KKU SCIENCE JOURNAL
Journal Home Page : https://ph01.tci-thaijo.org/index.php/KKUSciJ
Published by the Faculty of Science, Khon Kaen University, Thailand

การจัดกลุ่มคะแนนสอบ ESTS ของนักศึกษาคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเลย

Clustering of ESTS Test Scores of Students in the Faculty of Science and Technology at Loei Rajabhat University

สุไพลิน พิชัย[1] ดุลชาติ ศิริวัลลภ[1*] และ ทอแสง พิมพ์เบ้าธรรม[1]
Supailin Pichai[1], Dulachat Siriwallop[1*] and Thosang Phimbaotham[1]
[1]สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเลย 42000
[1]Computer Science major, Faculty of Science and Technology, Loei Rajabhat University, 42000, Thailand

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาเว็บแอปพลิเคชันสำหรับรายงานผลสอบ English for Science and Technology (ESTS) ของนักศึกษา คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเลย โดยมุ่งเน้นการนำเทคโนโลยีสารสนเทศมาใช้ในการวิเคราะห์ผลสอบและจัดกลุ่มข้อมูลนักศึกษาด้วยเทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) ผ่านกระบวนการจัดกลุ่มข้อมูล (Clustering) เพื่อสะท้อนภาพรวมของศักยภาพด้านภาษาอังกฤษได้อย่างมีประสิทธิภาพ ระบบดังกล่าวพัฒนาด้วยภาษา PHP ฐานข้อมูล MySQL และใช้ phpMyAdmin เป็นเครื่องมือจัดการฐานข้อมูล โดยได้นำเทคนิคการจัดกลุ่มข้อมูลด้วยอัลกอริทึม K-Means Clustering โดยทำการทดลองจำนวนกลุ่มตั้งแต่ 2 ถึง 20 กลุ่ม ใช้การสุ่มจุดศูนย์กลาง พร้อมการวัดระยะห่างด้วย Euclidean Distance และประเมินคุณภาพการจัดกลุ่มด้วยค่าดัชนี Davies-Bouldin (DBI) ผลการทดลองพบว่า การจัดกลุ่มที่ค่า K = 3 ให้ค่า DBI ต่ำที่สุดที่ 0.82 แสดงถึงประสิทธิภาพในการจำแนกกลุ่มข้อมูลได้อย่างชัดเจน ขณะเดียวกัน การเพิ่มจำนวนคลัสเตอร์ส่งผลให้ค่าเฉลี่ยระยะห่างภายในกลุ่มลดลง และค่า DBI ลดลงตาม ซึ่งบ่งชี้ถึงแนวโน้มของการจัดกลุ่มที่กะทัดรัดและมีความแตกต่างกันมากขึ้น จากการพัฒนาระบบนี้ ผู้ใช้งานสามารถเข้าถึงผลสอบในรูปแบบที่มีการจัดกลุ่มเชิงวิเคราะห์ได้สะดวก ส่งผลให้สามารถประเมินศักยภาพของนักศึกษาแต่ละกลุ่มได้อย่างครอบคลุม และสามารถนำข้อมูลไปใช้ในการวางแผนการพัฒนาทักษะภาษาอังกฤษได้อย่างเหมาะสมกับระดับความสามารถของผู้เรียน

ABSTRACT

This study aims to develop a web application for reporting English for Science and Technology (ESTS) examination results of students in the Faculty of Science and Technology, Loei Rajabhat University. The focus is on applying information technology to analyze test results and group student data using unsupervised learning via clustering, in order to effectively portray students' overall English proficiency. The system was developed in PHP with a MySQL database and managed through phpMyAdmin. K-means

*Corresponding Author, E-mail: dulachat.sri@lru.ac.th
Received date: 2 April 2025 | Revised date: 17 August 2025 | Accepted date: 25 August 2025
doi: 10.14456/kkuscij.2025.32

clustering was employed, experimenting with the number of clusters from 2 to 20, using random initialization and Euclidean distance. Clustering quality was evaluated with the Davies–Bouldin Index (DBI). The experiments show that K = 3 yielded the lowest DBI of 0.82, indicating clear cluster separation. At the same time, increasing the number of clusters reduced the average within-cluster distance and lowered the DBI, suggesting a trend toward more compact and distinct groupings. With this system, users can conveniently access analytically grouped score reports, enabling a comprehensive assessment of students' proficiency by cluster and supporting the planning of tailored English skill development aligned with learners' ability levels.

คำสำคัญ: การจัดกลุ่ม การจัดกลุ่มเคมีน ทักษะภาษาอังกฤษ
Keywords: Clustering, K-Mean Clustering, English Language Skill

INTRODUCTION

Faculty of Science and Technology, Loei Rajabhat University, has arranged for students to practice English language skills in order for them to have English proficiency according to the Common European Framework of Reference for Languages (CEFR) guidelines. The CEFR is an international standard used to describe language proficiency levels. It is widely accepted in Europe, and its adoption is increasing. Practice exercises and tests are available through the English for Science and Technology (ESTS) web application, with scores for each part of the test given to each student in the Faculty of Science and Technology. The results of these tests do not show the overall picture of the Faculty of Science and Technology and the fields of study, making the test data unable to be used to much benefit. As can be seen from the detailed display of the data report, each test takes some time to search. This makes it inconvenient for administrators, officials, or even faculty members in each subject to bring information to use at the desired time, including problems with the data presentation format. In addition, grouping student exam results data can be applied as important information in making decisions about planning work in many areas. For example, developing the skills of students in groups with little English proficiency in order to plan teaching and learning that increases the learning efficiency of student groups for those interested in research has not yet been responded to with methods and formats according to the concepts and principles of information in the Thailand 4.0 era.

Clustering techniques, together with classification techniques, can also be used in other areas, such as screening children with learning disabilities from the behavioral context (Baadel *et al*.,2020).

Data analysis and clustering techniques are widely used techniques. Data with similar characteristics will be arranged in the same group, and data with different characteristics will be arranged in different groups. Data clustering techniques are currently being used in various fields, such as K-Means Clustering and Hierarchical Clustering Algorithms, which are efficient algorithms for clustering data and are widely used in research.

Based on these factors, this investigation suggests the implementation of a web-based ESTS results reporting system that is integrated with clustering algorithms to categorize students according to their English proficiency levels. The system will produce actionable insights to support curriculum planning, targeted skill development, and strategic policy-making in alignment with both the objectives of the institution and the national goals set by the Ministry of Higher Education, Science, Research, and Innovation. These insights will be produced by comparing the performance of various algorithms and selecting an approach that is the most suitable for this dataset.

MATERIALS AND METHODS

Dataset Details

This study employed the English for Science and Technology (ESTS) examination results of 460 undergraduate students from the Faculty of Science and Technology at Loei Rajabhat University for the 2023 academic year to construct the dataset. The students represented twelve academic programs across diverse subjects, including Computer Science, Biology, Chemistry, Physics, and Environmental Science.

