VOL. 16, NO. 17, SEPTEMBER 2021 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

IMPLEMENTATION OF THE K-MEDOIDS METHOD IN THE OLD SCHOOL EXPECTATIONS IN INDONESIA BY UTILIZING EDUCATIONAL DATA MINING

Cecep Kustandi[1], Gita Widi Bhawika[2], Eri Susanto[3], Vina N. Van Harling[4] and Anita Dewi Ekawati[5]
[1]Universitas Negeri Jakarta, Jakarta, Indonesia
[2]Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[3]UIN Sunan Kalijaga Yogyakarta, Yogyakarta, Indonesia
[4]Politeknik Saint Paul Sorong, Indonesia
[5]Universitas Muhammadiyah Prof. DR. HAMKA, Jakarta, Indonesia
E-Mail: usurobbi85@zoho.com

ABSTRACT

The purpose of this study is to analyze whether the combination of data mining methods with clustering and classification techniques can be applied to the case of mapping the average number of years of schooling in Indonesia. The data source used in the study is secondary data obtained from the Central Statistics Agency (abbreviated BPS-RI) on the average length of school by province consisting of 34 records (2015-2019). The method used is a combination of k-medoids (clustering) and C4.5 (classification) methods where k-medoids are used to map clusters. The results of the cluster will be processed with C4.5 to see the value of the cluster in the form of a decision tree. The labels used in mapping clustering are high cluster for the average length of school (C1) and low cluster for the average length of school (C2) area. The average length of schooling is one indicator for the dimension of knowledge. The three dimensions are 1) Longevity and healthy living, 2) Knowledge and 3) Decent standard of living. These three dimensions are ways in which the population can access the results of development in obtaining income, health, education, and so on, which is called the Human Development Index. The results of cluster mapping mentioned that there were 9 provinces in the low cluster (26%). The low cluster is Kep. Bangka Belitung, Central Java, East Java, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, Gorontalo, West Sulawesi and Papua. Based on the decision tree value using the C4.5 method that the low cluster has values <8,763 and> 7,730. This means that for these low clusters the average length of schooling is to junior high school.

Keywords: data mining, k-medoids method, C4.5 method, average length of schooling, Indonesia.

1. INTRODUCTION

Data mining is one of the Unsupervised Learning techniques where the expected results cannot be known by anyone [1]. The results to be displayed only depend on the value of the weight that was compiled at the beginning of the construction of the system and classifying objects that are valued similar in a particular space or area [2], [3]. In other words, data mining is a learning method that is suitable for finding or classifying a pattern of many similar objects that are not completely the same [4]. One method is k-means, k-medoids which is a data mining method that is quite popular to use both in the business, academic, or industrial world [5]-[10]. The following illustration is the Unsupervised Learning technique with clustering technique as shown in the following image:

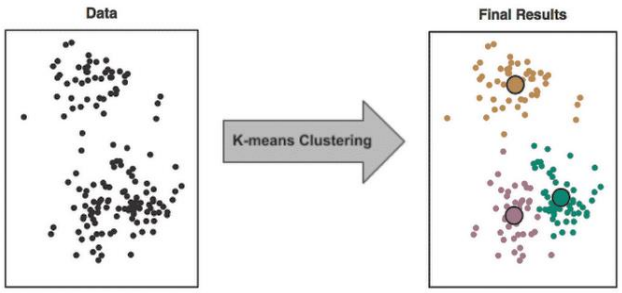

Figure-1. Unsupervised learning techniques in data mining clustering.

The average length of schooling is one indicator of the 3 parameters used to assess the human development index. These parameters are longevity and healthy living, knowledge and decent standard of living. The average length of schooling is part of the knowledge parameter. These three parameters will produce an education index (knowledge), a health index (longevity and healthy life) and an expenditure index (decent standard of living). The three indexes will determine the human development index which explains how the population can access the results of development in obtaining income, health, education, and so on [11]. Because Indonesia's national development places the people as the central point of development [5]-[10]. The following are illustrative

images of the human development index in Indonesia as shown in the following figure:

