Ministry of Education and Culture Accreditation,
No: 30/E/KPT/2019 (Sinta 4)
DOI: https://doi.org/10.31294/p.v24i2.1399

Paradigma,
Vol. 24, No. 2, September 2022, Page. 168-174
P-ISSN 1410-5063, E-ISSN: 2579-3500

Grouping Data in Predicting Infant Mortality Using K-Means and Decision Tree

Ridwansyah[1*], Verry Riyanto[2], Abdul Hamid[3], Sri Rahayu[4], Jajang Jaya Purnama[5]

[1,4,5]Universitas Nusa Mandiri
Jalan Jatiwaringin Raya No. 02, Cipinang Melayu, Makasar, Jakarta Timur, Indonesia

[2,3]Universitas Bina Sarana Informatika
Jl. Kramat Raya No. 98, Senen, Jakarta Pusat, Indonesia

e-mail: [1*]ridwansyah.rid@nusamandiri.ac.id

Abstrak - Tingginya angka kematian bayi merupakan hal yang utama dan harus pemerintah Indonesia prioritaskan. Banyaknya kematian bayi yang terjadi maka perlu adanya pengelompokan data kematian bayi dan memprediksi yang menyebabkan kematian bayi tersebut. Dengan adanya pengelompokan serta prediksi yang bertujuan untuk mengurangi data kematian bayi yang ada di Indonesia. Untuk pengelompokan data kematian bayi diperlukan suatu metode K-Means untuk menganalisis data dengan melakukan proses pemodelan data tanpa supervisi atau yang biasa disebut juga unsupervised learning. Jika data yang didapat dari hasil pengelompokan, maka akan di prediksi data tersebut dengan metode Decission Tree yang lebih handal dalam membuat sebuah keputusan dengan pohon keputusannya tersebut. Hasil dari K-Means dalam penentuan centroid dalam tahap awal algoritma k-means sangat berpengaruh pada hasil cluster yang dilakukan pada dataset kematian bayi dengan hasil centroid yang berbeda. Dari hasil clustering didapatkan empat label yang diuji Kembali menggunakan algoritma decision tree. Dari hasil yang didapat bahwa tingkat prediksi yang sangat bagus yang didapat. Dengan adanya metode K-Means Serta Decission Tree maka akan dipakai dan dievaluasi oleh pihak pemerintah atau bagian kesehatan guna mencegah kematian bayi yang sangat banyak.

***Kata Kunci**:* Decission Tree, K-Means, Kematian Bayi.

***Abstract** - The high infant mortality rate is the main thing and the Indonesian government must prioritize. The number of infant deaths that occur, it is necessary to group data on infant mortality and predict the cause of the infant's death. With the grouping and predictions that aim to reduce infant mortality data in Indonesia. For grouping infant mortality data, a K-Means method is needed to analyze data by carrying out a data modeling process without supervision or also known as unsupervised learning. If the data is obtained from the results of grouping, then the data will be predicted with the Decission Tree method which is more reliable in making a decision with the decision tree. The results of K-Means in determining the centroid in the early stages of the k-means algorithm greatly affect the results of clusters carried out on infant mortality datasets with different centroid results. From the clustering results, four labels were tested again using the decision tree algorithm. From the results obtained that a very good prediction rate is obtained. With the K-Means and Decission Tree methods, it will be used and evaluated by the government or the health department to prevent a lot of infant deaths.*

***Keywords**: Decission Tree, K-Means, Baby Death.*

INTRODUCTION

Death is something that we cannot avoid where, when and how death comes. Infant death is something we don't want, especially for newlywed couples or those who have been married for a long time but haven't gotten the baby they want.(Kohno et al., 2020). The high infant mortality rate is the main thing and the Indonesian government must prioritize, one of the government's efforts to reduce infant mortality is by conducting a surveillance program, namely PWS KIA where the program monitors maternal and infant health in the local area. Basically there are several infant deaths that have causes from the time of pregnancy, accidents, disasters, diseases or because it is destiny from God(Salina et al., 2019), For this reason, research is carried out in classifying infant mortality data(Junaedi et al., 2019). For grouping

Article Information	Received: August 22, 2022	Revised: September 13, 2022	Accepted: September 14, 2022

infant mortality data, a K-Means method is needed to analyze data by carrying out a data modeling process without supervision or also known as unsupervised learning.

