Jurnal Teknologi Informasi dan Pendidikan
Volume 16, No. 2, September 2023
https://doi.org/10.24036/jtip.v16i2.804

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

Clustering Analysis of Internal and External Factors Affecting Post-Pandemic Study Duration in XYZ Educational Institution Using the Orange Application

Imelda Muluk[1*]✉, Yohana Dewi Lulu Widyasari[2], Riska Amelia[3]
[1]Magister Terapan Teknik Komputer, Politeknik Caltex Riau, Riau, Indonesia
[2]Magister Terapan Teknik Komputer, Politeknik Caltex Riau, Riau, Indonesia
[3]Teknologi Rekayasa Komputer Jaringan, Universitas Bung Hatta, Padang, Indonesia
Corresponding Author: imelda21mttk@mhasiswa.pcr.ac.id

Article Information

Article history:

No. 804
Rec. October 18, 2023
Rev. November 23, 2023
Acc. November 25, 2023
Pub. December 19, 2023
Page. 191 – 202

Keywords:

- *Long-term study*
- *Silhouette Score*
- *clustering*

ABSTRACT

The COVID-19 epidemic has had a notable effect on Indonesia's education system as well as other economic sectors. As a result, significant changes have occurred in the educational landscape, such as the transition from traditional classroom instruction to online learning. This shift has helped all students in different ways, whether they come from academic or non-academic backgrounds. The results are displayed in the Education Quality Report for 2023. Character traits, numeracy, and reading comprehension have all declined this year. A character has declined by 4.54%, numeracy by 0.61%, and literacy by 32.43%. Thus, the proportion of these decreases is what has piqued our curiosity about the most important variables influencing long-term learning outcomes and how academic and non-academic elements affect students' academic performance. The K-Means algorithm, one of the Unsupervised Learning methods used in this study, applies machine learning techniques to investigate several reasons of learning results drop based on multiple clusters. Participants are given questionnaires to complete to gather data. 115 individuals will be chosen at random from a pool of 1400 at SMK Negeri 6 Pekanbaru in November 2023 for the study. To gather quantitative data, a Google Form with 19 questions will be used. Based on how much time students spend studying, the research generates a Silhouette Score that can vary from 0.5 to about 0.6, which affects the evaluation.

How to Cite:

Imelda, Widyasari, Y. D. L, & Amelia, R. (2023). Clustering Analysis of Internal and External Factors Affecting Post-Pandemic Study Duration in XYZ Educational Institution Using the Orange Application, 16(2), 191-202. https://doi.org/10.24036/jtip.v16i2.804

Jurnal Teknologi Informasi dan Pendidikan
Volume 16, No. 2, September 2023
https://doi.org/10.24036/jtip.v16i2.804

1. INTRODUCTION

The spread of the illness caused by the Coronavirus has caused an emergency known as the Covid-19 pandemic. This illness can kill people, though it shares the flu's typical early symptoms. The COVID-19 pandemic has been present in Indonesia for more than 2 (two) years, and it has had an effect on a number of areas including the economy, health, and education. [1] The definition of education is a conscious effort to realize cultural inheritance from one generation to another to another. [2] Education "is a conscious and planned effort to create a learning atmosphere and learning process so that students Individuals should develop their potential for spiritual strength, self-control, intelligence, and noble character, as well as the skills needed for themselves, society, and the nation. (Law No. 20 of 2003) [3]. Each learner has their character and potential, which are influenced by both internal and external factors, especially within the family environment, which plays a significant role in shaping internal factors. In the family environment, the role of parents, both mothers and fathers, and other family members, such as siblings, greatly influence attitudes, character, and learning patterns. According to Gunarsa (2009:6), [4]

A child's behavior and learning patterns are significantly impacted by environmental factors, such as their family life. This includes aspects like parental love and care, as well as the educational background of parents. Ihsan (2005:19) notes that family factors like parental attention and affection, role modeling, and harmony can affect a child's development. Gerungan (2002:185) [5] also highlights the influence of family environment on a child's development, including socioeconomic status, family integrity, parental attitudes and habits, and the child's status.