The ESTS examination was designed to evaluate students' English language proficiency by dividing it into four distinct sections, each concentrating on a certain linguistic skill. The sections comprised Part 1 (Vocabulary), Part 2 (Reading Comprehension), Part 3 (Listening Comprehension), and Part 4 (Grammar and Structure). The assessments for each section were recorded as numerical values on a scale ranging from 0 to 100.

The key attributes for the clustering analysis were the four section values, designated as Part1 (Vocabulary Score), Part2 (Reading Comprehension Score), Part3 (Listening Comprehension Score), and Part 4 (Grammar Score). To attain a more precise and nuanced classification of students according to their performance profiles, these criteria were chosen as they represent diverse yet complementary facets of English proficiency.

Table 1 Sample data of English for Science and Technology

		Part 1	Part 2	Part 3	Part 4	Total Score (400)	Percentage of score (100)
1	xxx	86.67	83.33	85.71	85.71	341.42	85.35
2	xxx	60	58.33	42.86	42.86	204.05	51.01
3	xxx	40	58.33	71.43	71.43	241.19	60.30
4	xxx	46.67	0	0	0	46.67	11.67
5	xxx	40	58.33	71.43	71.43	241.19	60.30

Data Preprocessing

Prior to conducting the clustering analysis, the dataset underwent several preprocessing steps to ensure data quality and consistency.

1. Data Filtering Only students who completed all four sections of the ESTS were retained in the dataset from 529 student records, and proper data filtering resulted in 460 valid records. Incomplete data filtering is important to avoid bias in clustering the results (Han *et al.*, 2011).

2. Handling Missing Values A thorough examination revealed no missing values in the filtered dataset, eliminating the need for imputation. Missing value treatment is a crucial step in data preprocessing to ensure validity.

3. The scores in all four sections are 100 points, giving a total score of 400. They are all normalized to a fixed scale from 0 - 100 using min - max normalization. This process ensures that all attributes participate equally in the clustering process and prevents large-scale attributes from dominating the distance calculations. (Jain, 2010).

4. Outlier detection using Z-score analysis ($|z| < 3$) No severe outliers were detected in the data set, confirming that all observed data fall within statistically acceptable limits. This method preserves the integrity of the clustering analysis, and the data can be grouped into each of the four dimensions.

These preprocessing steps ensured that the dataset was clean, consistent, and ready for subsequent clustering analysis.

Tools and Technologies

The data for this study were initially collected through a web-based application specifically developed for recording and reporting the English for Science and Technology (ESTS) examination results. The system was built on a MySQL relational database, managed via phpMyAdmin, to store and organize the scores of students across all four test sections. This centralized storage allowed for efficient retrieval, filtering, and preparation of the dataset for further analysis.

Once extracted from the database, the data were imported into Orange Data Mining, a visual analytics platform that supports workflow-based data processing and machine learning. Orange was used to design the initial clustering workflow, including preprocessing, normalization, and algorithm selection, through its drag-and-drop interface. To enhance analytical flexibility and precision, Python 3.11 was integrated into the process. Python’s extensive data science ecosystem, particularly the Pandas and NumPy libraries, was used for advanced data manipulation, while scikit-learn implemented the K-Means and Hierarchical Clustering algorithms. Matplotlib and Seaborn were employed to generate detailed visualizations such as cluster distribution plots and dendrograms, complementing Orange’s built-in visual outputs.

Data Mining Theory

Data mining, also known as Knowledge Discovery in Databases (KDD), is a technique for automatically discovering patterns in large amounts of data using algorithms from machine learning and pattern recognition. In another definition, Data mining is the process of working with large amounts of data to find patterns and relationships hidden in the dataset using statistics, recognition, machine learning, and many forms of mathematical knowledge obtained from data mining. It uses past data to find relationship patterns and new knowledge from the data (Phromma, 2013).

Cluster Analysis

Cluster analysis is an unsupervised learning technique aiming to group data with similar characteristics into groups (Tsai *et al*., 2011). Each group of data is called a cluster. Data analysis or classification can be divided into two types: hierarchical algorithms and non-hierarchical algorithms.

K-Means and Hierarchical Clustering were selected for this study due to their simplicity, scalability, and interpretability. K-Means is efficient for large datasets and works well with numerical data, while Hierarchical Clustering provides a tree-like structure that helps in understanding relationships among data points. Other clustering methods, such as DBSCAN or Gaussian Mixture Models, were not chosen due to their complexity or unsuitability for the dataset used in this research.

K-Means Clustering

K-Means Clustering (Yokkampol *et al.,* 2018) is an algorithm where each member within a group is closest to the center or representative of that group (Centroid). The process of grouping data using the K-Means Clustering Algorithm consists of initializing the number of initial groups, randomizing centroids, arranging each data into groups, and updating the centroid in each group (Rujasiri, 2009).

The method can be summarized into five steps as follows:

1. Initialize the number of groups (k) to be divided.

2. Randomize the centroid of each group.

3. Calculate the distance of each data point from the centroid of each group to determine which group the data point should belong to, based on the smallest distance.

4. Arrange the center of each group's data by calculating the average value of the data within the group (x_j) as shown in Equation (2) to be used as the new representative of the group.

5. Repeat steps 3 and 4 until all members of each group are unchanged to the other groups.

The centroid of the data or the representative of the group (x_j) according to step 4 can be calculated as follows:

$$x_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij} \quad ; i = 1,2,\dots,n_j \quad ; j = 1,2,\dots,k$$

Where n_j represents the number of data points in group j

x_{ij} represents the value of the variable or the observed value for unit i in group j

k represents the total number of groups.

In this research method, the process involves randomly selecting centroids that change according to randomly generated values, with the procedure repeated 100 times.

The K-Means Clustering method has the advantage of being simple, widely used, and capable of clustering data quickly.

An example of data clustering using the K-Means algorithm, when dividing the data into 3 groups, is shown in the image.

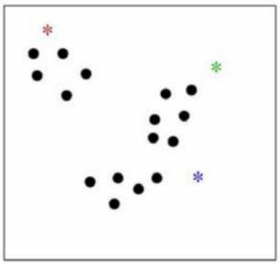

Figure 1 Initializing data points for use in K-Means Clustering.

From Figure 1, the black circle symbol represents any data point. The red, green, and blue asterisk symbols represent randomly selected data points that are used as the centroids of the data groups. In this case, the number of groups is set to 3 (k = 3).

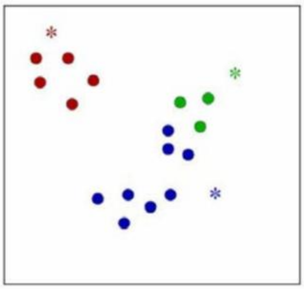

Figure 2 An example of data clustering by considering the distance between the data points and the centroids randomly selected in the first iteration.

From Figure 2, the red, green, and blue circle symbols represent the data points that have been clustered, based on the distance between the data points and the centroids randomly selected in the first iteration. The red, green, and blue asterisk symbols represent the centroids of the data groups that were randomly selected in the first iteration.

