Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.1, February 2026, pp. 115~124
e-ISSN: 2548-6861

A Fuzzy C-Means–Based Clustering Model for Analyzing TOEFL Prediction Scores in Higher Education

Filipus Mei Tri Boy Gulo, Rahmad Hidayat*, Hendrawaty, Rahmat Isma Hidayat, Muhammad Heikal Fasya, Syifaurrahman, Dea Syafira Ananda
Informatics Engineering Study Program, Department of Information Technology and Computers, Lhokseumawe State Polytechnic
*rahmad_hidayat@pnl.ac.id

Article Info

Article history:

Received 2025-10-09
Revised 2025-12-20
Accepted 2025-12-22

Keywords:

Clustering,
Language,
Fuzzy C-Means,
Learning Analytics,
Xie Beni Index,
Vocational Education.

ABSTRACT

In the era of digital transformation, the application of data mining in academic data management has become an important requirement for improving the quality of education. One crucial aspect is English proficiency. One of the tools for measuring English proficiency is the Test of English as a Foreign Language (TOEFL) Prediction test, which is administered at every university, including the State Polytechnic of Lhokseumawe. The management of TOEFL Prediction scores can utilize data mining as a basis for more in-depth learning analysis, as well as evaluation material. This study aims to design and develop a model for grouping the TOEFL scores of students at State Polytechnic of Lhokseumawe by applying the Fuzzy C-Means (FCM) algorithm. The research methods included observation and interviews, data collection and pre-processing, cluster model design, web-based system development, and system testing. Evaluation was conducted through Black Box and White Box testing for the system, as well as cluster quality validation using the Xie-Beni Index (XB) and Partition Coefficient. The results showed that the pre-test dataset of first-year students (651 data) produced three clusters with an XB value of 0.623, while the dataset of final-year students (826 data) produced six clusters with an XB value of 0.181. The developed model proved to be able to map students' English language abilities in a more structured manner and could be used as a basis for academic planning and skill improvement.

I. INTRODUCTION

In this modern era, various organizations across diverse sectors are required to adapt and improve their systems to remain competitive. Two key technologies driving this transformation are big data and artificial intelligence (AI)[1]. his digital transformation has fundamentally changed the paradigm of education, enabling smarter and more measurable academic data processing[2]. In the context of vocational higher education, English proficiency is a critical indicator in preparing globally competitive graduates, especially for International Mobility programs such as Erasmus+ and other overseas scholarships. The TOEFL ITP Test and Score Summary Report data for 2024 indicate that the average TOEFL score of Indonesian students (493) remains below the minimum standard for global universities (550), highlighting the need for improvement in the evaluation system[3]. State Polytechnic of Lhokseumawe is one of the state vocational colleges that requires new students and final year students to take the TOEFL test. This test aims to measure students' initial proficiency in English as a pre-test conducted by the Language Center. In addition to serving as a measure of competency, the administration of the TOEFL at State Polytechnic of Lhokseumawe is also used as one of the selection criteria for international study abroad programs, international internship programs, and as preparation for official tests such as the international TOEFL, TOEIC, or IELTS.

As a vocational education institution, State Polytechnic of Lhokseumawe (PNL) has not yet implemented a learning analytics system in the implementation and evaluation of TOEFL test results, particularly in the evaluation of English language proficiency, which still relies on manual data processing using spreadsheets (Excel). Based on this problem, a model is needed that can automatically and objectively group students' TOEFL scores based on their proficiency patterns. One clustering method that is considered effective is Fuzzy C-Means (FCM), because it has advantages in managing data with overlapping characteristics, such as TOEFL results that do not always have clear boundaries between proficiency levels.

A number of studies have explored the use of clustering algorithms in educational data analysis to improve the objectivity and interpretability of evaluation results. Previous studies, such as those mentioned in the article “Design and Application of the DPC-K-Means Clustering Algorithm for Evaluation of English Teaching Proficiency,” show that integrating clustering algorithms into evaluation systems can map English teaching abilities into clear categories that are useful in the academic planning process [4]. Then, A. R. (2023) applied FCM to group the quality levels of education in East Java and successfully identified two main clusters: high and low quality. The research is still theoretical without system implementation [5].

Putri et al. (2022) applied FCM and Fuzzy Possibilistic C-Means (PFCM) to cluster social media data, demonstrating the usefulness of fuzzy-based clustering methods in decision support systems[6]. Similarly, Anwar (2023) utilized FCM for scholarship recipient selection, while Yasir and Firmansyah (2024) integrated FCM and Analytical Hierarchy Process (AHP) in a web-based system for job promotion[7],[8]. Although these studies prove the reliability of FCM in various domains, none have specifically applied it in educational evaluation systems.

In the context of English language proficiency assessment, a previous study titled “Designing a TOEFL System for Analyzing Participants' Weaknesses Using the K-Means Clustering Algorithm” focused on grouping TOEFL scores using the K-Means algorithm, which is a hard clustering method[9]. Although effective in some contexts, hard clustering methods like K-Means have a fundamental limitation for this type of data. Language proficiency is often a continuum, where a student may fall between two levels (between 'intermediate' and 'advanced') and cannot be rigidly assigned to a single category. This natural characteristic creates 'soft boundaries' that K-Means, by its design, cannot properly handle. To overcome this limitation, the Fuzzy C-Means (FCM) algorithm offers a more flexible 'soft clustering' mechanism. By assigning membership degrees to each data point, FCM can accurately reflect the continuous and overlapping nature of language proficiency, making it a more suitable choice for this analysis.

This study introduces an integrated analytical system for clustering student TOEFL scores using the Fuzzy C-Means algorithm, developed in the form of a hybrid web application by combining the Laravel framework for system management and Python 3.11 for analytical computing. This integration utilizes Python libraries such as NumPy, Pandas, and Matplotlib to build a robust machine learning model for soft clustering and analytical visualization. This study aims to design and develop a TOEFL Score Clustering System for students at the State Polytechnic of Lhokseumawe, as well as to implement and evaluate the Fuzzy C-Means algorithm for automatic and data-based classification of student TOEFL scores. By introducing a hybrid analytical system that integrates web technology and fuzzy clustering, this research contributes to the field of educational data mining by offering a practical and scalable approach to evaluating language proficiency in vocational education environments.