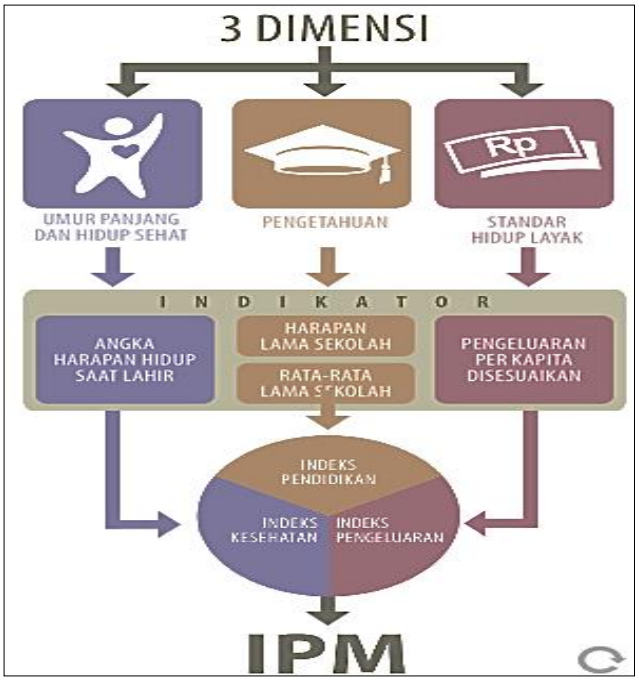

Figure-2. Human Development Index in Indonesia.

The research carried out focuses on data on the average length of schools managed by the Central Statistics Agency (abbreviated BPS-RI), where data will be processed using data mining techniques to map clusters to the average length of school in Indonesia. The technique used in this research is a combination of classification and classification methods. The clustering method used in this study is the k-medoids method which is very well known for its advantages [12]. One of the advantages of the method is the development of the k-means method [13]. Research conducted [11] on cluster analysis using the k-means method for grouping districts / cities in Maluku based on human development index indicators is a different study. These differences are found in the method used and the case studied. In this study, using Maluku Province to cluster there are areas based on the number of clusters (k) created. Whereas the research to be made is a combination of clustering and classification methods in mapping the average number of years of schooling in Indonesia. It is expected that the results of the research can increase knowledge in the field of data mining and provide information about mapping in the form of clusters to regions that have the lowest average length of school in Indonesia. Because after all the good quality of human resources is one indicator of the progress of a nation.

2. METHODOLOGY

2.1 Data Mining

Data mining is a process of discovering meaningful patterns, relationships, and new trends by filtering huge amounts of data stored in previously unknown storage [7], [9], [14]–[16]. Data mining processing consists of predictive classification, modeling, classification, and association [2], [17]–[19]. Clustering is often done as a first step in the data mining process. There are many clustering algorithms that have been used by previous researchers such as K-Means, Improved K-Means, K-Medoids (PAM), Fuzzy C-Means, DBSCAN, CLARANS and Fuzzy Substractive [20].

2.2 K-Medoids Method

The difference between the K-Medoids algorithm and the K-Means algorithm is that the K-Medoids method uses objects as a representative (medoid) cluster center for each cluster, while the K-Means method requires a mean value as the center of the cluster. In addition, the k-medoids method is more suitable for grouping data than the k-means method [21], [22].

2.3 Decision Tree Method (C4.5)

Decision trees are a well-known classification method that converts very large facts into decision trees that represent rules. Additionally decision trees are useful for exploring data, finding hidden relationships between a number of prospective variables input with a target variable [23], [24].

2.4 Data

The data used in this study is the average number of school years in 2015-2019 consisting of 34 records. The data comes from the statistical report of the Central Statistics Agency (BPS-RI) which can be accessed via the page https://www.bps.go.id. In addition the data obtained will be preprocessing data using Microsoft Excel software. Clean data will be analyzed using a combination of clustering and classification methods using the help of RapidMiner software. The following raw data and processed data as shown in the following table:

Table-1. Research data.