External factors, apart from the family, also include the learner's surrounding environment, such as playmates during leisure time before returning home or even on holidays. The reason why a person in a position like this doesn't give much thought to their age is because the person in that position is already learning how to update their identity. Due to this, older people with backgrounds who have should tend to enhance their communication skills and carry out coaching about the needs of children, both within and outside of the classroom. [6]Based on these internal and external factors, the author has compiled this journal with the hope that it will serve as a study for learners, parents, and especially educators in determining the appropriate methods to enhance learner motivation and achievement.

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

The author has put together this diary based on these internal and external aspects in the hopes that it would help students, parents, and educators in particular figure out how best to improve student motivation and accomplishment. Consequently, it will be possible to prevent the subsequent outcomes reported in the 2023 Education Report, which show a decline in literacy of 32.435%, numeracy of 0.61%, and character of 4.54% in the next years. The K-Means algorithm and machine learning techniques were used in this study. K-Means has the benefit of labeling each data point with the appropriate cluster, which makes it easy to recognize which groupings are already in place. Data was gathered from students in the eleventh and twelfth grades who completed questionnaires regarding internal and external.

2. RESEARCH METHOD

To analyze data, the author conducted a survey using Google Forms to gather quantitative data from students at SMK Negeri 6 Pekanbaru. Learning in SMK (Vocational Schools) seeks to develop graduates who are prepared to enter the business and industrial world immediately and who achieve the standards for the intermediate level. [7] The survey consisted of 19 questions adapted from a Kaggle dataset and was distributed to approximately 1400 students. Out of these, 115 students provided complete responses. The collected data included 19 features and 5 metadata, with no missing values.

The author's objective was to analyze the correlation between various internal and external factors that affect the duration of the study. Factors such as age, gender, residential area, health, attendance, dating status, number of siblings, family closeness, leisure time, going out with friends, parental education background, parents' occupations, reasons for choosing SMK Negeri 6 Pekanbaru, and support facilities including transportation to school and commuting time were used as attributes the use of machine learning is to allow computers or computer systems to learn from data and previous experiences without being explicitly programmed. Artificial intelligence's branch of machine learning is frequently employed to find solutions to a wide range of issues. [8] Machine learning is used to make predictions based on historical data. Machine learning can be used to classify data or groups based on the features found in the data. According to Poppy Meilina, 2015 [9] The following measures were taken throughout this research:

2.1. Data Verification (Process Of Cleaning Data)

Data cleansing is a crucial step in preparing data for correlation analysis. In this study, correlations between different data variables are looked for. Data cleansing is a technique used to turn low-quality data into high-quality data that may then be used in data mining. [10] Data cleansing is required to ensure that the data being processed is relevant to the needs because not all tables are utilized. Data is cleansed before processing because this is due to a variety of factors, some of which are listed below, including missing values

in attributes or other particular attributes that make something incomplete, having errors, or outlier numbers that differ from the rest of the data makes the data noisy. Inconsistent: Using names or codes inconsistently. Data warehouses require consistent data quality integration since good data quality is dependent on smart decisions. [11]

2.2. Process of Data Integration

Integrating data is combining information from different databases into a single new database. Finding related items in two or more databases and combining them into a single, non-redundant view is a multi-step process known as database integration. The requirement for users and applications to combine data from growing information resources has made database integration an active topic of research. [12] In this case, information is changed or converted into a numeric representation. Data transformation is the process of changing data from one format to another, often from that of a source system into that needed by a destination system. The bulk of data integration and management operations, including data manipulation and data warehousing, include data transformation as a step. [13] Similar to the stage before, data integration in this stage makes use of Microsoft Excel and Orange, two products that work well together.