II. METHOD

This research will be conducted through six main stages that are interrelated, as illustrated in Figure 1. The stages of designing the TOEFL Score Clustering System for Students begin with observation and interviews with the Head of the Language Department at the State Polytechnic of Lhokseumawe to understand the system requirements. Next, data on the TOEFL scores of active students from the 2021–2024 cohorts will be collected, followed by data preprocessing to clean and select relevant attributes, namely Listening, Reading, and Structure. In the clustering model design stage, the Python programming language and the Fuzzy C-Means (FCM) algorithm will be used to build a model capable of grouping data optimally. After that, the system design was carried out, including architecture design, use cases, activity diagrams, databases, and user interfaces (UI/UX). The next stage was system and model testing using the black box method to ensure that the functionality was working properly and the white box method to test the accuracy of the clustering model. Fine-tuning was carried out using the Xie-Beni Index and Partition Coefficient methods to optimize the clustering parameters. The final stage of the entire process was the implementation of the TOEFL score clustering system that had been developed and tested so that it was ready for official use at the State Polytechnic of Lhokseumawe Language Department as a tool for automatically analyzing and evaluating students' English language skills based on data.

A. Clustering

Clustering is a branch of data mining that involves analytical procedures for grouping objects based on similarity parameters, whereby each cluster exhibits high internal homogeneity and clear external heterogeneity.[10]. As an unsupervised learning approach, clustering works without initial labels and is used to identify hidden patterns in data. This method is commonly applied in data mining, machine learning, and statistics to find relationships in large and complex data sets.[11].

Clustering can be performed using various approaches, including hierarchical methods that organize data gradually, partitioning methods that divide data into specific groups, and density methods that are useful for finding clusters in dense and irregular data[12]. Clustering can be used as a method of analysis in the field of sales and supports the managerial decision-making process. This technique can also be combined with decision support system methods such as AHP, which is useful in the employee selection process[8]. In addition, clustering can be combined with predictive algorithms to analyze and predict specific data, such as in research on predicting palm oil harvest yields in North Aceh[13].

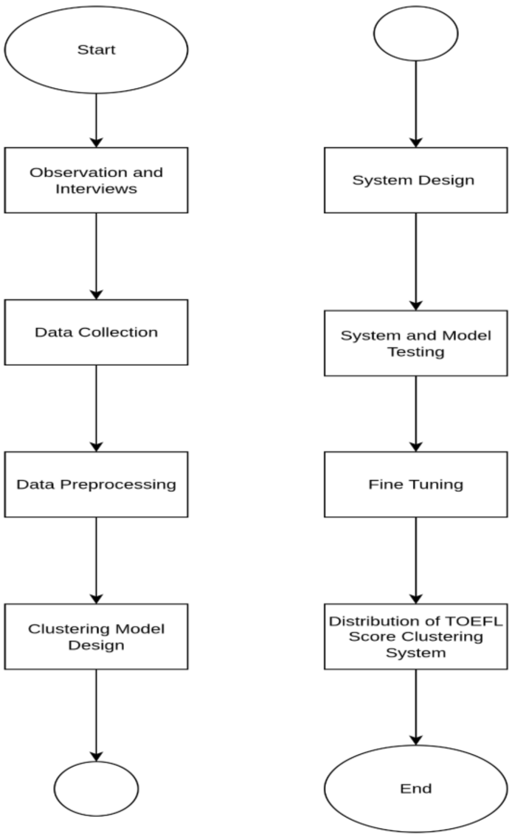

Figure 1. Flowchart of Implementation Stages

B. Fuzzy C-Means Algorithm

Fuzzy C-Means Clustering (FCM), also known as Fuzzy Isodata, is one of the most common soft clustering methods, serving as a flexible alternative to hard clustering techniques like K-Means. FCM uses a fuzzy clustering approach, where each data point can be a member of all classes or clusters with a membership level ranging from 0 to 1[13]. This membership degree indicates the extent to which a data point contributes to a cluster. In the process, this algorithm uses an iterative approach to gradually update the cluster center (centroid) and membership values until it reaches a state of convergence or minimum change. This approach is conceptually similar to the Expectation Maximization (EM) algorithm in statistics, which calculates the posterior probability for each observation, then allocates the observation to one of several groups to maximize the overall likelihood[14]. Similar to EM, FCM begins the process with a random initialization, and the final result is highly dependent on the initial parameters specified by the user, including the desired number of clusters. Therefore, selecting the right parameters is crucial to obtaining representative clustering results. Clusters in FCM are generally modeled in the form of geometric spaces such as circles (in two dimensions), which represent the dominance of a cluster based on data distribution. This method includes unsupervised clustering, allowing us to build fuzzy partitions from the data. The fuzzy clustering process is as follows:

1. The data input to be clustered in X is an n x m matrix (n = number of data samples, m = attributes). Xij = sample data to -i (i = 1,2,.....,n), attribute to -j (j = 1,2,...,n).
2. Determine
   a. Number of clusters = c
   b. Rank = w
   c. Maximum Iterations = MaxIter
   d. Expected minimum error
   e. Initial objective function, P0=0
   f. Initial iteration, t= 1
3. Generate random numbers, i = 1,2,......n; k = 1,2,....c; as elements of the initial partition matrix U:

$$Q1 = \sum_{k=1}^{c} \mu ik \quad (1)$$

With j = 1,2,.....,n

$$\mu ik = \frac{\mu ik}{Qi} \quad (2)$$

4. Calculate the distance between each data point and the centroid using Euclidean Distance:

$$Vkj = \frac{\sum_{i=1}^{n} ((\mu ik)^{w} \, x \, Xij)}{\sum_{i=1}^{n} (\mu ik)^{w}} \quad (3)$$

5. Calculate the Objective Function at iteration -t

$$Pt = \sum_{i=1}^{n} \sum_{k=1}^{n} \left(\left[\sum_{j=1}^{m} (Xij - Vkj)^{2} \right] (\mu ik)^{w} \right) \quad (4)$$