The province	2015	2016	2017	2018	2019
Aceh	9.32	9.36	9.42	9.46	9.59
North Sumatra	9.34	9.46	9.55	9.61	9.71
West Sumatra	8.85	8.97	9.02	9.1	9.22
Riau	8.89	8.97	9.06	9.11	9.35
Jambi	8.43	8.55	8.61	8.7	8.86
South Sumatra	8.26	8.32	8.41	8.48	8.6
Bengkulu	8.74	8.82	8.91	8.94	9.08
Lampung	8.01	8.1	8.19	8.29	8.36
Kep. Bangka Belitung	7.83	8.04	8.13	8.24	8.35
Kep. Riau	9.85	9.9	10	10	10.1
DKI Jakarta	10.9	10.9	11	11.1	11.1
West Java	8.31	8.41	8.46	8.61	8.79
Central Java	7.57	7.7	7.77	7.84	8.03
DI Yogyakarta	9.59	9.62	9.68	9.73	9.83
East Java	7.71	7.78	7.87	7.93	8.11
Banten	8.7	8.79	8.87	8.93	9.07
Bali	8.8	8.84	8.93	9	9.19
West Nusa Tenggara	7.51	7.57	7.64	7.69	7.98
East Nusa Tenggara	7.4	7.54	7.62	7.7	7.98
West Kalimantan	7.41	7.49	7.57	7.65	7.8
Central Kalimantan	8.4	8.52	8.59	8.66	8.83
South Borneo	8.14	8.28	8.37	8.45	8.59
East Kalimantan	9.52	9.55	9.62	9.63	9.88
North Kalimantan	8.67	9.01	9.1	9.18	9.24
North Sulawesi	9.19	9.31	9.4	9.51	9.63
Central Sulawesi	8.35	8.56	8.64	8.74	8.98
South Sulawesi	8.2	8.31	8.42	8.45	8.73
Southeast Sulawesi	8.74	8.86	8.93	9.03	9.25
Gorontalo	7.58	7.71	7.77	7.83	8.11
West Sulawesi	7.49	7.76	7.84	7.94	8.22
Maluku	9.54	9.69	9.74	9.78	10
North Maluku	8.81	8.96	9	9.07	9.32
West Papua	9.47	9.57	9.67	9.73	9.92
Papua	6.27	6.48	6.58	6.66	6.85

source: BPS-RI

Table-2. processed data.

The province	Average length of school
Aceh	9.43
North Sumatra	9.53
West Sumatra	9.03
Riau	9.08
Jambi	8.63
South Sumatra	8.41
Bengkulu	8.90
Lampung	8.19
Kep. Bangka Belitung	8.12
Kep. Riau	9.98
DKI Jakarta	10.99
West Java	8.52
Central Java	7.78
DI Yogyakarta	9.69
East Java	7.88
Banten	8.87
Bali	8.95
West Nusa Tenggara	7.68
East Nusa Tenggara	7.65
West Kalimantan	7.58
Central Kalimantan	8.60
South Borneo	8.37
East Kalimantan	9.64
North Kalimantan	9.04
North Sulawesi	9.41
Central Sulawesi	8.65
South Sulawesi	8.42
Southeast Sulawesi	8.96
Gorontalo	7.80
West Sulawesi	7.85
Maluku	9.76
North Maluku	9.03
West Papua	9.67
Papua	6.57

source: Processed data

3. RESULTS AND DISCUSSIONS

At this stage, the data presented in Table-1 will be processed. The first process is mapping using the k-means method. The mapping results will be classified using C4.5 to see the representation of the rules in the form of a decision tree, where the rules can be easily

understood in natural language. At the clustering (k-means) stage, the mapping labels used are 2 clusters namely high cluster (C1) and low cluster (C2). The following is the design of a combination model design (clustering and classification) using RapidMiner.

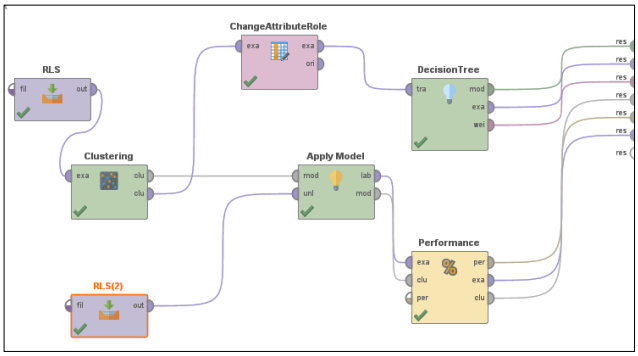

Figure-3. The RapidMiner model on average length of school (clustering and classification).

