JISS Journal of Indonesian Social Science

Vol. 5, No. 8, August 2024
E-ISSN: 2723-6692
P-ISSN: 2723-6595
http://jiss.publikasiindonesia.id/

Implementation of K-Means Algorithm for Clustering Scholarship Candidates at TPA Anak Sholeh

Nurwati, Yudi Santoso
Universitas Budi Luhur, Jakarta, Indonesia
Email: nurwati@budiluhur.ac.id, yudi.santoso@budiluhur.ac.id
Correspondence: nurwati@budiluhur.ac.id*

KEYWORDS
K-means algorithm; Scholarship; Qur'an Education Park

ABSTRACT
The Qur'an Education Park (TPA) plays an important role in improving children's education in the surrounding environment by providing facilities for learning to read and write the Quran, reading short prayers, and other positive activities. However, the criteria for scholarship recipients at TPA Anak Sholeh are still unclear, even though the provision of this scholarship aims to motivate students to be more active in learning and excelling. This study aims to determine more specific criteria for scholarship recipients, such as attendance, achievements, home conditions, and parents' income. Another objective is to validate and improve the accuracy of the K-Means Clustering model used in grouping the data of prospective scholarship recipients into three categories: accepted, considered, and rejected. The research method used is the K-Means Clustering algorithm. Algorithm performance evaluation is carried out through metrics such as accuracy, precision, recall, f-measure, and cross-validation. Testing was carried out on TPA student data as many as 80 student data. The results of the grouping using 3 clusters were obtained as many as 46 students received C1 scores, C2 scores were rejected by 12 students and C3 scores were considered for 22 students. The determination of the centroid distance has an effect on the value of the cluster results when conducting the scholarship attribute collateralization.

Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

1. Introduction

TPA (Alqur'an Education Park), one of the educational facilities close to residential areas and affordable in cost, is in great demand by the community, so many leave their children in landfills. TPA has an important role in children's education. One of the TPA programs is non-formal education, which includes reading and writing the Qur'an among children (Nurjayanti et al., 2020). The Qur'an Education Park (TPA) is a non-formal educational institution that has the leading position of directing the skills of reading and writing the Qur'an, also functions to increase the

growth of children's souls and form children's religious character (Guchi & Sunarti, 2021; Wandri et al., 2023).

As a landfill that was established 8 (eight) years ago, TPA Anak Sholeh has 7 (seven) teachers and students, totalling approximately 70 (seventy) children with its building as a result of the self-help of guardians, students, and donors from sharing parties. TPA accepts students from various walks of life. For 8 (eight) years of the learning process at TPA, many potential students have excelled in Qur'an reading, da'wah achievements, Qur'an memorization achievements and poetry humming (marawis). The Sholeh Children's Landfill already has a foundation that oversees the activities and supports the costs of the learning process. However, the terms and criteria for prospective scholarship recipients are still unclear: who and what is included in the recipient criteria. To motivate students at TPA, scholarships are planned to be provided to students who quickly accept lessons and have the potential to achieve achievements.

The scholarship provides financial assistance to individuals who aim to sustain the education pursued (Campbell & Neff, 2020; Manihuruk et al., 2020). The purpose of providing scholarships to TPA Anak Sholeh students is to appreciate the achievements that have been achieved and motivate students to study more actively at TPA. The scholarship assistance provided is in the form of full education fees while being a TPA student.

This study uses the K-Means algorithm to determine the criteria for grouping prospective scholarship recipients. The K-Means method partitions data into groups so that data with the same characteristics are entered into the same group and data with different characteristics are grouped into other groups. Data with a restrictive value equation in one group and data with differences in another group allows for the grouping of different data with a small degree of variation. The main principle of this technique is to compile K partitions/centroids/mean from a data set. This data grouping aims to minimize the objective functions set in the grouping process, which generally seeks to minimize variation within a group and maximize variation between groups (Hoban et al., 2020; Sulistiyawati & Supriyanto, 2021).

This algorithm groups prospective scholarship recipients into three categories: accepted, rejected, and considered. The K-Means Clustering Algorithm helps cluster the proposals of prospective scholarship recipients into categories that are worthy of being proposed and not worthy of receiving scholarships according to the criteria that have been determined. The K-means algorithm is used because there has been much research on grouping things using Clustering (Ahmed et al., 2020; Rahmah & Antares, 2022).