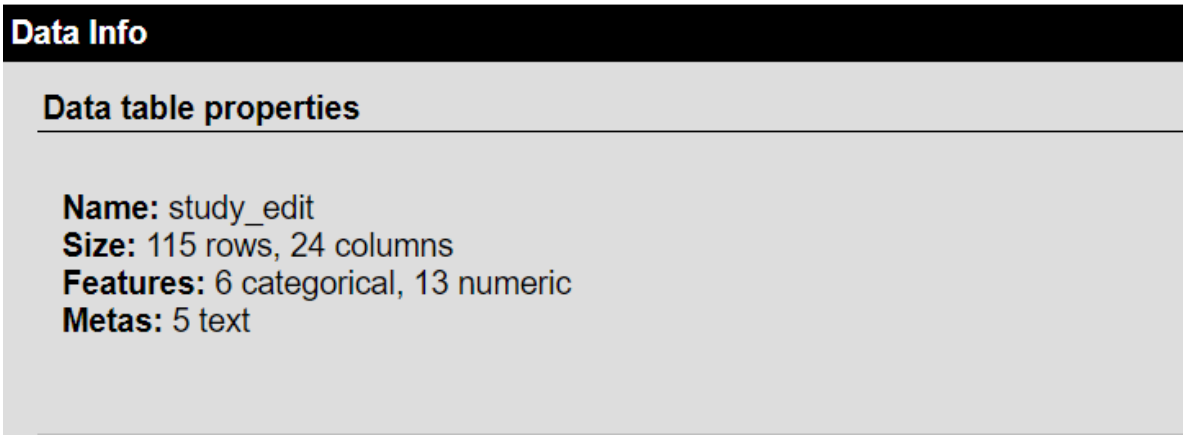
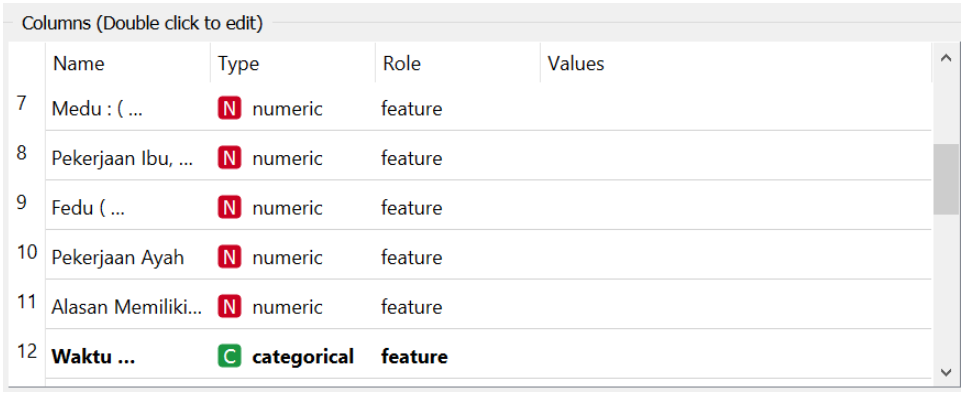


Fig 1. Dataset

2.3. Determine the Goal

You should characterize the data items and run several subtask operations on them before utilizing them for any operation to identify any potential issues that could arise from their use (for example, multicollinearity brought on by using highly correlated variables in the subsequent analysis). [14] In the data mining stage, where the target data serves as the input for the following operations, this stage describes the data that is prepared for use.

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id


Columns (Double click to edit)

	Name	Type	Role	Values
7	Medu : (...	N numeric	feature	
8	Pekerjaan Ibu, ...	N numeric	feature	
9	Fedu (...	N numeric	feature	
10	Pekerjaan Ayah	N numeric	feature	
11	Alasan Memiliki...	N numeric	feature	
12	**Waktu ...**	C **categorical**	**feature**	

Fig 2. Goal

2.4. Preprocessing

The author utilized the Orange application and several widgets to manipulate, predict, and visualize data. Since the dataset had no missing values or null values, data processing was straightforward. After processing the data, a linear regression analysis was conducted, and an intercept of -1.49659 was obtained. The analysis revealed that the student's age, father's occupation, and commuting time had the most significant impact and highest correlation with the duration of study for one week.

Often, preparation is necessary before data mining. Preparing raw data for further processes includes applying a variety of techniques to it. With the following signs, preprocessing in data mining aims to convert data into a format that is simpler and more useful for user needs: Getting more precise results, lowering the computing time for complex tasks, reducing the size of data values without changing the information they hold, Preprocessing can be done using a variety of techniques, including choosing a representative subset from a vast data pool by sampling, discretization, though a step in data reduction, has its importance, particularly for numerical data, remove values that are missing, persistence. Impute is the choice of features. Since there are no missing values, the data is converted into a numeric format for simpler and quicker data processing. Preprocessing Text is the earliest stage of turning text into data so that it can be processed further. [15]

Jurnal Teknologi Informasi dan Pendidikan
Volume 16, No. 2, September 2023
https://doi.org/10.24036/jtip.v16i2.804