In Figure-3 the data input process is explained using the read excel tool to enter data that has been prepared as shown in Table-2. The k-medoids and C4.5 models are entered to perform their respective tasks and functions, namely mapping clusters and classifications in the form a decision tree which will produce a rule that can provide information. In addition, performance tools are used to see the strength of the cluster formed. In this study using 2 cluster labels namely high cluster (C1) and low cluster (C2) on the average number of years of schooling in Indonesia.

Cluster Model

```
Cluster 0: 9 items
Cluster 1: 25 items
Total number of items: 34
```

Figure-4. Results of clustering with k-medoids.

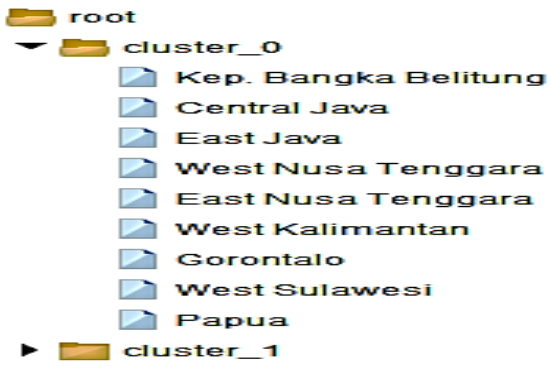

Figure-5. The low cluster (C2).

Following are the complete results of clustering that have been exported from RapidMiner to Excel as shown in Table-2 where the clusters are low (cluster_0) and cluster high (cluster_1).

Table-3. The results of the RapidMiner export file to Excel.

The province	label	Average length of school
Aceh	cluster_1	9.4
North Sumatra	cluster_1	9.5
West Sumatra	cluster_1	9.0
Riau	cluster_1	9.1
Jambi	cluster_1	8.6
South Sumatra	cluster_1	8.4
Bengkulu	cluster_1	8.9
Lampung	cluster_1	8.2
Kep. Bangka Belitung	cluster_0	8.1
Kep. Riau	cluster_1	10.0
DKI Jakarta	cluster_1	11.0
West Java	cluster_1	8.5
Central Java	cluster_0	7.8
DI Yogyakarta	cluster_1	9.7
East Java	cluster_0	7.9
Banten	cluster_1	8.9
Bali	cluster_1	9.0
West Nusa Tenggara	cluster_0	7.7
East Nusa Tenggara	cluster_0	7.6
West Kalimantan	cluster_0	7.6
Central Kalimantan	cluster_1	8.6
South Borneo	cluster_1	8.4
East Kalimantan	cluster_1	9.6
North Kalimantan	cluster_1	9.0
North Sulawesi	cluster_1	9.4
Central Sulawesi	cluster_1	8.7
South Sulawesi	cluster_1	8.4
Southeast Sulawesi	cluster_1	9.0
Gorontalo	cluster_0	7.8
West Sulawesi	cluster_0	7.9
Maluku	cluster_1	9.8
North Maluku	cluster_1	9.0
West Papua	cluster_1	9.7
Papua	cluster_0	6.6

In Table-3 we can explain the results of the mapping in the form of clusters in the average number of years of schooling in Indonesia where high cluster results (C1) were obtained around 73% (25 provinces) and 26% in low clusters (C2) or around nine provinces. The nine provinces are Kep. Bangka Belitung, Central Java, East Java, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, Gorontalo, West Sulawesi and Papua. Here are the final centroid values in the high cluster (cluster_1) and low cluster (cluster_0) as shown below:

Attribute	cluster_0	cluster_1
Average length of school	6.568	9.672

Figure-6. The final centroid results.

