2021

Clustering School Libraries in Indonesia using C4.5 and K-Means Algorithm

Dian Wardiana Sjuchro
Universitas Padjadjaran, Indonesia

Rulinawaty Rulinawaty
Universitas Terbuka, Indonesia

Fathurrahim Fathurrahim
Sekolah Tinggi Pariwisata Mataram, Indonesia

Danu Eko Agustinova
Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

Rima Herlina S Siburian
Universitas Papua, Papua Barat, Indonesia

See next page for additional authors

Sjuchro, Dian Wardiana; Rulinawaty, Rulinawaty; Fathurrahim, Fathurrahim; Agustinova, Danu Eko; Siburian, Rima Herlina S; Cakranegara, Pandu Adi; Ardianto, Ardianto; and Rahim, Robbi, "Clustering School Libraries in Indonesia using C4.5 and K-Means Algorithm" (2021). *Library Philosophy and Practice (e-journal)*. 5432.
https://digitalcommons.unl.edu/libphilprac/5432

Authors

Dian Wardiana Sjuchro, Rulinawaty Rulinawaty, Fathurrahim Fathurrahim, Danu Eko Agustinova, Rima Herlina S Siburian, Pandu Adi Cakranegara, Ardianto Ardianto, and Robbi Rahim

Clustering School Libraries in Indonesia using C4.5 and K-Means Algorithm

Dian Wardiana Sjuchro[1], Rulinawaty[2], Fathurrahim[3], Danu Eko Agustinova[4], Rima Herlina S Siburian[5], Pandu Adi Cakranegara[6], Ardianto[7], Robbi Rahim[8]

[1]Universitas Padjadjaran, Indonesia. Email: d.wardiana@unpad.ac.id
[2]Universitas Terbuka, Indonesia. Email: ruly@ecampus.ut.ac.id
[3]Sekolah Tinggi Pariwisata Mataram, Indonesia. Email: fathurrahim1102@gmail.com
[4]Universitas Negeri Yogyakarta, Yogyakarta, Indonesia. Email: danu_eko@uny.ac.id
[5]Universitas Papua, Papua Barat, Indonesia
[6]Universitas Presiden, Bekasi, Indonesia. Email: pandu.cakranegara@president.ac.id
[7]Institut Agama Islam Negeri Manado, Manado, Indonesia. Email: ardianto@iain-manado.ac.id
[8]Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia. Email: usurobbi85@zoho.com

Corresponding Email: usurobbi85@zoho.com

Abstract. This research is to analyze the data mining by analyzing clusters formed. This study was about determining the number of libraries available at the primary and junior high school levels in Indonesia. The main role of libraries is in the current implementation of quality education. The data are from the central statistics agency which contains 34 records indicating the number of libraries at the primary and junior high school levels. The proposed method is a combination of the C4.5 and k-means methods. clustering is performed based on the number of clusters (k). Thus, the results will be classified according to the C4.5 method. Using the number k = 2, the clusters obtained have average centroid values. (10786.75 and 3210.25). Meanwhile, there are thirty provinces in the low cluster (K-2) with centroid values . (1744.2 and 626.9). The results of the cluster group formed using the standard parameters (criterion = gain ratio; maximal depth = 20; confidence = 0.25 and minimum gain = 0.1) was accurately classified 97.5%. A combination of the C4.5 and k-means methods should be used.

Keywords: C4.5 Algorithm, k-means Algortihms, Indonesia Libraries, Clustering Library

1. INTRODUCTION

One of the most important learning resources is the library (Septiya Anggrairi et al. 2019). Implementation of libraries is also an aspect for schools and teachers to get better in teaching and learning process (Muslim 2011). One of the effort to improve the quality of life and competitiveness of society is through education as learning resources such as libraries in schools (Mann 2001). The function of school libraries are stated to be promoting of reading, training and schooling of the students (Shrestha 2008). According to this proposition, it is necessary to conduct a study and analysis of the number of libraries in Indonesia, especially at the primary and junior high schools.

There are several different machine learning algorithms that can perform this task. Few of them is data mining(Patil, Wadhai, and Gokhale 2010). Data mining is one of the supervised learning techniques that is able to extract and predict useful information from various databases(Bramer 2007). Data mining is needed because there is a large amount of useful data that can be used in banking, medicine, education and other fields. The purpose of this study is to combine classification and clustering methods to describe the relationship between the number of libraries in Indonesia.

