JPPIPA 10(4) (2024)

Jurnal Penelitian Pendidikan IPA

Journal of Research in Science Education

http://jppipa.unram.ac.id/index.php/jppipa/index

Prediction of Graduation Accuracy Using the K-Means Clustering Algorithm and Classification Decision Tree

Sri Rahmawati[1*], Sarjon Defit[1]

[1]Information Systems, Computer Science, Universitas Putra Indonesia YPTK Padang, Padang, Indonesia

Received: January 25, 2024
Revised: March 17, 2024
Accepted: April 25, 2024
Published: April 30, 2024

Corresponding Author:
Sri Rahmawati
sri_rahmawati@upiyptk.ac.id

DOI: 10.29303/jppipa.v10i4.7073

Abstract: Becoming a scholar at the right time for students is a very meaningful award for them if it is supported by seriousness and perseverance in their studies. Here, sample data was taken from 131 randomly taken in testing. Where there are still students who are not detected by the study program in completing their lectures, so research is carried out on clustering and classification with decision trees in determining the level of accuracy of lectures by clustering data, determining the initial centroid value and the centroid point. The results found were that there were 78 people grouped in cluster 0 and 53 people grouped in cluster 1, where those with potential for punctuality for their studies were in cluster 0 so they were students who could finish within the specified time. Meanwhile, students grouped in cluster 1 illustrate that these students need coaching and guidance both in the study program and with their supervisors. In the classification taken from the results of data clustering, two classes were obtained, namely class a and class b, with 73 and 58 data respectively, so that the results between clustering and classification did not differ too much in the data to predict the accuracy of a student's graduation.

Keywords: Centroid; Clustering; Decision tree K-Means; Random

Introduction

Punctuality for a student is success in completing education at the tertiary level in obtaining a bachelor's degree and also their accuracy in completing lectures within the specified time, but in reality, some students cannot carry out according to the specified time, namely during four years in education (Lo, 2023; Wirawan et al., 2019). Classifying, clustering, and association are models that exist in data mining (Chaudhry et al., 2023; Dol & Jawandhiya, 2023).

The technique for clustering data is a method of analyzing data whose aim is to group data according to the same data characteristics in the same place and different data characteristics in other places (Den Teuling et al., 2023; Ikotun et al., 2023). Text data can also be used as an indicator for data grouping (Al-Anazi et al., 2016). Analysis in grouping data can use the K-Means method into several data groups, and there is also a deficiency in determining the starting point of a clustering which is used randomly, which can cause the membership assessment to always change if repeated searches are carried out (Aldo, 2023; Asroni et al., 2020).

So, researchers are interested in doing research using the K Means method with elaboration using the Weka application (Loeng, 2020; Priyatna et al., 2018). This research is used to cluster students' level of accuracy and carry out planning in techniques that are very suitable for making decisions (Silva et al., 2021); (Rastrollo-Guerrero et al., 2020). Several other researchers assessed data on student grades and attendance (Asril, 2020; Márquez et al., 2023). Grouping data using the clustering method will produce unique characteristics (Mulyaningsih & Heikal, 2022; Oyewole & Thopil, 2023).

How to Cite:
Rahmawati, S., & Defit, S. (2024). Prediction of Graduation Accuracy Using the K-Means Clustering Algorithm and Classification Decision Tree. *Jurnal Penelitian Pendidikan IPA*, *10*(4), 2007–2013. https://doi.org/10.29303/jppipa.v10i4.7073

Method

The use of grouping methods in data mining is used to see the level of seriousness of students in their learning (Bharara et al., 2018; Ha et al., 2024). This research took random data from the semester achievement index scores of students in the Information Systems study program with 131 students as the research sample. The unsupervised technique in K-Means is part of the data grouping method which can divide the data into two or more groups (Abbas et al., 2023). Stages in solving existing problems in determining timeliness in student education; Determine the number of cluster centers "Centroid" that will be used randomly; Calculating the distance from the cluster center to calculate the distance between the data and the cluster center, Euclidian distance is used; Groups objects to determine cluster members based on distance.