Based on the data obtained, to find out which sub-districts and sub-districts can be grouped based on the same factors that affect the infant mortality rate in Jakarta Indonesia, in order to obtain good grouping results that can carry out more accurate and efficient handling of infant mortality, data processing is needed to determine the patterns from the data which then from the patterns obtained are taken hidden information from the data, then in the processing using the K-Means method which can analyze and classify a partition based on N objects with observations into groups of objects where the object group has the closest mean (Aditya et al., 2018) and performs data grouping with a partition system (Santiko et al., 2018). Data that has been grouped based on the same clusters and have the same characteristics are grouped into one cluster and clusters that have different characteristics are grouped into other clusters that have the same cluster in that cluster.(Suniantara et al., 2020). From the grouping data, it can be done by predicting infant mortality data using the Decisiion Tree algorithm (Arifin & Herliana, 2020)(Charbuty & Abdulazeez, 2021) which produces a decision tree that is flexible enough so that it is good at making decisions. (Syamsu et al., 2019).

This research focuses on data on infant mortality in the province of DKI Jakarta Indonesia in 2018 using the K-Means method which is then carried out using the Decisiion Tree model. Previously, in research on factors related to infant mortality, the quantitative nature of which used the cross sectional method and resulted in parental work and the cost of living a healthy life had an effect on infant mortality (Lengkong, G.T., Langi, F.L.F.G and Posangi, 2020). And distance is very influential on the death of the baby (Fitri et al., 2017). Neonatal deaths that occur with infants using the C.45 algorithm method have been successfully carried out and can support the results of the risk analysis on infant mortality (Junaedi et al., 2019). Existing research on the K-Means clustering method has not used existing data from data.go.id about infant mortality and there are no predictions in infant mortality. The novelty and contribution of the research is in the dataset used. The purpose of this study is that the community or government can reduce the infant mortality rate in the DKI Jakarta province of Indonesia (Wulan Sari et al., 2018).

RESEARCH METHODOLOGY

Metode non hierarchical Cluster(Oktavia et al., 2020) the way of working starts from determining the desired number of clusters, namely as many as four clusters in this study. After determining the number of clusters, determine the clusters in the infant mortality data without following the hierarchical process. The methodology in this study uses sample data taken from the global dataset of infant mortality data on data.go.id in 2018.

$$\mathrm{DL}_2(x_2, x_1) = ||x_2 - x_1||2 \quad \text{.....(1)}$$

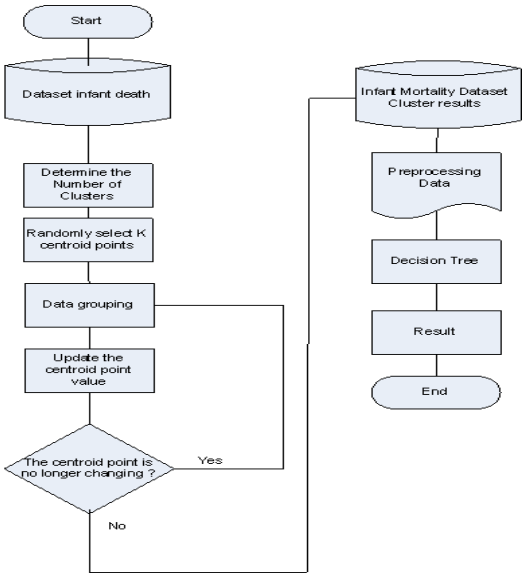
From the results of the cluster data will be used to predict infant mortality using a decision tree algorithm. The Decision Tree algorithm is one of the algorithms in the data mining process that makes a decision tree by having attributes and making it a root node and will make each branch for each value have the same class. Here's the formula for determining Gain and Entropy.

