Jurnal Teknik Informatika Vol. 17 No. 1, April 2024 (51-60)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v17i1.32132

JURNAL TEKNIK INFORMATIKA

Homepage : http://journal.uinjkt.ac.id/index.php/ti

A Comparative Study of Students Graduation Analysis Using Classification Methods in Undergraduate Electrical Engineering Tidar University

Damar Wicaksono[1], Sapto Nisworo[2] and Imam Adi Nata[3]

[1,2,3] Electrical Engineering, Faculty of Engineering, Tidar University
[1.2,3] Kapten Suparman No 39, Tuguran, Potrobangsan, North Magelang District, Magelang City, Central Java, 56116, Indonesia

Article:
Accepted: March, 21, 2024
Revised: January 04, 2024
Issued: April 30, 2024

***Correspondence Address:**
damar@untidar.ac.id

ABSTRACT

This research aimed to classify achievement factors for electrical engineering students at Tidar University using K-Means and Agglomerative Clustering classification algorithms. The goal was to understand if any parameters influence high-achieving student performance. The Indonesian government and private sector for university students provide significant education funds. Student scholarships are awarded based primarily on GPA and entry path, overburdening staff and causing confusion during distribution to eligible recipients. A system was needed to accommodate additional eligible criteria. The researcher selected factors to identify engineering student performance, including school origin, entry path, tuition fees, and GPA. These inputs could determine graduation status. The results compared calculation methods based on collected data accuracy, processing times, and characterizing clustered data to determine the best classification method. Agglomerative Hierarchical Clustering performed better. Accuracy testing on 600 training data points yielded 73.94% for improved K-means and 90.42% for AHC. The Average processing time was 674.92 seconds for improved K-means and 554.35 seconds for AHC. Silhouette testing also characterized calculation methods, with improved K-means scoring best at 0.654 and AHC at 0.787 using two clusters.

Keywords: *achievement; algorithms; scholarships; classification;*

Jurnal Teknik Informatika Vol. 17 No. 1, April 2024 (51-60)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v17i1.32132

1. INTRODUCTION

Each university has many educational assistance schemes for students who can demonstrate achievement in the academic field. [1]. Scholarships are a number of fees provided or borne by a party, and funding is given to recipients who are selected based on certain qualifications [2]. One kind usually offered is a scholarship with conditions for the underprivileged, according to accouterment regulation 48 of 2008 concerning education, chapter 27 paragraph, number 1 [3]. The involvement of many applicants, the weight of consideration of more than one criterion, as well as to avoid awarding scholarships that are not on target raises problems, namely the need for a method to select applicants with the establishment of specific criteria to provide recommendations or decision support that schools can use in selecting scholarships [4].

Tidar University, located in Central Java, is a private institution with five faculties. One faculty is engineering faculty, which is known as FT. An undergraduate major in the study, namely, electrical engineering. So far, the selection process for prospective scholarship recipients at Tidar University is still done manually, so the selection process is not objective. One way to make selecting scholarships for outstanding students run objectively is by clustering students among others: grade point average (GPA), gender, tuition fees, part-time work, organization, and university entrance paths.

The quality of students in tertiary institutions can be evaluated by learning analytics. It is a method of analyzing data collected through educational processes to enhance academic quality at institutions of higher learning. Using learning analytics methods in universities can potentially improve the education standard. Data mining can be automated or semi-automatically in process [6] [7].

Data mining can be applied to any existing data set that may yield valuable insights. It is a method that can automatically identify patterns or relationships within data relatively quickly and easily. Classification analysis is a common analytic method used for learning. Data classification is a commonly used method for organizing information, in which classification learning forms models to predict categorical labels on given data. Conducting such analysis offers organizations several potential benefits: curriculum development, graduation increase, time to accept jobs after graduation, improved lecturer performance, and increased research in the education field [8]. In the clustering process, the most important thing is to collect patterns into appropriate groups to find similarities and differences and order valuable conclusions [7]. Clustering helps analyze data patterns, grouping, and making decisions [9]. This research compares the classification of student study periods using two methods: K-Means and Agglomerative Clustering methods.

Previous research has been conducted in student graduation prediction with some methods. Rosmini [10] has researched the implementation of the K-Means method in selecting and mapping groups of students through lecture activity data using the variables above.

This study used k-means clustering with k = 2 to create two clusters of students: cluster A of those graduating on time and cluster B of those graduating not on time. This clustering of student data by graduation status provides input for faculty advisors to classify student achievement types.

