Application of the K-Means Algorithm to Evaluate Basic Programming Abilities of Undergraduate Students in Applied Software Engineering Technology

Made Pasek Agus Ariawan[1], Putu Indah Ciptayani[2],
and Ida Bagus Adisimakrisna Peling[3]

[1,2,3]Information Technology Department, Politeknik Negeri Bali, Bali, Indonesia
pasekagus@pnb.ac.id

Abstract. As official educational institutions, universities are faced with demands to produce graduates who are superior and competent in facing developments in science and technology. In this context, the Diploma IV Software Engineering Technology (D4 TRPL) study program aims to produce a quality workforce in software engineering technology. Research shows that the level of student success influences the quality of education and, in this digital era, the competitive ability of universities to utilize the resources they have. Basic programming skills are crucial, especially for students in study programs related to information technology. Clustering methods such as K-Means can evaluate students' basic programming abilities. The advantages of K-Means include its ability to group large objects and increase the speed of the grouping process. Previous research used the K-Means algorithm to group students based on their level of programming ability, with practical and effective results. Based on prior research, the researcher proposes further research by applying the clustering method to evaluate students' basic programming abilities. This research aims to provide a comprehensive picture of students' level of programming ability, assist lecturers in monitoring and guidance according to ability level, and support the development of more effective learning strategies through machine learning methods.

Keywords: Evaluation, Improving The Quality of Learning, K-Means Algorithm, Machine Learning

1 Introduction

Progress in science and technology encourages universities, as official educational institutions, to produce superior and competent graduates. Learning methods in higher education should be more innovative, creative, and responsive to job market demands (Fadrial, 2020). The Bachelor of Applied Software Engineering Technology Study Program, abbreviated as D4 TRPL, is one of the study programs under the Department of Information Technology. This study program is an Applied Undergraduate Study Program (S.Tr.Kom.), with the focus of graduates being to produce a quality workforce

A. A. N. G. Sapteka et al. (eds.), *Proceedings of the International Conference on Sustainable Green Tourism Applied Science - Engineering Applied Science 2024 (ICoSTAS-EAS 2024)*, Advances in Engineering Research 249,
https://doi.org/10.2991/978-94-6463-587-4_42

in software engineering technology/software development so that programming becomes a mandatory skill that students must master.

Research conducted by (Yudhistira & Andika, 2023) states that one of the factors that determines the quality of education is the high and low levels of student success in the learning assessment process. The world of education in this digital era is required to have the ability to compete by utilizing its resources, including facility resources, infrastructure resources, and human resources. Having good resources will support all operational learning activities and all activities in the decision-making process.

Basic programming skills are essential for students, especially those who major in or study programs related to information technology. These skills include understanding programming concepts such as variables, data types, operators, flow control, and functions.

Evaluation of students' basic programming abilities can be done in various ways using the clustering method. The clustering method groups objects based on similar characteristics (Handayani, n.d.). In evaluating students' basic programming skills, the clustering method can be used to group students based on their level of programming ability. The k-means clustering method is a clustering method that does not require training data to group or classify data objects. This technique is included in unsupervised machine learning (Abdurrahman & Ismawan, 2022).

According to (Dewi et al., 2022), the advantage of applying K-Means is that it can group large objects and increase the speed of the grouping process. The goal of this algorithm is to divide data into several clusters. Research conducted by (Putra & Yuniarti, n.d.) used the K-means algorithm in conducting student evaluations. Using the means method can make the process more practical and effective than the manual method.

Research conducted (Syam, 2017) grouped students based on clusters of students with achievements, clusters of students with potential for achievements, clusters of students with potential problems, and clusters of students with difficulties. Next, testing was carried out using the RapidMiner application, the results of which were the same as the K-Means Algorithm analysis calculations.

Another study (Suandi, 2021) used the AHC method to group academic history before college and student graduation data to analyze the number of student graduates so that there is no imbalance between the ratio of the number of lecturers and active students in the study program.

Based on the presentation of research that previous researchers have carried out, the author proposes research on the application of the clustering method to evaluate students' basic programming abilities to provide a more comprehensive picture of the level of students' programming abilities, helping lecturers to carry out monitoring and providing more appropriate guidance to students, with their level of ability. And help lecturers develop more effective learning strategies using machine learning methods.