These statistics are used to measure accuracy in clustering algorithm. A large number of studies uses both classification and clustering methods for research. This paper suggests two data mining techniques, k-means and C4.5. Cluster mapping of regions in Indonesia by provinces can be carried out using the k-means method. By using the C4.5 algorithm, the rules in the form of a decision tree have been shown. The survey studies found out that nine provinces are in high risk zone (C1 = red zone), three provinces are in alert zone (C2 = yellow zone) and twenty-two provinces are in low risk zone (C3 = green zone). On the basis of information in C4.5, the value obtained is 9236.85 for the tall cluster (C1 = red zone) based on C4.5. Next was performed by T. Ensemble Learning (EC3) (Chakraborty 2017). This essay proposes a new algorithm called EC3 for classification and grouping. EC3 combines several classifications and grouping techniques using optimization functions. This research also presents the variant of EC3, the iEC3. It seems that the approach performs significantly better than other methods. Besides, our method is faster (than the best baseline method) and more resistant to class noise and imbalance than the best baseline method. Integrating the classification algorithm with cluster algorithms allows the model to have better accuracy than just a mere clustering algorithm (Jurek et al. 2011; Agrawal et al. 2019). Integration of clustering with classification is carried out to counteract errors in calcifying existing classes (Arumugam and Christy 2018).

2. METHODOLOGY

Data Mining

Data Mining can be used to predict trends and new results of the old data. There are numerous examples of data mining tools in which classification, clustering, association and estimation are used (Ali, Ghoneim, and Saleh 2017; Bisandu, Prasad, and Liman 2019). Clustering techniques are often used in formative data processing work. There are a variety of clustering methods that have been used by past researchers; K-Means, Improved K-Means, Partitioning around Medoids, K-Medoids, Fuzzy C-Means, etc (Hossain et al. 2019).

Data mining refers to discovering patterns in large data sets that are not previously known to the algorithms which are being used. Ordinary data mining uses very large amounts of data. Usually big data use helps make research results more believable and data mining can be useful for making critically important decisions, especially in strategy.

Clustering Method

Clustering is a data mining technique to group objects. Some clustering needs include scalability, the ability to handle a variety of attributes, the ability to handle high dimensionality, handling data that has noise, and can be easily translated (Koren et al. 2019).

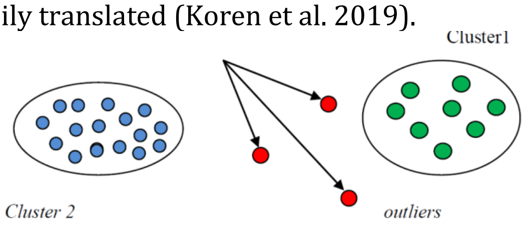


Figure 1. Clustering Example

The purpose of this data clustering is the objective function set in the clustering process, in which variations within a cluster are typically minimized and variations between clusters are maximized.

Classification Method

Classification is one job to sort things into different categories (Jumadi Dehotman Sitompul, Salim Sitompul, and Sihombing 2019; Ma, Gong, and Jiao 2011; Koren et al. 2019). In the Classifier, there are two main jobs which are (1) to construct a prototype model to be remembered and (2) to apply the prototype in performance of recognition, or classification, or prediction operation on another data objects so that it is known which class the object belongs to (Javadi et al. 2017).

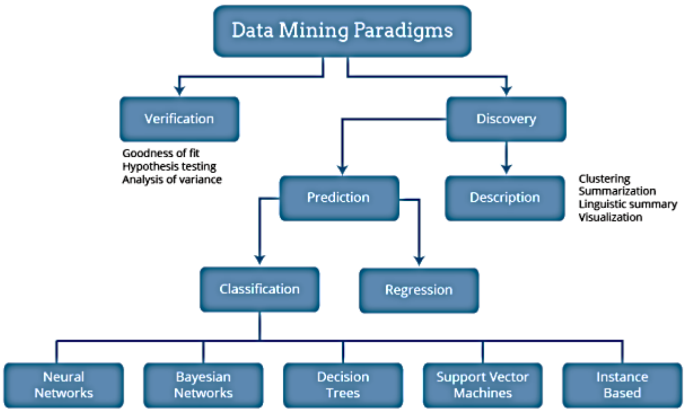


Figure 2. Data Mining Classification

Data

The data was sourced from the Ministry of Education and Culture, which was processed by the Central Statistics Agency (BPS). This data describes the number of public libraries in Indonesia by levels of elementary school and junior high school. The data used is the library per province, and the educational level per course for the 2018/2019 academic year. The following data is used in mapping the number of libraries in Indonesia.

