JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 4, September 2023, hlm. 715- 720 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i4.2770
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

PERFORMANCE ANALYSIS OF CLUSTERING MODELS BASED ON MACHINE LEARNING IN STUNTING DATA MAPPING

Masitah Handayani[1*], Mustika Fitri Larasati Sibuea[1]
Sistem Informasi, Sekolah Tinggi Manajemen Informatika dan Komputer Royal
email: [*]bungafairuz8212@gmail.com

Abstract: Stunting is one of the nutritional problems that the world pays the most attention to and a major nutritional problem in Indonesia. Stunting is a problem in toddler growth which is characterized by a toddler's height that is too short compared to toddlers of his age. In the research location, namely Asahan Regency, the mapping of areas prone to increased stunting rates has not been carried out optimally. The process of exploring the stunting data warehouse is useful for adding information that can assist the government in making policies. Therefore, the aim of this research is to map stunting-prone areas in Asahan district based on the number of stunting cases in Asahan district using the machine learning-based K-Means clustering model. Based on previous research reviews, the k-means clustering method used has not used the normalization process. In addition, distance measurement only uses Euclidean Distance. Meanwhile, in this research, clustering performance analysis was carried out using a more in-depth process, namely by applying data normalization at the beginning, using the elbow method to determine the best number of clusters (K), measuring distance using Euclidean Distance, Manhattan Distance and Minkowski Distance to obtain comparison results. better clusters. The analysis results show that the best number of clusters is cluster 2 which shows the mapping results into 2 groups with a DBI of 0.51290 and a silhouette_score of 0.71432.

Keywords: stunting; k-means clustering; machine learning

Abstrak: Stunting menjadi salah satu permasalahan gizi yang paling diperhatikan dunia dan permasalahan gizi yang utama di Indonesia. Stunting merupakan masalah pada pertumbuhan balita yang ditandai dengan tinggi badan balita yang terlalu pendek dibanding balita seusianya. Pada lokasi penelitian yaitu Kabupaten Asahan, pemetaan daerah rawan peningkatan angka stunting belum dilakukan dengan optimal. Proses eksplorasi gudang data stunting ini berguna untuk menambah informasi yang dapat membantu pemerintah dalam mengambil kebijakan. Maka dari itu, tujuan dari penelitian ini adalah pemetaan daerah rawan stunting di kabupaten Asahan berdasarkan jumlah kasus stunting di Kabupaten Asahan menggunakan model clustering metode K-Means berbasis machine learning. Berdasarkan tinjauan penelitian terdahulu, metode k-means clustering yang digunakan belum menggunakan proses normalisasi. Selain itu, pengukuran jarak hanya menggunakan Euclidean Distance. Sedangkan dalam penelitian ini, analisis kinerja clustering yang dilakukan dengan proses yang lebih mendalam yaitu dengan penerapan normalisasi data di awal, penggunaan elbow method untuk penentuan jumlah cluster (K) terbaik, pengukuran jarak dengan Euclidean Distance, Manhattan Distance dan Minkowski Distance untuk mendapatkan hasil perbandingan cluster yang lebih baik. Hasil analisis menunjukkan bahwa jumlah cluster terbaik yaitu cluster 2 yang menunjukkan hasil pemetaan menjadi 2 kelompok dengan DBI 0.51290 dan silhouette_score sebesar 0.71432.

Kata kunci: stunting; k-means clustering; machine learning

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 4, September 2023, hlm. 715- 720 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i4.2770
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

INTRODUCTION

Stunting is one of the world's most concerned nutritional problems and a major nutritional problem in Indonesia. Stunting is a problem in the growth of toddlers which is characterized by the toddler's height being too short compared to toddlers his age. Indonesia is ranked 34th out of 50 countries with the highest cases of stunting under five in the world, and is ranked 6th in Southeast Asia[1]. The results of the integration of the March 2019 Susenas with the 2019 Study on the Nutritional Status of Indonesian Toddlers (SSGBI) show that cases of stunted toddlers in Indonesia are 27.7%, this Image still does not reach the standard set by WHO, namely 20%. To prevent an increase in stunting rates, local governments need to map stunting-prone areas first so that stunting management programs can be more precise[2].

At the research location, namely Asahan Regency, mapping of areas prone to increasing stunting rates has not been carried out optimally. Studies on potential areas experiencing an increase in the number of stunting are still minimal. This is because the data warehouse owned by the district government regarding public health data, especially those related to stunting, has not been explored optimally. This data exploration process is useful for adding information that can help the government in making policies. Therefore, the aim of this research is to map stunting-prone areas in Asahan district based on the number of stunting cases in Asahan district using the machine learning-based K-Means clustering model.