6. Calculate the change in the matrix:

$$\mu ik = \frac{\left[\sum_{j=1}^{m} (Xij - Vkj)^{2} \right]^{\frac{-1}{w-1}}}{\sum_{k-1}^{c} \left[\sum_{j=1}^{m} (Xij - Vkj)^{2} \right]^{\frac{-1}{w-1}}} \quad .(5)$$

7. Check the stop condition:
   If : $(|Pt - Pt - 1| < 0) \; atau \; (t > MaxIter)$ then stop;

   If not: t = t-1, repeat step 4

In the implementation of this study, the parameters were set specifically based on the executed code. The Fuzziness

Coefficient (m) was set at 2, which is the de facto standard in the FCM literature for optimal fuzziness balance. The maximum iteration (maxiter) was set at 1000 iterations, and the stopping criterion or error threshold (error) was set at 0.0001 or (1e-4). As for the number of clusters (c), the value was not set a priori. Instead, the model was tested by varying the number of clusters (c=2 to 6) and evaluated using the Xie-Beni Index and Partition Coefficient to find the optimal number of clusters for the dataset.

C. Partition Coefficient

The cluster validity index plays an important role in assessing the quality of data clustering and helps determine the optimal number of clusters. This assessment is carried out through a quantitative approach based on data distribution patterns and the distance between groups. Methods that are often used include the partition coefficient, silhouette index, and Davies-Bouldin[15]. According to Bezdek and Dunn (1975), there are a number of validity indices specifically designed to evaluate clustering results in the context of fuzzy clustering. One recommended index is the Partition Coefficient (PC), which serves as a measure of the degree of overlap between clusters[16]. Partition Coefficient is one of the evaluation techniques in validity indices used to measure the extent to which clustering results can be considered optimal, particularly in determining the most appropriate number of clusters. The higher the Partition Coefficient value, the more definitive the division of data into specific clusters, indicating that the clustering results are better[16].

The following is the Partition Coefficient (PC) equation :

$$PC(c) = \frac{1}{n}\sum_{i=1}^{n} \sum_{k=1}^{c} (\mu ik)^{2} \quad (6)$$

The Partition Coefficient value is used as an indicator of the degree of data membership to a cluster; a value close to 1 indicates that the cluster division is stronger and more consistent[17].

D. Xie Beni Index

The Xie-Beni Index is a method for evaluating cluster validity that aims to determine the quality of clustering results produced by fuzzy algorithms, such as Fuzzy C-Means. This index fundamentally depends on the characteristics of the data used, the geometric separation between clusters, and the minimum distance between cluster centers (centroids)[18]. In addition, this index also considers the membership degree generated by the fuzzy clustering process. The Xie-Beni value is calculated as the ratio between the amount of dispersion within the cluster (intra-cluster dispersion) and the square of the minimum distance between cluster centers. Here is the Xie Beni Index formula:

$$XB = \frac{\sum_{i=1}^{c} \sum_{k=1}^{n} \mu ik^{\ w} ||Vi - Xj||^{\ 2}}{n * min\ i,j ||Vi - Vj||^{\ 2}} \quad (7)$$

This index value is calculated as the ratio between the level of compactness within a cluster (intra-class) and the separation distance between clusters (inter-class). In principle, the smaller the distance between members within a cluster, the more compact and uniform the cluster is considered to be. Conversely, the greater the distance between cluster centers, the more separated the clusters will be and the clearer their boundaries will be. Therefore, a low Xie-Beni index value indicates that the resulting cluster structure is of good quality, as it describes a dense distribution of data within clusters and optimal separation between clusters [19].

E. Common European Framework of Reference for Languages (CEFR)

The Common European Framework of Reference for Languages (CEFR) is a framework used to measure language proficiency and ability. It was introduced in 2001 and is used in 40 languages worldwide, including English.[20]. The conversion is displayed visually in Table I.

Based on a case study conducted at the Academic Support Unit (UPA) of State Polytechnic of Lhokseumawe, it was found that student language proficiency was assessed using a Paper-Based Test (PBT) with a score range of 310 to 677. Therefore, the conversion of PBT scores into the Common European Framework of Reference for Languages (CEFR) framework is expected to be a more objective and standardized reference in evaluating students' language proficiency levels. The Common European Framework of Reference for Languages (CEFR) has been widely used by educational institutions as a reference for assessing language proficiency.

TABLE I
TOEFL PBT TO CEFR CONVERSION TABLE (SOURCE: TITC INDONESIA, 2024)

The score obtained by the test taker	Proficiency Levels based on CEFR
310-420	Elementary - Pemula
420-480	Low Intermediate - Menengah ke bawah
480-520	High Intermediate - Menengah ke atas
525-677	Advance - Mahir

This framework serves to group new students based on their language proficiency levels, develop or adapt language testing instruments, and provide objective assessments of an individual's language skills.

Thus, the CEFR helps avoid differences in perception, such as when someone feels proficient in a language, but according to institutional standards is still at a basic level. The CEFR is also universal, measurable, and relevant to real-world needs, making it a widely accepted standard in various countries[20].

III. Results and Discussion

This section presents the results of experiments and a comprehensive analysis of the performance of the TOEFL score clustering system for students built using the Fuzzy C-Means (FCM) algorithm. The evaluation was conducted to assess the quality of the clustering results based on several key parameters, such as cluster accuracy, validation index value (Xie-Beni Index), Partition Coefficient, and comparison between the two tests to determine the most optimal cluster. This analysis aims to provide a quantitative and qualitative understanding of the model's effectiveness in grouping TOEFL score data accurately and objectively. Additionally, this analysis also discusses in detail the stages of dataset selection, data pre-processing, partition coefficient testing, Xie-Beni Index, comparison of partition coefficient and Xie-Beni Index testing, selection of the optimal number of clusters, and analysis of clustering results.

A. Datasets

This study utilizes the TOEFL score dataset of State Polytechnic of Lhokseumawe students from the class of 2021 as the main object in the analysis process. The dataset used consists of two data collection periods, namely the TOEFL test results in 2022, which were conducted as a pre-test for new students, and the TOEFL test results in 2025, which were taken prior to graduation as part of the graduation requirements.