Previous research on the K-means clustering algorithm was used to recommend scholarship recipients (Khomarudin et al., 2021). The results of the recommendation are in the form of placement of scholarship applicant data to each cluster group produced. The data used in this study are Final School Exam Scores (UAS), Report Card Scores, House Status, Home Condition, and Parents' Income. With a cluster (k=3) with 30 (thirty) scholarship recipients. Through attribute selection, K-means calculates each data into a predetermined cluster. Of the results of the calculations that have been processed, as many as 16% are accepted, 61% are considered, and 23% are rejected (Salam et al., 2020).

2. Materials and Methods

The research stages explain as follows:

1. Collect dataset data,

At this stage, collect the required data sets according to the research problem. The data set was obtained from the initial student registration documents registering to become students at the Anak Sholeh's landfill. A multiclass dataset is a unique dataset that has more than two labels. The performance of a method can be measured through testing curation, precision, recall and f-measure values, and by applying cross-validation to performance testing, the robustness of the performance can be seen (Azis et al., 2020).

2. Data processing by specifying,

After compiling the data set in stage 1. The next stage is the selection of the data set by determining what attributes are used. Data Mining preprocessing can improve the quality of processed data through the stages of data cleaning, data integration, data selection, and data transformation. This is done so that the processed data is of higher quality, meaning that the data is objective, representative, has a small sampling error, is updated and relevant (Alwendi; et al., 2023).

3. Get a new set of data,

The 3rd (three) stage is the result of the 2nd (two) stage, which is to obtain a new set of data with selected attributes.

4. Algorithm K-means (modelling)

The algorithm for performing K-Means clustering is as follows: (1) Pick K centroid points at random. (2) Group the data so that K clusters are formed, with the centroid point of each cluster being the centroid point that was selected previously. (3) Update the centroid point value. (4) Repeat steps 2 and 3 until the value of the centroid point no longer changes. Data can be grouped into clusters by calculating the closest distance from data to a centroid point. Minkowski's distance calculation can be used to calculate the distance between 2 pieces of data (Zuhal, 2022). To carry out the data processing process at each point of the centre of the cluster, namely with the Euclidean distance theory formulated (Harahap & Rambe, 2021), as follows:

K-means formula (Sulistiyawati & Supriyanto, 2021)

$$v_{ij} = \frac{1}{N_i}\sum_{k=0}^{N_i} X_{kj}$$

vij is the centroid or the ith cluster mean for the jth variable
This is the amount of data that is a member of the ith cluster
i,k is the index of the cluster
j is the index of the variable
xkj is the value of the kth data in the cluster for the jth variable.

5. Evaluation

Evaluation is carried out to determine whether all data has gone through the k-means algorithm. The evaluation was conducted to get the maximum results when testing the data set.

Data Validation and Testing

1. Data Validation: In data testing, cross-validation is used as the primary method to ensure that the developed model has consistent performance. This method divides the dataset into subsets, trains the model on a specific subset, and tests other subsets. This helps minimise model bias and ensures that the results can be generalized to new data.
2. Data Testing: Data is tested using several evaluation metrics, such as precision, recall, and f-measure. This metric measures how well the model correctly groups data based on predefined criteria. In addition, sensitivity analysis was also carried out to measure how sensitive the model is to small changes in the dataset.

3. Result and Discussion

Stages of collecting data sets

The research data source was obtained from TPA Anak Sholeh's registration documents. The data used are daily student attendance scores, home status, achievements, parental income from 80 students, and each student's assignment score.

Data processing by specifying

The data collected is as follows:

Tabel 1Tampilan atribut

Kode	Attribute Name
T03	Value attendance
T04	Home status
T05	Have achievements
T06	Parents' income
T07	Assign grades

Table 1 of the attribute display describes the attributes that will be used in calculating K-means clustering, namely with the code T03 for the student attendance value at the TPA. Code T04 for the status of the occupied student's house. Code T05 to have achievements. Code T06 for the student's parent's income. Code T07 for Grades of student assignments while attending lessons at TPA (the duration of study at TPA is approximately 2 hours per day from Monday to Friday).
The value of each code is shown in Table 2.