$$\mathrm{Gain}\ (\mathrm{S.A}) = \frac{Gain\ (S.A)}{Si\ (S.A)} \quad \text{.....(2)}$$

$$\mathrm{Si} = -\sum_i^n - pi \log 2pi \quad \text{.....(3)}$$

$$\mathrm{Entropy(S)} - \sin \sum \mathrm{Si/slog2\ Si/s\ c\ i} = 1 \quad \text{.....(4)}$$

The following is a model algorithm in infant mortality research using K-Means and Decisiion Tree which can be seen in figure 1.



Source: (Ridwansyah et al., 2022)

Figure 1. Research Steps

a. The collection of infant mortality datasets carried out by the DKI Jakarta provincial government and uploaded to data.go.id which can be downloaded for free which is then processed using datamining using k-means and decision trees.
b. The next stage is to determine the number of clusters from the infant mortality dataset, the determination of clusters does not specify how

many there are.

c. The number of selected clusters is 4 clusters which are randomly selected points from the infant mortality dataset which will later be used for grouping the data.
d. This data grouping from the data will be divided into four parts, data cluster1, data cluster2, data cluster3, data cluster4.
e. Update the cendroid point value until the value doesn't change anymore.
f. Repeat Steps d and e until the value of the centroid point does not change anymore.
g. The cluster results from the k-means test on the infant mortality dataset were again tested using the decision tree algorithm with the result labels from the cluster.
h. The infant mortality dataset is preprocessed first, such as normalizing etc
i. Testing the infant mortality dataset using a decision tree algorithm with the aim of getting high accuracy and AUC results to predict infant mortality.

RESULTS AND DISCUSSION

1. K-Means

The data processed in this study is data taken from the global dataset of infant mortality data in 2018. The infant mortality dataset consists of attributes of year, name of city, name of sub-district, name of village, gender and number. The data to be tested and grouped consists of 403 infant mortality data.

Tabel 1. Sample Data on infant mortality

Year	City	Districts	Ward	Sex	Number of Deaths
2018	1	1	1	1	1
2018	1	1	1	2	2
2018	1	1	2	1	2
2018	1	1	2	2	2
2018	1	1	2	2	5
2018	1	1	2	1	8
2018	1	2	3	2	1
2018	1	2	2	1	3
2018	1	2	3	2	3
2018	1	2	3	2	3
2018	1	2	3	1	3
2018	1	2	3	1	4
2018	1	2	2	2	5
2018	1	2	3	1	6
2018	1	2	3	1	8
2018	1	2	3	2	9

Source: (Ridwansyah et al., 2022)

From the data, it was tested and grouped using the K-Means method by determining the number of 4 clusters with the clusters being randomly selected which can be seen in table 2.

Table 2. Initial Cluster of infant mortality data

Group Cluster	Year	City	Districts	Ward	Sex	Number of Deaths
C1	2018	1	2	3	2	1
C2	2018	1	3	3	2	1
C3	2018	2	1	1	2	6
C4	2018	2	2	2	1	1

Source: (Ridwansyah et al., 2022)

From the initial cluster data, the distance between the data objects and the centroid is calculated by calculating the distance using L2 (Euclidean) distance space, from calculating the distance between two points it is calculated and can be seen in table 3.

Table 3. Sample Data The first distance to the center of the cluster in iteration 1 of infant mortality data

