undergraduate student perceptions of stress and mental health in engineering culture," *Int. J. STEM Educ.*, vol. 10, no. 1, pp. 30-, 2023, doi: 10.1186/S40594-023-00419-6.
[7] L. D. Wiranti, E. Budianita, A. Nazir, F. Insani, and R. Susanti, "Penerapan Algoritma K-Means Untuk Mengelompokkan Tingkat Stres Akademik Pada Mahasiswa," *Build. Informatics, Technol. Sci.*, vol. 7, no. 1, pp. 400–409, Jun. 2025, doi: 10.47065/BITS.V7I1.7410.
[8] N. S. Nurfadilah, E. Budianita, A. Nazir, F. Insani, and R. Susanti, "Pengelompokan Tingkat Stres Akademik Pada Mahasiswa Menggunakan Algoritma K-Medoids," *Build. Informatics, Technol. Sci.*, vol. 7, no. 1, pp. 344–353, Jun. 2025, doi: 10.47065/BITS.V7I1.7409.
[9] R. Z. Alfaiza, E. Budianita, S. K. Gusti, and I. Afrianty, "Pengelompokkan Tingkat Stres Akademik Pada Mahasiswa Menggunakan Algoritma Fuzzy C-Means," *TIN Terap. Inform. Nusant.*, vol. 6, no. 5, pp. 516–527, 2025, doi: 10.47065/tin.v6i5.8460.
[10] J. Wijaya, T. Magdalena, A. Januaviani, and K. Kunci, "Clustering Faktor Stres Pada Mahasiswa Aktif Menggunakan Algoritma K-Means dan K-Modes," *Mutiara Multidiciplinary Sci.*, vol. 2, no. 2, pp. 907–917, Mar. 2024, doi: 10.57185/MUTIARA.V2I2.137.
[11] A. Trifani, S. Tunas Bangsa, A. Perdana, W. Stikom, T. Bangsa, and H. Qurniawan, "Penerapan Data Mining Klasifikasi C4.5 dalam Menentukan Tingkat Stres Mahasiswa Akhir," *J. Ris. RUMPUN ILMU Tek.*, vol. 1, no. 2, pp. 91–105, Oct. 2022, doi: 10.55606/JURRITEK.V1I2.414.
[12] K. P. Simanjuntak and U. Khaira, "Pengelompokkan Titik Api di Provinsi Jambi dengan Algoritma Agglomerative Hierarchical Clustering," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 1, no. 1, pp. 7–16, Mar. 2021, doi: 10.57152/MALCOM.V1I1.6.
[13] R. Kusumastuti, E. Bayunanda, A. M. Rifa'i, M. R. G. Asgar, F. I. Ilmawati, and K. Kusrini, "Clustering Titik Panas Menggunakan Algoritma Agglomerative Hierarchical Clustering (AHC)," *CogITo Smart J.*, vol. 8, no. 2, pp. 501–513, Dec. 2022, doi: 10.31154/COGITO.V8I2.438.501-513.
[14] E. Widodo, P. Ermayani, L. N. Laila, and A. T. Madani, "Pengelompokkan Provinsi di Indonesia Berdasarkan Tingkat Kemiskinan Menggunakan Analisis Hierarchical Agglomerative Clustering," *Semin. Nas. Off. Stat.*, vol. 2021, no. 1, pp. 557–566, Nov. 2021, doi: 10.34123/SEMNASOFFSTAT.V2021I1.971.
[15] A. Sujjada, G. P. Insany, and S. Noer, "Analisis Clustering Data Penyandang Disabilitas Menggunakan Metode Agglomerative Hierarchical Clustering dan K-means," *J. Teknol. dan Manaj. Inform.*, vol. 10, no. 1, pp. 1–12, Jun. 2024, doi: 10.26905/JTMI.V10I1.10654.
[16] T. Abdulpatah, B. N. Sari, U. S. Karawang, J. H. S. R. Waluyo, T. Timur, and J. Barat, "Perbandingan Algoritma K-Means dan Agglomerative Hierarchical Clustering untuk Pengelompokan Daerah Penghasil Padi di Indonesia," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 3, pp. 2830–7062, 2025, doi: 10.23960/JITET.V13I3.7251.
[17] L. Husna, D. Hamdhana, and M. Ula, "Analisis Perbandingan Kinerja Algoritma Agglomerative Hierarchical Clustering dan K-Medoids untuk Klasterisasi Jenis Penyakit Pasien Rawat Inap," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 10, no. 2, pp. 1355–1368, 2025, doi: 10.36341/RABIT.V10I2.6554.
[18] C. Tjipta *et al.*, "Comparison of K-Means++ and Agglomerative Hierarchical Methods in Clustering Healthcare Workers," *INOVTEK Polbeng - Seri Inform.*, vol. 10, no. 2, pp. 717–728, 2025, doi: 10.35314/PCBRS043.
[19] W. R. Murdhiono and V. Vidayanti, "Examining Academic Stress and Its Source Among Nursing Professional Students (Ners) Using the Modified Perception of Academic Stress Scale (PAS)," *Indones. Nurs. J. Educ. Clin.*, vol. 7, no. 1, p. 2, 2022, doi: 10.24990/INJEC.V7I1.441.