Table 1. Research data

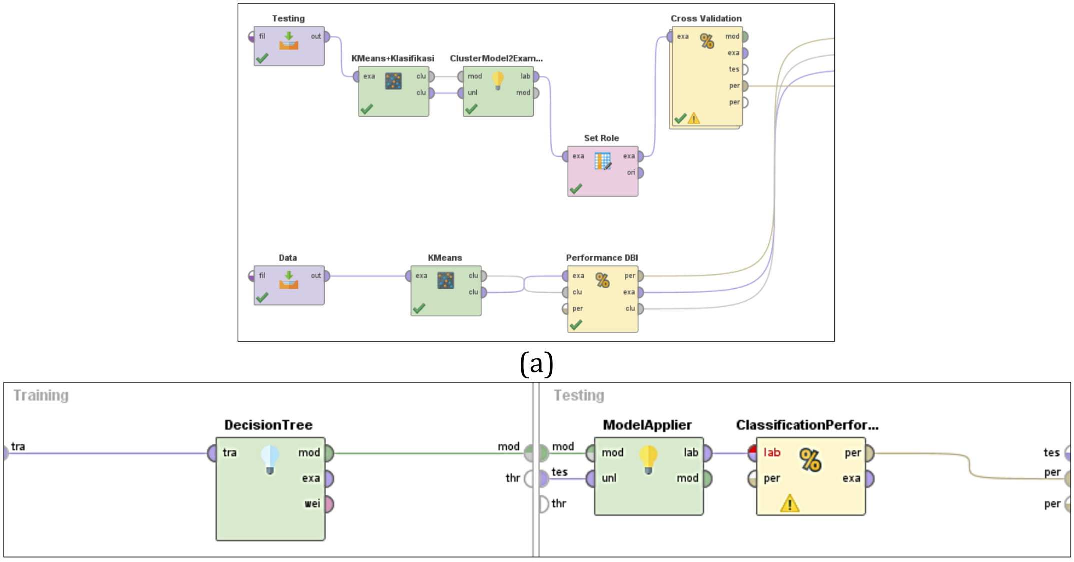
Province	Primary School	Junior High
Aceh	2765	957
North Sumatra	5898	2072
West Sumatra	2886	712
Riau	2162	861
Jambi	1705	541
South Sumatra	318	1047
Bengkulu	1057	357
Lampung	3000	1091
Kep Bangka Belitung	798	197
Kep Riau	688	270
DKI Jakarta	2003	1060
West Java	10958	3999
Central Java	13649	3107
DI Yogyakarta	1647	449
East Java	12642	3663
Banten	2612	1122
Bali	2033	392
West Nusa Tenggara	2347	663
East Nusa Tenggara	3476	1185
West Kalimantan	2951	1022
Central Kalimantan	1588	595
South Borneo	2049	572
East Kalimantan	1238	516
North Kalimantan	279	128
North Sulawesi	1500	634
Central Sulawesi	1864	588
South Sulawesi	5066	1349
Southeast Sulawesi	1683	630
Gorontalo	795	250
West Sulawesi	801	268

Maluku	1058	423
North Maluku	762	284
West Papua	415	220
Papua	780	425

Source: BPS

3. RESULTS AND DISCUSSION

The clustering and classification integration process uses the RapidMiner software. The cluster method used is k-means. While the method of classification used is the tree of deception (C4.5). Using Rapid Miner software for analysis, the model design is made using 2 cluster labels in the k-means method. The process of determining cluster labels is carried out by looking at the optimum value of the Davies Bouldin Index (DBI) parameter. While the results of the number of clusters formed are further classified with C4.5, the accuracy value as shown below is as follows:

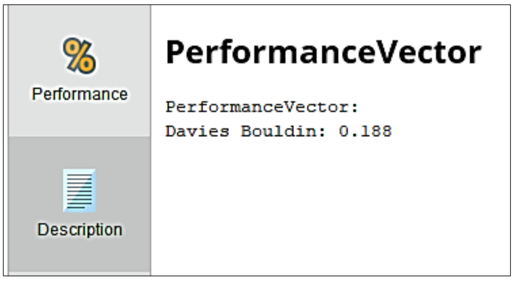


(a)

(b)

Figure 3. k-means and C4.5 combination model with Rapid Miner

At the same time, the stage shown in Figure 3 (a) applies methods that use excel formulas for inputting data. Determining the number of clusters during cluster process is achieved by using "Performance Cluster Distance" operator. This operator has a very practical solution that makes economic sense. The DBI method is used to locate optimal number of clusters. It takes the value of each point by the sum of the components of the fractal dimension divided by the distance between the two cluster center points. The smaller DBI indicated the best number of clusters. Therefore DBI values can form the basis for clustered models. The cluster analysis will be conducted using the C4.5 method. C4.5 uses different parameters (criterion = gain ratio; maximal depth = 20; confidence = 0.25 and minimum gain = 0.1). In order to measure the classification performance using the "Performance Binominal Classification" operator, we will use the "Accuracy", "Precision", "Recall" and "Area Under the Curve" parameters. From the series of trials, the maximum DBI result was the following.