The selection of these two periods aims to provide a comparative picture of the development or changes in students' English language skills during their studies at the university. By comparing the initial and final data, it is hoped that more in-depth information can be obtained regarding the effectiveness of English language learning during the lecture period. To be clear, the clustering analysis was conducted using three specific input features (attributes): the scores for Listening, Structure, and Reading.

TABLE II
EXCERPT FROM THE TOEFL SCORE DATASET OF LHOKSEUMAWE STATE POLYTECHNIC STUDENTS

No	Name	Listening Score	Structure Score	Reading Score
1.	Alda Mauliza	42	26	39
2.	Alfi Syahrin	41	38	43
3.	Alviona Putika Husna	45	26	28
4.	Andiny	45	36	35
…	…………….	…………	………..	………..

The *total score* was intentionally excluded as an input feature for the clustering process because it is a derived value from these three components; the total score was only calculated *after* the clusters were formed to aid in their semantic interpretation. In its implementation, TOEFL score data—containing these three features—can be inputted into the system by uploading files in .csv or .xls format. The system is designed to be able to flexibly read data from both formats without changing the basic data structure. A sample of the TOEFL scores of State Polytechnic of Lhokseumawe students is shown in Table 2.

The data set contains the pre-test scores of the 2021 cohort of students for the 2022 academic year, with a total of 651 students. This data is original data obtained from UPA. Bahasa State Polytechnic of Lhokseumawe. For comparison, the dataset of pre-test scores for students enrolled in 2021 in the 2022 academic year with a total of 826 students. The comparison between these two datasets aims to evaluate the consistency of the Fuzzy C-Means (FCM) clustering model's performance against data variations and observe differences in the distribution patterns of TOEFL scores of the 2021 cohort during lectures.

B. Partition Coefficient Testing

One of the evaluation methods used in this study is the Partition Coefficient (PC). This method aims to measure the extent to which data membership in each cluster is definite or clear. The following presents the results of testing the clustering of TOEFL scores using the Partition Coefficient to evaluate the performance and consistency of the model built. Figure 2 is a graph of the Partition Coefficient test results for the dataset of the 2021 pre-test scores when the students were new in the 2022 test year.

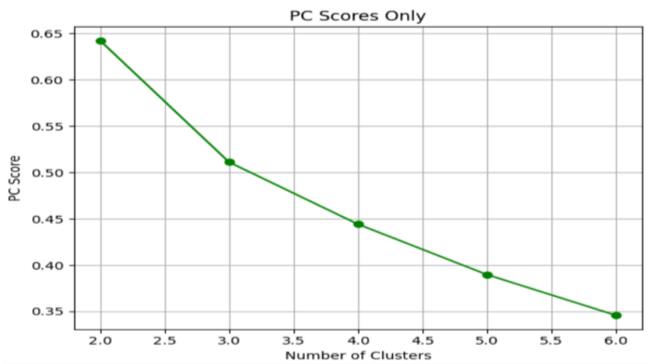

Figure 2. Partition Coefficient Test Results for the 2021 student pre-test dataset in 2022

The optimal number of clusters is 2 with a Partition Coefficient value of 0.641, which, as revealed in the theory of Partition Coefficient, the higher the value and the closer it is to 1, the better the determination of the number of clusters. Thus, the results of this test indicate that the two clusters formed from the dataset are able to represent the distribution of students' English language abilities quite clearly.

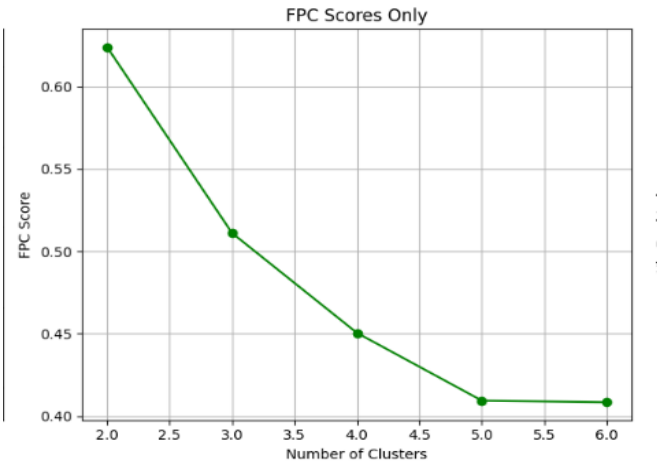

Figure 3. Partition Coefficient Test Results for the 2021 student pre-test dataset in 2025

Figure 3 is a graph of the Partition Coefficient test results for the dataset of the 2021 pre-test student test scores when the new students were tested in 2025, with an optimal number of clusters of 2 and a Partition Coefficient value of 0.623.

C. *Xie Beni Index Testing*

In addition to the Partition Coefficient, the clustering results were also evaluated using the Xie-Beni Index (XB Index). Unlike the Partition Coefficient, which tends to produce biased results for small clusters, the Xie-Beni Index provides a more realistic and balanced picture of the number of clusters and the quality of the clustering results. The results of the Xie Beni Index test are presented in Figures 6 and 7.

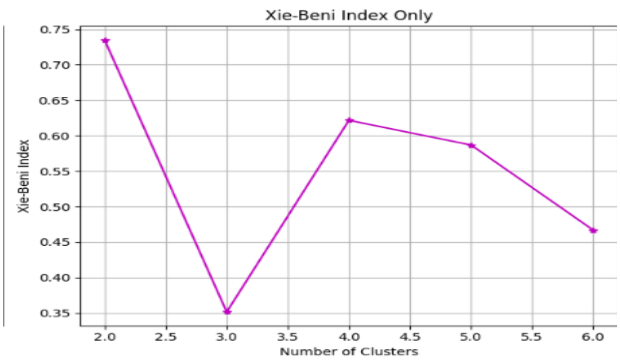

Figure 4. Results of the Xie Beni Index test for the pre-test dataset of the 2021 cohort of students in 2022

Figure 4 is a graph of the Xie Beni Index test results for the dataset of the 2021 pre-test scores of students when they were new students in the 2022 test year, with an optimal number of clusters of 3 and a Xie-Beni Index of 0.351. This value indicates good clustering quality, considering that in the Xie-Beni Index theory, it is stated that the smaller the index value, the better the cluster separation results produced by the algorithm. A value close to zero indicates that the data in each cluster tends to be homogeneous and the distance between cluster centers is quite large, so that the clusters formed can be significantly distinguished. Figure 5 is a graph of the Xie-Beni Index test results for the dataset of the 2021 pre-test scores when the students were freshmen in the 2025 test year, with an optimal number of clusters of 6 and a Xie-Beni Index of 0.181.