data to	Year	City	Districts	Ward	sex	Number of Deaths	Result of Cluster 1	Result of Cluster 2	Result of Cluster 3	Result of Cluster 4	Group Cluster
1	2018	1	1	1	1	1	7.071068	3	4.123106	2	4
2	2018	1	1	1	2	2	2.44949	3	4.123106	2.236068	4
3	2018	1	1	2	1	2	1.414214	2.44949	4.123106	2	1
4	2018	1	1	2	2	2	1.732051	2.44949	5.291503	2	1
5	2018	1	1	2	2	5	2.44949	3.872983	2.645751	4.472136	1
6	2018	1	1	2	1	8	5.196152	6.403124	1.414214	7.28011	3
7	2018	1	2	3	2	1	3.162278	2	8.124038	1.414214	4
8	2018	1	2	2	1	3	2.236068	1.732051	2.44949	2.44949	2
9	2018	1	2	3	2	3	3.162278	2	3	2.236068	2
10	2018	1	2	3	2	3	5.09902	1.732051	3.162278	2.236068	2
11	2018	1	2	3	1	3	6.082763	2	3.162278	2.645751	2
12	2018	1	2	3	1	4	6.324555	2.44949	3.872983	3	2
13	2018	1	2	2	2	5	5.09902	3.316625	4.123106	3.316625	4
14	2018	1	2	3	1	6	5.477226	4.123106	5.291503	4.242641	2
15	2018	1	2	3	1	8	16.12452	6.164414	7.211103	6.244998	2
16	2018	1	2	3	2	9	8.3666	6.403124	8.124038	7.071068	2
17	2018	1	2	1	1	10	9.110434	7.141428	9.273618	8.306624	2
18	2018	1	2	1	2	10	9.165151	7.211103	9.273618	8.3666	2
19	2018	1	2	1	2	11	10.0995	8.124038	9.055385	8.124038	4
20	2018	1	2	1	1	24	23.04344	21.04757	22.06808	21.04757	4
21	2018	1	3	1	1	1	2.236068	2.236068	2.828427	3.316625	4
22	2018	1	3	2	1	1	1.414214	2.236068	2	3.464102	1
23	2018	1	3	2	1	1	1.414214	2.236068	2.236068	3.464102	1
24	2018	1	3	2	2	1	1	2.236068	2	4.358899	1
25	2018	1	3	2	1	1	1.414214	2.236068	2.645751	5.291503	1
26	2018	1	3	3	2	1	1.414214	3.605551	3.316625	5.656854	1
27	2018	1	3	3	2	1	1.414214	3.741657	2.828427	2.236068	1
28	2018	1	3	2	2	2	1	2.236068	2	1.732051	1
29	2018	1	3	2	2	2	1.414214	2	2	1.414214	4
30	2018	1	3	2	2	2	1	3.162278	2	1.414214	1
31	2018	1	3	2	2	2	1.732051	3.316625	2	1.732051	4
32	2018	1	3	3	1	2	1	3.741657	1.732051	1.732051	1

Source: (Ridwansyah et al., 2022)

Data that has been placed in the form of the nearest cluster and can be calculated back to the center of the new cluster based on the average of the members in the nearest cluster. With the results of the calculations, the new centroids of cluster 1, cluster 2, cluster 3, cluster 4 are obtained, a new center point is obtained from each cluster, then recalculate the data with the new cluster center and can be repeated until the last pattern of the same cluster in the cluster is obtained. previous iteration that has not moved. In the study of infant mortality data, the data was calculated at the 10th iteration (Ten), in the 10th iteration the cluster data did not change and there was no more data moving from one cluster to another which can be seen in table 4.

Table 4. Calculation Results and New Centroid of infant mortality data

Group Cluster	Number of Cluster	Calculation resut of Cluster	New Centroid				
			City	Districts	Kelurahan	Sex	Number of Deaths
C1	43	3.930233	2.4	3	2.6	1.6	4.6
C2	126	4.230159	2	3	2.5	2	6
C3	45	1	2	1	3	1.333333	8
C4	189	2.31746	1.776081	2.05598	1.966921	1.483461	2.86514

Source: (Ridwansyah et al., 2022)

From the calculation of the cluster center and the results of the new centroid, it can be seen the results and the last pattern of the distance between the centroid distance and the center of the cluster.

