Application of the Cluster Classification Data Mining Method to Child Illiteracy in Indonesia

Muhammad Arifin[1], Gita Widi Bhawika[2], M.A. Muazar Habibi[3], Winci Firdaus[4], Danu Eko Agustinova[5], Robbi Rahim[6*]

[1]Universitas Muria Kudus, Kudus, Indonesia. Email: arifin.m@umk.ac.id
[2]Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. Email: gita@mmt.its.ac.id
[3]University of Mataram, Indonesia. Email: muazar.habibi@unram.ac.id
[4]Badan Pengembangan dan Pembinaan Bahasa, Indonesia. Email: wincifirdaus1@gmail.com
[5]Universitas Negeri Yogyakarta, Yogyakarta, Indonesia. Email: danu_eko@uny.ac.id
[6]Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia. Email: usurobbi85@zoho.com

Corresponding Email: usurobbi85@zoho.com

Abstract – The objective of this study is to cluster and classify data using a combination of the k-means and C4.5 methods. The process involves clustering and subsequent classification. The classification process uses k-folds = 10 and samples = stratified sampling. In this study, analphabets in Indonesia of a minimum age of 15 years (15+) were evaluated. The data are the percentage of analogs between 2017 and 2019. The dataset was obtained from https://www.bps.go.id and is accessible at https://osf.io/crwug. In this study, the Davies Bouldin index (DBI) was used to determine the number of clusters with an optimal DBI value of k = 2, namely, 0,121. The results of the cluster maps in Indonesian territories demonstrate low clustering (C 0 = 22 provinces) and high clustering (C 1 = 11 provinces) for children with k = 2 analphabets. Then, the clustering results were classified, and an accuracy of 97.50 was realized, along with a recall of 90.91%, a precision of 100.00%, and an AUC (optimistic) of 0.95 (excellent classification).

Keywords –Data Mining, Classification, C4.5 Algorithm, K-Means, Child Illiteracy

1. Introduction

Illiteracy is one of the factors that limits human resource quality. Illiteracy in the community must be eradicated to improve the quality of human resources[1]. According to UNESCO, through the Dakkar Declaration 2013, illiteracy is a global problem. Most illiterate people live in developing countries, and Indonesia is in this development category [2]. The government is always committed to developing the education sector, which has a strategic role [3]. Educational development is undertaken to increase human resources to promote successful development. However, in the past few years, the government has not been successful in reducing analphabet rates in all regions [4]. Thus, the percentage of illiterate children in Indonesia requires cluster mapping. Illiteracy renders a nation's younger generation poorer. Various studies have been conducted on illiteracy [5]. This paper proposes a k-medoid clustering method and utilizes it to map the proportion of illiterate individuals between 2009 and 2017 using a dataset. The dataset that is used was collected in 2017-2019 on analphabets aged 15 years and older (15+). The Davies Bouldin index (DBI) algorithm determines the optimal number of clusters before the result is obtained. After the mapping results are obtained, classification is performed to evaluate the accuracy, precision, recall, and AUC of the clustering results. This study differs substantially from previous studies. Several studies have been conducted on the combination of clustering and classification [6]–[9]. In [10], opportunities for placement in education are examined. This paper proposes a combination of clustering and classification methods for future management students (K-means, SVC and naive Bayes) when selecting the specialization of a master's degree in business administration (MBA), such as finance, marketing, human resources, or operations. This combination proved to be the best forecast algorithm for the cluster model, with a precision of 83.05%.

2. Methodology

2.1. Dataset

A simulation was conducted to calculate the cluster mapping of illiterate children in Indonesia using K-means and C4.5 algorithm combinations. It utilized 33 regional data items for the 15+ age group in 2017-2019. The dataset can be accessed from bps.go.id/ at https://osf.io/crwug. The analysis process uses the Rapid Miner software. The phases of this study include the following:

a) Dataset preparation and determination of the optimal number of clusters using the Davies Bouldin index (DBI) algorithm;
b) Clustering using cross-validation after determining the number of clusters;
c) Simulation to obtain the clustering results and use of these results as input for the classification process to evaluate the accuracy, accuracy, recall and AUC of the clusters and to draw conclusions.