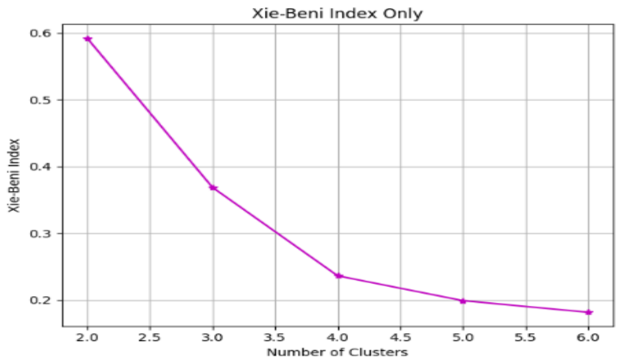

Figure 5. Results of the Xie Beni Index test for the pre-test dataset of the 2021 cohort of students in 2025

D. *Comparison of Partition Coefficient Testing and Xie Beni Index Testing*

To obtain optimal clustering results, this study used two validation methods, namely the Partition Coefficient and the Xie-Beni Index. These two methods were used complementarily because each has a different approach and focus of analysis in assessing the quality of clustering results. By comparing the results of these two methods, the study was able to obtain a more comprehensive picture of the stability, accuracy, and clarity of the boundaries between clusters.

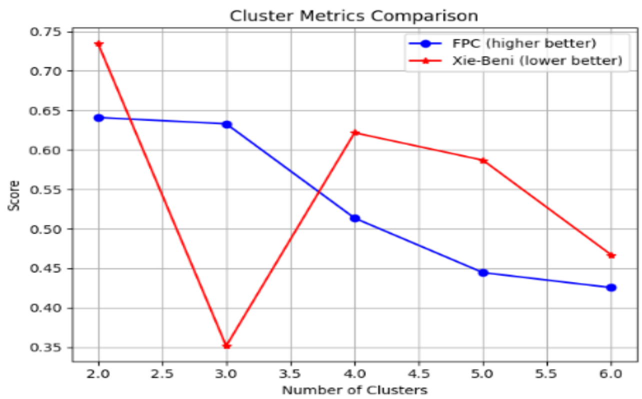

Figure 6. Comparison of Partition Coefficient and Xie Beni Index Testing for the 2021 student pre-test dataset in 2022

This approach also ensures that the number of clusters selected truly represents the natural structure of the students' TOEFL scores, so that the clustering results can be used effectively in analyzing English language proficiency at the State Polytechnic of Lhokseumawe. Figure 6 presents a comparative graph between the Partition Coefficient (PC) and Xie-Beni Index (XB) for the pre-test dataset of new students enrolled in 2021 for the 2022 test. Based on the test results, the PC method shows an optimal value in 2 clusters with a score of 0.641 (equivalent to 64.11%). This value reflects a cluster structure with a fairly good membership level, where most of the data has been clearly classified into each cluster, although there is still room for improvement in terms of separating the boundaries between clusters. Conversely, the XB method shows optimal results in 3 clusters with a value of 0.351 (equivalent to 64.9%). In the context of XB, the smaller the value obtained, the better the clustering quality. This value indicates that the cluster structure formed has a fairly clear separation and minimal overlap between clusters.

Considering both evaluation results, the Xie-Beni Index (XB) method was chosen as the main reference in determining the optimal number of clusters in this study. This selection is based on the superiority of XB in handling fuzzy clustering characteristics, as this index not only considers the level of data compactness within clusters, but also takes into account the distance between cluster centers to measure the quality of separation between clusters. Therefore, in the context of this study, the selection of three clusters based on the lowest XB value is considered the most representative and appropriate for the distribution of student TOEFL scores.

Figure 7. Comparison of Partition Coefficient and Xie Beni Index Testing for the 2021 student pre-test dataset in 2025

Figure 7 presents a comparative graph between the Partition Coefficient (PC) and Xie-Beni Index (XB) for the pre-test dataset of new students enrolled in 2021 for the 2025 test. Based on the test results, the PC method shows an optimal value in 2 clusters with a score of 0.623 (equivalent to 62.03%). This value reflects that most of the data has been relatively clearly classified into specific clusters, although there is still the possibility of overlap between clusters. Conversely, the XB method shows optimal results in 6 clusters with a value of 0.181 (equivalent to 81.9%). In the context of XB, the smaller the value obtained, the better the clustering quality. This value indicates that the clustering process can be considered successful in forming clusters with a stable structure, separate, and minimal membership ambiguity.

Considering these two results, the Xie-Beni Index method was chosen as the main reference in determining the optimal number of clusters. This is based on the superiority of XB over PC and its advantage in handling fuzzy clustering because it considers the distance between cluster centers and the quality of separation between clusters. In addition, XB produces spatially distinct clusters with minimal overlap, making the Xie-Beni Index superior and more stable than PC, which tends to be biased toward fewer clusters and decreases in value as the number of clusters increases. Thus, the selection of six clusters based on the lowest XB value is considered the most representative in the clustering process in this study.

E. Cluster Results Analysis

This section discusses the analysis of TOEFL score clustering results obtained through the application of the Fuzzy C-Means (FCM) algorithm. The analysis was conducted to identify patterns in the distribution of students' English language abilities based on three main assessment components, namely Listening, Reading, and Structure. In addition, this analysis also aims to evaluate the effectiveness of the FCM model in grouping data objectively by considering the membership degree of each data point to each cluster. The clustering results are then interpreted to understand the characteristics of each group of student abilities and their relevance to the learning process and academic policies at the State Polytechnic of Lhokseumawe.