The following is a mapping image in the form of scattered plots by region on the average length of school as shown in the following figure:

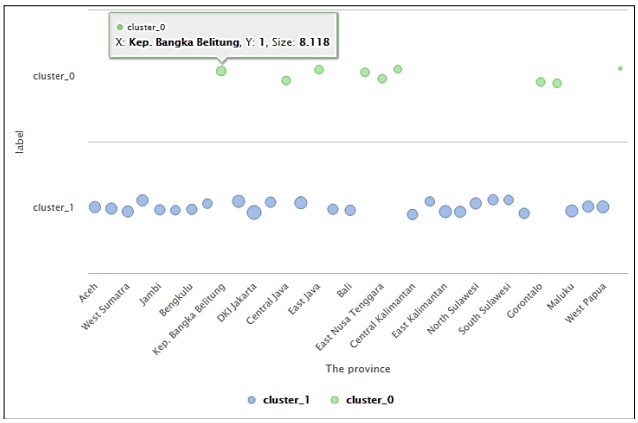

Figure-7. Visualization of clustering results with scatter plotter.

The results of the mapping in the form of a cluster on the average number of years of school, will be classified using the C4.5 method to see the value of the rules contained in the decision tree as shown in the following figure:

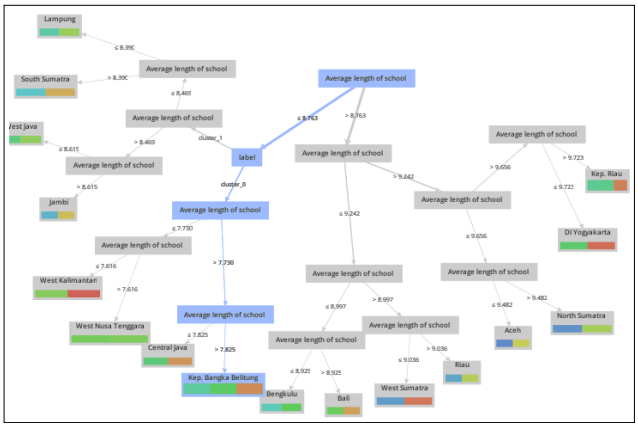

Figure-8. Decision tree results from pure participation rates.

Based on the value of decision trees using the C4.5 method that low clusters have values < 8,763 and > 7,730. This means that for this low cluster, the average length of schooling is junior high school. In the clustering results created, the validity test is used to see the relationship of the clustering by using Davies-Bouldin tools. Testing performed on the number of clusters (k=2) with a value = 0.576 as shown in the results of the image with the RapidMiner software.