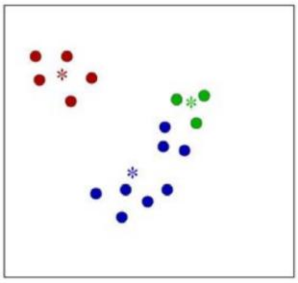

Figure 3 An example of adjusting the centroid of each group to a position that best covers the nearby data points within the same group, by calculating the average value of the data within the group.

From Figure 3, the red, green, and blue circle symbols represent the data points, while the red, green, and blue asterisk symbols represent the centroids of the data groups that have been adjusted to positions that better encompass the data within the same group.

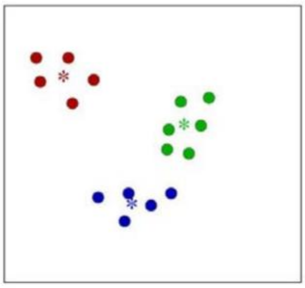

Figure 4 The completed clustering, where the resulting centroids are positioned as the representatives of each data group.

From Figure 4, the red, green, and blue circle symbols represent the data points, while the red, green, and blue asterisk symbols represent the centroids, which are the most suitable representatives of the data in each group.

Hierarchical Clustering

Hierarchical Clustering is a technique that groups data based on their similarity, using measures of similarity or dissimilarity such as Euclidean, Cityblock, Mahalanobis, and Cosine (Oranuch, 2005). The output of Hierarchical Clustering is displayed in the form of a tree, where each class node consists of child nodes. This technique can be divided into two types of tree-building methods: Agglomerative (Bottom-Up) and Divisive (Top-Down) (Oranuch, 2005).

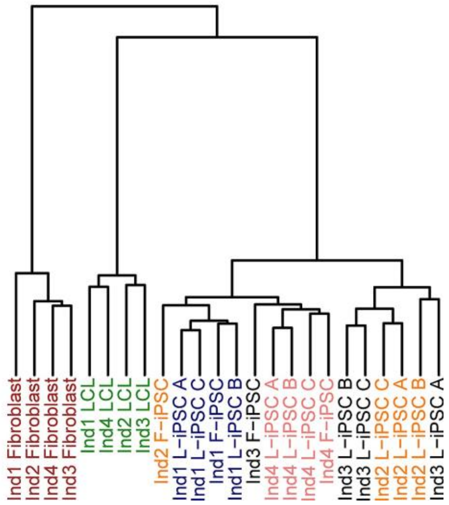

Figure 5 Hierarchical Clustering of gene expression data. (Burrows *et al*., 2016)

Hierarchical Clustering can be displayed in the form of a tree diagram called a dendrogram, where the number of groups can be determined by drawing a line across the dendrogram at a certain distance or similarity value.

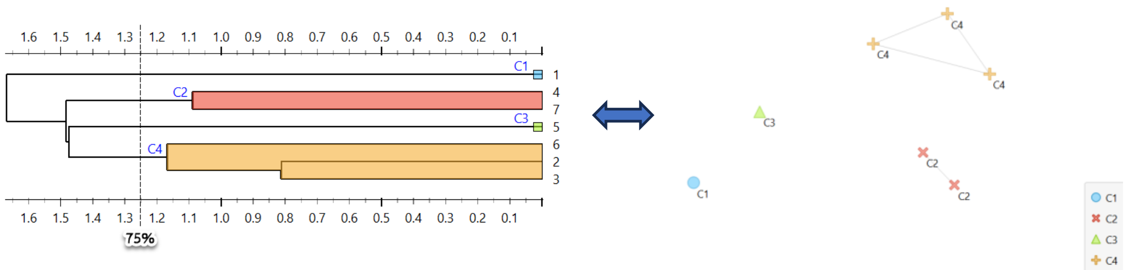

Figure 6 Example of creating and Hierarchical Clustering using a dendrogram.

From Figure 6 on the left, when the distance is set to 2.5 and a line is drawn across the dendrogram, the data can be grouped into three clusters, as shown in Figure 6 on the right.

Hierarchical Clustering can be divided into two types:

1. Agglomerative works by starting with the data grouped into n clusters, where each of the n data points is an external node. Then, it merges the two clusters with the smallest distance between them, combining two clusters at a time. This process is repeated until all the data points are merged into a single cluster, which becomes the root node.

2. Divisive works similarly to the agglomerative method, but with the key difference that it operates in the opposite direction. That is, the clustering starts from the root node and divides down to the external nodes, or from top to bottom. This method is not as popular because it is more complex and requires more computational time compared to the agglomerative method. (Köhn and Hubert, 2014).

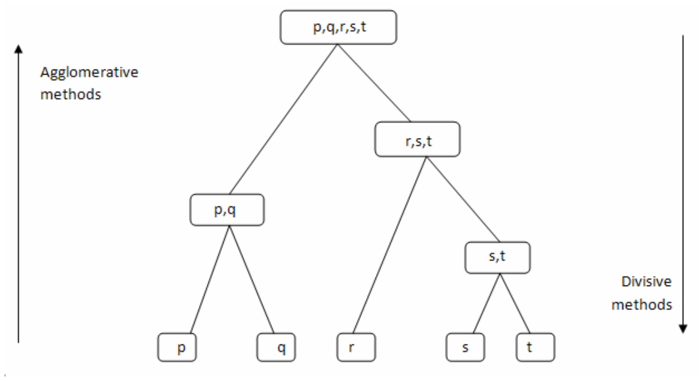

Figure 7 Examples of group formation: Agglomerative and Divisive methods. (Erman *et al.* 2015)

Model Performance Comparison

Performance testing of clustering algorithms: This research selects algorithms for data classification to test and compare their performance. The two algorithms used for comparison are K-Means Clustering (KMC) and Hierarchical Clustering Algorithms (HCA). When both algorithms are applied to cluster the data, their performance can be evaluated.

The performance of the data clustering methods in this research is measured by the Correctly Clustered Instances (CCI), which compares the clusters obtained from the data under study with those formed using the Hierarchical Clustering and K-Means Clustering methods. Here, m_i represents the number of data points correctly clustered in the respective group, and n is the total number of data points in that dataset. The accuracy of the clustering method is calculated as follows:

$$CCI = \frac{\sum_{i=1}^{k} m_i}{n} x100$$

This study employed K-Means and Hierarchical Clustering algorithms to group students based on their English for Science and Technology (ESTS) test scores, which include four components: vocabulary, reading comprehension, listening, and grammar.

The K-Means algorithm was selected for its simplicity and effectiveness in handling numerical data. It works by partitioning the dataset into k clusters, minimizing the intra-cluster variance through iterative centroid updates. The value of k was tested from 2 to 20.

To evaluate clustering quality, the Davies–Bouldin Index (DBI) was used. DBI measures the average similarity between clusters, where a lower DBI value indicates better separation and compactness among the clusters. The optimal number of clusters was determined by selecting the k that produced the lowest DBI value.

In addition, Hierarchical Clustering using the agglomerative approach and Euclidean distance was used as a comparative method. A dendrogram was generated to support visual interpretation of the cluster structure.