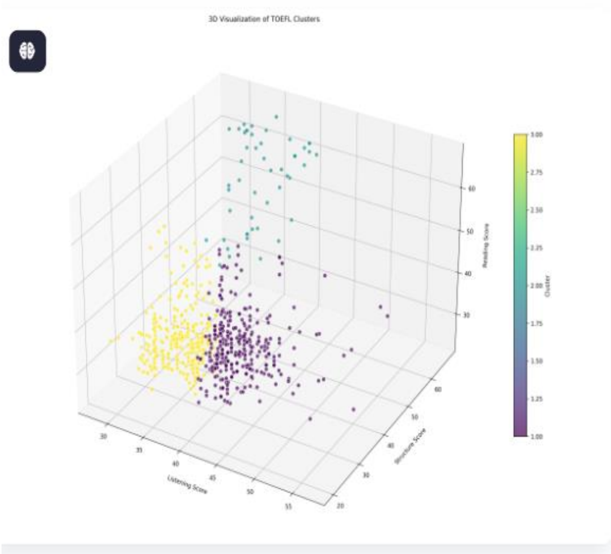

Figure 8. Cluster distribution of the 2021 student pre-test dataset in 2022

Figure 8 illustrates the distribution of the 2021 student pre-test dataset (2022), which forms three clusters: the yellow cluster represents students with low Listening and Reading scores (30–45), the purple cluster indicates those with moderate abilities but inconsistent performance across components (45–55), and the green cluster represents high-achieving students with strong Reading and Structure skills, showing potential for international programs.

Cluster	Characteristic and Recomendations
1 Cluster 1	Listening Average: 41.4 Structure Average: 35.1 Reading Average: 34.3 Total Score: 369.5 (CEFR: A2–Elementary) Dominant Skill: Listening Area of Improvement: Reading
2 Cluster 2	Listening Average: 36.8 Structure Average: 56.2 Reading Average: 58.5 Total Score: 504.9 (CEFR: B2–High Intermediate) Dominant Skill: Reading Area of Improvement: Listening
3 Cluster 3	Listening Average: 34.3 Structure Average: 31.9 Reading Average: 33.9 Total Score: 333.8 (CEFR: A2–Elementary) Dominant Skill: Reading Area of Improvement: Listening

Figure 9. Cluster analysis results and the characteristics of the 2021 student pre-test dataset in 2022

For more details on the cluster analysis results and the characteristics of each cluster in the 2021 student pre-test dataset for 2022, see Figure 9. Figure 9 shows the results of clustering the pre-test dataset of 651 students from the 2021 cohort in 2022, divided into three clusters. Cluster 1 consists of 315 students (48.4%), with the following characteristics: average Listening score of 41.4, Structure score of 35.1, and Reading score of 34.3, resulting in a total TOEFL score of 369.5. Based on the CEFR scale, this score is at the A2 (Elementary) level. The dominant skill of students in this cluster is Listening, while the area that needs improvement is Reading. Cluster 2 included 51 students (7.8%), with the following characteristics: average Listening score of 36.8, Structure score of 56.2, and Reading score of 58.5, resulting in a total TOEFL score of 504.9. Based on the CEFR, this score is at the B2 (High Intermediate) level. The dominant skill of students in this cluster is Reading, while the area for improvement is Listening.

Cluster 3 includes 285 students (43.8%), with an average Listening score of 34.3, Structure score of 31.9, and Reading score of 33.9, resulting in a total TOEFL score of 333.8. This score is also categorized at level A2 (Elementary) according to the CEFR. Students in this cluster demonstrate dominant abilities in reading, with the main area for improvement being listening. The TOEFL passing standard applied at the State Polytechnic of Lhokseumawe is a minimum total score of 400. Based on this standard, from the pre-test dataset of new students in the 2021 cohort, 105 students (16.1%) passed, and 546 students (83.9%) failed.

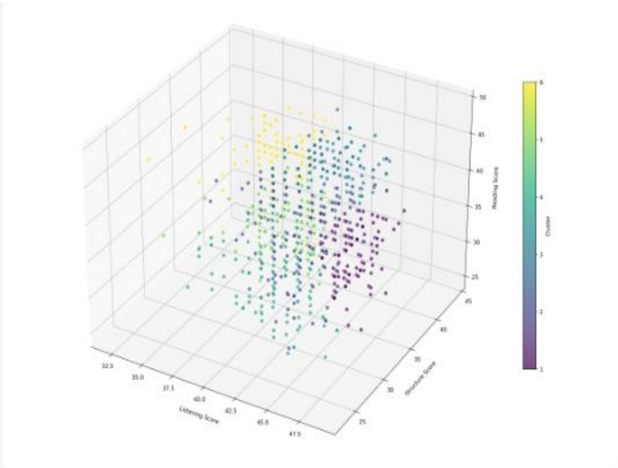

Figure 10. Cluster distribution of the pre-test dataset for the 2021 student cohort in 2025

The data distribution in the figure 10 shows that clusters are formed based on a combination of three main attributes, namely Listening (X-axis), Structure (Y-axis), and Reading (Z-axis), rather than just one specific dimension of ability. Clusters that show strong Reading abilities, located in Clusters 2, 3, and 6, appear to be concentrated in areas with higher Z values. Conversely, clusters with weaknesses in Structure, namely Clusters 2 and 4, tend to have lower Y values. In addition, the overlapping zones between clusters show that there are students with mixed profiles, for example, those with intermediate Listening skills but high Reading skills.

This condition demonstrates the superiority of the Fuzzy C-Means (FCM) algorithm, which is able to maintain information in the form of dual membership degrees between clusters. From the overall distribution, Cluster 3 stands out as the group with the best performance because it has an average total TOEFL score above 400, making it an important indicator in assessing students' readiness to participate in international programs. For more details on the cluster analysis results and the characteristics of each cluster in the 2021 student pre-test dataset for 2022, see Figure 11.