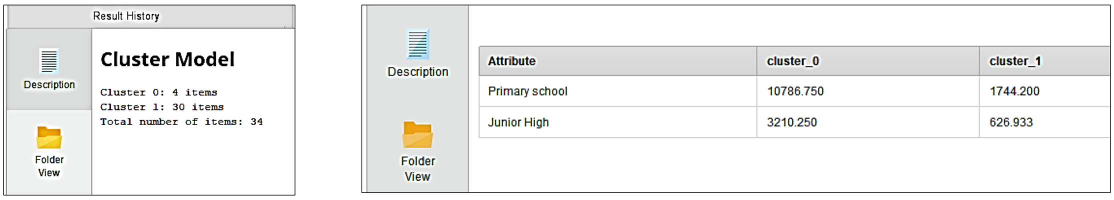


PerformanceVector

```
PerformanceVector:
Davies Bouldin: 0.188
```

Figure 4. Optimal DBI value = 0.188

The k-means approach was used to reach the optimal decision-making possibilities in Figure 4. The descriptive statistics used were k = 2, max runs = 10, measures of type = Bregman Divergences, and type = Mahalanobis Distance. The clusters formed by using attributes are as shown in the above figure:



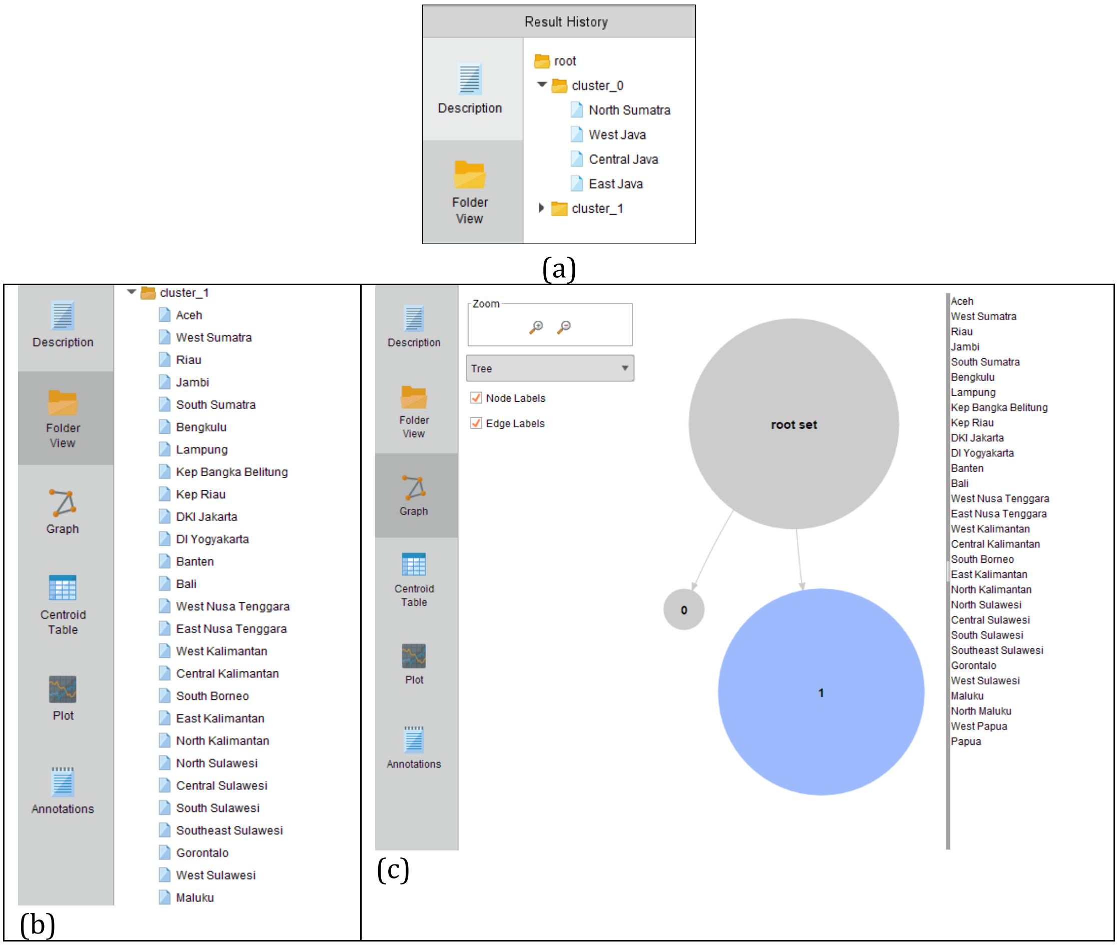
Cluster Model

```
Cluster 0: 4 items
Cluster 1: 30 items
Total number of items: 34
```

Attribute	cluster_0	cluster_1
Primary school	10786.750	1744.200
Junior High	3210.250	626.933