$$Qe = \sqrt{(x_i - s_i)^2 + (y_i - t_i)} \quad (1)$$

Research Stages

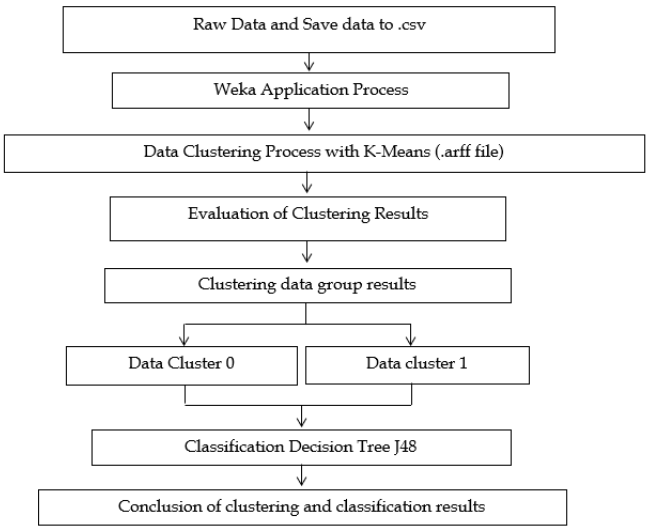

Figure 1. Research stages

The stages of this research are shown in Figure 1 which explains the stages of data search and storage which will then be processed using the Weka application which will then group and classify the data.

Research Scope

The discussion in this research is about student accuracy in their studies in higher education by taking data samples from 131 information systems study program students and analyzing them using Weka software to cluster data on accuracy in their studies.

Table 1. Student Achievement Index Data

Name					Class
	1	2	3	4	5
Ronaldi Putra	2.90	3.35	3.00	3.06	2.75
Degi Syah Putra	3.2	3.29	3.28	3.33	3.60
Ayu Winanda	3.00	3.00	3.00	3.11	3.40
Nadya Dwi Yasra	3.25	3.35	3.11	3.06	3.30
Randi Sulaeman	3.50	3.05	3.44	3.78	3.65
Shinta Amelia Ananda	3.40	3.20	3.33	2.94	3.40
Chalil Gibram	2.20	2.25	3.17	3.11	3.30
Inggi Pangestu	3.15	2.95	2.67	2.56	2.83
Muhammad Rozi Alfarabi	2.20	2.25	3.17	3.11	3.30
Nadia Eka Safitri	3.30	2.90	3.17	2.72	3.22
Retchi Puspita	3.30	3.50	3.28	3.28	3.25
Suryanto	3.80	3.20	2.90	3.4	3.85
Teguh Lendra Akbar	3.80	3.35	3.50	3.28	3.40
Vince Kris Hiburan	3.25	3.35	3.17	2.83	3.05
Baihaqi Azizi	3.45	3.15	3.44	3.56	3.60
Dian Carina	3.30	3.20	3.28	3.22	3.65
Gian Edri Sandi	3.20	3.10	3.17	3.22	3.45
M Rafan	2.75	2.35	2.73	2.39	2.95
Yolan Ananda Putri	3.40	3.10	3.17	3.72	3.20

Based on the results of the raw data above, by carrying out a data grouping process, we will find students who have the potential to complete their studies on time and see to what extent students are still considered lacking in their learning so that they need guidance in their studies and coaching for students who are predicted to be late in their studies.

Result and Discussion

In the dataset shown in Table 1 previously, a clustering process will be carried out on the data using the Weka application, wherein in the initial stage we will prepare the raw data which will be converted into .csv, and in the application used, we will save it with the extension. Arff.

Predicting the level of on-time graduation for students in the Information Systems Study Program. The process of predicting the educational level according to a certain time using data mining techniques (Salim et al., 2020); (Suwitno & Wibowo, 2019). Looking at the results obtained from a sample of students currently studying, the grades are taken from semester 1 to semester 5. The next step is that the study program analyzes the students who are late less than being on time, the data taken does not represent students as a whole. In clustering, data obtained such as social studies from semesters 1 to 5 can be used (Ahuja & Kankane, 2017); (Maziah Wan Ab Razak et al., 2019).

Manual Completion

Determine the number of cluster centers "Centroid" which will be used randomly.

Is known the value of K = 2, namely to determine the decision "On Time" and "Late"; Data in sequence 18 is used as the first grouping center, namely = 3.2, 3.1, 3.17, 3.22, 3.45; 130th Data as the Center of the 2nd Cluster, namely = 3.65, 3.25, 3.5,3.89,3.3 Carried out for further data.