Sunaryanto et al. [11] implemented a K-means clustering algorithm on student data from the Informatics Engineering program. K-Means is an algorithm used to classify objects based on their proximity to predefined clusters. The value of k, representing the number of clusters, cannot exceed the number of training examples and must be an odd number greater than one. The training samples are represented as vectors in a multidimensional feature space, where each dimension reflects an attribute of the data. This space is then partitioned based on the cluster assignments of the training samples.

Februariyanti and Santoso's research [12] utilized agglomerative clustering methods to group student theses based on thesis title variables. The clustering process resulted in five distinct groups (k = 5). Analysis of the thesis titles within each cluster identified five common topics pursued by students for their theses: Information Systems, Semarang Information Systems, Web-based expert systems, Expert systems for diagnosing diseases in plants, and design of teaching tools for children.

An enhanced k-means algorithm with meliorated initial center, constructed by Chen Guang-ping and Wang Wen-peng, provided the

basis for earlier research. This study uses a dataset taken from the KDD cup. The research was conducted to find out the differences between traditional k-means and improved K-means. The steps taken are to find the center point (center point) by finding the furthest distance between the data. In the next stage, the center point is used in calculations with the k-means algorithm. This study concludes that calculations with improved km-means produce better center points and have a higher accuracy level than traditional k-means [13].

The K-Means methods are clustering methods usually used by several previous researchers, and they have been implemented especially in the education field [12]-[13]. Meanwhile, agglomerative methods have also been applied to group data by a number of earlier researchers [16]-[18], which, in certain case studies performs and is more accurate than K-Means [19]. To date, no comparative study has been conducted examining the use of various data mining methods to forecast graduation outcomes for engaged students based on certain characteristics or criteria [20].

The clustering method makes it easier for researchers to determine student recipients of scholarships and seek results in determining recipients of outstanding scholarships because the clustering process is based on GPA, college admissions path, tuition fees, parental income, and number of parental dependents.

This study compares several clustering algorithms to predict the performance of high-achieving students. This research aims to determine if predicting which students are eligible for achievement scholarships could allow academics to implement activities that minimize the number of students at risk of delayed graduation, such as tutoring programs, short-term coursework, or other beneficial initiatives. These exercises could enhance the learning outcomes for students. Additionally, a number of data mining methods are used in this study to segment data in order to group student academic records. The objective is to identify the method that delivers the highest quality clusters.

2. METHODS

2.1. Data Collection

The entire procedure in this study was able to move forward through several successive steps. The various steps followed in the study allowed for a systematic approach to be taken.

2.1.1. Interview

An interview is a direct data collection method with participants to obtain information related to research problems and requirements for the intended algorithm.

2.1.2. Library studies

Data collection methods involve studying, researching, and reading various sources such as online books, internet information, journals, theses, and dissertations related to the algorithms system that will be discussed.

2.1.3. Observation

Observation, or direct observation of the research subject and ongoing activities, is also used. Observation is typically conducted with structured observation by preparing a list of required data points and sources. Observational data can be obtained by reviewing the staffing and student affairs sections.

2.2. Research Methodology

Data collected from literature reviews and student records in the Tidar University electrical engineering department will be analyzed to group graduates into three clusters. The first cluster will contain students who graduated early, the second who graduated on time, and the third who did not. Clustering analysis methods will segment the students based on attributes from their academic records and background characteristics obtained through informants. The research methodology is depicted in Figure 1.

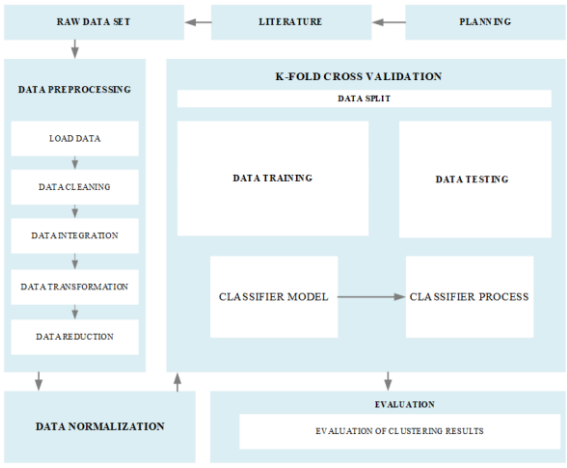

Figure 1. The process of classifying data in research

Jurnal Teknik Informatika Vol. 17 No. 1, April 2024 (51-60)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v17i1.32132

2.3. Classification stages

Each method has different classification stages, and the algorithm used to determine a student's study period is based on test results that are compared to more test examples. In this study, the authors employ two distinct methods in the classification process.