2.2. Cross-Validation

Cross-validation is a method for evaluating algorithm performance. K folds of the data are utilized iteratively to obtain a data precision value for the algorithm [11].

2.3. Davies Bouldin Index (DBI)

DBI [12]–[14] is a measure that is used to assess the clustering performance. DBI is utilized as a clustering metric due to the general clustering approach [15], [16]. Validation consists of two key categories, namely, external validation and internal validation, for evaluating the performance [17].

2.4. Model Evaluation

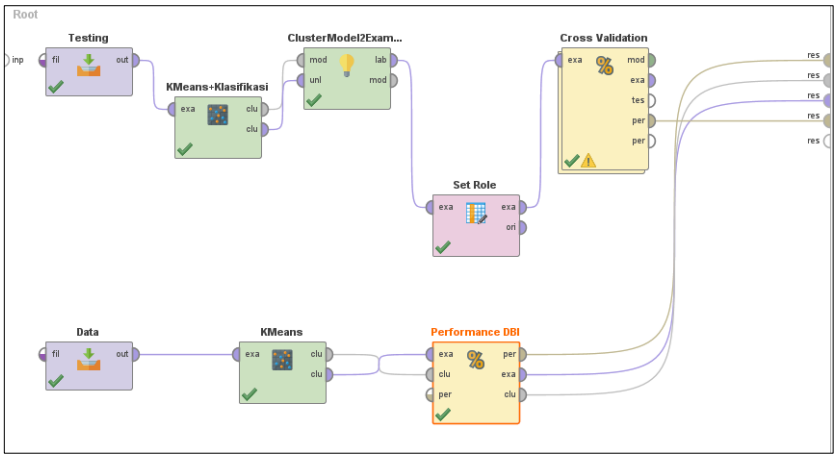
As an accuracy indicator, the area under the curve (AUC) should increase with the convergence between experiments. The guidelines for AUC accuracy classification are presented in Table 1 [18].

Table 1. AUC values

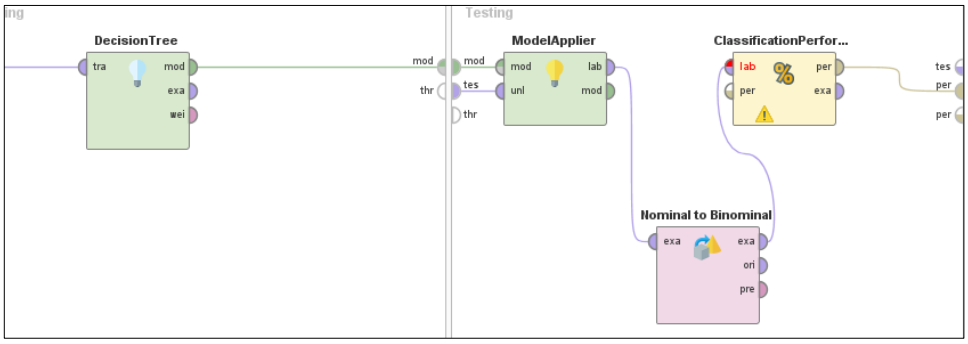
AUC	Meaning
0.90 - 1.00	Excellent Classification
0.80 - 0.90	Good Classification
0.70 - 0.80	Fair Classification
0.60 - 0.70	Poor Classification
< 0.60	Failure

3. Results and Discussion

This stage of the test procedure uses the Rapid Miner software to apply the k-means and C4.5 clustering methods in combination on an illiterate child dataset from Indonesia. The model for combining the two methods with the Rapid Miner software is illustrated in Figure 1.