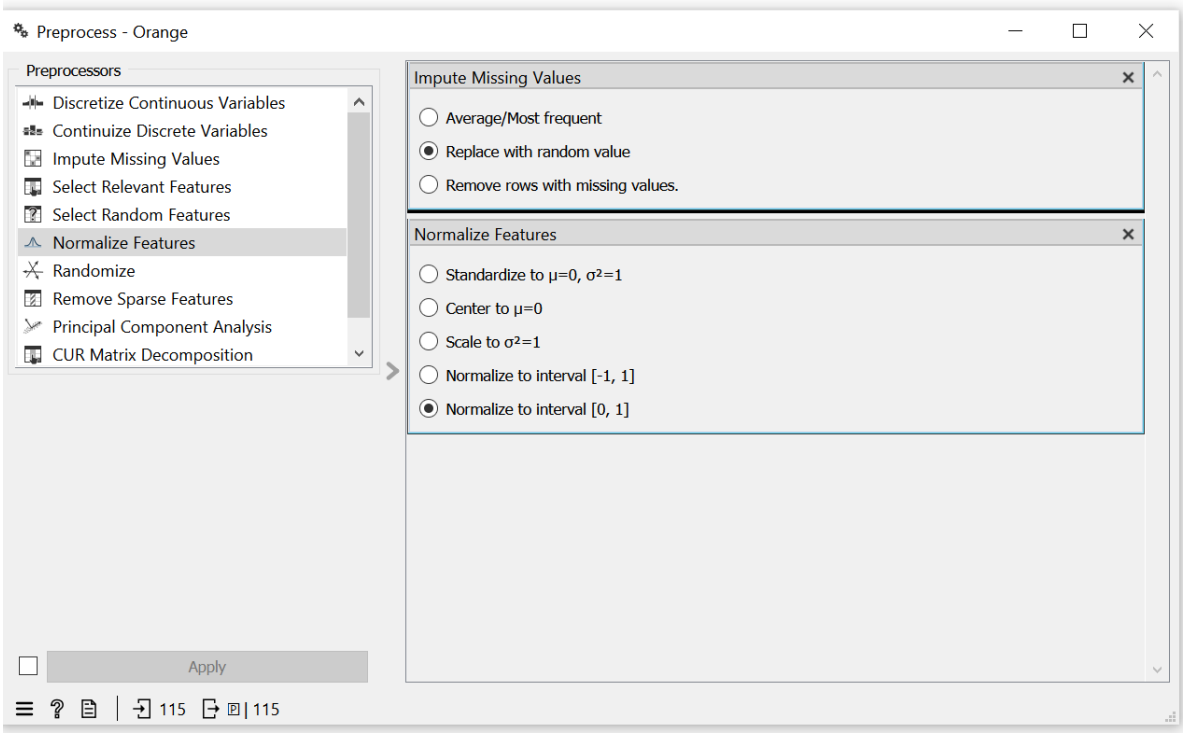


Fig 3. Preprocessing

2.5. Visualization

The following stage is data visualization, which entails visually viewing data using graphical representations like Histograms, distribution diagrams, scatter plots, or location descriptions like mean, median, mode, quartiles, and percentiles. For the preliminary study, data mining software methods like visualization (which plots data and establishes associations) and cluster analysis (which identifies which variables work well together) are helpful. Tools Initial association rules can be created using techniques like generalized rule induction. As data knowledge increases (typically through pattern recognition), models with greater detail, and identification are triggered by observing model output. [16] A scatter diagram representing the relationship between two numerical variables can be created using the Scatter Plot widget.