Figure 5. Cluster Results and Final Centroid Value

Figure 5 shows the results of the cluster formed by the number of libraries in Indonesia at the primary and junior high school levels. Out of 34 records, we've got 4 items in cluster 0 and 30 in cluster 1. Cluster 0 is a high cluster (K-1) as seen from the centroid value, namely: elementary school attributes = 10786,750 and junior high school attributes = 3210,250. While Cluster 1 is the low cluster (K-2) as seen from the centroid value, namely: elementary school attributes = 1744.2 and junior high school attributes = 626.933. The following are details of the items for each cluster as shown in the following image:



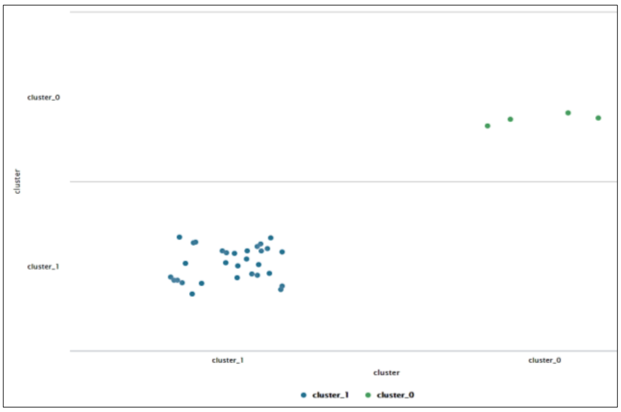
(a)
- root
  - cluster_0
    - North Sumatra
    - West Java
    - Central Java
    - East Java
  - cluster_1

(b)
- cluster_1
  - Aceh
  - West Sumatra
  - Riau
  - Jambi
  - South Sumatra
  - Bengkulu
  - Lampung
  - Kep Bangka Belitung
  - Kep Riau
  - DKI Jakarta
  - DI Yogyakarta
  - Banten
  - Bali
  - West Nusa Tenggara
  - East Nusa Tenggara
  - West Kalimantan
  - Central Kalimantan
  - South Borneo
  - East Kalimantan
  - North Kalimantan
  - North Sulawesi
  - Central Sulawesi
  - South Sulawesi
  - Southeast Sulawesi
  - Gorontalo
  - West Sulawesi
  - Maluku

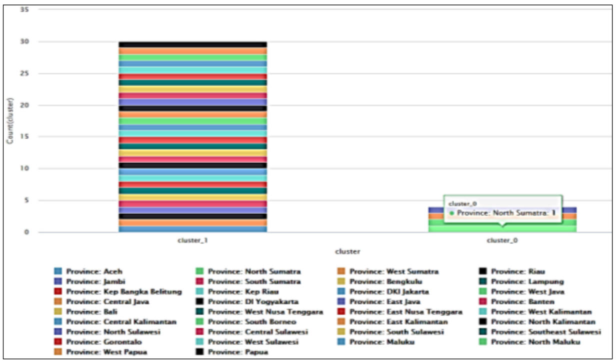
(c)
Aceh
West Sumatra
Riau
Jambi
South Sumatra
Bengkulu
Lampung
Kep Bangka Belitung
Kep Riau
DKI Jakarta
DI Yogyakarta
Banten
Bali
West Nusa Tenggara
East Nusa Tenggara
West Kalimantan
Central Kalimantan
South Borneo
East Kalimantan
North Kalimantan
North Sulawesi
Central Sulawesi
South Sulawesi
Southeast Sulawesi
Gorontalo
West Sulawesi
Maluku
North Maluku
West Papua
Papua

Figure 6. Results of the cluster with k-means method

The mapping results shown in Figure 6 are clusters formed using the value of k = 2, where only four provinces are in a high cluster of libraries, namely North Sumatra, West Java, Central Java and East Java. The following visualization is shown in the form of a bar and a scatter as shown in the image below.



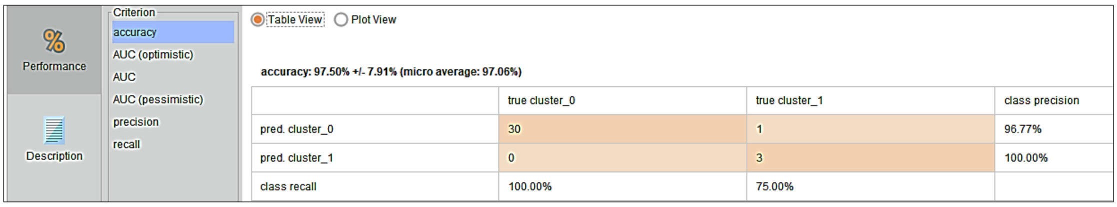
(a)



(b)