Table 5. Final Results and Patterns of Centroid Distance and Cluster Center of Infant Mortality Data

Data to	Number of Cluster	Result of Cluster 1	Result of Cluster 2	Result of Cluster 3	Result of Cluster 4	Group Cluster
199	5	22.11473	13.75364	9.871129	5.91608	4
200	5	25.26976	15.31066	11.07821	5.656854	4
201	6	16.50074	12.62511	10.93391	6.78233	4
202	6	17.29346	12.78835	11.22346	6.78233	4
203	7	19.80581	14.20427	11.38086	7.483315	4
204	7	11.93935	11.27444	7.170376	8	3
205	8	12.56038	11.43508	8.139648	8.888194	3
206	9	14.71839	13.15543	8.237813	9.949874	3
207	1	12.33711	9.241189	3.316625	3.316625	4
208	1	12.66722	9.112921	3.741657	3.741657	4
209	1	16.36145	10.44198	3.316625	3.316625	4
256	4	16.25909	7.090231	5.830952	5.830952	4
257	4	15.5617	7.290888	6.244998	6.244998	4
258	6	12.81593	7.198604	7.549834	7.549834	2
259	6	15.2791	7.581655	7.874008	7.874008	2
260	1	16.63674	9.200774	2.828427	2.828427	4
261	1	14.24109	8.575569	3.316625	3.316625	4
341	1	16.06112	9.719569	2.828427	2.828427	4
342	1	9.798758	5.830952	5.830952	5.830952	4
343	1	9.857216	6.244998	6.244998	6.244998	4
344	1	6.275309	6.557439	6.557439	6.557439	1
345	2	6.148128	6.78233	6.78233	6.78233	1
346	2	6.170949	7.141428	7.141428	7.141428	1
347	2	7.160592	4	4	4	4
348	1	7.871604	3.605551	3.605551	3.605551	4
397	2	8.055905	6.324555	6.324555	6.324555	4
398	1	5.101021	6.63325	6.63325	6.63325	1
399	2	4.968365	6.855655	6.855655	6.855655	1
400	1	6.772116	6.708204	6.708204	6.708204	4
401	1	4.898979	4.898979	4.898979	4.898979	4
402	2	5.196152	5.196152	5.196152	5.196152	4
403	2	5.196152	5.196152	5.196152	5.196152	4

Source: (Ridwansyah et al., 2022)

The data grouped in cluster 1 amounted to 43 deaths, in cluster 2 there were 126 deaths, in cluster 3 there were 45 deaths and in cluster 4 there were 189 deaths. With the data criteria that cluster 4 has the highest infant mortality criteria, cluster 2 is ranked 2nd, cluster 3 is ranked 3rd and cluster 1 has the last ranking criteria.

From the cluster data obtained, which were tested with the Decission Tree algorithm to produce accurate and correct prediction results in predicting infant mortality data in sub-districts and urban villages in the DKI Jakarta province.