Calculate the distance to the cluster center:

=$\sqrt{(3.2-$2.9) ^2 + (3.1-3.35) ^2 + (3.17-3) ^2 +(3.22-3.06) ^2 + (3.45-2.75) ^2

= 0.83

C2=$\sqrt{(3.65-2.9)}$^2 + (3.25-3.35) ^2 + (3.5-3) ^2 + (3.89-3.06) ^2 + (3.3-2.75) ^2

= 1.35

The process will continue to be repeated until we get the distance from the 131st data to the cluster center. The first column shows the calculation of the distance to the center of the initial cluster. The second column shows the data's distance value to the next cluster's center. The complete distance values are:

Table 2. Distance Values

C1	C2
0.83	1.35
0.81	0.29
0.31	1.16
0.34	1.01
0.72	0.45
0.4	1
1.33	1.95
1.05	1.74
1.33	1.95
…	…
0.59	0.48

Group objects to determine cluster members based on distance.

Table 3. Cluster members based on distance

C1	C2
0	1
1	0
0	1
0	1
1	0
0	1
0	1
0	1
0	1
…..	….
1	0

Completion of Clustering and Classification with the Weka Application

In Figure 2, the results of the clustering test using the K-Means method show the results: Cluster 0 = 78 people (60%) and Cluster 1 = 53 people (40%).

```
Cluster 0: 3.4,4,4,3.83,3.75
Cluster 1: 3.15,3.4,3.44,4,3.9

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute     Full Data             0          1
                (131.0)       (78.0)     (53.0)
=============================================
IPS1            3.2195        3.4064     2.9443
IPS2            3.1869        3.4315     2.8268
IPS3            3.2118        3.4895     2.8032
IPS4            3.4026        3.6451     3.0457
IPS5            3.3368          3.55      3.023



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       78 ( 60%)
1       53 ( 40%)
```

Figure 2. Clustering Test Results

In the next stage, we will look at the data from the clustering results in Figure 3 and we will find a grouping of each existing data, namely groups for cluster 0 and cluster 1.

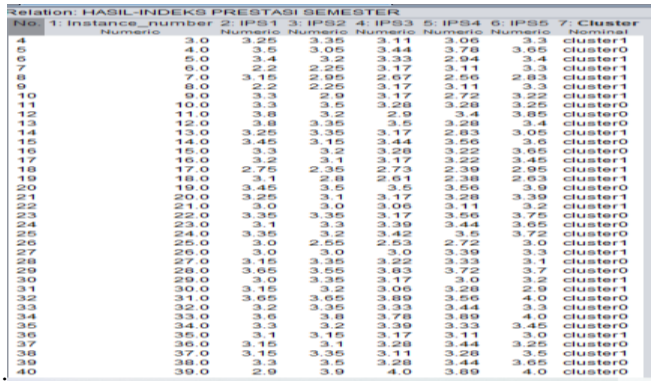

Figure 3. Clustering Results Data

The visualization of cluster assignments in Figure 4 tells about the distribution of data, in blue are the results of cluster 0, and in red are the results of cluster 1.