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

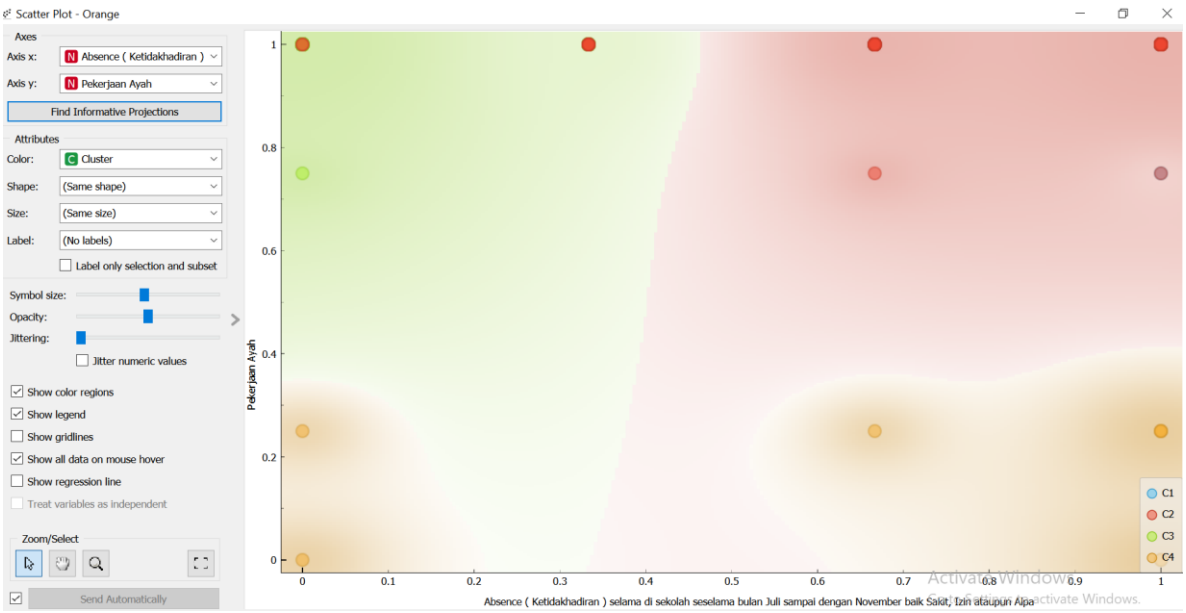


Fig 4. Scatter Plots

2.6. Process of Data Mining

This stage involves creating decision trees, regression models, and classification models. For machine learning systems to be transparent and trustworthy, this is essential. We can do categorization, prediction, estimate, and other helpful operations on a big data collection using data mining, which can be used if it is appropriate for the data type, dividing up info into modeling also requires training and test sets. [17]Machine learning is the practice of applying mathematical and computer algorithms through a learning process drawn from data and producing future predictions. [18]. Decision trees, regression models, and classification models are created at this phase. It is imperative to guarantee openness and reliability in machine learning systems. If appropriate for the type of data, we can use data mining to do classification, prediction, estimation, and other helpful operations on a large dataset. Training and test sets are also necessary for the segmentation of data into modeling. However, we use unsupervised learning techniques with the K-Means technique because the study time data does not have a specified aim. Machine learning is the process of using computer and mathematical algorithms to learn from data and produce predictions for the future. K-Means is a machine learning technique that clusters data into several subgroups.

The following are the steps in K-means: Ascertain the quantity of clusters. Sort the data into clusters at random. Determine the data's mean for each cluster. Step 3 should be repeated based on the designated threshold value. Determine the distance using K-means

clustering between the data points and the centroid. The distance between centroids and data points can be computed using distance space. The Manhattan/City Block distance is one example of a frequently used distance computation forecast. K-Means is a machine-learning technique that clusters data into several subgroups.