Table 2 Attribute Values

No.	Code	Attribute name	Name Value attribute	Value
1	T04	Home status	Own	0
2	T04	Home status	Contracting	1
3	T05	Have achievements	Achievement	1
4	T05	Have achievements	No Achievement	0
5	T06	Parents' income	<=2.000.000	1
6	T06	Parents' income	>2.000.000	0

Table 2 regarding attribute values. The attribute value view contains 0 and 1 for each attribute.

Get a new set of data

This stage calculates the weight of each attribute's value. In the previous stage, it was determined that there are three clusters: accepted, rejected, and considered. The next step is to analyze the data and then process it with the K-means clustering algorithm. Table 3 of the overall result display of student data is shown below.

Table 3 Display of Overall Results of Student Data

No.	Student Name	Attendance Value	Home Status	Have achievements	Parents' income	Assign grades
1	Bagas	100	Contracting	2	1.900.000	100
2	Fauzan	100	Contracting	3	1.000.000	100
3	Syauqi	100	Own	1	4.000.000	100
4	Aisyah	100	Own	1	5.000.000	100
5	Shanum	97	Contracting	1	1.800.000	100
6	Nadia	100	Contracting	2	1.950.000	100
7	Arraya	100	Own	2	5.000.000	100
8	Darrel	89	Contracting	1	1.200.000	100
9	Juan	100	Own	1	3.000.000	100
10	Cantika	98	Contracting	2	1.980.000	100
…	….	…	…	…	…	…
…	…	…	…	…	…	…
80	Arjuna	99	Contracting	3	1.890.000	100

Table 3 shows the overall results of this student data that will be processed and grouped into 3 clusters.

Algoritma K-means (modelling)

For the modeling process using the K-means algorithm, a centroid value is generated from the data obtained with the provision that the desired clustering is 3 (Manihuruk et al., 2020). The analysis results table is shown in table 4 of the analysis data display below.

Table 4 Display of Analysis Results Data

No.	Student Name	Attendance Value	Home Status	Have achievements	Parents' income	Assign grades
1	Bagas	100	1	1	1	100
2	Fauzan	100	1	1	1	100
3	Syauqi	100	0	0	0	100
4	Aisyah	100	0	0	0	100
5	Shanum	97	1	0	1	100
6	Nadia	100	1	1	1	100
7	Arraya	100	0	1	0	100
8	Darrel	89	1	0	1	100
9	Juan	100	0	0	0	100
10	Cantika	98	1	1	1	100
…	….	…	…	…	…	…
…	…	…	…	…	…	…
80	Arjuna	99	1	1	1	100

Table 4 shows the results of the analysis of all the collected data. The initial cluster point is determined by taking the largest value for the accepted cluster (C1), the smallest value for the rejected cluster (C2), and the average value for the considered cluster (C3). Table 5 shows the table of the initial cluster points (Initial centroid values) below.

Table 5 Early centroid values

Cluster	Value attendance	Home status	Have achievements	Parents' income	Assign grades
C1	100	1	1	1	100
C2	89	0	0	0	80
C3	98,4375	0,7875	0,6625	0,725	95,125

Table 5 shows the initial centroid values of K-means processing. Then, using the initial centroid value, it is obtained into 3 clusters. This cluster process is processed by taking the closest distance from each processed data. From the student data of 80 students, grouping was obtained in iterations 1 to 3 clusters. The calculation of the cluster centre distance is shown in Table 6 below.

Table 6 Cluster centre distance calculation

No	Student Name	Attendance Value	Home Status	Have achievements	Parents' income	Assign grades	Average	C1	C2	C3	Closest Distance
1	Bagas	100	1	1	1	100	40,6	0	22,9	5,2	0
2	Fauzan	100	1	1	1	100	40,6	0	22,9	5,2	0
3	Syauqi	100	0	0	0	100	40	1,7	22,8	5,3	1,7
4	Aisyah	100	0	0	0	100	40	1,7	22,8	5,3	1,7
5	Shanum	97	1	0	1	100	39,8	3,2	21,6	5,2	3,2
6	Nadia	100	1	1	1	100	40,6	0	22,9	5,2	0
7	Arraya	100	0	1	0	100	40,6	1	22,8	5,3	1
8	Darrel	89	1	0	1	100	38,2	11	20,0	10,6	10,6
9	Juan	100	0	0	0	100	40	2	22,8	5,3	2
10	Cantika	98	1	1	1	100	40,2	2	22,0	4,9	2
...	....	...	...	...	...	...	...	..	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...
80	Arjuna	99	1	1	1	100	40,4	1	22,4	5,0	1

Table 6 displays the calculation of the cluster centre distance, obtaining the values of C1, C2, C3, and the nearest distance. The next step is to group the results shown in Table 7 below.