2.4. Improved K-Means clustering

Clustering is a method for organizing data when the class label is known. It operates by classifying data objects into a predetermined number of clusters, including patterns, entities, events, units, results, and observations. Put differently, this method divides data into multiple categories based on specific attributes. The method of grouping a data collection into clusters based on similarities is carried out through clustering. K-means clustering, where k is the number of clusters, is one method for locating clusters in the data [19].

If class c is the most common classification among the k closest neighbors of a point in this space, then that point is classified as class c. Proximity of neighbors is generally determined based on Euclidean distance, which can be defined as the straight-line distance between two points in geometric space as follows:

$$d_{euc}(x, y) = \sqrt{\sum_{j=1}^{d}(x_j - y_j)^2} \qquad (1)$$

Where:

x_j dan y_j = the j attribute's value

This algorithm provides a simple way to classify a given data set into a specified number of clusters (k clusters). The first step is to define k initial centroids, one for each cluster. As the location of centroids can impact results, care should be taken to strategically place them as far apart as possible for optimal separation of clusters.

The next step is to assign each data point to the nearest centroid, creating an initial clustering. This first step is complete once all data points have been assigned and no unassigned points remain. New centroids must then be calculated as the centers of mass of the clusters from the previous step.

Following the calculation of the new k centroids, another assignment must be done where each data point is allocated to the nearest of the new centroids. A loop is created whereby centroid locations may change on some iterations as assignments are made until no further changes occur - at this point, the centroids are stable in their final positions [21].

Clusters are created by grouping data according to how similar its qualities are. It is possible to assess similarity by using distance measures. This study's distance computation methodology is as follows: [12]:

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\| x_i^{(j)} - c_j \right\|^2 \qquad (2)$$

Where:

J = obj function
k = number on clusters
n = number on cases
x_i = case on i
c_j = centroid on cluster j

Calculate the distance between objects xi and xj with object data in cluster M. If object data in cluster M has a minimum distance to

object datax_i , then the data is included in cluster X. If the object data in cluster M has a minimum distance to the object datax_j, then the data is included in cluster Y [13].

2.5. Agglomerative hierarchical clustering (AHC)

Hierarchical clustering is a method for organizing data objects into a hierarchical structure. [21]. There are two main strategies for hierarchical clustering. The first is agglomerative hierarchical clustering, which follows a bottom-up approach that initially considers each object as a separate cluster and then iteratively merges them into larger clusters. The second is divisive hierarchical clustering, which takes a top-down approach that considers all objects as part of one cluster and then iteratively splits them into smaller clusters. Several clustering algorithms can be used to perform hierarchical clustering, including agglomerative algorithms that merge the pair of clusters with the most similar objects into a new cluster and divisive algorithms that successively split clusters with the greatest dissimilarity [22].

2.5.1. Single linkage

Based on the smallest distance between objects, the single linkage algorithm clusters objects. This grouping procedure first identifies the smallest distance value in the distance matrix and combines the corresponding objects

Jurnal Teknik Informatika Vol. 17 No. 1, April 2024 (51-60)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v17i1.32132

into the matrix $D = \{dij\}$. Then, the clusters combine similar objects, like V and U (UV).

The next step is finding process the distance(UV) With other clusters, for example, W. The equation may be rewritten in the following manner:

$$d_{(UV)W} = min(d_{UW}, d_{VW}) \quad (3)$$

d_{UW} is how far away the closest neighbor is from the cluster U and W thend$_{VW}$ Is the nearest neighbor's distance from *cluster V* and W.

2.5.2. Complete linkage

The full linkage algorithm is an agglomerative clustering method based on the greatest distance between items. In order to link the two clusters with the least maximum distance between each component member, this method starts by determining the largest distance between objects. Until all objects are connected into a single cluster in matrix $D = \{dij\}$, the procedure is repeated, joining the clusters with the least maximum inter-cluster distance each time, then merging the related objects e.g., U and V to get the clusters (UV). The next step is finding the distance between (UV) With the other clusters, for example on the formulation W can be written as follows [23]:

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

d_{UW} is the farthest neighbor distance from the cluster U and W then d_{VW} is the farthest distance of neighbor from *cluster V* and W.