RESULTS AND DISCUSSION

The study of the clustering algorithm using K-Means Clustering for the ESTS exam results of the Faculty of Science and Technology, Loei Rajabhat University, involved using a Performance Vector to determine the optimal number of clusters. The resulting clusters were validated by comparing them with the data labels. Clustering involves grouping the data by selecting the best value for k (the number of clusters) based on the Performance Vector, which is calculated by measuring the distance between the data points and the centroid using Euclidean Distance. The experiment conducted clustering for k = 2, 3, ... 20, and the most suitable value for k was determined. The test was conducted under the assumption that there was no prior knowledge of the number of data groups. The K-Means algorithm was chosen to cluster the data into different groups, with k representing the number of clusters. The values of k were selected from k = 2, 3, ... 20, and the best value for k was chosen based on different evaluation techniques. Once the optimal value for k was selected, it was compared with the actual number of groups to identify the errors.

The results of the K-Means algorithm with different values of k involve performing the Performance Cluster Distance, where the average within-cluster distance is calculated using the average distance between each data point and the centroid as the criterion. This process is used to determine when to stop the iterative process of the K-Means algorithm after reaching the maximum number of 100 iterations. Several values are reported as follows:

1. Avg_within_centroid_distance is the calculation of the average within-cluster distance, using the average distance between the centroids of all clusters.

2. The Davies-Bouldin index is a criterion used to measure the quality of clustering, which is used in the analysis for data partitioning. The calculation of the Davies–Bouldin index is the ratio between the sum of the dispersion of data within the clusters and the distance between the clusters. It can be seen that for good clustering, the dispersion within the clusters should be low, and the distance between the clusters should be large.

Clustering was tested using the K-Means Clustering method to find the best clustering. Based on the ESTS exam scores, it was found that there were 529 participants from 12 fields of study. After excluding participants who did not complete all parts of the exam, the total number of participants was 460 from 12 fields of study. Clustering was performed using the K-Means method with the number of clusters set from 2 to 20, with 100 iterations, random centroids selected from the sample data, and the distance measured using Euclidean Distance. The average distance between the data points and the centroids of each cluster, as well as the Davies–Bouldin index, are shown in Table 2.

From the table and graph, it can be observed that as the number of clusters K increases, the average distance between the data points and the centroids of each cluster decreases. When considering the Davies–Bouldin index, it is found that at K = 3, the Davies-Bouldin index is the lowest at 0.82. This indicates that clustering with K = 3 is the best clustering solution. This suggests that students' ESTS scores can be meaningfully categorized into three distinct groups, which are interpreted as follows: Cluster 2 (High-performance group): Students with consistently high scores in all four parts. Cluster 1 (Moderate-performance group): Students with average scores, showing strengths in some sections but weaknesses in others. Cluster 0 (Low-performance group): Students with below-average scores across all test components. When considering the clustering with K = 3, it was found that the number of members in each cluster is as shown in the table.

Table 2 The average distance between the data points and the centroids of each cluster

Cluster NO	Avg. within centroid distance_cluster	Davies–Bouldin
2	825.15	1.02
3	519.46	**0.82**
4	442.12	1.07
5	378.23	1.09
6	332.72	1.11
7	303.81	1.09
8	280.94	1.19
9	262.57	1.21
10	245.01	1.20
11	229.08	1.17
12	215.79	1.16
13	202.19	1.10
14	197.08	1.08
15	184.84	1.07
16	177.39	1.11
17	169.45	1.10
18	164.93	1.10
19	158.86	1.05
20	152.92	1.06

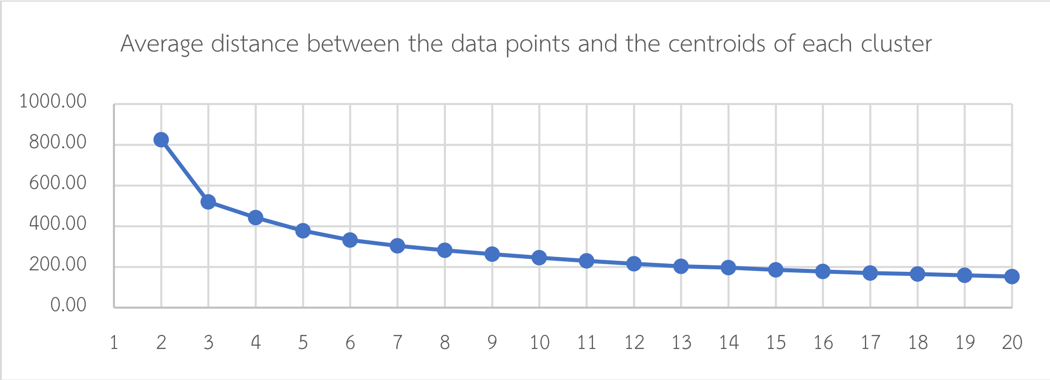

Figure 8 Graph of the average distance between the data points and the centroids of each cluster.

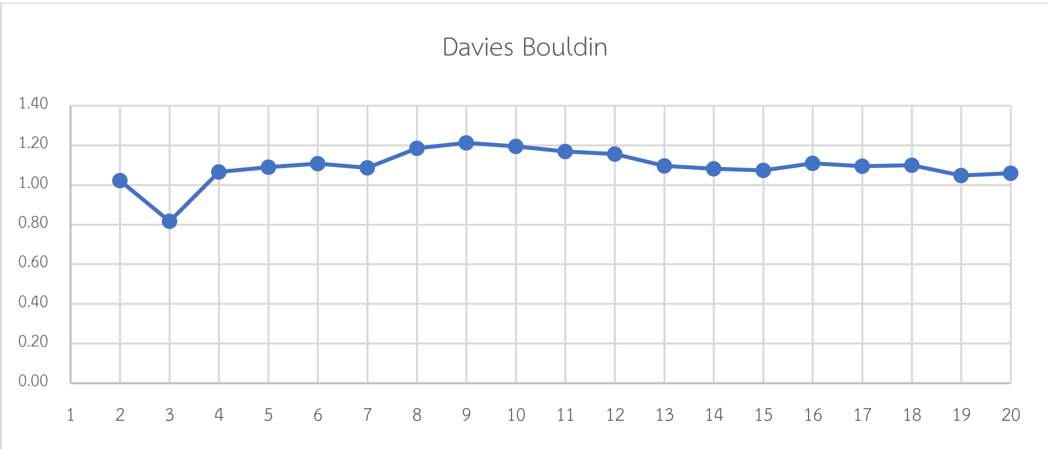

Figure 9 Graph showing the Davies–Bouldin index.