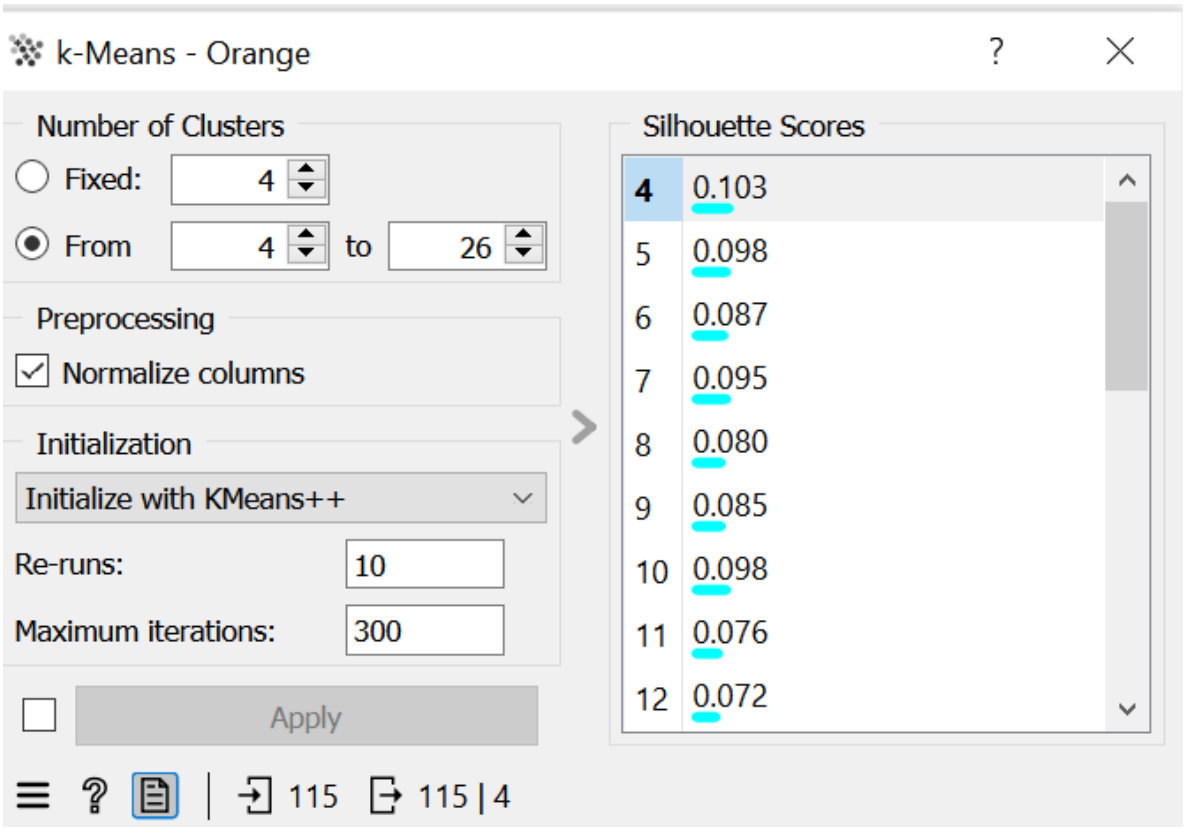


Fig 5. K - Means

There following are the steps involved in k-means machine learning: Fixed: Ascertain the required number of clusters. Three clusters are to be utilized in this instance. From x to y: Using the silhouette score—which is calculated by comparing the average distance between an element in one cluster and an element in another—calculate how many clusters there are. Make columns normal: The columns will be normalized (mean centered at 0 and standard deviation scaled at 1) if this option is selected. Initialization: This refers to how the algorithm starts the clustering process. There are two choices. The first is K - Means which uses a random process to calculate the initial values of the centroid and then selects points from the remaining set with probabilities based on the square of the distance from the closest center. In the second method, known as random initialization, clusters are first selected at random and subsequently updated in successive repetitions. Reruns: Count the number of times the algorithm is executed starting at random points in time. The

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

outcome with the lowest sum of squares across all clusters will be employed. Maximum iterations: You can manually change the maximum number of iterations that an algorithm can execute. Apply Automatically: If this option is selected in the Orange Data Mining software's Preprocess widget, then any modifications made to the preprocessing procedures will be instantly applied to the data the widget processes.

3. RESULTS AND DISCUSSION

According to the correlation table, there is a connection between study time and variables like age, the father's profession, and having a companion after returning from school. [19] Compared to other factors, the age factor has the greatest impact on study time."