2.5.3. Average linkage

The average linkage algorithm is a hierarchical clustering methodology that utilizes the average distance between clusters. Specifically, average linkage clustering commences by calculating the distance matrix between all objects, which records the dissimilarity between each pair of objects. Subsequently, clusters are created by combining the most similar pairs of clusters, with the distance between two clusters being determined by averaging the distances between every pair of objects in each of the distinct clusters. This merging process is repeated until all objects are clustered into a single matrix $D = \{dij\}$ to get the closest object, for example U and V, then these objects are combined into clusters (UV) and then the distance between (UV) with other clusters W, so it can be written as follows [24]:

$$d_{(uv)w} = \frac{d(uw)+d(vw)}{n(uv)nw} \quad (5)$$

$n_{(UV)}$ is the members number in the cluster(UV) and n_w is the members number in the cluster W.

2.5.4. d. Silhouette coefficient for clustering evaluation method

The average linkage algorithm is a hierarchical clustering methodology that utilizes the average distance between clusters. Specifically, average linkage clustering commences by calculating the distance matrix between all objects, which records the dissimilarity between each pair of objects. Clusters are then formed sequentially by merging the closest pair of clusters, where the distance between two clusters is defined as the average of the distances between all pairs of objects in the different clusters. This merging process is repeated until all objects are clustered into a single cluster [23].

Testing occurs after convergence is reached at stage 0, where the results of the latest grouping are identical to the preceding grouping. In other words, no data shifts clusters. Testing employs the silhouette coefficient equation. Finding the average distance between the i^{th} data point and every other data point in the same cluster is the first step in calculating the silhouette coefficient, assuming the i^{th} data belongs to cluster A. [24]. The formula for a(i) can be written in this following equation.

$$a(i) = \frac{\pi r^2}{|A|-1} \sum j \in A, j \neq 1 \; d(i,j) \quad (6)$$

where A = the amount of data located in cluster A

The next step is to calculate the value of b(i) which represents the minimum average distance between the i^{th} data point and all data points assigned to different clusters. Assuming the data mean belongs to either cluster A or cluster C, b(i) is computed as the average distance between the i^{th} data and all the data in cluster C can be written in the following equation:

$$d(i,C) = \frac{1}{|C|} \sum j \in C \; d(i,j) \quad (7)$$

C = The total volume of data contained within cluster C.

Once d(i,C) has been calculated for each cluster C that does not equal A, choose the value of b(i) to be the minimum distance.

$$b(i) = \min_{C \neq A} d(i,j) \quad (8)$$

Assuming a minimum distance for cluster B, then $d(i,B)\ equal\ b(i)$ which represents the neighbor of the i^{th} data and denotes the second-best cluster for the i^{th} data after cluster A. With $a(i)$ and $b(i)$ identified, the last step is to calculate the silhouette coefficient.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i) - b(i)\}} \quad (9)$$

The variable s(i) can range from -1 to 1 inclusive, with each value having the following interpretation [25]:

a. $s(i)\ equal\ with$ 1 means the i data is well classified (in cluster A)
b. $s(i)\ equal\ with$ 0 means the i^{th} data between the two clusters (A and B)
c. $s(i)\ equal\ with\ -1$ means the i^{th} data classified as weak (nearer to cluster B than A)

The interpretation of the silhouette coefficient value is shown in Table 1 [24][25].

Table 1. *Silhouete coefficient* Interpretation

Silhouette coefficient	Interpretation
0.71 - 1.00	0.71 - 1.00
The resulting structure is strong 0.51 – 0.70	The resulting structure is strong 0.51 – 0.70
The resulting structure is good	The resulting structure is good

3. RESULTS AND DISCUSSION

3.1. Analysis Data Collection with Criteria and Alternatives

Students in the Electrical Engineering Department at Tidar University who were accepted through the Mandiri (SMUT) in 2015 and 2018 provided the data used in this study, SNMPTN / SBMPTN, and PMDK-report, which have been declared passed. The data were obtained from the FT Tidar University Administrative office, and there was a total of 631 data. The following is the data that has been obtained through several stages of pre-processing data, from data cleaning data integration to data reduction, as shown in Figure 1, namely the 2015 and 2018 Electrical Engineering student datasets can be seen in Table 2.