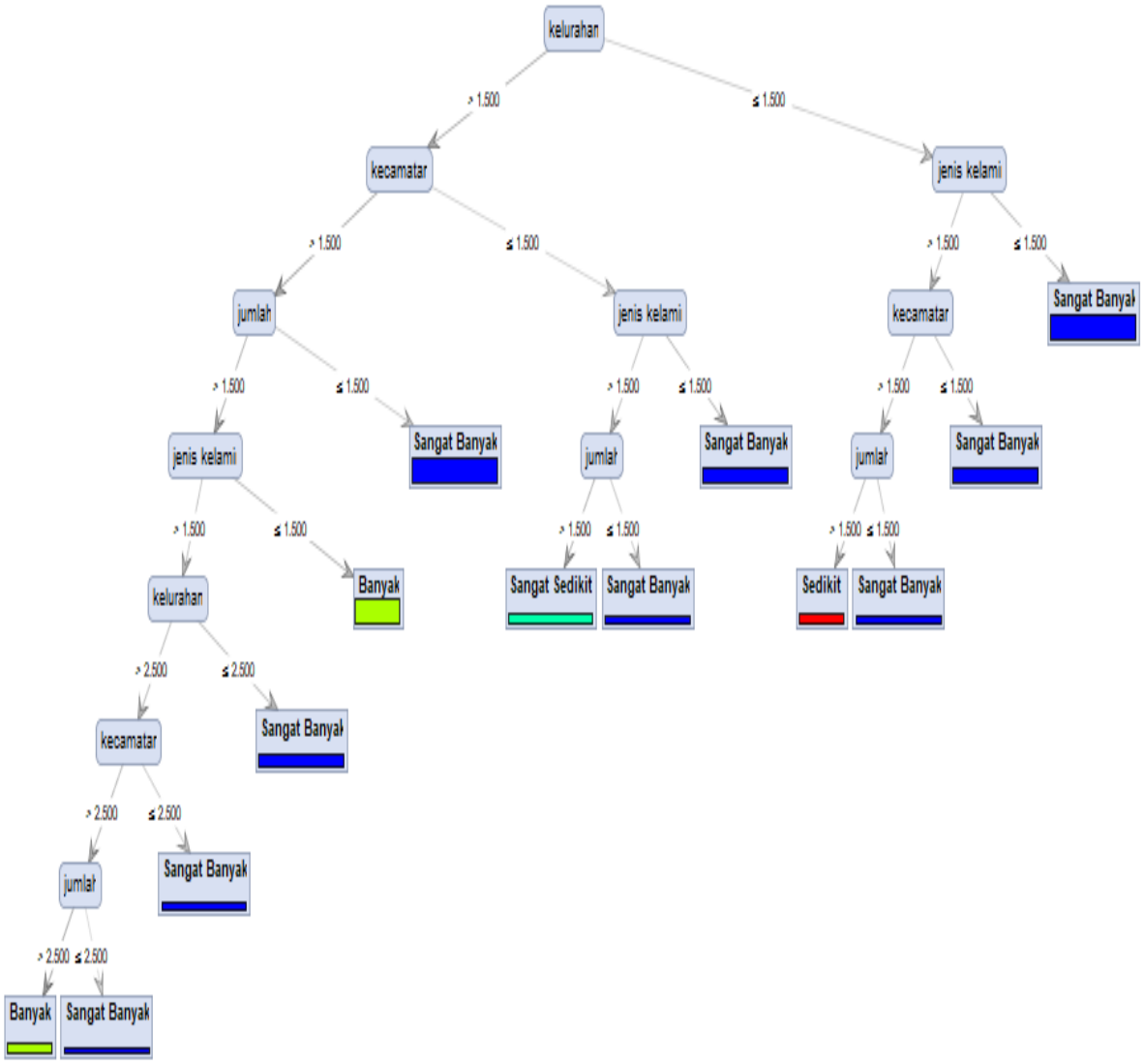
2. Decission Tree

From the test results by grouping infant mortality data using K-Means, it will be tested using a decision tree algorithm by calculating first with the entropy and gain formulas to determine the attribute to be used as the root node and determine other attributes to become the next node.

Table 6. Entropy and Gain Value

Atribut	Entropy	Gain
Total Roots	1.2815	
City		
1	1.4091	
2	1.1402	
3	1.0214	
		0.0627
District		
1	0.6136	
2	1.3202	
3	1.0388	
		0.0303
Warp		
1	0.5970	
2	1.1252	
3	1.3250	
		0.2717
Sex		
Male	0.9277	
Female	1.3138	
		0.1660

From table 6 it can be seen that the kelurahan attribute has the highest value for the gain value of 0.2717 and can be used as an attribute for the root in determining the prediction of infant mortality with the results of the decision tree which can be seen in Figure 2.



Sumber: (Ridwansyah et al., 2022)

Figure 2. Infant Death Decision Tree

After getting the decision tree, the accuracy of the decision tree algorithm will be obtained with an accuracy value of 99.75%. From these results, it can be used by the community and the government to overcome the high infant mortality and infant mortality in the highest urban village must be considered first.