Cluster 1 consists of 126 students (15.3%). The characteristics of this cluster are: an average Listening score of 45.0, Structure score of 35.6, and Reading score of 33.6, with a total TOEFL score of 381.0. Based on the CEFR scale, this result falls into the A2 (Elementary) category. The dominant skill of students in this cluster is Listening, while the area that needs improvement is Reading. Cluster 2 consists of 128 students (15.5%), with an average Listening score of 44.3, Structure score of 28.1, and Reading score of

40.7, and a total score of 376.8. This score also falls into the A2 (Elementary) category. Students in this cluster have a dominant ability in Reading and need improvement in Structure.

Cluster	Characteristic and Recomendations
1 Cluster 1	Listening Average: 45.0 Structure Average: 35.6 Reading Average: 33.6 Total Score: 381.0 (CEFR: A2-Elementary) Dominant Skill: Listening Area of Improvement: Reading
2 Cluster 2	Listening Average: 44.3 Structure Average: 28.1 Reading Average: 40.7 Total Score: 376.8 (CEFR: A2-Elementary) Dominant Skill: Reading Area of Improvement: Structure
3 Cluster 3	Listening Average: 43.7 Structure Average: 35.2 Reading Average: 44.7 Total Score: 411.7 (CEFR: A2-Elementary) Dominant Skill: Reading Area of Improvement: Listening
4 Cluster 4	Listening Average: 41.6 Structure Average: 29.8 Reading Average: 32.4 Total Score: 346.2 (CEFR: A2-Elementary) Dominant Skill: Listening Area of Improvement: Reading
5 Cluster 5	Listening Average: 39.8 Structure Average: 35.9 Reading Average: 34.0 Total Score: 365.8 (CEFR: A2-Elementary) Dominant Skill: Structure Area of Improvement: Reading
6 Cluster 6	Listening Average: 38.8 Structure Average: 36.0 Reading Average: 44.2 Total Score: 396.6 (CEFR: A2-Elementary) Dominant Skill: Reading Area of Improvement: Listening

Figure 11. Cluster analysis results and the characteristics of the 2021 student pre-test dataset in 2022

Cluster 3 includes 161 students (19.5%). The average Listening score is 43.7, Structure is 35.2, and Reading is 44.7, with a total score of 411.7. This score falls into the A2 (Elementary) category. Reading is the dominant skill, while Listening is the area for improvement. Cluster 4 also consists of 161 students (16.5%), with an average Listening score of 41.6, Structure score of 29.8, and Reading score of 32.4, with a total score of 346.2. This cluster is also at the A2 (Elementary) level. The dominant skill is Listening, while the area that needs improvement is Reading.

Cluster 5 consists of 148 students (17.9%), with an average Listening score of 39.8, Structure score of 35.9, and Reading score of 34.0, resulting in a total score of 365.8. This cluster shows a dominance in Structure, with Reading as the area that needs improvement. Cluster 6 consists of 127 students (15.4%), with an average Listening score of 38.8, Structure score of 36.0, and Reading score of 44.2, resulting in a total score of 396.6. This score is still in the A2 (Elementary) category, with Structure as the dominant skill and Reading as the area for improvement.

Based on the TOEFL passing standard at State Polytechnic of Lhokseumawe, which is a minimum total score of 400, out of 826 students, 219 students (26.5%) passed, while 607 students (73.5%) failed. Cluster 1 consisted of 126 students (15.3%). The characteristics of this cluster are: an average Listening score of 45.0, Structure score of 35.6, and Reading score of 33.6, with a total TOEFL score of 381.0. Based on the CEFR scale, this result falls into the A2 (Elementary) category. The dominant skill of students in this cluster is Listening, while the area that needs improvement is Reading.

IV. Conclusion

The stages of designing a TOEFL score clustering model for students using the Fuzzy C-Means algorithm include data preprocessing, determining the optimal number of clusters using the Xie Beni Index and Partition Coefficient methods, clustering with Fuzzy C-Means, and evaluating the clustering results. The clustering results for the pre-test dataset of 651 new students enrolled in 2021 show that the optimal number of clusters is 3, with an XB Index value of 0.623 (equivalent to 62.03%) and the following cluster distribution: cluster 1 with 315 students (48.4%), cluster 2 with 51 students (7.8%), and cluster 3 with 285 students (43.8%). The passing standard for the State Polytechnic of Lhokseumawe in this dataset was 105 students (16.1%), while 546 students (83.9%) failed. The clustering results for the final-year student test dataset for the class of 2021, comprising 826 students, indicate that the optimal number of clusters is 6, with an XB Index of 0.181 (equivalent to 81.9%). The cluster distribution is as follows: cluster 1 has 126 students (15.3%), cluster 2 has 128 students (15.5%), cluster 3 has 161 students (19.5%), cluster 4 has 161 students (16.5%), cluster 5 has 148 students (17.9%), and cluster 6 has 127 students (15.4%). For the graduation standard of State Polytechnic of Lhokseumawe in this dataset, 219 students (26.5%) passed, and 607 students (73.5%) did not pass. This clustering model has practical implications for academic policy at the Language Center of the Lhokseumawe State Polytechnic. Based on the semantic interpretation of the clusters (which the system automatically highlights to identify specific areas of weakness, such as Reading, Listening, or Structure, for each group) and the dominant abilities of each group, the institution can design targeted interventions. Specifically, this model enables the Language Center to transition from a general (one-size-fits-all) training method to tailored training programs based on recommendations and the unique weakness profiles of each cluster. This ensures that the allocation of resources and educational interventions—such as intensive workshops in Listening or Structure—can be carried out precisely, making the learning process more adaptive, efficient, and data-driven.

ACKNOWLEDGEMENTS

The author would like to express his gratitude to the Directorate of Learning and Student Affairs (Belmawa), Ministry of Education, Culture, Research, and Technology through the Student Creativity Program (PKM) for the funding support provided so that this research could be carried out properly. Gratitude is also extended to the State Polytechnic of Lhokseumawe, particularly the Language Department, for the facilities, data, and opportunities provided in the implementation of this research. The support from various parties has been crucial to the success of this research activity and the preparation of this scientific work.

REFERENCES

[1]. Z. Waznah and M. I. P. Nasution, "Peran Teknologi Terbaru: Big Data dan Kecerdasan Buatan dalam Mengoptimalkan Sistem Informasi Manajemen Organisasi," Neraca Manajemen, Ekonomi, vol. 10, no. 9, 2024. DOI: 10.8734/mnmae.v1i2.359.