Table 3 The number of members in each cluster

Cluster	Number of members	Percentage
Cluster 0	70	14.93
Cluster 1	61	13.01
Cluster 2	338	72.07
Total	**469**	**100**

Table 4 Average scores of each exam part and cluster

Attribute	Cluster 0 (70 Persons)	Cluster 1 (61 Persons)	Cluster 2 (338 Persons)
Part 1	42.00	78.47	77.43
Part 2	50.36	18.71	87.45
Part 3	54.08	85.01	78.82
Part 4	46.53	81.26	76.63

Considering the exam scores for each part, it was found that Cluster 0 has an average score range for each part between 42 and 54.08. The part with the highest average score is Part 3, with an average score of 54.08, while the part with the lowest average score is Part 1, with an average score of 42. Cluster 1 has an average score range for each part between 18.71 and 85.01. The part with the highest average score is Part 3, with an average score of 85.01, while the part with the lowest average score is Part 1, with an average score of 78.47. Cluster 2 has an average score range for each part between 77.43 and 87.45. The part with the highest average score is Part 2, with an average score of 87.45, while the part with the lowest average score is Part 1, with an average score of 87.45. When further examining the details of each group, Cluster 0 was analyzed by looking at the average scores of each part to determine the score range within the group. It was found that in Part 1, the average score is 42. When dividing the score range into a histogram with a width of 10 points, most of the scores fall between 43.33 and 53.33. The graph of the Part 1 scores for Cluster 0 is shown in Figure 10.

Further analysis of the score distributions within each cluster revealed key characteristics: Cluster 2 (High-performing): These students scored above 75 in most sections, indicating a strong foundation in English, particularly in grammar and vocabulary. They likely require advanced or enrichment instruction to maintain engagement. Cluster 1 (Mid-level): Students in this group had moderate scores, often struggling in Part 2 (Reading) while performing better in Listening. This suggests a need for targeted support in reading comprehension and academic vocabulary. Cluster 0 (Low-performing): Students consistently scored below 50 across all sections. This group would benefit from remedial instruction focusing on basic grammar and listening skills, with an emphasis on building confidence in using English.

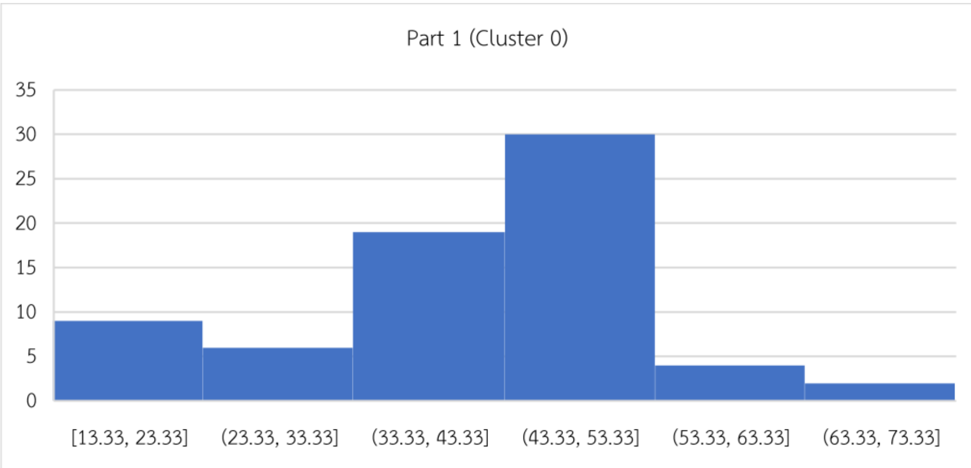

Figure 10 Graph of the Part 1 scores for Cluster 0.

In Part 2, the average score is 50.36. When dividing the score range into a histogram with a width of 10 points, most of the scores fall between 48.33 and 58.33. The graph of the Part 2 scores for Cluster 0 is shown in the figure.

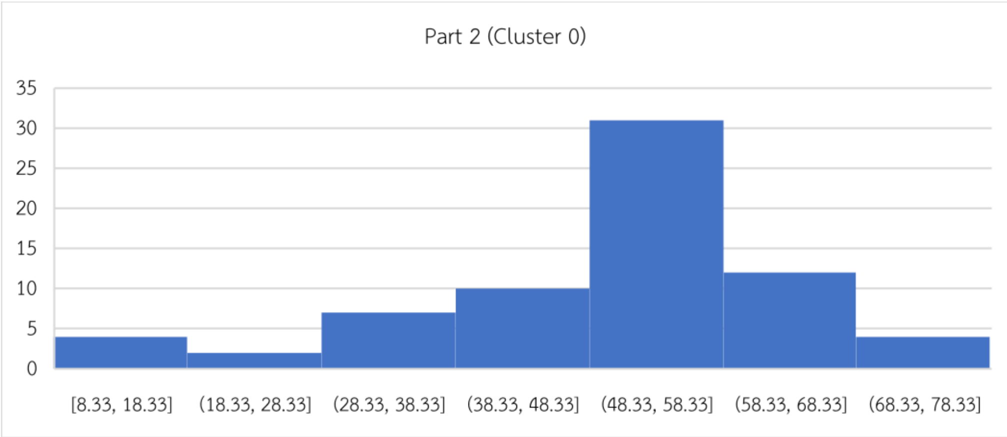

Figure 11 Graph of the Part 2 scores for Cluster 0.

In Part 3, the average score is 54.08. When dividing the score range into a histogram with a width of 10 points, most of the scores fall between 65.29 and 82.29. The graph of the Part 3 scores for Cluster 0 is shown in Figure 12.

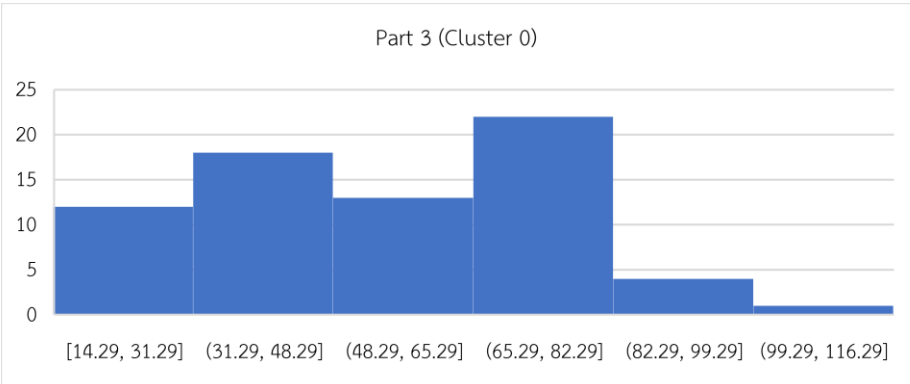

Figure 12 Graph of the Part 3 scores for Cluster 0.

In Part 4, the average score is 46.53. When dividing the score range into a histogram with a width of 10 points, most of the scores fall between 14.29 and 30.29. The graph of the Part 4 scores for Cluster 0 is shown in Figure 13.

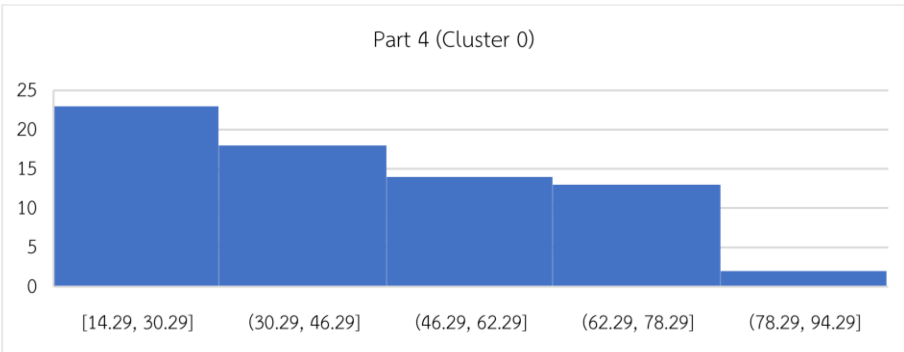

Figure 13 Graph of the Part 4 scores for Cluster 0.