Table 2. *Dataset student as training data*

Student number	High school	Regency	Entrance paths	Tuition Fee	GPA	Graduation status
1510501002	SMA	Magelang	SBMPTN	k25	3.49	late
1510501003	SMA	Wonosobo	SMUT	k24	3.34	on time
1510501012	MAN	Temanggung	SBMPTN	k24	3.40	early
1510501020	SMK	Weleri	SBMPTN	k24	3.49	on time
1510501022	SMK	Magelang	SBMPTN	k24	3.44	late
1510501050	SMA	Wonosobo	SBMPTN	k25	3.37	late
1510501051	SMK	Magelang	SBMPTN	k22	3.49	late
1510501063	SMA	Purworejo	SMM-UNTIDAR	k5	3.34	late
….	…	…	….	….	….	…
1510501066	SMA	Magelang	SMM-UNTIDAR	k5	3.40	on time

The obtained data in this study were then verified and tested by taking 82 sample data in the 2019-year Electrical Engineering major using k-fold cross-validation through the data splitting stage, and predictions and comparisons were made using the K-means method and agglomerative hierarchical clustering. The acquired dataset is displayed in Table 3 below.

Table 3. *Dataset* student as testing data

Student number	High school	Regency	Entrance paths	Tuition fee	GPA	Graduation status
1910501023	SMK	Magelang	SBMPTN	k25	3.49	late
1910501028	SMA	Wonosobo	SMUT	k24	3.34	late
1910501034	SMK	Temanggung	SBMPTN	k24	3.40	early
1910501048	SMK	Wonosobo	SBMPTN	k24	3.49	on time
1910501054	SMA	Wonosobo	SBMPTN	k24	3.44	late
1910501058	SMA	Magelang	SBMPTN	k25	3.37	late
1910501065	SMK	Wonosobo	SBMPTN	k22	3.49	late
1910501070	SMA	Purworejo	SMM-UNTIDAR	k5	3.34	late
….	…	….	….	….	….	…
1910501108	SMA	Magelang	SMM UNTIDAR	k5	3.40	on time

3.2. Test results and analysis of prediction results

3.2.1. The results of testing the accuracy of the number of training data

A comparative analysis evaluated the average predictive accuracy of the K-means and agglomerative hierarchical clustering algorithms. Both algorithms were tested using the same datasets containing 100, 200, 300, 400, 500, and 600 observations for model training purposes. For the K-means algorithm, 5

Jurnal Teknik Informatika Vol. 17 No. 1, April 2024 (51-60)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v17i1.32132

independent trials were run for each training data size to account for random initialization effects. The average accuracy across the 5 trials was then computed. Regarding the agglomerative clustering approach, the hierarchical clustering process was repeatedly performed until convergence was reached, defined as the number of clusters KL plus one equals the square root of the total number of observations N or KL + 1 to √N. The predictive accuracy results for each algorithm and training data size configuration are provided in tabular form as Table 4 below.

Table 4. *The results of evaluating both algorithms' training data variants' accuracy level*

Clustering method	Data Amount	Amount of predictive label data	Amount of predictive label data is not the same as the original label (200 test data)	Average level of accuracy
Improved K-means	100	185	17	75.55%
	200	195	14	64,56%
	300	185	22	78,66%
	400	175	14	75,50%
	500	185	18	75,50%
	600	185	18	75,50%
Average		**185**	**17**	**73,94%**
Agglomerative hierarchical clustering	100	154	45	90,50%
	200	134	86	85,50%
	300	165	56	90,56%
	400	165	75	91,75%
	500	154	65	91,55%
	600	134	43	92,66%
Average		**151**	**62**	**90,42%**

3.2.2. The results of testing the length of processing time on the number of data variants in the K-means and agglomerative clustering methods.

The second test aimed to determine the convergence and average time comparison of the two algorithms. Similar to the last test, 500 training data points are utilized, broken down into five test sections: 100, 200, 300, 400, and 500 training data points. The average time for one convergence in the agglomerative clustering method is calculated by dividing the total time for one process by the number of iterations required to find convergence. In contrast, the K-Means method will be evaluated five times, as indicated in Table 5 below, with the average convergence being determined once from the average time in each segment of the training data (100, 200, 300, 400, and 500 training data).

Table 5. *The number of training data variants in both algorithms converges to the table of the one-time test results*

Clustering method	Data *Training*	Amount of predictive label data	Amount of predictive label data is not the same as the original label (200 test data)	Average processing time in seconds (on average 1 time converges)
Improved K-means	100	185	17	155,56
	200	195	14	256,72
	300	185	22	562,55
	400	175	14	784,33
	500	185	18	1045,78
	600	185	18	1244,55
Average		**185**	**17**	**674,92**
Agglomerative hierarchical clustering	100	154	45	125,56
	200	134	86	210,45
	300	165	56	489,78
	400	165	75	658,67
	500	154	65	854,78
	600	134	43	986,88
Average		**151**	**62**	**554,35**

Based on the data presented in Table 5, it can be concluded that increasing the training data size will result in a longer computation time required for each algorithm to reach convergence. This happens because processing more significant amounts of data takes longer for each algorithm's sub-processes.