Nama	Kelas	Jurusan	Cluster	Silhouette	aki - Laki, 1 = Pe	Age (Usia)	di Pinggiran = 0	uarga, >3=0, ku	alnya) : T (Ting	atan) 1:
M. Luthfiandra	11	Rpl 1	C1	0.579023	0	0.0	0	0	1	
R.Daffa	12	Rekayasa Peran...	C1	0.564452	0	0.250	0	1	1	
Arya ariri	XI	Rekayasa peran...	C1	0.557213	0	0.0	0	0	1	
Raihan Maulan...	12 RPL 1	Rekayasa Peran...	C1	0.54747	0	0.250	0	1	1	
Chelsea	XI RPL 1	Rekayasa Peran...	C1	0.541269	1	0.0	0	0	1	
Nabilah Bilquis ...	12	Rekayasa peran...	C1	0.538774	1	0.250	0	0	1	
Raihan Andriya...	XI	RPLG 1	C1	0.535763	0	0.0	0	0	1	
Habib Fadhlan ...	XII	RPL	C1	0.535758	0	0.250	0	0	1	
Jonathan natha...	11(RPL 2)	Rekayasa peran...	C1	0.530629	0	0.0	0	0	1	
Rayhan Qolbi R...	XI	RPLG 1	C1	0.523859	0	0.0	0	0	1	
Adelia Putri	XII	Rekayasa Peran...	C1	0.516473	1	0.250	1	0	1	
Fadil naik	X1	RPLG	C1	0.509607	0	0.0	1	1	1	
Arya Yudha Pra...	XI	REKAYASA PER...	C1	0.509606	0	0.250	0	0	1	
Sonia Husni	XI RPL 1	REKAYASA PER...	C1	0.490709	1	0.0	0	0	1	
Prisca Tri Aulia	12 RPL 1	REKAYASA PER...	C1	0.480642	1	0.250	0	0	0	
Idris khoidir	11	Rekayasa Peran...	C2	0.559042	0	0.5	0	0	1	
SANDIISHAQ	11	RPL2	C2	0.554477	0	0.5	0	1	1	
Eduard Pranata S	XII RPL 1	Rekayasa Peran...	C2	0.551622	0	0.5	1	0	1	
M. Fahri Alfarabi	XI	RPL 2	C2	0.550697	0	0.0	0	0	1	
Mifta khudin p...	XII RPL?	Rekayasa peran...	C2	0.549972	1	0.250	0	0	1	
Taufik Ilham	XI	RPL?	C2	0.546892	0	0.250	0	0	1	
Faadhil Rahman	XI RPL 2	Rekayasa peran...	C2	0.544664	0	0.0	1	1	1	
Rifky fadillah	XI	RPL?	C2	0.544103	0	0.250	1	0	1	
Ridho abby ram...	11	Rpl 2	C2	0.542807	0	0.250	1	1	0	
Chelsi hendryanti	12 RPL 2	Rekayasa peran...	C2	0.542226	1	0.250	0	0	0	
Dafa Arraafi Put...	XI	Rekayasa Peran...	C2	0.542158	0	0.250	0	0	1	
M.RISKY BATUB...	12 RPL 2	Rekayasa Peran...	C2	0.541827	0	0.5	0	0	1	
M.devan nuzul ...	XI	Rpl2	C2	0.540737	0	0.5	0	0	1	

(a)

Nama	Kelas	Jurusan	Cluster	Silhouette
INTAN YULIA N...	XII RPL 2	Rekayasa peran...	C3	0.537202
ALIYAH TIRAH...	12	REKAYASA PER...	C3	0.535897
Hana Leonika Si...	11	Rekayasa peran...	C3	0.535218
M. Rizky firman...	XII Rpl 1	Rpl	C3	0.534264
M ZIKRI NOFRI...	XII (12)	Rekayasa peran...	C3	0.533741
Muhammad Fe...	XII	RPL 1	C3	0.5333
zeniati	12	rpl	C3	0.531308
Loren Valery	XI	RPL?	C3	0.531287
RICKY SANDIKO	XIRPL?	Rekayasa peran...	C3	0.530803
Dameria Siman...	12	Rekayasa Peran...	C3	0.526785
Dhea amanda	XI	RPLG2	C3	0.526084
Dino Adriano B...	11	Rekayasa Peran...	C3	0.525305
Gandawa Ripo ...	XI rpl 1	Rekayasa Peran...	C3	0.524968
Gusmiati Gea	XII	Rekayasa peran...	C3	0.520769
Eli Sadrat Sitorus	12	Rekayasa peran...	C3	0.519272
Nashua Revana	12	Rekayasa Peran...	C3	0.519177
Putri Melati Br ...	XI RPL1	Rekayasa peran...	C3	0.51857
Nabil Saputra	12 rpl 1	Rekayasa peran...	C3	0.51766
Dea Enjel	XI	RPL 1	C3	0.514115
Muhammad Ra...	XI	Rpl2	C3	0.513083
Refaldi Julidinsy...	XII	Rekayasa Peran...	C3	0.511441
Rafandi Nova Fi...	12 RPL 1	Rekayasa Peran...	C3	0.511202
Yasnimar Wati L...	IX	Rekayasa peran...	C3	0.508607
Sadan Walliyan...	12 rpl 1	RPL	C3	0.507708
Muhammad Ra...	X1	XRPLG 1	C4	0.609591
Muhammad Ra...	X1	XRPLG 1	C4	0.598519
Sunnatul 'Aini ...	XI	RPLG 2	C4	0.579249

(b)

Fig 6. (a)(b) Result Of K - Means

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

In clustering, positive values are possible; if the silhouette score is close to 1, it means that the objects in the cluster are neatly arranged and kept apart from neighboring clusters. This implies that the objects have been successfully separated by the clustering.