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -0.873
Avg. within centroid distance_cluster_0: -1.354
Avg. within centroid distance_cluster_1: -0.699
Davies Bouldin: -0.576
```

Figure-9. Performance vector results.

4. CONCLUSIONS

Based on the results of the study can be obtained that the application of data mining can be done on the mapping and classification of the average number of years of school in Indonesia. The results state that there are nine provinces in the low cluster (C2) which means that the average length of schooling for the region is up to junior high school. Because after all the good quality of human resources is one indicator of the progress of a nation.

REFERENCES

[1] B. Supriyadi, A. P. Windarto, T. Soemartono, and Mungad. 2018. Classification of natural disaster prone areas in Indonesia using K-means. Int. J. Grid Distrib. Comput. 11(8): 87-98.

[2] A. P. Windarto et al. 2019. Analysis of the K-Means Algorithm on Clean Water Customers Based on the Province. J. Phys. Conf. Ser. 1255(1).

[3] Sudirman, A. P. Windarto and A. Wanto. 2018. Data mining tools | rapidminer: K-means method on clustering of rice crops by province as efforts to stabilize food crops in Indonesia. IOP Conf. Ser. Mater. Sci. Eng. 420: 012089.

[4] E. H. S. Atmaja. 2019. Implementation of k-Medoids Clustering Algorithm to Cluster Crime Patterns in Yogyakarta. Int. J. Appl. Sci. Smart Technol. 1(1): 33-44.

[5] K. F. Irnanda, F. N. Arifah, M. R. Raharjo, A. Arifin and A. P. Windarto. 2019. The selection of Calcium Milk Products that are appropriate for advanced age

using PROMETHEE II Algorithm. J. Phys. Conf. Ser. 1381(1).

[6] D. Hartama, A. Perdana Windarto and A. Wanto. 2019. The Application of Data Mining in Determining Patterns of Interest of High School Graduates. J. Phys. Conf. Ser. 1339(1).

[7] N. Rofiqo, A. P. Windarto and D. Hartama. 2018. Penerapan Clustering Pada Penduduk Yang Mempunyai Keluhan Kesehatan Dengan Datamining K-Means. KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer). 2(1): 216-223.

[8] W. Katrina, H. J. Damanik, F. Parhusip, D. Hartama, A. P. Windarto and A. Wanto. 2019. C.45 Classification Rules Model for Determining Students Level of Understanding of the Subject. J. Phys. Conf. Ser. 1255(012005): 1-7.

[9] H. Pratiwi *et al*. 2020. Sigmoid Activation Function in Selecting the Best Model of Artificial Neural Networks. J. Phys. Conf. Ser. 1471(1).

[10] M. Widyastuti, A. G. Fepdiani Simanjuntak, D. Hartama, A. P. Windarto, and A. Wanto. 2019.vClassification Model C.45 on Determining the Quality of Custumer Service in Bank BTN Pematangsiantar Branch. J. Phys. Conf. Ser. 1255(012002): 1-6.

[11] M. W. Talakua, Z. A. Leleury and A. W. Talluta. 2017. Analisis Cluster Dengan Menggunakan Metode Provinsi Maluku Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2014 Cluster. J. Ilmu Mat. dan Terap. 11(2): 119-128.

[12] P. Arora, Deepali and S. Varshney. 2016. Analysis of K-Means and K-Medoids Algorithm for Big Data. Phys. Procedia. 78(no. December 2015): 507-512.

[13] S. Defiyanti, M. Jajuli and N. Rohmawati. 2017. K-Medoid Algorithm in Clustering Student Scholarship Applicants. Sci. J. Informatics. 4(1): 27-33.

[14] F. Rahman, I. I. Ridho, M. Muflih, S. Pratama, M. R. Raharjo and A. P. Windarto. 2020. Application of Data Mining Technique using K-Medoids in the case of Export of Crude Petroleum Materials to the Destination Country Application of Data Mining Technique using K-Medoids in the case of Export of Crude Petroleum Materials to the Destination C.

[15] P. Alkhairi and A. P. Windarto. 2019. Penerapan K-Means Cluster Pada Daerah Potensi Pertanian Karet Produktif di Sumatera Utara. Semin. Nas. Teknol. Komput. Sains. pp. 762-767.

[16] W. M. Sari *et al*. 2020. Improving the Quality of Management with the Concept of Decision Support Systems in Determining Factors for Choosing a Cafe based on Consumers. J. Phys. Conf. Ser. 1471(1).

[17] A. P. Windarto. 2017. Implementation of Data Mining on Rice Imports by Major Country of Origin Using Algorithm Using K-Means Clustering Method. Int. J. Artif. Intell. Res. 1(2): 26-33.

[18] M. G. Sadewo, A. P. Windarto and D. Hartama. 2017. Penerapan Datamining Pada Populasi Daging Ayam Ras Pedaging Di Indonesia Berdasarkan Provinsi Menggunakan K-Means Clustering. InfoTekJar (Jurnal Nas. Inform. dan Teknol. Jaringan). 2(1): 60-67.

[19] A. P. Windarto. 2017. Penerapan Data Mining Pada Ekspor Buah-Buahan Menurut Negara Tujuan Menggunakan K-Means Clustering. Techno.COM. 16(4): 348-357.

[20] D. Marlina, N. Lina, A. Fernando and A. Ramadhan. 2018. Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak. J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf. 4(2): 64.

[21] R. W. Sari, A. Wanto and A. P. Windarto. 2018. Implementasi Rapidminer Dengan Metode K-Means (Study Kasus: Imunisasi Campak Pada Balita Berdasarkan Provinsi). KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer). 2(1): 224-230.

[22] T. M. Kodinariya and P. R. Makwana. 2013. Review on determining number of Cluster in K-Means Clustering. International Journal of Advance Research in Computer Science and Management Studies.

[23] D. R. S. P, A. P. Windarto, D. Hartama and I. S. Damanik. 2020. Penerapan klasifikasi c4.5 dalam meningkatkan sistem pembelajaran mahasiswa. KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer). 3: 593-597.

[24] A. Wanto *et al*. 2020. Data Mining: Algoritma dan Implementasi. Medan: Yayasan Kita Menulis.