CONCLUSION

From the results of the study, the goal obtained is that with the successful trial of the K-Means and Decision tree methods. So with these data it can be overcome which area has the greater infant mortality, so that infant mortality in the blood can be reduced. Performed in the application of data mining models in the grouping of infant mortality, the pattern obtained is re-implemented using a decision tree algorithm, from the results of the analysis it is known that:

a. Determination of the closest distance in making the k-means pattern using the Euclidean distance.
b. Determination of the closest distance is more optimal than using Mahattan distance and chbchep distance in classifying infant mortality.
c. In determining the centroid in the early stages of the k-means algorithm, it is very influential on the results of the cluster carried out on the infant mortality dataset taken from data.go.id with different centroid results.
d. The results of the clustering model pattern that can be evaluated by the government or the Health department to prevent infant mortality.
e. From the clustering results, four labels were tested using the decision tree algorithm.

REFERENCE

Aditya, K. B., Setiawan, Y., & Puspitaningrum, D. (2018). Sistem Informasi Geografis Pemetaan Faktor-Faktor Yang Mempengaruhi Angka Kematian Ibu (Aki) Dan Angka Kematian Bayi (Akb) Dengan Metode K-Means Clustering (Studi Kasus: Provinsi Bengkulu). *Jurnal Teknik Informatika*, *10*(1), 59–66. https://doi.org/10.15408/jti.v10i1.6817

Arifin, T., & Herliana, A. (2020). Optimizing decision tree using particle swarm optimization to identify eye diseases based on texture analysis. *Jurnal Teknologi Dan Sistem Komputer*, *8*(1), 59–63. https://doi.org/10.14710/jtsiskom.8.1.2020.59-63

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

Junaedi, I., Nuswantari, N., & Yasin, V. (2019). PERANCANGAN DAN IMPLEMENTASI ALGORITMAC4.5 UNTUK DATAMININGANALISIS TINGKATRISIKO KEMATIAN NEONATUM PADA BAYI. *Journal of Information System, Informatics and Computing*, *3*(1).

Kohno, A., Techasrivichien, T., Pilar Suguimoto, S., Dahlui, M., Nik Farid, N. D., & Nakayama, T. (2020). Investigation of the key factors that influence the girls to enter into child marriage: A meta-synthesis of qualitative evidence. *PLoS ONE*, *15*(7 July), 1–20. https://doi.org/10.1371/journal.pone.0235959

Oktavia, R., Hardinata, J. T., & Irawan, I. (2020). Penerapan Metode Algoritma K-means Dalam Pengelompokan Angka Harapan Hidup Saat Lahir Menurut Provinsi. *Kesatria: Jurnal Penerapan …*, *1*(4), 154–161. http://tunasbangsa.ac.id/pkm/index.php/kesatria/article/view/41

Ridwansyah, Riyanto, V., Hamid, A., Rahayu, S., & Purnama, J. J. (2022). *Laporan Akhir Penelitian:Pengelompokan Dalam Memprediksi Data Kematian Bayi Menggunakan K-Means dan Decision Tree*.

Salina, F. H., Almeida, I. A. de, & Bittencourt, F. R. (2019). *Renewable Energy and Sustainable Buildings*.

Santiko, I., Kurniawan, D., & Astuti, T. (2018). Clustering Berat Bayi Lahir Rendah Berdasarkan Anthropometri Menggunakan Metode K-Mean. *Citisee*, 121–124.

Suniantara, I. K. P., Hendayanti, N. P. N., & Suwardika, G. (2020). K-Means Bootstrap Analysis in the Birth Weight Classification of the Born Babies. *International Journal of Advances in Scientific Research and Engineering*, *06*(04), 99–105. https://doi.org/10.31695/ijasre.2020.33794

Syamsu, S., Muhajirin, M., & Wijaya, N. S. (2019). Rules Generation Untuk Klasifikasi Data Bakat dan Minat Berdasarkan Rumpun Ilmu Dengan Decision Tree. *Inspiration: Jurnal Teknologi Informasi Dan Komunikasi*, *9*(1), 40. https://doi.org/10.35585/inspir.v9i1.2495