[2]. Direktorat Pembelajaran dan Kemahasiswaan, Direktorat Jenderal Pendidikan Tinggi Riset dan Teknologi dan Kementerian Pendidikan Kebudayaan Riset dan Teknologi. 2024. *Buku Panduan Merdeka Belajar-Kampus Merdeka 2024*. Jakarta.

[3]. *TOEFL ITP Test and Score Data Summary Report, January–December 2024*. ETS, 2025. [Online]. Tersedia: https://www.ets.org/toefl/itp

[4]. M. Niu, "Design and Application of the DPC-K-Means Clustering Algorithm for Evaluation of English Teaching Proficiency," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 8, 2024, doi: 10.14569/ijacsa.2024.0150851.

[5]. [5] N. Ulinnuha dan D. C. R. Novitasari, "Penerapan *Fuzzy C-Means* Untuk Pengelompokkan Tingkat Kualitas Pendidikan Di Jawa Timur," Simetris J. Tek. Mesin Elektro Dan Ilmu Komput., vol. 14, no. 2, hlm. 419–426, Nov 2023, doi: 10.24176/simet.v14i2.9442.

[6]. G. N. S. Putri, D. Ispriyanti, dan T. Widiharih, "Implementasi Algoritma *Fuzzy C-Means* Dan Fuzzy Possibilistics C-Means Untuk Klasterisasi Data Tweets Pada Akun Twitter Tokopedia," *J. Gaussian*, vol. 11, no. 1, hlm. 86–98, Mei 2022, doi: 10.14710/j.gauss.v11i1.33996.

[7]. A. N. Anwar, "Implementasi Fuzzy C-Mean (FCM) untuk Menentukan Penerima Beasiswa" Jurnal Ilmu Komputer, vol. 6, no. 1, pp. 84–87, 2023.

[8]. A. Yasir and A. U. Firmansyah, "Implementasi metode *Fuzzy C-Means* dan metode AHP dalam pemilihan promosi jabatan karyawan berbasis web (Studi kasus: PT. Tunas Dwipa Matra Sekampung)," Journal of Science and Social Research, vol. 7, no. 4, pp. 1616–1619, 2024.

[9]. A. R. N. Nabella, Hani Zulfia Zahro', dan Yosep Agus Pranoto, "Rancang Bangun Sistem TOEFL Untuk Analisis Kelemahan Peserta Dengan Penerapan Algoritma K-Means Clustering," *Infotek J. Inform. Dan Teknol.*, vol. 8, no. 1, hlm. 94–103, Jan 2025, doi: 10.29408/jit.v8i1.28260.

[10]. Y. P. Putra and R. Nuari, "Application of K-Means algorithm to cluster students' *Reading* patterns in the digital age," INOVTEK Polbeng - Seri Informatika, vol. 10, no. 1, pp. 320–331, 2025.

[11]. M. A. Septianto, A. Faqih, dan A. R. Rinaldi, "Klasterisasi Data Produksi Pertanian Di Kabupaten Cirebon Dengan Algoritma K-Means," *J. Inform. Dan Tek. Elektro Terap.*, vol. 13, no. 2, Apr 2025, doi: 10.23960/jitet.v13i2.6174.

[12]. M. Cui et al., "Introduction to the k-means clustering algorithm based on the elbow method," Accounting, Auditing and Finance, vol. 1, no. 1, pp. 5–8, 2020.

[13]. M. Ula, G. Perdinanta, R. Hidayat, dan I. Sahputra, "Analyze the Clustering and Predicting Results of Palm Oil Production in Aceh Utara," *IJCCS Indones. J. Comput. Cybern. Syst.*, vol. 17, no. 2, Apr 2023, doi: 10.22146/ijccs.83195.

[14]. D. Krasnov, D. Davis, K. Malott, Y. Chen, X. Shi, dan A. Wong, "Fuzzy C-Means Clustering: A Review of Applications in Breast Cancer Detection," *Entropy*, vol. 25, no. 7, hlm. 1021, Jul 2023, doi: 10.3390/e25071021.

[15]. R. K. Verma, R. Tiwari, dan P. S. Thakur, "Partition Coefficient and Partition Entropy in Fuzzy C Means Clustering," *J. Sci. Res. Rep.*, vol. 29, no. 12, hlm. 1–6, Des 2023, doi: 10.9734/jsrr/2023/v29i121812.

[16]. D. Nurmin, M. N. Hayati, dan R. Goejantoro, "Penerapan Metode *Fuzzy C-Means* pada Pengelompokan Kabupaten/Kota di Pulau Kalimantan Berdasarkan Indikator Kesejahteraan Rakyat Tahun 2020," *Laboratorium Statistika Terapan dan Statistika Komputasi, FMIPA, Universitas Mulawarman*, 2020

[17]. S. F. Octavia dan M. Mustakim, "Penerapan K-Means dan Fuzzy C-Means untuk Pengelompokan Data Kasus Covid-19 di Kabupaten Indragiri Hilir," *Build. Inform. Technol. Sci. BITS*, vol. 3, no. 2, hlm. 88–94, Sep 2021, doi: 10.47065/bits.v3i2.1005.

[18]. M. A. Mallik, N. F. Zulkurnain, M. K. Nizamuddin, dan A. Kc, "An Efficient Fuzzy C-Least Median Clustering Algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1070, hlm. 012050, Feb 2021, doi: 10.1088/1757-899x/1070/1/012050.

[19]. H. Y. Wang, J. S. Wang, dan L. F. Zhu, "A new validity function of FCM clustering algorithm based on intra-class compactness and inter-class separation," *J. Intell. Fuzzy Syst.*, vol. 40, no. 6, hlm. 12411–12432, Jun 2021, doi: 10.3233/jifs-210555.

[20]. B. Rais dan L. Awwalia, "What is the student's level? Analyzing Students' English Proficiency Levels in High Education Level," *SALEE Study Appl. Linguist. Engl. Educ.*, vol. 6, no. 1, hlm. 305–316, Feb 2025, doi: 10.35961/salee.v6i1.1762.