[20] N. F. Hadi and N. K. Afandi, "Literature Review is A Part of Research," *Sulawesi Tenggara Educ. J.*, vol. 1, no. 3, pp. 64–71, 2021, doi: 10.54297/SEDUJ.V1I3.203.
[21] A. M. Hudin and M. S. Budiani, "Hubungan antara Workplace Well-Being dengan Kinerja Karyawan pada PT. X di Sidoarjo," *Character J. Penelit. Psikol.*, vol. 8, no. 4, pp. 257–267, 2021, doi: 10.26740/CJPP.V8I4.41192.
[22] D. Chicco, L. Oneto, and E. Tavazzi, "Eleven quick tips for data cleaning and feature engineering," *PLOS Comput. Biol.*, vol. 18, no. 12, p. e1010718, 2022, doi: 10.1371/JOURNAL.PCBI.1010718.
[23] M. S. I. Lubis *et al.*, "Analisis Pengelompokan Wilayah Kepolisian Daerah di Indonesia menggunakan Algoritma Hierarchical Clustering," *Innov. J. Soc. Sci. Res.*, vol. 5, no. 3, pp. 6646–6663, Jun. 2025, doi: 10.31004/INNOVATIVE.V5I3.19790.
[24] E. K. Tokuda, C. H. Comin, and L. da F. Costa, "Revisiting agglomerative clustering," *Phys. A Stat. Mech. its Appl.*, vol. 585, p. 126433, 2022, doi: 10.1016/J.PHYSA.2021.126433.
[25] E. U. Oti and M. O. Olusola, "Overview of Agglomerative Hierarchical Clustering Methods," *Br. J. Comput. Netw. Inf. Technol.*, vol. 7, no. 2, pp. 14–23, 2024, doi: 10.52589/BJCNIT-CV9POOGW.
[26] T. Wismarini, S. Eniyati, E. Lestariningsih, S. Soelistijadi, E. Ardhianto, and W. Handoko, "Identifikasi Pola Konflik Lahan Perkebunan di Lingkungan PTPN Group Berbasis Data Hukum Menggunakan Hierarchical Clustering dengan Algoritma Agglomerative," *J. FASILKOM*, vol. 14, no. 3, pp. 654–666, Dec. 2024, doi: 10.37859/JF.V14I3.7915.
[27] A. Karna and K. Gibert, "Automatic identification of the number of clusters in hierarchical clustering," *Neural Comput. Appl. 2021 341*, vol. 34, no. 1, pp. 119–134, 2021, doi: 10.1007/S00521-021-05873-3.
[28] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (Ny).*, vol. 622, pp. 178–210, 2023, doi: 10.1016/J.INS.2022.11.139.
[29] X. Wang, Z. Shao, Y. Shen, and Y. He, "Research on fast marking method for indicator diagram of pumping well based on K-means clustering," *Heliyon*, vol. 9, no. 10, p. e20468, 2023, doi: 10.1016/J.HELIYON.2023.E20468.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 425−438
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10265

[30] D. F. Surianto and D. F. Surianto, "Enhancing K-Means Clustering for Journal Articles using TF-IDF and LDA Feature Extraction," *Brill. Res. Artif. Intell.*, vol. 4, no. 2, pp. 964–972, 2024, doi: 10.47709/BRILLIANCE.V4I2.5547.
[31] S. Saraçlı and M. Akşit, "Comparison of Hierarchic Clustering Methods with Cophenetic Correlation Coefficient in Big Data," *Afyon Kocatepe Üniversitesi Fen Ve Mühendislik Bilim. Derg.*, vol. 22, no. 3, pp. 552–559, 2022, doi: 10.35414/AKUFEMUBID.1018302.
[32] A. Gere, "Recommendations for validating hierarchical clustering in consumer sensory projects," *Curr. Res. Food Sci.*, vol. 6, p. 100522, 2023, doi: 10.1016/J.CRFS.2023.100522.
[33] D. Chicco, A. Campagner, A. Spagnolo, D. Ciucci, and G. Jurman, "The Silhouette coefficient and the Davies-Bouldin index are more informative than Dunn index, Calinski-Harabasz index, Shannon entropy, and Gap statistic for unsupervised clustering internal evaluation of two convex clusters," *PeerJ Comput. Sci.*, vol. 11, p. e3309, 2025, doi: 10.7717/PEERJ-CS.3309/TABLE-18.
[34] A. Anfossi and D. Chicco, "An easy guide to the Davies-Bouldin index for unsupervised internal clustering evaluation," *Discov. Comput. 2026 291*, vol. 29, no. 1, pp. 212-, 2026, doi: 10.1007/S10791-026-10094-0.
[35] F. Pascoal, P. Branco, L. Torgo, R. Costa, and C. Magalhães, "Definition of the microbial rare biosphere through unsupervised machine learning," *Commun. Biol.*, vol. 8, no. 1, 2025, doi: 10.1038/S42003-025-07912-4.