Journal of Computer Science Advancements, 1(5) - October 2023 314-319

Fuzzy C-Means and K-Means Methods Comparison for the Detection of Diabetes

Roy Efendi Subarja[1], Billy Hendrik[2]
[1]*Universitas Graha Nusantara Padangsidimpuan, Indonesia*
[2]*Universitas Putra Indonesia YPTK Padang, Indonesia*

Corresponding Author: Roy Efendi Subarja, E-mail; subardjaroy@ugn.ac.id

Article Information:
Received October 29, 2021
Revised November 6, 2021
Accepted November 10, 2021

ABSTRACT

The necessity for precise and effective disease detection techniques has increased due to the rising incidence of diabetes. The main objective of this study is to assess how well the fuzzy C-Means and K-Means clustering algorithms detect diabetes. Based on pertinent medical data, the study attempts to examine how well these two clustering approaches identify cases of diabetes. For testing, a dataset with a variety of health and diagnostic indicator variables was used. Metrics including sensitivity, specificity, accuracy, and F1-score were used to evaluate the detection performance of the Fuzzy C-Means and K-Means algorithms that were used to cluster the dataset. Based on several evaluation criteria, the results show that both clustering approaches have promising potential for diabetes identification. However, their performance varies. This study sheds light on the advantages and disadvantages of clustering algorithms and advances our understanding of their suitability for diabetes identification. Improved diagnosis precision and early diabetes management intervention could result from more optimization and validation of these algorithms

Keywords: *Diabetes, Fuzzy Inference System, Fuzzy C-Means, Fuzzy K-Means Clustering*

Journal Homepage https://journal.ypidathu.or.id/index.php/jcsa

How to cite: Subarja, E, R., & Hendrik, B. (2023). Fuzzy C-Means and K-Means Methods Comparison for the Detection of Diabetes. *Journal of Computer Science Advancements*, 1(5). 314-319 https://doi.org/10.70177/jsca.v1i5.620
Published by: Yayasan Pendidikan Islam Daarut Thufulah

INTRODUCTION

Diabetes is a serious issue for world health. Its prevalence is steadily rising and poses a serious threat to both people's quality of life and the health system (Abdel-Wahab, 2022; Ali dkk., 2020). Reducing the disease's harmful effects and complications requires early detection and adequate treatment (Agrawal, 2022). Much research has been done to create better, more precise, and effective techniques in an

attempt to enhance diabetes early detection. Data grouping methods, such as fuzzy C-Means and K-Means clustering, have been employed as a strategy (Alonso-Silverio, 2021). These methods can be used to segregate people who are at a high risk of developing diabetes and find hidden patterns in clinical data (Ballal, 2021). The performance and efficacy of these two clustering algorithms in the particular context of illness detection still need to be thoroughly assessed, notwithstanding efforts to utilize data clustering for diabetes detection (Aqeel, 2020; Berg, 2022). This assessment will assist in determining the benefits and drawbacks of each technique and offer direction to medical professionals in choosing the best course of action (Lee, 2022). Thus, the purpose of this work is to assess diabetes diagnosis performance using fuzzy C-Means and K-Means clustering algorithms (Haviluddin, 2022). It is intended that this research will help develop more precise and successful early detection strategies for the treatment of diabetes by delving deeper into the capacity of these two approaches to identify diabetes-related patterns in clinical data. As a result, data that satisfies the conditions and belongs to one cluster but not another will be grouped by the K-Means algorithm (Shantsila, 2023). Nevertheless, the data is determined with the maximum degree of membership in the Fuzzy C-Means Clustering approach; that is, the data may belong to multiple clusters (Huang, 2021). In light of the need for multiple blood sugar tests to obtain reliable results and the constraints faced by medical professionals in identifying diabetes based on multiple indicators and large amounts of data, this study will develop a diabetes detection application utilizing fuzzy C-Means method clustering and K-Means clustering (Marin, 2020).

RESEARCH METHODOLOGY

To answer research questions and accomplish specific goals, researchers employ a range of processes, techniques, and strategies known as research methodology (Lion, 2021). These are utilized in the collection, analysis, and interpretation of data. A methodical framework for research technique directs the entire process of conducting research, from planning to carrying it out and analyzing the findings (Tsuda, 2019). Fuzzy C-Means and K-Means clustering, an empirical research methodology, will be utilized to assess the effectiveness of diabetes diagnosis (Chimbunde, 2023). This strategy will entail a number of crucial processes for data collection, analysis, and evaluation. A thorough explanation of the research techniques that will be used is provided below:

a. Data Collection: Appropriate sources, including clinic databases, patient medical records, and other relevant sources, will be used to gather pertinent medical data. Numerous health metrics, including blood glucose levels, body weight, blood pressure, family history, and other risk factors that may be involved in the identification of diabetes, will be included in this data.
b. Preprocessing of the Data: In order to guarantee consistency and quality, the obtained data will be treated. Preprocessing procedures include removing outliers that could affect the analysis, normalizing data to convert the variable

scale to a uniform one, and addressing missing values by filling in the gaps in the data.