The use of a machine learning-based clustering model was chosen because clustering has good capabilities in grouping data [3][4]–[6]. In the process, the clustering model groups data based on the level of data similarity based on the amount of data for each factor or criterion [7], [8]. The elbow method is a method that is often used to determine the number of clusters to be used in k-means clustering[9]–[11]. The clustering method that will be used in this research is K-Means clustering by optimizing the number of clusters through normalization and the elbow method using Python programming.

METHOD

Problem solving through performance analysis of machine learning-based clustering models is carried out using several processes. Starting with carrying out a data normalization process which aims to equalize the data dimensions for each factor to make the cluster process easier. The process continues by determining the best number of clusters (K) using the Elbow method which aims to minimize iterations so that the cluster process becomes faster.

Next, the analysis process using the K-Means clustering method is carried out by comparing several data distance measurement techniques. Several data distance formulas that will be compared are Euclidean Distance, Minkowski Distance, and Manhattan Distance. This aims to measure performance using the Davies Bouldin Index for the performance of each data distance so that the best data cluster results can be obtained and become the final result of this research. The programming used is Python.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 4, September 2023, hlm. 715- 720 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i4.2770
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Data normalization uses Min-Max normalization.

The min-max normalization method converts a data set into a scale ranging from 0 (min) to 1 (max)[12], [13].

$$X_{norm} = \frac{x' - \min(x)}{\max(x) - \min(x)} (new_{max}(x) - new_{min}(x)) + new_{min}(x) \quad (1)$$

Information:
x = data attributes
min(x) and max(x) = minimum and maximum absolute values of x
x' = old value of each entry in data
new_max(x) and new_min(x)

Elbow Method

Elbow method to determine the most optimal number of clusters by calculating the SSE (Sum of Squares Error) value of each cluster [14], [15].

$$SSE = \sum_{k-1}^{k} \sum_{xieSk} ||Xi - Ck||_2^2 \quad (2)$$

Information:
Xi = attribute value from the ith data
Ck = value of the i-th Cluster center point attribute

Steps in the K-Means Clustering method [16], [17] are determines the initial centroid and calculate the distance between the centroid point and the point of each object.

In this research, a comparative study of 3 cluster distances will be carried out, namely:

a. ***Euclidean Distance***

$$d(x,y) = |x - y| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (3)$$

Information:
d = the distance between x and y
x = cluster center data
y = data on attributes
i = every data
n = amount of data
xi = data at the center of the i cluster
yi = data on each i data

b. ***Manhattan Distance***

$$d(x,y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \quad (4)$$

Information:
d = the distance between x and y
x = cluster center data
y = data on attributes
i = every data
n = amount of data,
xi = data at the center of the i cluster
yi = data on each i data
p = power

c. ***Minkowski Distance***

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i| \quad (5)$$

Information:
d = the distance between x and y
x = cluster center data
y = data on attributes

1. Grouping objects to determine cluster members is by taking into account the minimum distance of the object.
2. Return to stage 2, repeat until the resulting centroid value is constant.

RESULTS AND DISCUSSION

In the data processing process, data on community health centers will be given a letter code to disguise the actual location, especially when it is published. This is because the data is private data that is not to be shared widely. The analysis process using Python starts from installing the library and importing the data.

```
#import dataset
import matplotlib.pyplot as plt
from pandas.plotting import
scatter_matrix
import pandas as pd
```

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 4, September 2023, hlm. 715- 720 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i4.2770
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

df=pd.read_excel ('datastuntingok.xlsx')
df

with the following data distribution results.

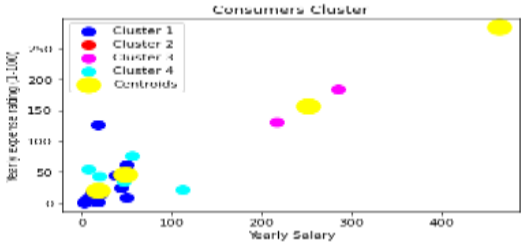

Image 1. Distribution of Stunting data

Next, an elbow method search process will be carried out to obtain the best K information that will be selected in the cluster process. So we get the elbow method as follows

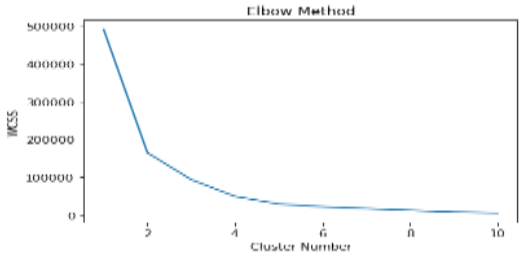

Image 2. elbow method

Based on the elbow method in Image 2, there are 3 possible optimal clusters, namely K=2, K=3 and K=4. The food will be analyzed to see the performance of each cluster.