2 Methodology

This section will explain the stages in the research that will be carried out. The stages carried out in this research, as shown in Figure 1, can be explained as follows

2.1 Related Research

Firzada and Yunus' research uses the K-Medoids algorithm to predict students' timely study periods. The results produced two clusters: students who graduated on time and those who did not. This algorithm has proven effective in helping campuses evaluate and make policies to increase graduation rates (Firzada & Yunus, 2021).

Suandi's research aims to predict students' timely graduation and provide recommendations to those at risk of not graduating. The hierarchical clustering method is used to analyze academic data. In conclusion, data mining techniques are effective for predicting graduation and helping faculty guide at-risk students (Suandi & Zahrotun, 2021).

Research conducted by (Bahri & Midyanti, 2023) shows that dropout is a problem in universities, so universities need to minimize the number of students dropping out. Data mining methods, such as the K-Medoids algorithm, can be used to predict students who are at risk of dropping out. The results of this research can help universities provide early intervention to students at risk of dropping out.

Suraya et al.'s research uses the K-Means algorithm to group students' academic achievements based on GPA for semesters 1 to 6. The results produce two clusters: students with exemplary and less good achievements, helping with mapping and mentoring for students in need (Suraya et al., 2023).

Research related to the k-means method was carried out (Hartanti, 2022) with the background that evaluation materials were needed to determine the level of student understanding. The evaluation results are used to determine whether or not students are eligible to continue to the next semester. The method used in this research is the K-Means clustering algorithm with five criteria, including assignment scores, UTS, UAS, Final Project, and attendance, with 3 cluster results, namely the highest average score, sufficient average score, and average score. Lowest average.

2.2 Clustering

According to Han et al., clustering is grouping data into several groups based on similar attributes or distance calculations. Clustering is also known as unsupervised classification because it is more exploratory (Han et al., 2012). Cluster analysis divides data objects into several subsets using an algorithm so that objects in one cluster are similar and different from objects in other clusters. Clustering is useful for finding hidden groups in data (Sholeh et al., 2022).

2.3 K-Means Algorithm

According to Eko, the K-Means algorithm is a distance-based clustering technique that partitions data into several clusters based on characteristics. The center of the cluster, the centroid, is chosen randomly from the data (Eko, 2012). K-Means then groups the data to the nearest centroid based on distance. This process repeats until no data changes groups. This algorithm effectively separates data based on similar characteristics and centroid distance (Yudha et al., 2020).

Data collection. The data collection technique of giving tests to respondents is commonly used to obtain information about the respondent's knowledge, skills, or abilities. Tests can be taken in various formats, such as: a) Written test: This test consists of multiple-choice, true-false, or short essay questions. A written test is created to determine students' theoretical scores; b) Performance test: This test asks respondents to create a console program. It can measure students' skills and speed in developing programs.

Initial data processing. Initial data processing is an essential step in the data analysis process. The goal is to clean, transform, and summarize data so that it is ready for analysis. In this research, a data transformation process will be carried out using the Minmax normalization method. The normalization method is carried out by subtracting each data from the value of the feature with the minimum value of the feature and dividing the results by dividing the maximum value minus the minimum value of the feature (Azzahra Nasution et al., 2019).

Implementation of the K-Means method. At this stage, a system prototype will be created using the K-Means method in the clustering process. The tool used at this stage is Matlab.

Testing. The performance test uses the Dunn and Silhoutte method to test the performance of the k value used. Dunn looks for the highest value, meaning the cluster increasingly differs from other clusters (Monalisa & Kurnia, 2019). Silhouette looks for the highest value because it shows that the degree of confidence regarding the placement of data in the cluster is getting higher (Monalisa, 2018).

Data visualization. The clustering detection results are then displayed in graphical form to make it easier to read the data. Data analysis in the form of visualization can be in the form of pie charts, histograms, scatters, and others.

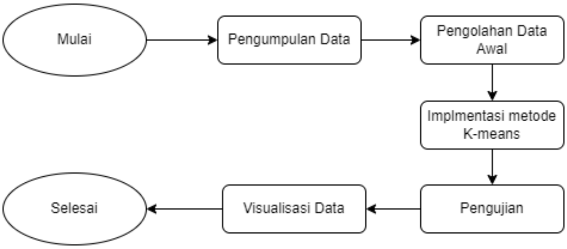


Figure 1. Research Stages (in Indonesia language)

3 Results and Discussion

3.1 Results

Data collection. Collecting data through tests on respondents is an essential method in this research to measure the basic programming abilities of Bachelor of Applied Software Engineering Technology students. The K-Means algorithm is then applied to group students based on their abilities, allowing for deeper analysis and identification of patterns.