(a)



(b)

Figure 1. Combination model of k-means and C4.5 (a)(b)

In Figure 1(a), Excel format is used (.xls). In the design, the optimal DBI value is obtained using performance operators (k = 2, 3, 4). The optimal DBI results are used as a reference in clustering. The clustering process uses the sampling type cross-validation test (k = 10) with stratified sampling. Then, the classification process in Figure (b) uses C4.5 with standard parameters. The output of the classification is used to determine the precision, accuracy, precision and AUC to evaluate the operator's clustering (classification) performance. The results of the calculation of the DBI value in the clustering process are as follows:

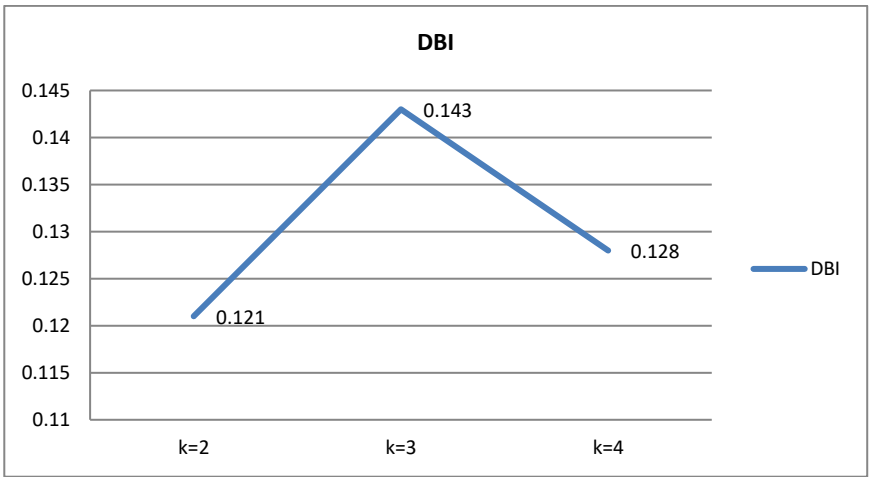


Figure 2. DBI value

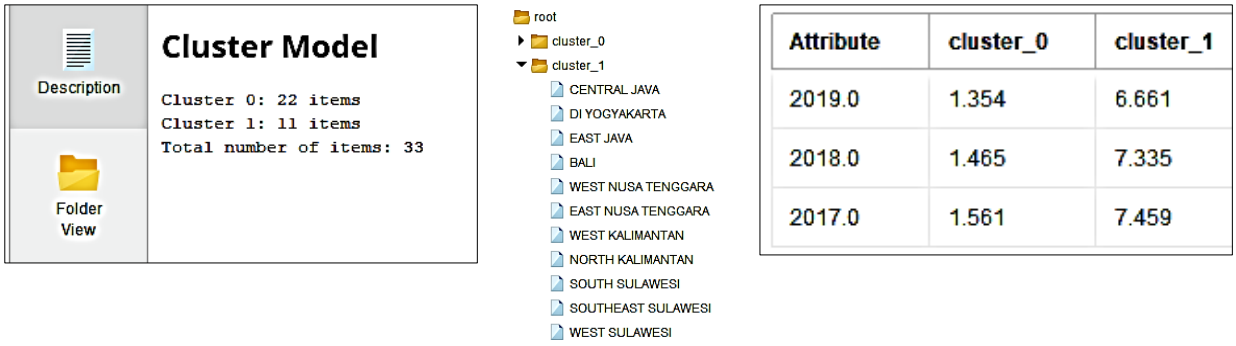
In the test results, the optimal number of clusters is 2 (k = 2), which yields a DBI value of 0.121.

3.1. Results of the Clustering and Classification Analysis

Clustering was conducted with k = 2 and the following parameters:

a) Measure type: MixedMeasure;
b) Mixed measure: MixedEuclideanDistance.

The obtained results are presented in the following figure:



Attribute	cluster_0	cluster_1
2019.0	1.354	6.661
2018.0	1.465	7.335
2017.0	1.561	7.459

Figure 3. Clustering results

As shown in Figure 3, the results of mapping the numbers of clusters for analphabets in Indonesia (k = 2) demonstrate that the high cluster consists of 11 provinces (cluster 1) and that the low cluster includes 22 provinces (cluster 0). The high and low clusters are determined by examining the final results (Figure 3). Figure 4 shows the mapping diagram of the clustering results.