Figure 7. Graph mapping visualization with bars and scatter

The results of the mapping of clusters are shown in Figure 7, then classified using the C4.5 (decision tree) method to see the accuracy value generated from existing clusters. The following results are shown using the Rapid Miner Software.



accuracy: 97.50% +/- 7.91% (micro average: 97.06%)

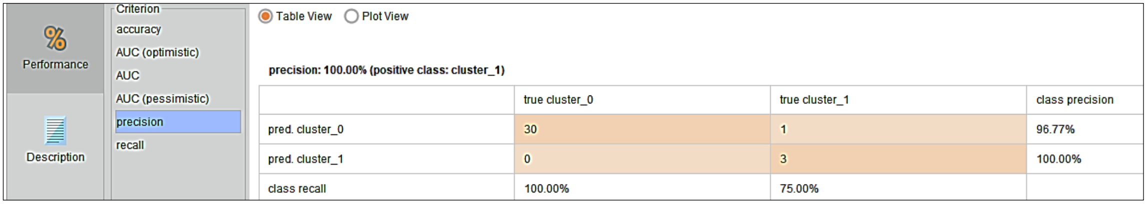
	true cluster_0	true cluster_1	class precision
pred. cluster_0	30	1	96.77%
pred. cluster_1	0	3	100.00%
class recall	100.00%	75.00%	

Figure 8. The resulting accuracy value

The accuracy is the ratio of the true positives to the overall data. Accuracy in answering the question "What percentage of the correct cluster is predicted (cluster 0 and cluster 1) of the total data". The formula used:

$$\text{accuracy} = (TP + TN) / (TP+FP+FN+TN) \qquad (1)$$

So that the accuracy = (30+3) / (30+1+0+3) = 33/34= 97%.



precision: 100.00% (positive class: cluster_1)

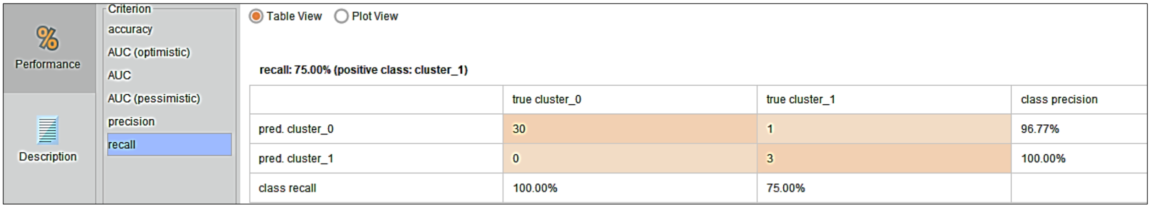
	true cluster_0	true cluster_1	class precision
pred. cluster_0	30	1	96.77%
pred. cluster_1	0	3	100.00%
class recall	100.00%	75.00%	

Figure 9. The precision values that are formed

Precision is having the correct proportion of positive predictions. Precision is the ratio of positive truth to the overall positive prediction. Precision answers the question "What percentage of clusters are correct (cluster 0 of the total cluster predicted cluster 1)?" The formula for this question is:

$$\text{Precision} = (TP) / (TP + FN) \qquad (2)$$

So that the precision = 30/(30+0) = 30/30 =100%.



● Table View ○ Plot View

recall: 75.00% (positive class: cluster_1)

	true cluster_0	true cluster_1	class precision
pred. cluster_0	30	1	96.77%
pred. cluster_1	0	3	100.00%
class recall	100.00%	75.00%	

Figure 10. The recall value formed

Recall is known as the true positive rate, or the proportion of positive cases that are seen quickly. Recall value is the ratio of predictions that are correct to the total number of predictions. Recall answers the question "What percentage of clusters is predicted cluster 0 compared to the total cluster data that is actually cluster 1". The procedure works as follows:

$$\text{Recall} = (TP) / (TP+FP) \qquad (3)$$

So that the Recall = 30 / (30+1) =30/31 = 96,77%.

In Figure 8, the accuracy value is very good to use as a reference for algorithm performance if the dataset has very close numbers of false negative and false positive data (Symmetric). In the example above, the FP and FN values are close, so that accuracy can be used as a reference for performance measurement.