Table 1. Student primary programming test results

NIM	Value of theory	Practice value	Theory time	Practice time
2315354001	Very good	Good	On-time	On-time
2315354004	Good	Very good	On-time	On-time
2315354005	Enough	Good	On-time	On-time
2315354009	Good	Good	On-time	On-time
2315354029	Not enough	Enough	Not on time	Not on time
2315354033	Enough	Enough	On-time	Not on time
2315354037	Good	Good	On-time	On-time
2315354040	Very good	Enough	On-time	Not on time
2315354049	Good	Enough	On-time	On-time
2315354052	Enough	Very less	On-time	On-time
2315354053	Enough	Very less	On-time	On-time
2315354057	Enough	Enough	Not on time	Not on time

In this research, two tests were carried out: a theory test and a practical test. To maintain the test results' objectivity, programming practitioners created questions and assessed tests. Table 1 shows the results of the test. The features used in cluster analysis using the K-Means method are obtained based on the test results. These features include Theory Test Scores, Practical Test Scores, and processing time

Initial Data Processing. The results are still ordinal data at this stage, so they must be processed further into nominal data. Table 2 shows a data transformation from ordinal to nominal data.

Table 2. Data set transformation

NIM	Value of theory	Practice value	Time average
2315354001	1	0.8	1
2315354004	0.8	1	1
2315354005	0.6	0.8	1
2315354009	0.8	0.8	1
2315354037	0.8	0.8	1
2315354040	1	0.6	0.75
2315354041	0.8	0.8	1
2315354045	0.2	0.8	1
2315354049	0.8	0.6	1
2315354052	0.6	0.2	1
2315354053	0.6	0.2	1
2315354057	0.6	0.6	0.5

K-Means Method testing. Table 3 shows the result of the K-means method test, which was carried out ten times. This needs to be done due to the different centroid initializations in K-means. The K-means method uses a randomly selected centroid starting point. Different initializations can produce different clusters, so running the algorithm several times helps find a more optimal solution. Repeated testing can also avoid Local Solutions: The K-means algorithm can get stuck in a local minimum, where the algorithm stops at a non-optimal solution; by doing repeated testing, there is a greater chance of finding a global minimum or a better solution. In this research, the fourth test result was an optimal solution with a Dunn value of 0.390360 and a Silhouette value of 0.56223. Table 3 displays more detailed results from testing the K-Means method.

Table 3. Cluster testing results

Test	Dunn	SI
1	0.227921	0.206147
2	0.390360	0.458920
3	0.262049	0.407095
4	0.390360	0.562239
5	0.323381	0.467258
6	0.312348	0.475902
7	0.298142	0.522771
8	0.298142	0.545293
9	0.323381	0.443940
10	0.323381	0.435928

Dunn: Dunn's index value for each test or cluster. Dunn's index is a metric used to assess the quality of clusterization by considering the distance between clusters and the distance within clusters. Higher values indicate better clusters, with more considerable distances between clusters and smaller distances within clusters.

SI: Silhouette Index value for each test or cluster. The Silhouette Index is a metric that measures how similar an object is to its cluster compared to other clusters.

Silhouette values range from -1 to 1, with higher values indicating that the object matches its cluster more closely and less closely to different clusters.

Data visualization. Data visualization is carried out at this stage using the k-means clustering method. Table 4 is an initial explanation of cluster visualization based on the number of members in each cluster.

Table 4. Number of members of each cluster

Clusters	Number of members	Percentage(%)
Cluster 1	6	7%
Cluster 2	41	49%
Cluster 3	16	19%
Cluster 4	15	18%
Cluster 5	5	6%