The process of selecting convergent values will likewise take longer with more diverse training data. Agglomerative clustering outperforms K-means clustering in terms of average computation time for every variation of the training data, according to a comparison of the findings. This is due to the algorithm's instability caused by the K-means method's random selection of the initial cluster centers. On the other hand, hierarchical agglomerative clustering has benefits such as not being affected by outliers and precisely calculating the right number of clusters. This enables it to determine an initial cluster center value that is more ideal. The processing durations of the agglomerative clustering method will be better with a more ideal starting point.

Jurnal Teknik Informatika Vol. 17 No. 1, April 2024 (51-60)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v17i1.32132

3.2.3. Characterization of data clustering

The K-Means clustering process tested various cluster values and determined that two (2) clusters best fit the experimental data, as depicted in Figure 2. The agglomerative clustering method was also applied to the experimental data, with each linkage approach - single, complete, and average - producing the results as shown in Figure 3a until 3c, respectively. The optimal number of clusters was evaluated ten times across a range of cluster counts. According to the tests, two (2) clusters were ideal for the full and average linkage methods, while three (3) clusters worked best for the single linkage strategy. This experiment, which made use of a specified set of training and test data, may thus be stated to have typically produced the best fit with three (3) clusters using the agglomerative clustering approach. The amount of previously processed training and test data is definitely utilized in this investigation. Different values were obtained in the K-Means clustering procedure, and two clusters showed the best result, as shown in Figure 2.

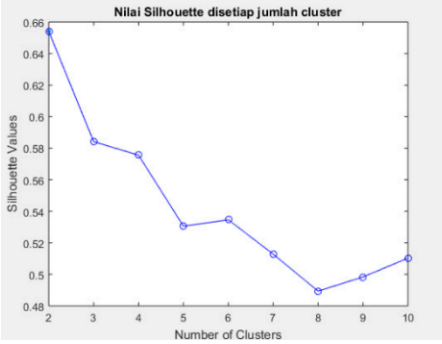

Figure 2. *Silhouette values from K-Means*

The experimental data underwent a clustering process utilizing an agglomerative clustering method. As depicted in Figures 3a, 3b, and 3c, each approach yielded the following results. The optimal number of clusters was evaluated ten times across a varying number of cluster configurations. The test findings indicated that three clusters were the ideal number for the single linkage method, but two clusters were the ideal number for the full and average linkage approaches. Thus, it can be concluded that three clusters are the ideal cluster value when employing agglomerative clustering. Undoubtedly, the quantity of pre-processed training and test data was utilized in this investigation.

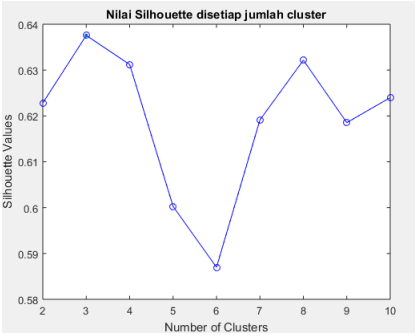

Figure 3. *(a) Single-linkage agglomerative clustering silhouette values*

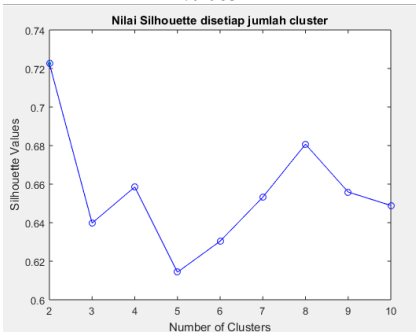

Figure 3. *(b) Average-linkage agglomerative clustering silhouette values*

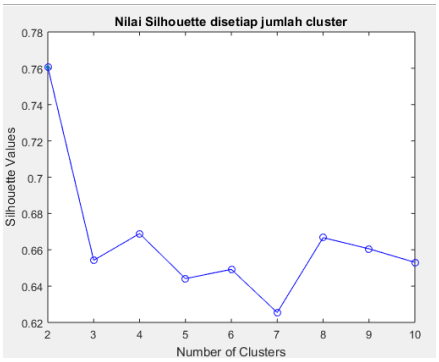

Figure 3. *(c) Complete-linkage agglomerative clustering silhouette values*

Table 6 below compares silhouette values from clustering results using predetermined methods.