In Cluster 1, when examining the average scores of each part to determine the score range within the group, it was found that in Part 1, the average score is 78.47. When dividing the score range into a histogram with a width of 10 points, most of the scores fall between 63.33 and 73.33. The graph of the Part 1 scores for Cluster 1 is shown in Figure 14.

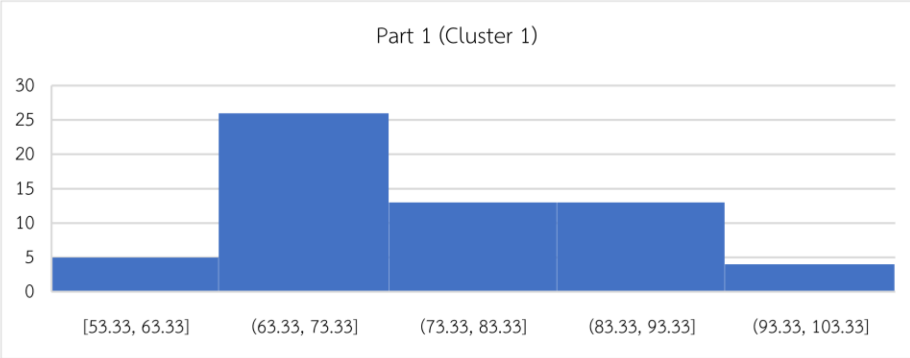

Figure 14 Graph of the Part 1 scores for Cluster 1.

In Part 2, the average score is 18.71. When dividing the score range into a histogram with a width of 10 points, most of the scores fall between 8.13 and 18.33. The graph of the Part 2 scores for Cluster 1 is shown in Figure 15.

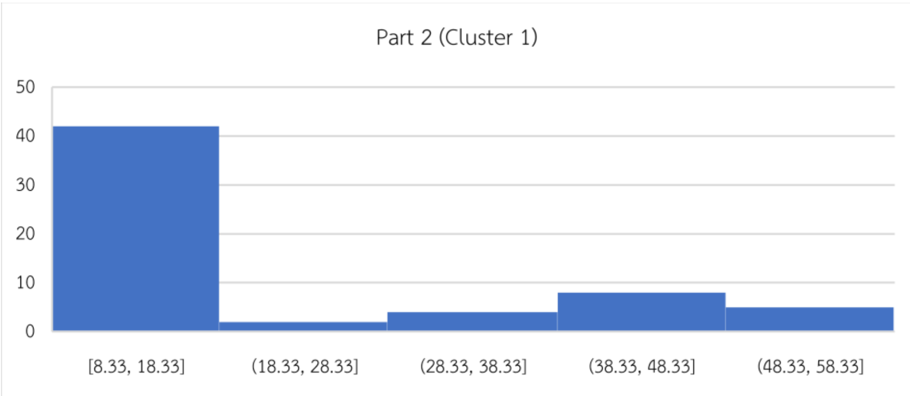

Figure 14 Graph of the Part 2 scores for Cluster 1.

In Part 3, the average score is 85.01. When dividing the score range into a histogram with a width of 10 points, most of the scores fall between 71.43 and 81.43. The graph of the Part 3 scores for Cluster 1 is shown in Figure 15.

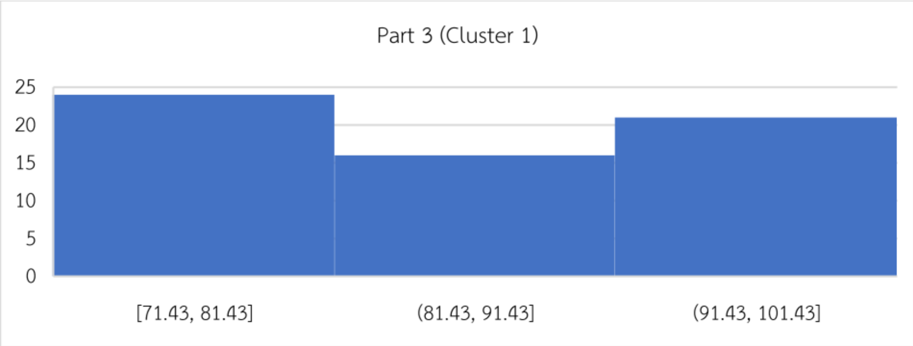

Figure 15 Graph of the Part 3 scores for Cluster 1.

In Part 4, the average score is 81.26. When dividing the score range into a histogram with a width of 10 points, most of the scores fall between 71.43 and 81.43. The graph of the Part 4 scores for Cluster 1 is shown in Figure 16.

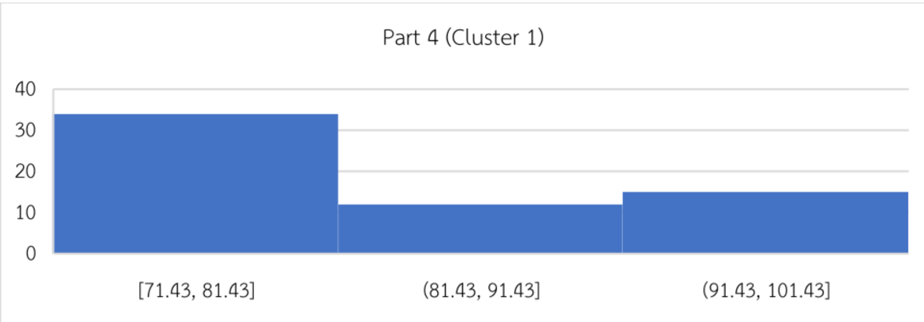

Figure 16 Graph of the Part 4 scores for Cluster 1.

In Cluster 1, when examining the average scores of each part to determine the score range within the group, it was found that in Part 1, the average score is 78.47. When dividing the score range into a histogram with a width of 10 points, most of the scores fall between 63.33 and 73.33. The graph of the Part 1 scores for Cluster 1 is shown in Figure 16. In Part 2, the average score is 18.71, when dividing the score range into a histogram with a width of. 10 points, most of the scores fall between 8.33 and 18.33. The graph of the Part 2 scores for Cluster 1 is shown in Figure 17.