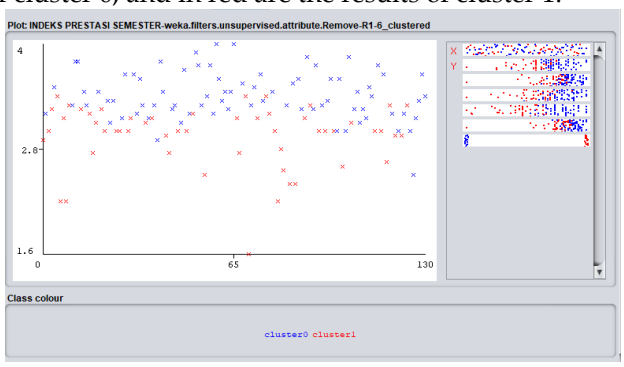

Figure 4. Visualization of cluster assignments

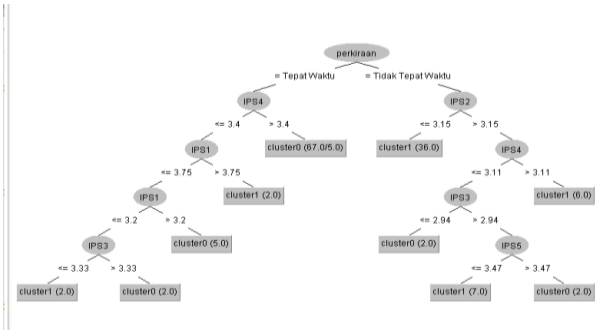

Figure 5. Decision Tree

J48 pruned tree

---

estimated = On Time
| IPS4 <= 3.40
| | IPS1 <= 3.75
| | | IPS1 <= 3.20
| | | | IPS3 <= 3.33: cluster1 (2.0)
| | | | IPS3 > 3.33: cluster0 (2.0)
| | | IPS1 > 3.2: cluster0 (5.0)
| | IPS1 > 3.75: cluster1 (2.0)
| IPS4 > 3.4: cluster0 (67.0/5.0)
Estimation = Not on time
| IPS2 <= 3.15: cluster1 (36.0)
| IPS2 > 3.15
| | IPS4 <= 3.11
| | | IPS3 <= 2.94: cluster0 (2.0)
| | | IPS3 > 2.94
| | | | IPS5 <= 3.47: cluster1 (7.0)
| | | | IPS5 > 3.47: cluster0 (2.0)
| | IPS4 > 3.11: cluster1 (6.0)

In the results of the classification, 2 classes were formed, namely a and b. Class a = 73 data but in class a there are 67 and b there are 6 data. Class B = 58 data, but there are 12 data classified as class A and 46 data in class B.

K-means is included in partitioning clustering, that is, every data must be entered in a particular cluster and allows for any data included in the cluster at a certain stage of the process, at the next stage it moves to the same cluster other (Meng et al., 2018); (Jasinska-Piadlo et al., 2023). K-means separates data into k separate regions, where k is positive integer number (Galluccio et al., 2012); (Ali et al., 2022). The K-Means algorithm is an iterative clustering algorithm partition the data set into a number of K clusters that have been implemented in beginning (Pérez-Ortega et al., 2020). The K-Means algorithm is simple to implement and run, relatively fast, easy to adapt, commonly used in practice. By Ashari et al. (2022); Yuan & Yang (2019), K-Means has been one of the most important algorithms in the field data mining.

K-Means is a method for searching and grouping datahave similar characteristics (similarity) between one data and other data (Putra & Dharma, 2023); (Iyohu et al., 2023). So each clu ster will contain similar data. Clustering is a data mining method that is unsupervised and undirected (Liu & Barahona, 2020). This means that this method does not involve any training exercises and does not require targets output. In data mining there are two types of clustering methods used for data grouping, namely Hierarchical clustering and Non-hierarchical clustering (Syafiyah et al., 2022).

Table 4. Stratified cross-validation

Correctly Instances	Classified	113	86.25%
Incorrectly Instances	Classified	18	13.74 %
Kappa statistic		0.71	
Mean absolute error		0.19	
Root mean squared error		0.35	
Relative absolute error		38.97 %	
Root relative squared error		71.46%	
Total Number of Instances		131	

Table 5. Detailed Accuracy by Class

Correctly Classified Instances	113	86.25 %
Incorrectly Classified Instances	18	13.74 %
Kappa statistic	0.71	
Mean absolute error	0.19	
Root mean squared error	0.35	
Relative absolute error	38.97 %	
Root relative squared error	71.46%	
Total Number of Instances	131	

Decision trees are a method commonly used to make informal or simple decisions (Tan et al., 2023). However, according to (Petropoulos et al., 2022); (Van De Schoot et al., 2021), quite a few people use it to predict results systematically. One example is in data analysis. Decision trees are a very accurate method (Zhang & Gionis, 2023). However, there are still advantages and disadvantages to decision trees that you need to consider. If you are confident with the decision tree method, now you can make it manually or digitally using a computer. So, of course making it will be easier because there is already a choice of templates (Nowell et al., 2017).