Table 7 Grouping Results

No	Name	C1	C2	C3
1	Bagas	1		
2	Fauzan	1		
3	Syauqi	1		
4	Aisyah	1		
5	Shanum	1		
6	Nadia	1		
7	Arraya	1		
8	Darrel			1
9	Juan	1		

10	Cantika	1		
...	....	...	...	...
...	....	...	...	...
80	Arjuna	1		
	Total	46	12	22

In table 7, the grouping results obtained that the C1 score was accepted by 46 students, the C2 score was rejected by 12 students and the C3 score was considered for 22 students. Table 7 is equipped with the graph shown in figure 2 of the grouping result graph below.

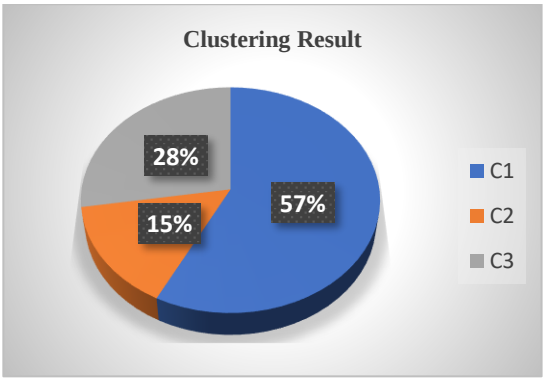

Figure 1 Graph of Clustering Results

The study's results using the K-Means algorithm for data grouping at the Sholeh Children's Landfill have practical implications, including improving the student registration process and decision-making related to scholarship acceptance.

With accurate data grouping, TPA can more efficiently identify eligible students to receive scholarships based on predetermined criteria, such as academic achievement and economic conditions. This allows TPA to distribute assistance to those who need it and have the potential to achieve further.

Furthermore, this grouping method can also be applied in a broader context, such as in other educational institutions that want to implement a more structured and data-driven scholarship selection system. In addition, the results of this grouping can also be used as a reference in improving the quality of education at TPA, for example, by providing special attention or additional learning programs for groups of students who are in the category of "considered" or "rejected," so that they can improve their achievements.

4. Conclusion

Based on the results of the research obtained, it is concluded that the K-means clustering method can produce recommendations for scholarship recipients by involving 5 attributes, namely the attribute of attendance value, the attribute of home status, the attribute of having achievements, the attribute of parental income, and attribute of assignment value. The determination of the centroid or central point value affects the value of cluster results such as testing the data of the Sholeh's children TPA students. There are 3 clusters, namely C1 for the cluster accepted, C2 for the

cluster rejected and C3 for the cluster considered. This study used 80 student data. The results of data grouping obtained that the C1 score was accepted by 46 students, the C2 score was rejected by 12 students and the C3 score was considered for 22 students. The results obtained from the research can be input to TPA Anak Sholeh in providing scholarships using 3 criteria. Research Recommendations, Specific Scholarship Criteria: Clarify the criteria for scholarship recipients based on attributes such as attendance, achievements, home conditions, and parental income, to improve the accuracy of the grouping. This ensures that scholarships are given to those who need it the most and have maximum potential. Accumulate the K-Means Model: Advanced validation and adjustment of the number of centroids can improve clustering results. Evaluate the performance of the algorithm with various metrics to ensure consistency and accuracy of the grouping data.

5. References

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics*, *9*(8), 1295. https://doi.org/10.3390/electronics9081295

Alwendi;, Mandopa, A. S., & Hasibuan, E. A. (2023). Aplikasi Data Mining Untuk Menentukan Masa Studi Mahasiswa Menggunakan Metode Association Rule Data Mining Application to Determine Student Study Period Using Association Rule Method. *Jurnal Pendidikan Dewantara*, *2*(1), 1–6.