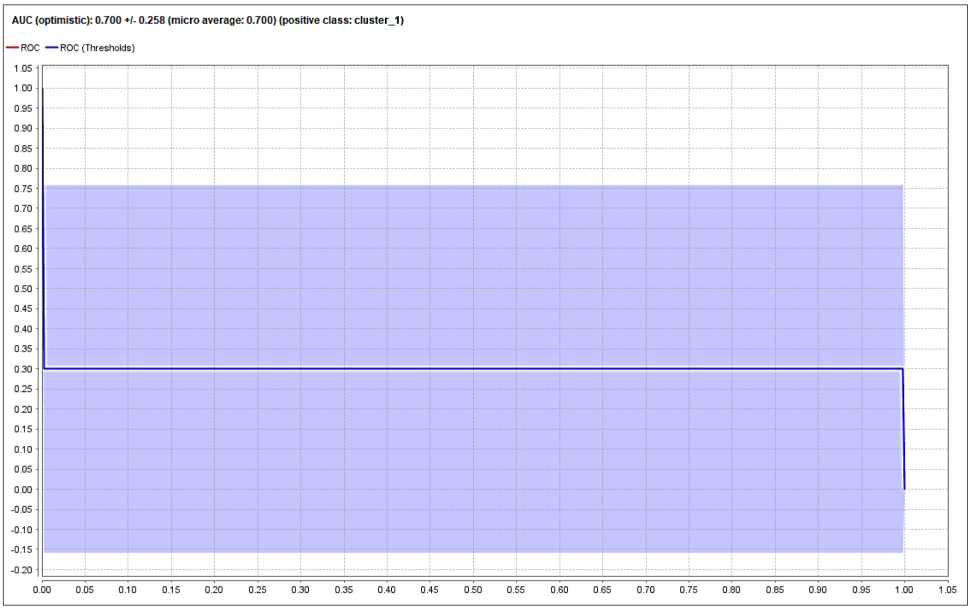


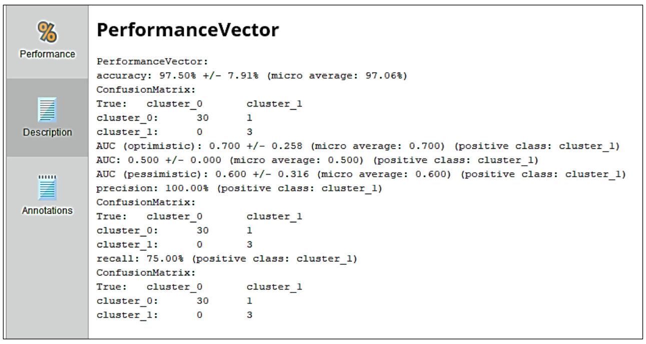
Figure 11. ROC Curve

The ROC curve is divided into two parameters, with Y axis being TP level and X axis being FP level. Such statistical models (the Red Queenian fitness model) that is interpreted as probability [9]. The AUC is used to measure discriminatory performance by estimating the probability of the output of a randomly selected sample from a positive or negative population. The value of AUC always falls between 0 and 1. Table 2 is a guide to the accuracy of classification using the AUC.

Table 2. AUC classification

Performance	classification
0,90 – 1,00	The best
0,80 – 0,90	Good
0,70 – 0,80	Fair or equal
0,60 – 0,70	Low
0,50 – 0,60	Failed

In Figure 11, by using the ROC curve through the RapidMiner software, the AUC test results reach a maximum result of 0.7 (Fair). Following are the complete results of the performance (Confusion Matrix) as shown in the following figure 12:



PerformanceVector

```
PerformanceVector:
accuracy: 97.50% +/- 7.91% (micro average: 97.06%)
ConfusionMatrix:
True:   cluster_0       cluster_1
cluster_0:      30      1
cluster_1:      0       3
AUC (optimistic): 0.700 +/- 0.258 (micro average: 0.700) (positive class: cluster_1)
AUC: 0.500 +/- 0.000 (micro average: 0.500) (positive class: cluster_1)
AUC (pessimistic): 0.600 +/- 0.316 (micro average: 0.600) (positive class: cluster_1)
precision: 100.00% (positive class: cluster_1)
ConfusionMatrix:
True:   cluster_0       cluster_1
cluster_0:      30      1
cluster_1:      0       3
recall: 75.00% (positive class: cluster_1)
ConfusionMatrix:
True:   cluster_0       cluster_1
cluster_0:      30      1
cluster_1:      0       3
```

Figure 12. Performa Vector

Overall, the number of libraries in Indonesia at the primary and the junior secondary level can be combined with the clustering. The feasibility of the number of libraries is still very small from the cluster that was created and the cluster's accuracy was determined, namely 97.5%. Currently, Java island appears to be there. In the meantime, funding for libraries in all levels of education also needs the support of government.

4. CONCLUSION

According to research, incorporation of clustering and classification techniques yielded better results in the numbers of public schools in Indonesia. Total number of books in primary and junior secondary school library. Grouping and classifying algorithms has an accuracy of 97.5 percent; precision is 100 percent and recall is 75 percent. The k-means method uses the optimal number of clusters by examining the Davies-Bouldin Index (DBI) value is 0.188.