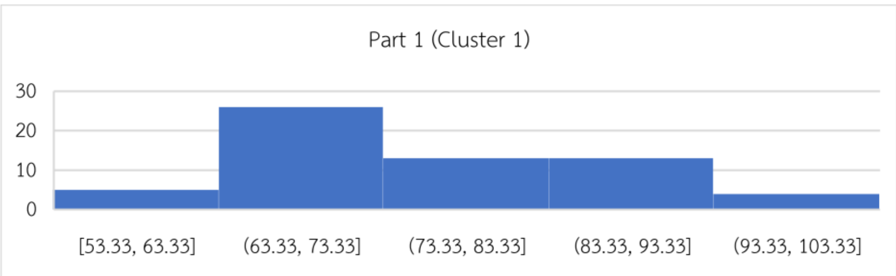

Figure 16 Graph of the Part 1 scores for Cluster 1.

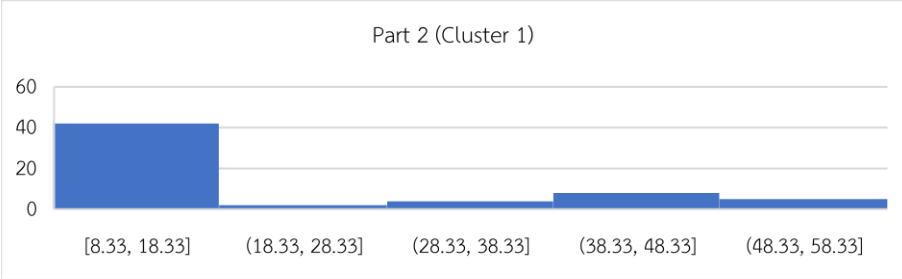

Figure 17 Graph of the Part 1 scores for Cluster 1.

In Part 3, the average score is 85.01, when dividing the score range into a histogram with a width of 10 points, most of the scores fall between 71.43 and 81.43. The graph of the Part 3 scores for Cluster 1 is shown in Figure 17.

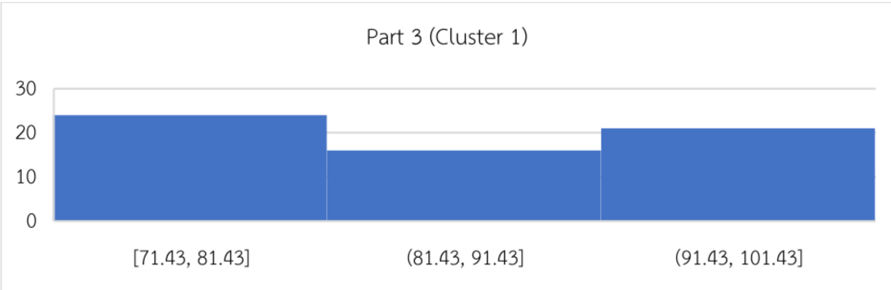

Figure 17 Graph of the Part 3 scores for Cluster 1.

In Part 4, the average score is 81.26, when dividing the score range into a histogram with a width of 10 points, most of the scores fall between 71.43 and 81.43. The graph of the Part 4 scores for Cluster 1 is shown in Figure 18.

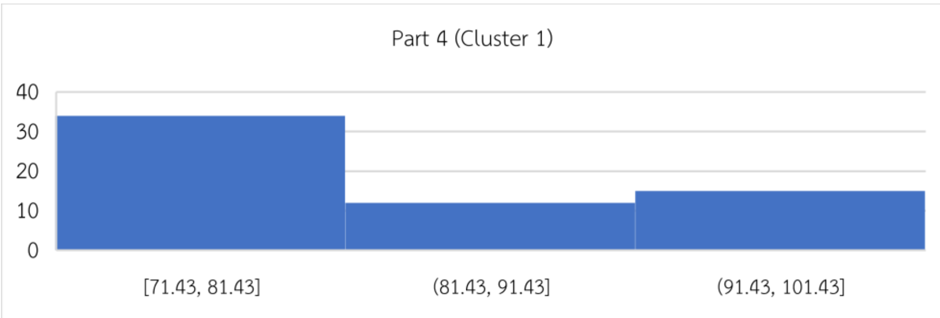

Figure 18 Graph of the Part 4 scores for Cluster 1.

In Cluster 2, when examining the average scores of each part to determine the score range within the group, it was found that in Part 1, the average score is 77.43. When dividing the score range into a histogram with a width of 10 points, most of the scores fall between 63.33 and 73.33. The graph of the Part 1 scores for Cluster 2 is shown in Figure 19.

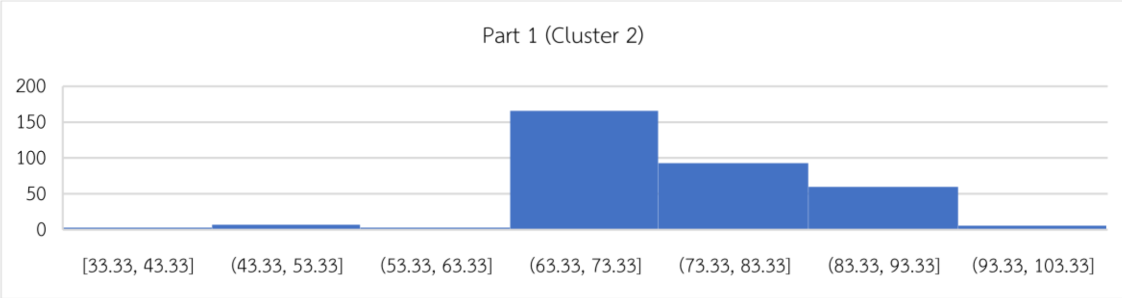

Figure 19 Graph of the Part 1 scores for Cluster 2.

In Part 2, the average score is 87.45. When dividing the score range into a histogram with a width of 10 points, most of the scores fall between 88.33 and 98.33. The graph of the Part 2 scores for Cluster 2 is shown in Figure 20.

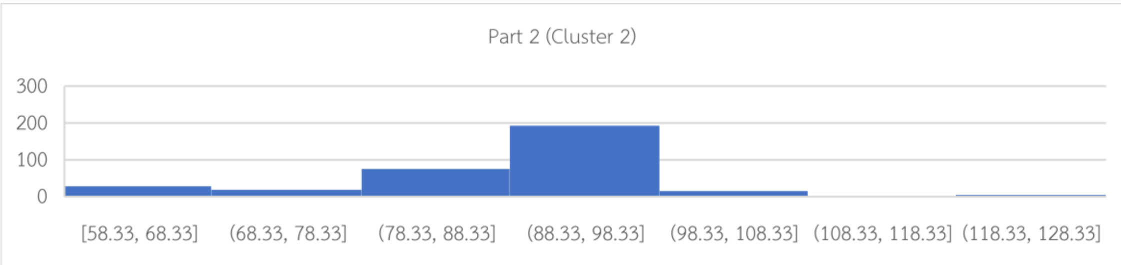

Figure 20 Graph of the Part 2 scores for Cluster 2.

In Part 3, the average score is 78.82. When dividing the score range into a histogram with a width of 10 points, most of the scores fall between 67.14 and 77.14. The graph of the Part 3 scores for Cluster 2 is shown in Figure 21.