Azis, H., Tangguh Admojo, F., & Susanti, E. (2020). Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah. *Techno.Com*, *19*(3), 286–294. https://doi.org/10.33633/tc.v19i3.3646

Campbell, A. C., & Neff, E. (2020). A Systematic Review of International Higher Education Scholarships for Students From the Global South. *Review of Educational Research*, *90*(6), 824–861. https://doi.org/10.3102/0034654320947783

Guchi, G. A., & Sunarti, V. (2021). Relationship Between Parenting Style and Establishment of Religious Characters at Taman Pendidikan Alquran (TPA) Masjid Alfurqon Desa Sikuliek Kecataman Koto Tangah Kota Padang. *SPEKTRUM: Jurnal Pendidikan Luar Sekolah (PLS)*, *9*(2), 195. https://doi.org/10.24036/spektrumpls.v9i2.112400

Harahap, B., & Rambe, A. (2021). Implementasi K-Means Clustering Terhadap Mahasiswa yang Menerima Beasiswa Yayasan Pendidikan Battuta di Universitas Battuta Tahun 2020/2021 Studi Kasus Prodi Informatika. *Informatika*, *9*(3), 90–97. https://doi.org/10.36987/informatika.v9i3.2185

Hoban, S., Bruford, M., D'Urban Jackson, J., Lopes-Fernandes, M., Heuertz, M., Hohenlohe, P. A., Paz-Vinas, I., Sjögren-Gulve, P., Segelbacher, G., Vernesi, C., Aitken, S., Bertola, L. D., Bloomer, P., Breed, M., Rodríguez-Correa, H., Funk, W. C., Grueber, C. E., Hunter, M. E., Jaffe, R., … Laikre, L. (2020). Genetic diversity targets and indicators in the CBD post-2020 Global Biodiversity Framework must be improved. *Biological Conservation*, *248*, 108654. https://doi.org/10.1016/j.biocon.2020.108654

Khomarudin, A. N., Zakir, S., Novita, R., Endrawati, E., Mat, M. Z. bin A., & Maiyana, E. (2021). K-Mean Clustering Algorithm in Grouping Prospective Scholarship Recipients. *Journal of Physics: Conference Series*, *1779*(1), 012007. https://doi.org/10.1088/1742-6596/1779/1/012007

Manihuruk, N. A., Zarlis, M., Irawan, E., & Tambunan, H. S. (2020). Penerapan Data Mining Dalam Mengelompokkan Calon Penerima Beasiswa Dengan Menggunakan Algoritma K-Means. *KOMIK*

(Konferensi Nasional Teknologi Informasi Dan Komputer), *4*(1), 29–34. https://doi.org/10.30865/komik.v4i1.2575

Nurjayanti, D., Pudyaningtyas, A. R., & Dewi, N. K. (2020). Penerapan Program Taman Pendidikan Alquran (Tpa) Untuk Anak Usia Dini. *Kumara Cendekia*, *8*(2), 183. https://doi.org/10.20961/kc.v8i2.34631

Rahmah, S. A., & Antares, J. (2022). Klasterisasi Seleksi Mahasiswa Calon Penerima Beasiswa Yayasan Menggunakan K-Means Clustering. *I N F O R M a T I K a*, *13*(2), 25. https://doi.org/10.36723/juri.v13i2.282

Salam, A., Adiatma, D., & Zeniarja, J. (2020). Implementasi Algoritma K-Means Dalam Pengklasteran untuk Rekomendasi Penerima Beasiswa PPA di UDINUS. *JOINS (Journal of Information System)*, *5*(1), 62–68. https://doi.org/10.33633/joins.v5i1.3350

Sulistiyawati, A., & Supriyanto, E. (2021). Implementasi Algoritma K-means Clustring dalam Penetuan Siswa Kelas Unggulan. *Jurnal Tekno Kompak*, *15*(2), 25. https://doi.org/10.33365/jtk.v15i2.1162

Wandri, R., Arta, Y., Hanafiah, A., & Oktaviaani, R. (2023). Prediction of Student Scholarship Recipients Using the K-Means Algorithm and C4. 5. *Indonesian Journal of Computer Science*, *12*(1).

Zuhal, N. K. (2022). Study comparison K-Means clustering dengan algoritma hierarchical clustering. *Prosiding Seminar Nasional Teknologi Dan Sains*, *1*, 200–205.