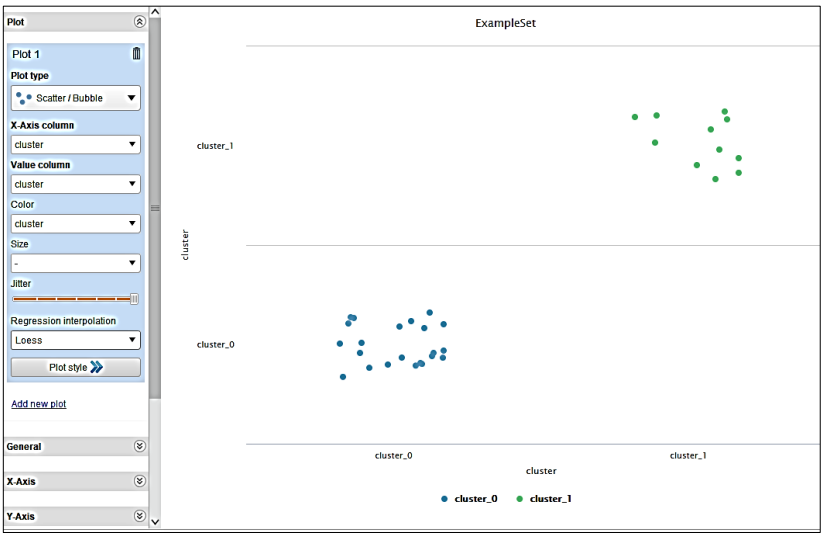
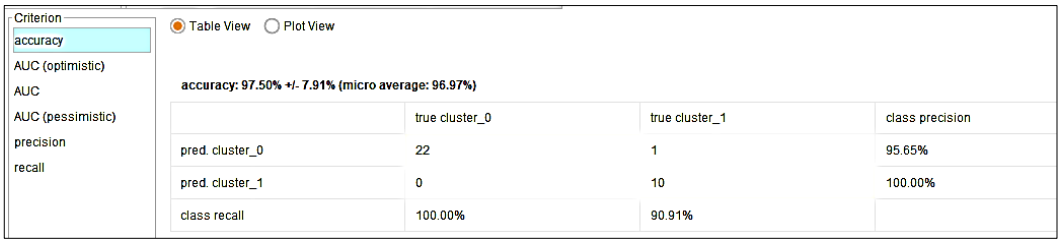
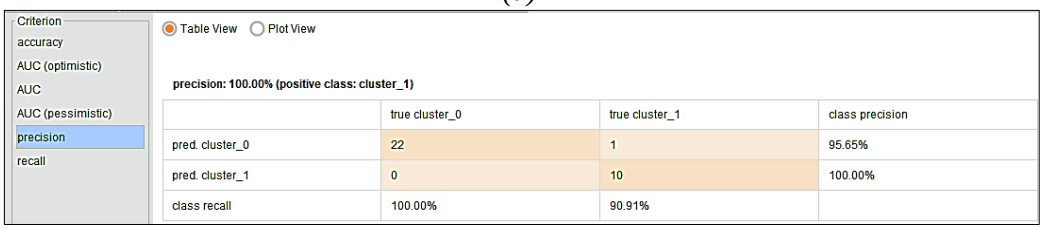


Figure 4. Graph of the cluster mapping results

The results from Figure 4 are evaluated using the C4.5 method to determine the precision, accuracy, recall and AUC. The test results with the classification are presented in the following figure.