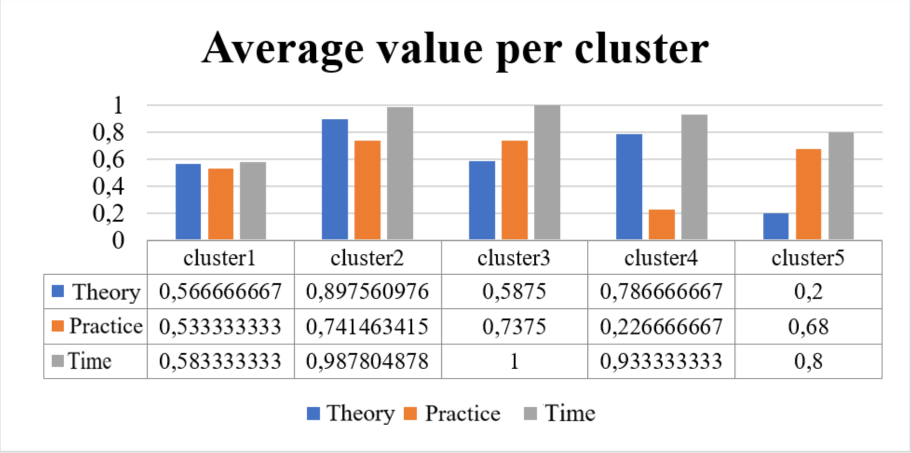
Cluster 1: 6 members. This cluster has a relatively small number of members compared to other clusters. This suggests that the data in this cluster may have more unique or rare characteristics.

Cluster 2: 41 members. This is the largest cluster in the data, with 41 members. Compared to other clusters, it may include the most common data or have the most features in common.

Cluster 3: 16 members. This cluster has a moderate number of members. This may reflect a data group with fairly common characteristics but fewer than those in cluster 2.

Cluster 4: 15 members. Like cluster 3, cluster 4 also has a moderate number of members. This indicates the existence of two data groups that are almost the same size, perhaps with similar features that differ from cluster 3.

Cluster 5: 5 members. This cluster is the smallest, having only five members. This may reflect precise data or outliers not found much in the dataset.



	cluster1	cluster2	cluster3	cluster4	cluster5
Theory	0,566666667	0,897560976	0,5875	0,786666667	0,2
Practice	0,533333333	0,741463415	0,7375	0,226666667	0,68
Time	0,583333333	0,987804878	1	0,933333333	0,8

Figure 2. Cluster data visualization graph (in Indonesia language)

Figure 2 is a descriptive Trend Analysis from the table provided, which can be done by comparing the average values for each category (Theory, Practice, and Time) in each cluster. The following is a comparative analysis: a) Cluster 1 has a relatively balanced average score between theory and practice, with a slightly higher score on the time aspect; b) Cluster 2 has high average scores in all categories, especially time, indicating excellent performance overall; c) Cluster 3 shows better practical performance than theory and has perfect timing. d) Cluster 4 significantly differs between theoretical and practical values, with low practical values but high time values; e) Cluster 5 had lower theory scores but higher practice and time scores, indicating a greater focus on practice.

The general conclusion from the analysis results is that Cluster 2 and 3 stand out with good performance in all categories, especially in time. Cluster 4 has a significant imbalance between theory and practice. Cluster 5 has low theoretical value but is good in practice and time. Cluster 1 has a balanced value but is relatively lower than the other clusters, except for Cluster.

4 Conclusion

Based on the results of the research that has been carried out, it can be concluded that: The K-Means Method Can speed up the process of grouping student data for evaluating each student's basic programming abilities. The application of the K-Means method in grouping data is influenced by several factors, namely determining the initial centroid, determining the number of clusters to be formed, and labeling the results obtained.

References

Abdurrahman, A., & Ismawan, F. (2022). Model Machine Learning klasifikasi data sekolah TK berdasarkan status dan Kabupaten/Kota administrasi Provinsi DKI Jakarta. *15*(2), 1979–276. https://doi.org/10.30998/faktorexacta.v15i2.13211.

Nasution, D. A., Khotimah, H. H., & Chamidah, N. (2019). Perbandingan normalisasi data untuk klasifikasi wine menggunakan Algoritma K-NN. *Journal of Computer Engineering System and Science*, *4*(1), 2502–7131.

Bahri, S., & Midyanti, D. M. (2023). Penerapan metode K-Medoids untuk pengelompokan mahasiswa berpotensi Drop Out. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *10*(1). https://doi.org/10.25126/jtiik.20231016643.

Eko, P. (2012). *Data Mining Konsep dan Aplikasi Menggunakan Matlab*. Andi Yogyakarta.

Fadrial, Y. E. (2020). Klasterisasi hasil evaluasi akademik menggunakan Metode K-Means: Studi kasus Fakultas Ilmu Komputer Unilak. *Prosiding-Seminar Nasional Teknologi Informasi & Ilmu Komputer (SEMASTER)*, *1*(1), 53–65.