Table 6. Detailed Accuracy by Class

			Weighted Avg.
TP Rate	0.91	0.79	0.86
FP Rate	0.20	0.08	0.15
Precision	0.84	0.88	0.86
Recall	0.91	0.79	0.86
F-Measure	0.88	0.83	0.86
MCC	0.72	0.72	0.72
ROC Area	0.81	0.81	0.81
PRC Area	0.77	0.78	0.78
Class	Cluster 0	Cluster 1	

Table 7. Classification Results

A –Cluster 0	B- Cluster 1
67	6
12	46

Conclusion

Based on the application of the grouping technique using the K-Means method and classification decision tree J48, a conclusion can be drawn, that it is known that the sample consists of 131 students taken randomly, in clustering the value k = 2 is determined, namely to determine timely and late decisions using achievement index data. the semester from semesters 1 to 5 for students to get decision results, clustering results are obtained, namely cluster 0 = 78 people with 60% and cluster 1 = 53 people with 40%. Guidance will be provided for the future process so that you graduate on time. Where those included in Cluster 0 are students who have the potential to be on time for their graduation, and those included in Cluster 1 are students who need to be further developed in their learning with the assistance of the study program and academic supervisors in each lecture they take, so that later they will be able to graduate on time. From the results carried out manually, we also got the same results with the same cluster values. The classification results show that two classes are formed, namely a= 73 and b= 58 data. In clustering and continuing with the classification of the clustering results, there is not too much difference in predicting student graduation.

Acknowledgments
The author would like to thank all forms of participation and support in writing the journal. We are also aware that there are still many shortcomings and errors, so we need suggestions and input for the perfection of this article.

Author Contributions
Conceptualization, S. R and S. D., methodology, S. R.; validation, S. R.; formal analysis, S. D..; investigation, S. R.; resources, S. D. and. S. R; data curation, S. D: writing—original draft preparation, S. R. and S. D.; writing—review and editing, S. R.: visualization, S. R. and S. D. All authors have read and agreed to the published version of the manuscript.

Funding
This research was independently funded by researchers.

Conflicts of Interest
The authors declare no conflict of interest.

Reference

Abbas, K. A., Gharavi, A., Hindi, N. A., Hassan, M., Alhosin, H. Y., Gholinezhad, J., Ghoochaninejad, H., Barati, H., Buick, J., Yousefi, P., Alasmar, R., & Al-Saegh, S. (2023). Unsupervised machine learning technique for classifying production zones in unconventional reservoirs. *International Journal of Intelligent Networks*, *4*, 29–37. https://doi.org/10.1016/j.ijin.2022.11.007

Ahuja, R., & Kankane, Y. (2017). Predicting the probability of student's degree completion by using different data mining techniques. *2017 Fourth International Conference on Image Information Processing (ICIIP)*, 1–4. https://doi.org/10.1109/ICIIP.2017.8313763

Al-Anazi, S., AlMahmoud, H., & Al-Turaiki, I. (2016). Finding Similar Documents Using Different Clustering Techniques. *Procedia Computer Science*, *82*, 28–34. https://doi.org/10.1016/j.procs.2016.04.005

Aldo, D. (2023). Data Mining Sales of Skin Care Products Using the K-Means Method. *Sinkron*, *8*(1), 295–304. https://doi.org/10.33395/sinkron.v8i1.12007

Ali, I., Rehman, A. U., Khan, D. M., Khan, Z., Shafiq, M., & Choi, J.-G. (2022). Model Selection Using K-Means Clustering Algorithm for the Symmetrical Segmentation of Remote Sensing Datasets. *Symmetry*, *14*(6), 1149. https://doi.org/10.3390/sym14061149

Ashari, I. F., Banjarnahor, R., Farida, D. R., Aisyah, S. P., Dewi, A. P., & Humaya, N. (2022). Application of Data Mining with the K-Means Clustering Method and Davies Bouldin Index for Grouping IMDB Movies. *Journal of Applied Informatics and Computing*, *6*(1), 07–15. https://doi.org/10.30871/jaic.v6i1.3485

Asril, T. (2020). Prediction of Students Study Period using K-Nearest Neighbor Algorithm. *International Journal of Emerging Trends in Engineering Research*, *8*(6), 2585–2593. https://doi.org/10.30534/ijeter/2020/60862020