REFERENCES

Agrawal, Utkarsh, Daniele Soria, Christian Wagner, Jonathan Garibaldi, Ian O. Ellis, John M.S. Bartlett, David Cameron, Emad A. Rakha, and Andrew R. Green. 2019. “Combining Clustering and Classification Ensembles: A Novel Pipeline to Identify Breast Cancer Profiles.” *Artificial Intelligence in Medicine* 97: 27–37. https://doi.org/10.1016/j.artmed.2019.05.002.

Ali, Doaa S., Ayman Ghoneim, and Mohamed Saleh. 2017. “Data Clustering Method Based on Mixed Similarity Measures.” *ICORES 2017 - Proceedings of the 6th International Conference on Operations Research and Enterprise Systems* 2017-Janua (Icores): 192–99. https://doi.org/10.5220/0006245601920199.

Arumugam, P., and V. Christy. 2018. “Analysis of Clustering and Classification Methods for Actionable Knowledge.” *Materials Today: Proceedings* 5 (1): 1839–45. https://doi.org/10.1016/j.matpr.2017.11.283.

Bisandu, Desmond Bala, Rajesh Prasad, and Musa Muhammad Liman. 2019. “Data Clustering Using Efficient Similarity Measures.” *Journal of Statistics and Management Systems* 22 (5): 901–22. https://doi.org/10.1080/09720510.2019.1565443.

Bramer, Max. 2007. *Principles of Data Mining*. *Springer*. https://doi.org/10.1007/978-1-4471-4884-5_1.

Chakraborty, Tanmoy. 2017. “EC3: Combining Clustering and Classification for Ensemble Learning.” *Proceedings - IEEE International Conference on Data Mining, ICDM* 2017-Novem (9): 781–86. https://doi.org/10.1109/ICDM.2017.92.

Hossain, Md Zakir, Md Nasim Akhtar, R. Badlishah Ahmad, and Mostafijur Rahman. 2019. “A Dynamic K-Means Clustering for Data Mining.” *Indonesian Journal of Electrical Engineering and Computer Science* 13 (2): 521–26. https://doi.org/10.11591/ijeecs.v13.i2.pp521-526.

Javadi, S., S. M. Hashemy, K. Mohammadi, K. W.F. Howard, and A. Neshat. 2017. “Classification of Aquifer Vulnerability Using K-Means Cluster Analysis.” *Journal of Hydrology* 549: 27–37. https://doi.org/10.1016/j.jhydrol.2017.03.060.

Jumadi Dehotman Sitompul, Bernad, Opim Salim Sitompul, and Poltak Sihombing. 2019. “Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm.” In *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/1235/1/012015.

Jurek, Anna, Yaxin Bi, Shengli Wu, and Chris Nugent. 2011. “Classification by Cluster Analysis: A New Meta-Learning Based Approach.” *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6713 LNCS: 259–68. https://doi.org/10.1007/978-3-642-21557-5_28.

Koren, Oded, carina Antonia Hallin, nir Perel, and Dror Bendet. 2019. “Decision-Making Enhancement in a Big Data Environment: Application of the K-Means Algorithm to Mixed Data.” *Journal of Artificial Intelligence and Soft Computing Research* 9 (4): 293–302. https://doi.org/10.2478/jaiscr-2019-0010.

Ma, Jingjing, Maoguo Gong, and Licheng Jiao. 2011. “Evolutionary Clustering Algorithm Based on Mixed Measures.” *International Journal of Intelligent Computing and Cybernetics* 4 (4): 511–26. https://doi.org/10.1108/17563781111186770.

Mann, Thomas. 2001. “The Importance of Books, Free Access, and Libraries as Places - And the Dangerous Inadequacy of the Information Science Paradigm.” *Journal of Academic Librarianship* 27 (4): 268–81. https://doi.org/10.1016/S0099-1333(01)00214-2.

Muslim, Ahmad. 2011. “Peranan Konsumsi Dalam Perekonomian Indonesia Dan Kaitannya Dengan Ekonomi Islam.” *Al-Azhar Indonesia Seri Pranata Sosial* 1 (2): 70–82.

Patil, Dipti D., V.M. Wadhai, and J.A. Gokhale. 2010. “Evaluation of Decision Tree Pruning Algorithms for Complexity and Classification Accuracy.” *International Journal of Computer Applications* 11 (2): 975–8887.

Septiya Anggrairi, Egga, Ali Subagyo, Denny Oktavina Radianto, and Politeknik Perkapalan Negeri Surabaya. 2019. “Analisis Pengaruh Fasilitas Pendidikan Terhadap Tingkat Pengangguran Dan Kemiskinan Di Wilayah Indonesia Tahun 2018.” *Riset Dan Konseptual* 3 (2): 109–15.

Shrestha, Nina. 2008. “A Study on Student ’s Use of Library Resources and Self-Efficacy.” *Kathmandu: Central Department of Library and Information Science*, no. December: 133–42.