(a)



(b)

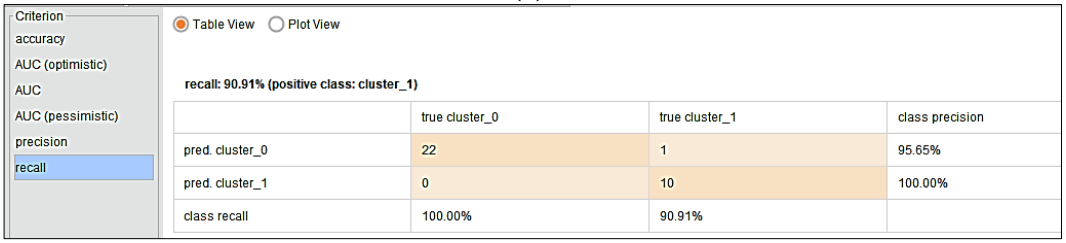


Figure 5. Accuracy, recall and precision values (a)(b)(c)

As presented in Figure 5, the obtained accuracy value is 97.0%, the recall is 90.91% and the precision is 100%. The AUC value is used to measure the discrimination performance based on the probability of the sample results being selected randomly from a positive or negative population.

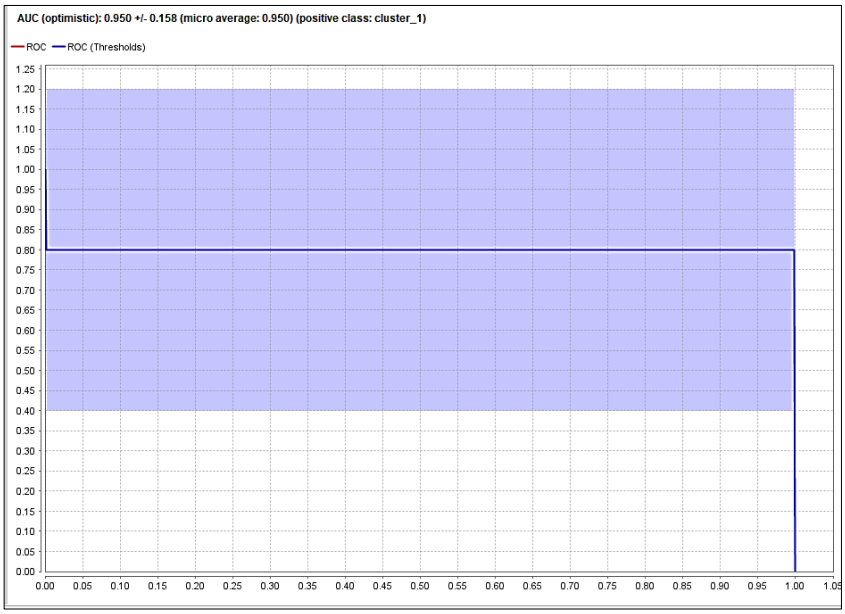


Figure 6. AUC (Area under the ROC curve)

The best AUC value in Figure 6 was 0.95; hence, the results were categorized as "excellent classification" results (Table 1).

3.2. Discussion

The analysis that was conducted using a combination of clustering and classification methods for analphabets in Indonesia yielded excellent clustering results through the combination of a series of clustering results (Figure 6).

4. Conclusions

According to results of this study and the related discussions, for mapping illiterate children in Indonesia, a combination of clustering and classifying methods (k-means and C4.5) has satisfied objectives such as the satisfactory formation of clusters based on the optimum number of clusters that was determined using the Davies Bouldin Index (DBI). The clustering results were evaluated via classification using cross-validation (k-folds = 10). According to the AUC (area under the ROC curve) test, the clustering results belong to the "Excellent Classification" category.

References

[1] V. Jessica, A. Halis, D. W. Ningsi, G. F. Virginia, and . Syahidah, "Pemberantasan Buta Aksara untuk Peningkatan Kualitas Sumber Daya Manusia Masyarakat Sekitar Hutan Desa Manipi, Kecamatan Pana, Kabupaten Mamasa," *Agrokreatif J. Ilm. Pengabdi. Kpd. Masy.*, vol. 3, no. 2, p. 136, 2017.