4. CONCLUSION

The main distinction between clustering and classification is that the former lacks a target variable. The goal of clustering is not to categorize, estimate, or forecast a target variable's values. Rather, the goal of clustering algorithms is to separate all the data into many groups that show proximity or similarity (homogeneity), where records in one group have a higher degree of similarity with each other than with records in other groups. "A silhouette score ranging from 0.5 to 0.6 with 4 clusters is produced when using the k-means algorithm for clustering. We can therefore conclude that family closeness and the educational attainment of the mother and father have a major impact on school absenteeism and outside-the-home leisure time, both of which have an impact on academic achievement."

REFERENCES

[1] A. F. Yogananti, "College Student's Perception toward " Peduli Lindungi" Application through the Usability Scale Method," *urnal Teknologi Informasi dan Pendidikan*, vol. 15, no. 2, pp. 73-83, 2023.

[2] A. Rahman, "Pengertian Pendidikan, Ilmu Pendidikan dan Unsur-Unsur Pendidikan," *Al Urwatul Wutsqa: Kajian Pendidikan Islam*, vol. 2, no. 1, pp. 1-8, 2022.

[3] D. Pristiwanti, "Pengertian Pendidikan," *Jurnal Pendidikan Dan Konseling (JPDK)*, vol. 4, no. 6, 2022.

[4] W. A. S. Halasan Simanullang1), "Peran Lingkungan Keluarga Dalam Meningkatkan Prestasi Belajar Siswa".

[5] N. O. Afriyani, "Peranan Keluarga Terhadap Prestasi Siswa Pada Mata Pelajaran Ilmu Pengetahuan Sosial".

[6] S. M. Johnson, The next generation of teachers: Who enters, who stays, and why, 2008.

[7] B. Azra, "the Development of Engineering Drawing E-Module for Grade X At Smk (Vocational School)," *Jurnal Teknologi Informasi dan Pendidikan*, no. 460, 2021.

[8] A. Roihan, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper," *IJCIT (Indonesian Journal on Computer and Information Technology)*, vol. 5, no. 1, pp. 75-82, 2020.

[9] P. Meilina, "Penerapan Data Mining dengan Metode Klasifikasi Menggunakan Decision Tree dan Regresi".

[10] A. G. Lazuardy, "Proceeding SINTAK 2019," pp. 1-6, 2019.

[11] A. Hemmati-Sarapardeh, "Chapter 1 - Introduction," 2020, pp. 1-22.

[12] I. Almahy, "Database integration: Importance and approaches," *Journal of Theoretical and Applied Information Technology*, vol. 54, pp. 150-154, 2013.

[13] S. Pandey, "Data Integration and Transformation using Artificial Intelligence," in *2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, 2023.

[14] R. Nisbet, "Chapter 18 - A Data Preparation Cookbook," 2018, pp. 727-740.

[15] T. J. Melmambessy, "Analysis of the Opinion Students about The Online Learning System During the Pandemic Using The K-NN and Naïve Bayes Methods," *Jurnal Teknologi Informasi dan Pendidikan*, vol. 16, no. 1, pp. 75-85, 2023.

[16] D. L. Olson, "Data Mining Process BT - Advanced Data Mining Techniques," 2008, pp. 9-35.

[17] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5 Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD) . Jurnal Edik Informatika," *Jurnal Edik Informatika*, vol. 2, no. 213-219, p. 2, 2019.

[18] A. Roihan, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper," *JCIT (Indonesian Journal on Computer and Information Technology)*, vol. 5, no. 1, pp. 75-82, 2020.

[19] F. L. V. Voorhis, "The impact of family involvement on the education of children aged 3 to 8," *Mdrc*, no. 10, p. 229, 2013.

[20] F. S. Tehrani, "Teknik Clustering Dengan Algoritma K-Medoids Untuk Menangani Strategi Promosi Di Politeknik Tedc Bandung," *Jurnal Teknologi Informasi dan Pendidikan*, vol. 12, no. 2, pp. 1-7, 2019.

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id