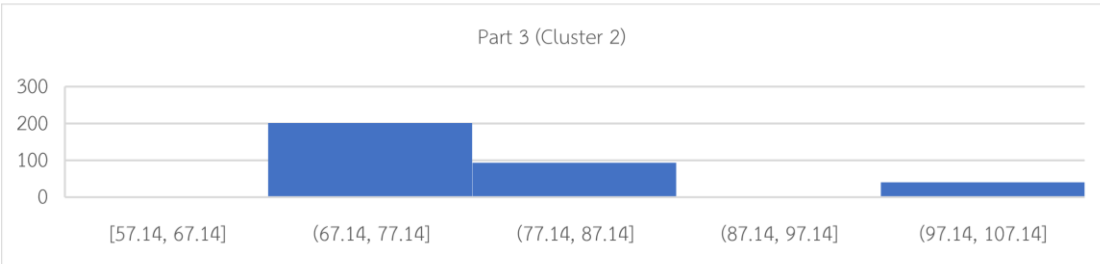

Figure 21 Graph of the Part 3 scores for Cluster 2.

In Part 4, the average score is 76.63. When dividing the score range into a histogram with a width of 10 points, most of the scores fall between 68.57 and 78.57. The graph of the Part 4 scores for Cluster 2 is shown in Figure 22.

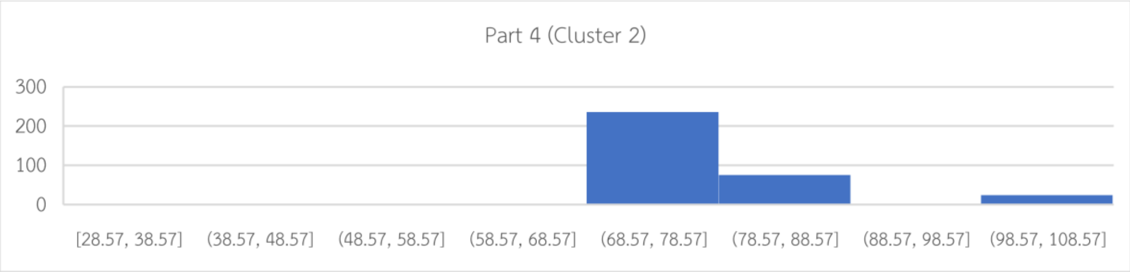

Figure 22 Graph of the Part 4 scores for Cluster 2.

Practical Implications

This study lays the groundwork for tailoring English instruction in the Faculty of Science and Technology. Different skill levels can be catered for in the curriculum design and extracurricular activities. Students who are at risk can get extra help and their progress can be checked in on a regular basis. These methods are especially useful for educational institutions that want to improve language skills in science and technology fields.

Limitations of the Study

The study only looked at ESTS results in vocabulary, reading, listening, and grammar. It didn't look at other skills or motivational factors. The dataset was only from one school, which means it might not be applicable to other situations. It was up to the person interpreting clusters to make decisions, and the data only showed one point in time, so it wasn't possible to see how things changed over time. Selecting K = 3 effectively categorizes students into high, moderate, and low proficiency groups, enabling more targeted curriculum design and support strategies.

CONCLUSIONS

According to the clustering test using the K-Means Clustering method to find the best clustering, it was found that from the ESTS exam scores of the Faculty of Science and Technology, there were a total of 529 participants from 12 fields of study. After excluding the data of participants who did not complete all parts of the exam, the total number of participants remaining was 460 from 12 fields of study.

From the clustering using the K-Means Clustering method, with the number of clusters set from 2 to 20, 100 iterations were performed, with random centroids selected from the sample data. The distance was measured using Euclidean Distance, and the average distance between the data points and the centroids of each cluster, along with the Davies–Bouldin index, were calculated. It was found that for clustering with K = 3, the Davies–Bouldin index was the lowest at 0.82, meaning that clustering with K = 3 is the best clustering solution.

The results of the K-Means Clustering method show that the average within-centroid distance and the Davies–Bouldin index were measured for different numbers of clusters, ranging from 2 to 20. The lowest average within-centroid distance was 152.92, while the lowest Davies–Bouldin index was 1.05, which occurred with 20 clusters. This suggests that increasing the number of clusters generally leads to a decrease in the average within-centroid distance and a reduction in the Davies–Bouldin index, indicating that the data can be partitioned into distinct and more compact clusters.

The novelty of this research lies in integrating clustering analysis with domain-specific English test data to support targeted curriculum development. Future research might expand the dataset, include additional language skill measures, and apply hybrid clustering techniques to increase classification accuracy.

REFERENCES

Baadel, S., Thabtah, F. and Lu, J.(2020). A clustering approach for autistic trait classification. Informatics for Health and Social Care 45(3): 309 - 326. doi: 10.1080/17538157.2019.1687482.

Burrows, C.K., Banovich, N.E., Pavlovic, B.J., Patterson, K., Gallego Romero, I., Pritchard, J.K. and Gilad, Y. (2016). Hierarchical clustering and principal components analysis. PLOS Genetics doi: 10.1371/journal.pgen.1005793.g002.

Chaimuen, O. (2015). A comparative study of the division of Thai handicraft customer data using the 2-step method SOM with K-Mean Algorithm and Hierarchical with K-Mean Algorithm. Master's thesis, Kasetsart University.

Erman, N., Korosec, A. and Suklan, J. (2015). Performance of selected agglomerative hierarchical clustering methods. Innovative Issues and Approaches in Social Sciences 8(1): 180 - 204. doi: 10.12959/issn.1855-0541.IIASS-2015-no1-art11.

Han, J., Pei, J. and Kamber, M. (2011). Data mining: Concepts and techniques (3rd ed.). U.S.A.: Morgan Kaufmann.

Jain, A.K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters 31(8): 651 - 666. doi: 10.1016/j.patrec.2009.09.011

Köhn, H. F. and Hubert, L. (2014). Hierarchical Cluster Analysis. In Wiley StatsRef: Statistics Reference Online 1 - 13. doi: 10.1002/9781118445112.STAT02449.PUB2.

Phalitsorn, W., Ditsawan, N. and Sangpom, N. (2016). A development management information system for project and research using cluster analysis techniques: Case study Computer Engineering Division in Rajamangala University of Technology Suvarnabhumi. Suvarnabhumi Rajamangala University of Technology.

Phromma, K. (2013). Comparison of data clustering efficiency using 2-step clustering algorithm. Master's thesis, Chulalongkorn University.

Rujasiri, P. (2009). Comparison of clustering techniques for cluster analysis. Master's thesis, Kasetsart University. Bangkok. 378 - 388.

Tsai, C.F., Tsai, C.T., Hung, C.S. and Hwang, P. S. (2011). Data Mining Techniques for Identifying Students at Risk of Failing a Computer Proficiency Test Required for Graduation. Australasian Journal of Educational Technology 27(3): 481 - 498.

Yokkampol, U., Rattanalertnusorn, A. and Jaroengeratikun, U. (2018). A comparison of the efficiency of data clustering using hierarchical methods and k-means methods for mixed data with clusters and numerical data. In: 2018 National Applied Statistics and Information Technology Academic Conference. National Institute of Development Administration. ISBN(e-book): 978-974-231-999-1.

❑❑❑❑❑