c. Number of Clusters (k) Selection: The number of clusters (k) must be decided upon before to using the clustering process. This might be accomplished using techniques like the Elbow Method, Silhouette Analysis, or domain expertise.
d. Fuzzy C-Means and K-Means Algorithm Implementation: A programming language or data analysis software will be used to implement both algorithms. Fuzzy C-Means will be used to assess the preprocessed data in order to determine whether data belong in each cluster, and K-Means will be used to arrange the data into exclusive clusters.
e. Performance assessment: F1 score, sensitivity, specificity, accuracy, and other performance assessment metrics will be used to assess the outcomes of both algorithms. Based on currently available medical data, these metrics will give a broad picture of the algorithm's capacity to categorize diabetic cases.
f. Cross-Validation: A cross-validation technique will be used to guarantee that the results are broadly applicable. To test the algorithm on non-training data, the data will be split into subsets for training and testing.
g. Comprehensive analysis and interpretation of the results will be done using data from the cross-validation and performance evaluation processes. The comparison of Fuzzy C-Means and K-Means, together with an analysis of the relevance of the findings, will be showcased.

Discussion and Conclusions: The research findings will be examined in light of prior discoveries and other scientific literature. Based on the results analysis and their implications for the application of the two algorithms in diabetes detection, conclusions will be made.

RESULT AND DISCUSSION

The experiment with Random Data 4 yielded the highest percentage value when it came to the K-Means Clustering technique findings. The degree of truth was 73.438%, and there were seven iterations generated (Wang, 2021). With two iterations and an accuracy score of 61.979%, studies employing zero data produced the lowest results. Table 4 below displays the results of experiments using the K-Means Clustering technique on all data.

Table 1. presents the K-Means method test results.

No	Data	Jumlah Iterasi	Tingkat Kebenaran
1	Data Kelurahan	4	66%
2	Data Nol	2	62%
3	Data Non Nol	11	67%
4	Data Ambigu	5	72%
5	Data Non Ambigu	8	72%
6	Data Acak 1	7	64%
7	Data Acak 2	9	68%
8	Data Acak 3	10	71%
9	Data Acak 4	7	73%

Table 2. presents the Fuzzy-Means method test results.

No	Data	Exponen (w)	Jumlah Iterasi	Tingkat Kebenaran
1	Data Keseluruhan	2	43	65.89%
2	Data Nol	2	31	69.53%
3	Data Non Nol	2	56	68.49%
4	Data Ambigu	2	51	69.13%
5	Data Non Ambigu	2	40	72.77%
6		2	59	65.63%
7		3	54	66.15%
8	Data Acak 1	4	45	66.15%
9		5	40	66.67%
10		6	39	67.18%
11		2	35	72.92%
12		3	36	72.92%
13	Data Acak 2	4	40	72.92%
14		5	40	73.96%
15		6	53	73.44%
16		2	36	70.16%
17		3	37	71.73%
18	Data Acak 3	4	36	73.30%
19		5	37	74.87%
20		6	36	75.92%
21		2	56	71.88%
22		3	48	55.21%
23	Data Acak 4	4	48	79.17%
24		5	45	81.77%
25		6	38	82.81%

The following settings were used for the Fuzzy C-Means technique experiments: 500 for the Minimum Iteration and 0.000005 for the Maximum Error. The range of exponent values used in random experimental data 1 through 4 is 2 to 6. The experiment employs Random Data 4 with an exponent value of 6, which yields the highest percentage value for the level of truth when it is completed. There are 38 iterations generated, and the degree of accuracy is 82.812%. With 43 iteration results and an

accuracy level of 65.885%, the experiments employing Overall Data produced the lowest results.

CONCLUSION

In this study, the fuzzy C-Means and K-Means clustering techniques were used to assess the effectiveness of diabetes detection. Finding out how well these two approaches categorize people as diabetes or non-diabetics based on medical data was the primary goal of the study. The research's results and findings, when interpreted in accordance with the planned methodological stages, lead to numerous significant conclusions.

1. Combining Methods: More thorough information on diabetes detection can be obtained by combining the outcomes of the two methods, fuzzy C-Means and fuzzy K-Means. Diagnostic accuracy can be improved by combining the specificity of K-Means with the sensitivity of Fuzzy C-Means.
2. While K-Means has the benefit of generating more exclusive groups, fuzzy C-Means has the ability to detect diabetes cases with higher sensitivity.
3. Using fuzzy C-Means and K-Means clustering algorithms to evaluate diabetes diagnosis performance offers important new insights into the application of data analysis techniques in medicine.
4. On random data, the clustering findings in both approaches demonstrate superior performance.