Table 6. *Comparative silhouette coefficient*

Amount of Cluster	Enhanced K-Means	*Silhouette coefficient* AHC *Single Linkage*	*Average Linkage*	*Complete Linkage*
2	**0.654**	0.623	0.723	**0. 787**
3	0.541	0.638	0.64	0.656
4	0.631	0.631	0.659	0.661
5	0.509	0.600	0.614	0.653
6	0.502	0.587	0.63	0.649
7	0.513	0.619	0.653	0.625
8	0.644	0.632	0.681	0.644
9	0.567	0.619	0.656	0.661
10	0.540	0.624	0.649	0.625

The full linkage strategy had the best results, yielding a value of 0.787 (78.70%) compared to the other two agglomerative clustering approaches. We thus compare the agglomerative clustering method to the K-Means method using the entire linkage methodology. As can be shown in Table 3 above, the K-Means method's clustering findings yielded a value of 0.654 (65.40%). Comparing these results to the agglomerative clustering method's clustering yields lesser results. Two clusters are determined to be the most optimal number of clusters using the agglomerative clustering method with the complete linkage approach. Consequently, the agglomerative clustering approach outperformed the K-Means method in this instance.

The result of this research also has better accuracy than previous research [14], which only got 63.80% in the K-Means method, different with [19], in which the result showed that K-Means got a better result than the AHC method but in a small dataset-only. The better result accuracy was found in [21], with a result of 37.79% in K-Means and 47.24% in AHC.

CONCLUSION

This report details the evaluation of two predictive algorithms and their ability to accurately forecast and categorize active students who are likely to graduate from Tidar University Two predictive algorithms can be evaluated for their ability to forecast and categorize active students who are likely to graduate. Four varieties of data attributes may serve as variables for student projections, including: school origin, tuition fees, university entrance paths, and grade point average (GPA). AHC generates a silhouette value of 0.787 after evaluating the whole linkage approach model, and in this case study, the agglomerative clustering method has also been utilized to group data that has superior performance, accuracy, and processing time than the K-Means, which is 554.35 in seconds.

REFERENCES

[1] B. Bertaccini, S. Bacci, and A. Petrucci, "A graduates' satisfaction index for the evaluation of the university overall quality," *Socio-Economic Planning Sciences*, vol. 73, p. 100875, Feb. 2021, doi:https://doi.org/10.1016/j.seps.2020.100875

[2] BAN-PT, Buku VI Matriks Penilaian Instrumen Akreditasi Program Studi Sarjana. Jakarta, 2008.

[3] C. Aina and G. Casalone, "Early labor market outcomes of university graduates: Does time to degree matter," Socioecon. Plann. Sci., p. 100822, Mar. 2020. doi: 10.1016/j.seps.2020. 100822

[4] X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with internet usage behaviors using machine learning algorithms," Comput. Human Behav., vol. 98, pp. 166–173, Sep. 2019. doi:10.1016/j.chb.2019.04.015

[5] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Computers & Education*, vol. 113, pp. 177–194, Oct. 2019, doi: https://doi.org/10.1016/j.compedu.2017.05.007

[6] J. M. Borwn, "Predicting Math Test Scores Using K- Nearest Neighbor," *IEEE Integr. STEM Conf.*, 2017. doi:10.1109/ISECon.2017. 7910221

[7] S. A. D. Syarif, "Trending Topic Prediction by Optimizing K-Nearest Neighbor Algorithm," *IEEE*, 2020.

[8] D. Kabakchieva, "Student Performance Prediction by Using Data Mining Classification Algorithms," Int. J. Comput. Sci. Manag. Res., vol. 1, no. 4, pp. 686–690, 2020. doi: 10.5772/intechopen.91449

[9] Y. Palumpun and S. N. Alam, "Pengelompokan Tingkat Kelulusan Mahasiswa Menggunakan Algoritma K-Means," no. November, pp. 98–102, 2019.