[2] Mariyono, "Strategi Pemberantasan Buta Aksara Melalui Penggunaan Teknik Metastasis Berbasis Keluarga," *Pancaran*, vol. 5, no. 1, pp. 55–66, 2016.

[3] R. Anisykurillah, "Evaluasi Pembangunan Pendidikan KEAKSARAAN (Studi pada Program Pendidikan Non-Formal di Kota Malang)," *J. Kebijak. Pembang.*, vol. 15, no. 1, pp. 25–36, 2020.

[4] Syamsiah, Hidayah Quraisy, and R. Babo, "Pemberdayaan Masyarakat Desa Yang Buta Huruf," *J. Equilib. Pendidik. Sosiol.*, vol. 3, no. 2, pp. 213–222, 2016.

[5] S. R. Ningsih, I. S. Damanik, A. P. Windarto, and H. Satria, "Analisis K-Medoids Dalam Pengelompokkan Penduduk Buta Huruf Menurut Provinsi," no. September, pp. 721–730, 2019.

[6] T. Chakraborty, "EC3: Combining clustering and classification for ensemble learning," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, vol. 2017-November, no. 9, pp. 781–786, 2017.

[7] P. Arumugam and V. Christy, "Analysis of Clustering and Classification Methods for Actionable Knowledge," *Mater. Today Proc.*, vol. 5, no. 1, pp. 1839–1845, 2018.

[8] A. Jurek, Y. Bi, S. Wu, and C. Nugent, "Classification by cluster analysis: A new meta-learning based approach," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6713 LNCS, pp. 259–268, 2011.

[9] U. Agrawal *et al.*, “Combining clustering and classification ensembles: A novel pipeline to identify breast cancer profiles,” *Artif. Intell. Med.*, vol. 97, pp. 27–37, 2019.
[10] S. J. Rahmathulla, “A Combined Clustering and Classification Approach for Predicting Placement Chance,” vol. 4, no. 27, pp. 1–4, 2016.
[11] A. C. Rumahorbo and K. A. Sekarwati, “Penerapan Data Mining dengan Menggunakan C4.5 pada Klasifikasi Fasilitas Kesehatan Provinsi di Indonesia,” *J. Ilm. KOMPUTASI*, vol. 19, no. 1, pp. 27–38, 2020.
[12] W. A. Tol, I. H. Komproe, M. J. D. Jordans, D. Susanty, and J. T. V. M. De Jong, “Developing a function impairment measure for children affected by political violence: A mixed methods approach in Indonesia,” *Int. J. Qual. Heal. Care*, vol. 23, no. 4, pp. 375–383, 2011.
[13] M. Li, D. Xu, D. Zhang, and J. Zou, “The seeding algorithms for spherical k-means clustering,” *J. Glob. Optim.*, vol. 76, no. 4, pp. 695–708, 2020.
[14] Z. kai Feng, W. jing Niu, R. Zhang, S. Wang, and C. tian Cheng, “Operation rule derivation of hydropower reservoir by k-means clustering method and extreme learning machine based on particle swarm optimization,” *J. Hydrol.*, vol. 576, no. April, pp. 229–238, 2019.
[15] B. Supriyadi, A. P. Windarto, T. Soemartono, and Mungad, “Classification of natural disaster prone areas in Indonesia using K-means,” *Int. J. Grid Distrib. Comput.*, vol. 11, no. 8, pp. 87–98, 2018.
[16] Z. R. S. Elsi *et al.*, “Utilization of Data Mining Techniques in National Food Security during the Covid-19 Pandemic in Indonesia,” *J. Phys. Conf. Ser.*, vol. 1594, no. 1, 2020.
[17] J. Xiao, J. Lu, and X. Li, “Davies Bouldin Index based hierarchical initialization K-means,” *Intell. Data Anal.*, vol. 21, no. 6, pp. 1327–1338, 2017.
[18] R. S. Wahono, N. S. Herman, and S. Ahmad, “A comparison framework of classification models for software defect prediction,” *Adv. Sci. Lett.*, vol. 20, no. 10–12, pp. 1945–1950, 2014.