Plot and Performance at K=2 (2 Clusters)

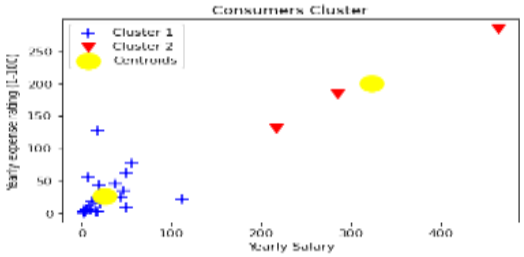

Image 3. Distribution of K=2 data

With performance values as follows:

silhouette_score 2 Cluster =
==========
0.7143247034059921
davies_bouldin_score 2 Cluster =
==========
0.5129107070094454

With the condition that silhouette_score is close to 1, the best, close to min 1, the worst, davies_bouldin_score is close to 0, which is better, where the - value is ignored.

Plot and Performance at K=3 (3 Clusters)

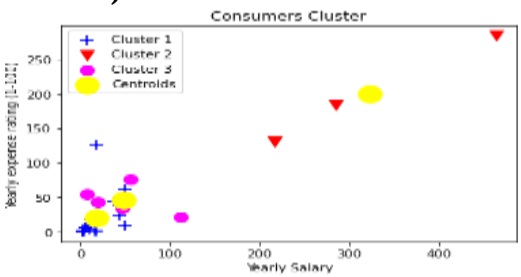

Image 4. Distribution of data K=3

silhouette_score 3 Cluster =
==========
0.5782753964820787
davies_bouldin_score 3 Cluster =
==========
0.6197873736962533

With the condition that the silhouette_score is close to 1, the best, the worst is close to 1, the Davies Bouldin score is close to 0, the better, where the - value is ignored.

Plot and Performance at K=4 (4 Clusters)

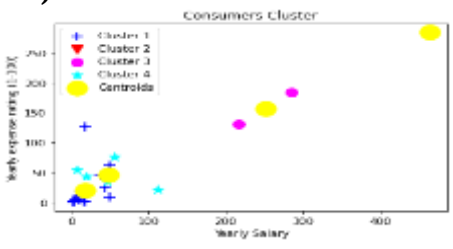

Image 5. Data Distribution K=4

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 4, September 2023, hlm. 715- 720 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i4.2770
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

silhouette_score 4 Cluster =
==========
0.5859845679738441
davies_bouldin_score 4 Cluster =
==========
0.47775947595692403

With the condition that silhouette_score is close to 1, the best, close to min 1, the worst, davies_bouldin_score is close to 0, which is better, where the - value is ignored.

Table 2. Cluster Performance

Cluster/Index	Silhouette	Davies_Bouldin
K=2	**0.71432**	**0.51290**
K=3	0.57827	0.61978
K=4	0.58598	0.477759

Based on the three analysis processes carried out, the results of which are shown in table 2 above, it can be concluded that the best cluster produced is cluster 2.

CONCLUSION

Based on the three analysis processes carried out, the results of which are shown in table 2 above, it can be concluded that the best cluster produced is cluster 2 which shows the results of mapping into 2 groups with a DBI of 0.51290 and a silhouette_score of 0.71432. With the condition that silhouette_score is close to 1, the best, close to min 1, the worst, davies_bouldin_score is close to 0, which is better, where the - value is ignored.

ACKNOWLEDGEMENT

The researcher would like to express his deepest gratitude to the Directorate of Research, Technology and Community Service of the Ministry of Education, Culture, Research and Technology for supporting the implementation of this research activity, especially in providing funding.

BIBLIOGRAPHY

[1] A. Fadilah, M. N. Pangestu, S. Lumbanbatu, and S. Defiyanti, "Pengelompokan Kabupaten/Kota Di Indonesia Berdasarkan Faktor Penyebab Stunting Pada Balita Menggunakan Algoritma K-Means," *JIKO (Jurnal Inform. dan Komputer)*, vol. 6, no. 2, p. 223, 2022, doi: 10.26798/jiko.v6i2.581.

[2] A. N. H. Yuni Nur'afiah, "Progr-am ' Gebrak Stunting ' sebagai Upaya Pencegahan Stunting di Desa Sukasenang Kecamatan Sindangkasih," *J. Kependudukan, Keluarga, dan Sumber Daya Mns.*, vol. 3, no. 1, pp. 1–13, 2022, doi: 10.37269/pancanaka.v3i1.106.