REFERENCES

Abdel-Wahab, M. (2022). Comparison of a Pure Plug-Based Versus a Primary Suture-Based Vascular Closure Device Strategy for Transfemoral Transcatheter Aortic Valve Replacement: The CHOICE-CLOSURE Randomized Clinical Trial. *Circulation*, *145*(3), 170–183. https://doi.org/10.1161/CIRCULATIONAHA.121.057856

Agrawal, H. (2022). Machine learning models for non-invasive glucose measurement: Towards diabetes management in smart healthcare. *Health and Technology*, *12*(5), 955–970. https://doi.org/10.1007/s12553-022-00690-7

Ali, Md. M., Hamid, M. O., & Hardy, I. (2020). Ritualisation of testing: Problematising high-stakes English-language testing in Bangladesh. *Compare: A Journal of Comparative and International Education*, *50*(4), 533–553. https://doi.org/10.1080/03057925.2018.1535890

Alonso-Silverio, G. A. (2021). Toward non-invasive estimation of blood glucose concentration: A comparative performance. *Mathematics*, *9*(20). https://doi.org/10.3390/math9202529

Aqeel, M. M. (2020). Temporal Dietary Patterns Are Associated with Obesity in US Adults. *Journal of Nutrition*, *150*(12), 3259–3268. https://doi.org/10.1093/jn/nxaa287

Ballal, P. (2021). Warfarin use and risk of knee and hip replacements. *Annals of the Rheumatic Diseases*, *80*(5), 605–609. https://doi.org/10.1136/annrheumdis-2020-219646

Berg, T. W. van de. (2022). Serial thrombin generation and exploration of alternative anticoagulants in critically ill COVID-19 patients: Observations from Maastricht Intensive Care COVID Cohort. *Frontiers in Cardiovascular Medicine*, *9*(Query date: 2024-05-26 20:47:26). https://doi.org/10.3389/fcvm.2022.929284

Chimbunde, E. (2023). Machine learning algorithms for predicting determinants of COVID-19 mortality in South Africa. *Frontiers in Artificial Intelligence*, *6*(Query date: 2024-05-26 20:47:26). https://doi.org/10.3389/frai.2023.1171256

Haviluddin. (2022). Naïve Bayes and K-Nearest Neighbor Algorithms Performance Comparison in Diabetes Mellitus Early Diagnosis. *International journal of online and biomedical engineering*, *18*(15), 202–215. https://doi.org/10.3991/ijoe.v18i15.34143

Huang, H. K. (2021). Risk of developing diabetes in patients with atrial fibrillation taking non-vitamin K antagonist oral anticoagulants or warfarin: A nationwide cohort study. *Diabetes, Obesity and Metabolism*, *23*(2), 499–507. https://doi.org/10.1111/dom.14243

Lee, K. B. (2022). Stroke and Systemic Thromboembolism according to CHA2DS2-VASc Score in Contemporary Korean Patients with Atrial Fibrillation. *Yonsei Medical Journal*, *63*(4), 317–324. https://doi.org/10.3349/ymj.2022.63.4.317

Lion, M. (2021). Implementation and evaluation of a multivariate abstraction-based, interval-based dynamic time-warping method as a similarity measure for longitudinal medical records. *Journal of Biomedical Informatics*, *123*(Query date: 2024-05-26 20:47:26). https://doi.org/10.1016/j.jbi.2021.103919

Marin, J. G. (2020). Prescription Patterns in Dialysis Patients: Differences Between Hemodialysis and Peritoneal Dialysis Patients and Opportunities for Deprescription. *Canadian Journal of Kidney Health and Disease*, *7*(Query date: 2024-05-26 20:47:26). https://doi.org/10.1177/2054358120912652

Shantsila, A. (2023). Contemporary management of atrial fibrillation in primary and secondary care in the UK: the prospective long-term AF-GEN-UK Registry. *Europace*, *25*(2), 308–317. https://doi.org/10.1093/europace/euac153

Tsuda, T. (2019). Effect of hypertrophic cardiomyopathy on the prediction of thromboembolism in patients with nonvalvular atrial fibrillation. *Heart Rhythm*, *16*(6), 829–837. https://doi.org/10.1016/j.hrthm.2018.11.029

Wang, Y. (2021). A Comparison of Machine Learning Algorithms in Blood Glucose Prediction for People with Type 1 Diabetes. *ACM International Conference Proceeding Series*, *Query date: 2024-05-26 20:47:26*, 351–360. https://doi.org/10.1145/3500931.3500993