[10] R. Rosmini, A. Fadlil, and S. Sunardi, "Implementasi Metode K-Means Dalam Pemetaan Kelompok Mahasiswa Melalui Data Aktivitas Kuliah," It J. Res. Dev., vol.3, no. 1, p.22,2018.doi:10.25299/itjrd.2018.vol3(1).1773

[11] H. Sunaryanto, M. A. Hasan, and G. Guntoro, "Classification Analysis of Unilak Informatics Engineering Students Using Support Vector Machine (SVM), Iterative Dichotomiser 3 (ID3), Random Forest and K-Nearest Neighbors (KNN)," *IT Journal Research and Development*, vol. 7, no. 1, pp. 36–42, Aug. 2022, doi: https://doi.org/10.25299/itjrd.2022.8912

[12] H. Februariyanti and D. B. Santoso, "Hierarchical Agglomerative Clustering Untuk Pengelompokan Skripsi Mahasiswa," Pattern Recognition, 2019, doi: 10.1016/0031-3203(79)90049-9.

[13] Chen Guang-ping and Wang Wen-peng, "An improved K-means algorithm with meliorated initial center," Jul. 2019, doi: https://doi.org/10.1109/iccse.2019.6295047

[14] Y. H. Chrisnanto and G. Abdullah, "The uses of educational data mining in academic performance analysis at higher education institutions (case study at UNJANI)," *Matrix : Jurnal Manajemen Teknologi dan Informatika*, vol. 11, no. 1, pp. 26–35, Mar. 2021, doi: https://doi.org/10.31940/matrix.v11i1.2330

[15] Bora, D.J., dan Gupta, A.K., 2019, Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab, (IJCSIT) *International Journal of Computer Science and Information Technologies*, Vol. 5 (2), 2014, 2501-2506.

[16] Y. Christian and J. Jimmy, "Perancangan Model Predıksı Performa Akademık Mahasıswa Menggunakan Algorıtma K-Means Clusterıng (Studı Kasus: Unıversıtas Xyz)," *CoMBInES - Conference on Management, Business, Innovation, Education and Social Sciences*, vol. 1, no. 1, pp. 643–649, Mar. 2021.

[17] Sri Dewi, Sarjon Defit, and Y. Yunus, "Akurasi Pemetaan Kelompok Belajar Siswa Menuju Prestasi Menggunakan Metode K-Means (Studi Kasus SMP Pembangunan Laboratorium UNP)," *Jurnal Sistim Informasi dan Teknologi*, Sep. 2020, doi: https://doi.org/10.37034/jsisfotek.v3i1.98

[18] D. Exasanti and A. Jananto, "Analisa Hasil Pengelompokan Wilayah Kejadian Non-Kebakaran Menggunakan Agglomerative Hierachical Clustering di Semarang," *Jurnal Tekno Kompak*, vol. 15, no. 2, p. 63, Aug. 2021, doi: https://doi.org/10.33365/jtk.v15i2.1166

[19] L. Zahrotun, "Analisis Pengelompokan Jumlah Penumpang Bus Trans Jogja Menggunakan Metode Clustering K-Means Dan Agglomerative Hierarchical Clustering (AHC)," *Jurnal Informatika*, vol. 9, no. 1, Jan. 2019, doi: https://doi.org/10.26555/jifo.v 9i1.a2045

[20] K. K. Raihana, S. M. K. Rishad, T. Sadia, S. Ahmed, M. S. Alam, and R. M. Rahman, "Identifying Flood Prone Regions In Bangladesh By Clustering," Proc. - 17th IEEE/ACIS Int. Conf. Comput. Inf. Sci. ICIS 2018, pp. 556–561, 2018, doi: 10.1109/ICIS.2018.8466533.

[21] Z. Arifin, S. Santosa, and M. A. Soeleman, "Klasterisasi Genre Cerpen Kompas Menggunakan Agglomerative Hierarchical Clustering- Single Linkage," *J. Cyberku*, vol. 13, no. 2, pp. 2–2, Dec. 2019.

[22] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, ASA- SIAM Series on Statistics and Applied Probability. Society for Industrial and Applied Mathematics, Alexandria, VA, 2007.

[23] I. H. Witten, E. Frank, dan M. A. Hall, *Data mining: practical machine learning tools and methods*, 3rd ed. Burlington, MA: Morgan Kaufmann, 2011. doi:10.1016/C2009-0-19715-5

[24] R. A. Johnson dan D. W. Wlchern, *Applied Multivarate Statistical Analysis (Sixth Edition)*, 6 ed. Pearson Prentice Hall : New Jersey, 2007.doi: 10.1007/978-3-642-17229-8

[25] A. Struyf, M. Hubert, and P. J. Rousseeuw, "Clustering in an Object-Oriented Environment," *Journal of Statictical Software*, vol 1, Issue 4, pp. 1-30, 1997. doi: 10.18637/jss.v001.i04