Asroni, A., Kurniasari, D., & Kurnianti, A. (2020). The Implementation of Clustering Method With K-Means Algorithm In Grouping Data of Students' Course Scores at Universitas Muhammadiyah

Yogyakarta. *Emerging Information Science and Technology*, *1*(3), 75–83. https://doi.org/10.18196/eist.v1i3.13172

Bharara, S., Sabitha, S., & Bansal, A. (2018). Application of learning analytics using clustering data Mining for Students' disposition analysis. *Education and Information Technologies*, *23*(2), 957–984. https://doi.org/10.1007/s10639-017-9645-7

Chaudhry, M., Shafi, I., Mahnoor, M., Vargas, D. L. R., Thompson, E. B., & Ashraf, I. (2023). A Systematic Literature Review on Identifying Patterns Using Unsupervised Clustering Algorithms: A Data Mining Perspective. *Symmetry*, *15*(9), 1679. https://doi.org/10.3390/sym15091679

Den Teuling, N. G. P., Pauws, S. C., & Van Den Heuvel, E. R. (2023). A comparison of methods for clustering longitudinal data with slowly changing trends. *Communications in Statistics - Simulation and Computation*, *52*(3), 621–648. https://doi.org/10.1080/03610918.2020.1861464

Dol, S. M., & Jawandhiya, P. M. (2023). Classification Technique and its Combination with Clustering and Association Rule Mining in Educational Data Mining—A survey. *Engineering Applications of Artificial Intelligence*, *122*, 106071. https://doi.org/10.1016/j.engappai.2023.106071

Galluccio, L., Michel, O., Comon, P., & Hero, A. O. (2012). Graph based k-means clustering. *Signal Processing*, *92*(9), 1970–1984. https://doi.org/10.1016/j.sigpro.2011.12.009

Ha, W., Ma, L., Cao, Y., Feng, Q., & Bu, S. (2024). The effects of class attendance on academic performance: Evidence from synchronous courses during Covid-19 at a Chinese research university. *International Journal of Educational Development*, *104*, 102952. https://doi.org/10.1016/j.ijedudev.2023.102952

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Iyohu, L. R., Ismail Djakaria, & La Ode Nashar. (2023). Perbandingan Metode K-Means Clustering dengan Self-Organizing Maps (SOM) untuk Pengelompokan Provinsi di Indonesia Berdasarkan Data Potensi Desa. *Jurnal Statistika Dan Aplikasinya*, *7*(2), 195–206. https://doi.org/10.21009/JSA.07208

Jasinska-Piadlo, A., Bond, R., Biglarbeigi, P., Brisk, R., Campbell, P., Browne, F., & McEneaney, D. (2023). Data-driven versus a domain-led approach to k-means clustering on an open heart failure dataset. *International Journal of Data Science and Analytics*, *15*(1), 49–66. https://doi.org/10.1007/s41060-022-00346-9

Liu, Z., & Barahona, M. (2020). Graph-based data clustering via multiscale community detection. *Applied Network Science*, *5*(1), 3. https://doi.org/10.1007/s41109-019-0248-7

Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences*, *13*(4), 410. https://doi.org/10.3390/educsci13040410

Loeng, S. (2020). Self-Directed Learning: A Core Concept in Adult Education. *Education Research International*, *2020*, 1–12. https://doi.org/10.1155/2020/3816132

Márquez, J., Lazcano, L., Bada, C., & Arroyo-Barrigüete, J. L. (2023). Class participation and feedback as enablers of student academic performance. *SAGE Open*, *13*(2), 215824402311772. https://doi.org/10.1177/21582440231177298

Maziah Wan Ab Razak, W., Alia Syed Baharom, S., Abdullah, Z., Hamdan, H., Ulfa Abd Aziz, N., & Ismail Mohd Anuar, A. (2019). Academic Performance of University Students: A Case in a Higher Learning Institution. *KnE Social Sciences*, *3*(13), 1294. https://doi.org/10.18502/kss.v3i13.4285

Meng, Y., Liang, J., Cao, F., & He, Y. (2018). A new distance with derivative information for functional k-means clustering algorithm. *Information Sciences*, *463–464*, 166–185. https://doi.org/10.1016/j.ins.2018.06.035