[3] E. Retnoningsih and R. Pramudita, "Mengenal Machine Learning Dengan Teknik Supervised Dan Unsupervised Learning Menggunakan Python," *Bina Insa. Ict J.*, vol. 7, no. 2, p. 156, 2020, doi: 10.51211/biict.v7i2.1422.

[4] S. Sharma and P. Chaudhary, "Machine learning and deep learning," *Quantum Comput. Artif. Intell. Train. Mach. Deep Learn. Algorithms Quantum Comput.*, pp. 71–84, 2023, doi: 10.1515/9783-110791402-004.

[5] J. Wei *et al.*, "Machine learning in materials science," *InfoMat*, vol. 1, no. 3, pp. 338–358, 2019, doi: 10.1002/inf2.12028.

[6] Mahesh Batta, "Machine Learning

Algorithms - A Review," *Int. J. Sci. Res.*, no. October, 2020, doi: 10.21275/ART20203995.

[7] R. Muliono and Z. Sembiring, "Data Mining Clustering Menggunakan Algoritma K-Means Untuk Klasterisasi Tingkat Tridarma Pengajaran Dosen," *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 4, no. 2, pp. 2502–714, 2019, doi: https://doi.org/10.24114/cess.v4i2.13620.

[8] R. R. Syoer and Y. Wahyudin, "Analisis Kelompok Dengan Algoritma Fuzzy Clustering (Studi Kasus Pengelompokkan Desa Di Provinsi Kalimantan Timur)," *BESTARI Bul. Statisitika dan Apl. Terkini*, vol. 1, pp. 1–11, 2021, [Online]. Available: https://bestari.bpskaltim.com/index.php/bestari-bpskaltim/article/view/1

[9] H. Humaira and R. Rasyidah, "Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm," 2020, doi: 10.4108/eai.24-1-2018.2292388.

[10] M. Cui, "Introduction to the K-Means Clustering Algorithm Based on the Elbow Method," *Clausius Sci. Press*, vol. 1, no. 1, pp. 5–8, 2020, doi: 10.23977-/accaf.2020.010102.

[11] E. Schubert, "Stop using the elbow criterion for k-means and how to choose the number of clusters instead," *ACM SIGKDD Explor. Newsl.*, vol. 25, no. 1, pp. 36–42, 2023, doi: 10.1145/360627-4.3606278.

[12] S. K. Dirjen *et al.*, "Terakreditasi SINTA Peringkat 2 Analisis Pengaruh Data Scaling Terhadap Performa Algoritme Machine Learning untuk Identifikasi Tanaman," *Masa Berlaku Mulai*, vol. 1, no. 3, pp. 117–122, 2017.

[13] S. Z. H. Rukmana, A. Aziz, and W. Harianto, "Optimasi Algoritma K-Nearest Neighbor (Knn) Dengan Normalisasi Dan Seleksi Fitur Untuk Klasifikasi Penyakit Liver," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 2, pp. 439–445, 2022, doi:https://doi.org/10.36040/jati.v6i2.4722.

[14] F. Sutomo *et al.*, "Optimization Of The K-Nearest Neighbors Algorithm Using The Elbow Method On Stroke Prediction," *J. Tek. Inform.*, vol. 4, no. 1, pp. 125–130, 2023, doi:https://doi.org/10.20884/1.jutif.2023.4.1.839.

[15] A. Winarta and W. J. Kurniawan, "Optimasi Cluster K-Means Menggunakan Metode Elbow pada Data Pengguna Narkoba dengan Pemrograman Python," *J. Tek. Inform. Kaputama*, vol. 5, no. 1, pp. 113–119, 2021, [Online]. Available:http://jurnal.kaputama.ac.id/index.php/JTIK/article/view/466

[16] A. Sulistiyawati and E. Supriyanto, "Implementasi Algoritma K-means Clustring dalam Penetuan Siswa Kelas Unggulan," *J. Tekno Kompak*, vol. 15, no. 2, p. 25, 2021, doi: 10.33365/jtk-.v15i2.1162.

[17] S. Handoko, F. Fauziah, and E. T. E. Handayani, "Implementasi Data Mining Untuk Menentukan Tingkat Penjualan Paket Data Telkomsel Menggunakan Metode K-Means Clustering," *J. Ilm. Teknol. dan Rekayasa*, vol. 25, no. 1, pp. 76–88, 2020, doi: 10.35760/tr.2020.v25i1.2677.