Firzada, F., & Yunus, Y. (2021). Klasterisasi tingkat masa studi tepat waktu mahasiswa menggunakan Algoritma K-Medoids. *Jurnal Sistim Informasi Dan Teknologi*. https://doi.org/10.37034/jsisfotek.v3i3.146.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and techniques. *San Francisco, CA, itd: Morgan Kaufmann*. https://doi.org/10.1016/B978-0-12-381479-1.00001-0.

Handayani, F. (n.d.). Aplikasi Data Mining menggunakan Algoritma K-Means Clustering untuk mengelompokkan mahasiswa berdasarkan gaya belajar. *Jurnal Teknologi Dan Informasi*. https://doi.org/10.34010/jati.v12i1.

Suandi, B. H. T., & Zahrotun, L. (2021). Penerapan Data Mining dalam mengelompokkan data riwayat akademik sebelum kuliah dan data kelulusan mahasiswa menggunakan Metode Agglomerative Hierarchical Clustering. *JTIKA*, *3*(1), 62–71. http://jtika.if.unram.ac.id/index.php/JTIKA.

Hartanti, N. T. (2022). Mengukur tingkat pemahaman mahasiswa pada mata kuliah Pemrograman dengan Algoritma K-Means Clustering. *Jurnal Sisfotenika*, *12*(1).

Putra, B. J. M., & Yuniarti, D. A. F. (n.d.). Analisis hasil belajar mahasiswa dengan clustering menggunakan Metode K-Means. *Jurnal POROS TEKNIK*, *12*(2), 49–58.

Monalisa, S. (2018). Klusterisasi customer lifetime value dengan Model LRFM menggunakan Algoritma K-Means. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(2). https://doi.org/10.25126/jtiik.201852690.

Monalisa, S., & Kurnia, F. (2019). Analysis of DBSCAN and K-Means Algorithm for evaluating outlier on RFM model of customer behaviour. *Telkomnika (Telecommunication Computing Electronics and Control)*, *17*(1). https://doi.org/10.12928/TELKOMNIKA.v17i1.9394.

Dewi, F. P., Aryni, P. S., & Umaidah, Y. (2022). Implementasi Algoritma K-Means Clustering seleksi siswa berprestasi berdasarkan keaktifan dalam proses pembelajaran. *Jurnal Informatika Sunan Kalijaga)*, *7*(2), 111–121.

Sholeh, M., Andayati, D., & Rachmawati, R. Y. (2022). Data Mining Model klasifikasi menggunakan Algoritma K-Nearest Neighbor dengan normalisasi untuk prediksi penyakit diabetes. *TeIKa*, *12*(02). https://doi.org/10.36342/teika.v12i02.2911.

Suandi, B. H. T. (2021). Penerapan penerapan data mining dalam mengelompokkan data riwayat akademik sebelum kuliah dan data kelulusan mahasiswa menggunakan Metode Agglomerative Hierarchical Clustering (AHC). *Jurnal Teknologi Informasi, Komputer, Dan Aplikasinya (JTIKA)*, *3*(1). https://doi.org/10.29303/jtika.v3i1.130.

Suraya, S., Sholeh, M., & Andayati, D. (2023). Penerapan Metode Clustering dengan Algoritma K-Means pada pengelompokan indeks prestasi akademik mahasiswa. *SKANIKA*, *6*(1). https://doi.org/10.36080/skanika.v6i1.2982.

Syam, F. A. (2017). Implementasi metode klastering K-Means untuk mengelompokan hasil evaluasi mahasiswa. *Jurnal Ilmu Komputer Dan Bisnis*, *8*(1), 1857–1864.

Yudha, N., Ernawati, E., & Putri, P. E. (2020). Pemanfaatan citra penginderaan jauh untuk pemetaan klasifikasi tutupan lahan menggunakan metode Unsupervised K-Means berbasis Web GIS: Studi kasus Sub-DAS Bengkulu Hilir). *Jurnal Rekursif*, *8*(1), 100–110.

Yudhistira, A., & Andika, R. (2023). Pengelompokan data nilai siswa menggunakan metode K-Means Clustering. *Journal of Artificial Intelligence and Technology Information (JAITI)*, *1*(1), 20–28. https://doi.org/10.58602/jaiti.v1i1.22.