Mulyaningsih, S., & Heikal, J. (2022). K-Means Clustering Using Principal Component Analysis (PCA) Indonesia Multi-Finance Industry Performance Before and During Covid-19. *Asia Pacific Management and Business Application*, *011*(02), 131–142. https://doi.org/10.21776/ub.apmba.2022.011.02.1

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods*, *16*(1), 160940691773384. https://doi.org/10.1177/1609406917733847

Oyewole, G. J., & Thopil, G. A. (2023). Data clustering: Application and trends. *Artificial Intelligence Review*, *56*(7), 6439–6475. https://doi.org/10.1007/s10462-022-10325-y

Pérez-Ortega, J., Nely Almanza-Ortega, N., Vega-Villalobos, A., Pazos-Rangel, R., Zavala-Díaz, C., & Martínez-Rebollar, A. (2020). The *K* -Means Algorithm Evolution. *Introduction to Data Science and Machine Learning*, 69-90. https://doi.org/10.5772/intechopen.85447

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C.,

Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., … Ziel, F. (2022). Forecasting: Theory and practice. *International Journal of Forecasting*, *38*(3), 705–871. https://doi.org/10.1016/j.ijforecast.2021.11.001

Priyatna, R. D., Tulus, & Ramli, M. (2018). K-Means algorithm and modification using gain ratio. *IOP Conference Series: Materials Science and Engineering*, *420*, 012133. https://doi.org/10.1088/1757-899X/420/1/012133

Putra, I. G. K. K., & Dharma, I. G. W. S. (2023). Application of The K-Means Clustering Method To Search For Potential Tourists of Bendesa Hotel. *TIERS Information Technology Journal*, *4*(1), 8–15. https://doi.org/10.38043/tiers.v4i1.4297

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. *Applied Sciences*, *10*(3), 1042. https://doi.org/10.3390/app10031042

Salim, A. P., Laksitowening, K. A., & Asror, I. (2020). Time Series Prediction on College Graduation Using KNN Algorithm. *2020 8th International Conference on Information and Communication Technology (ICoICT)*, 1–4. https://doi.org/10.1109/ICoICT49345.2020.9166238

Silva, M. D. B., De Oliveira, R. D. V. C., Da Silveira Barroso Alves, D., & Melo, E. C. P. (2021). Predicting risk of early discontinuation of exclusive breastfeeding at a Brazilian referral hospital for high-risk neonates and infants: A decision-tree analysis. *International Breastfeeding Journal*, *16*(1), 2. https://doi.org/10.1186/s13006-020-00349-x

Suwitno, S., & Wibowo, A. (2019). On-Time Graduation Prediction System Using Data Mining Classification Method. *Proceedings of the Proceedings of the 1st Workshop on Multidisciplinary and Its Applications Part 1, WMA-01 2018, 19-20 January 2018, Aceh, Indonesia,* 1-9 https://doi.org/10.4108/eai.20-1-2018.2281900

Syafiyah, U., Puspitasari, D. P., Asrafi, I., Wicaksono, B., & Sirait, F. M. (2022). Analisis Perbandingan Hierarchical dan Non-Hierarchical Clustering Pada Data Indikator Ketenagakerjaan di Jawa Barat Tahun 2020. *Seminar Nasional Official Statistics*, *2022*(1), 803–812. https://doi.org/10.34123/semnasoffstat.v2022i1.1221

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. *Applied Sciences*, *13*(7), 4550. https://doi.org/10.3390/app13074550

Van De Schoot, R., De Bruin, J., Schram, R., Zahedi, P., De Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3*(2), 125–133. https://doi.org/10.1038/s42256-020-00287-7

Wirawan, C., Khudzaeva, E., Hasibuan, T. H., Karjono, & Lubis, Y. H. K. (2019). Application of Data mining to Prediction of Timeliness Graduation of Students (A Case Study). *2019 7th International Conference on Cyber and IT Service Management (CITSM)*, 1–4. https://doi.org/10.1109/CITSM47753.2019.8965425

Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *J*, *2*(2), 226–235. https://doi.org/10.3390/j2020016

Zhang, G., & Gionis, A. (2023). Regularized impurity reduction: Accurate decision trees with complexity guarantees. *Data Mining and Knowledge Discovery*, *37*(1), 434–475. https://doi.org/10.1007/s10618-022-00884-7