Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 24, No. 1, November 2024, pp. 61∼72
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v24i1.4043

Cluster Validity for Optimizing Classification Model: Davies Bouldin Index – Random Forest Algorithm

Prihandoko[1], **Deny Jollyta**[2], **Gusrianty**[2], **Muhammad Siddik**[2], **Johan**[2]
[1]Universitas Gunadarma, Depok, Indonesia
[2]Institut Bisnis dan Teknologi Pelita Indonesia, Pekanbaru, Indonesia

Article Info

Article history:

Received May 08, 2024
Revised July 16, 2024
Accepted September 23, 2024

Keywords:

Classification
Davies Bouldin index
Maternal health risk
Optimal cluster
Random forest algorithm

ABSTRACT

Several factors impact pregnant women's health and mortality rates. The symptoms of disease in pregnant women are often similar. This makes it difficult to evaluate which factors contribute to a low, medium, or high risk of mortality among pregnant women. The purpose of this research is to generate classification rules for maternal health risk using optimal clusters. The optimal cluster is obtained from the process carried out by the validity cluster. The methods used are K-Means clustering, Davies Bouldin Index (DBI), and the Random Forest algorithm. These methods build optimum clusters from a set of k-tests to produce the best classification. Optimal clusters comprising cluster members with strong similarities are high-dimensional data. Therefore, the Principal Component Analysis (PCA) technique is required to evaluate attribute value. The result of the research is that the best classification rule was obtained from k-tests = 22 on the 20th cluster, which has an accuracy of 97% to low, mid, and high risk. The novelty lies in using DBI for data that the Random Forest will classify. According to the research findings, the classification rules created through optimal clusters are 9.7% better than without the clustering process. This demonstrates that optimizing the data group has implications for enhancing the classification algorithm's performance.

Corresponding Author:

Deny Jollyta, 08127585546
Faculty of Computer Science, Informatics,
Institut Bisnis dan Teknologi Pelita Indonesia, Pekanbaru, Indonesia,
Email: deny.jollyta@lecturer.pelitaindonesia.ac.id

How to Cite:
P. Prihandoko, D. Jollyta, G. Gusrianty, M. Siddik, and J. Johan, "Cluster Validity for Optimizing Classification Model: Davies Bouldin Index – Random Forest Algorithm", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 24, No. 1, pp. 60-70, 2024.

1. INTRODUCTION

Pregnancy, a natural progression towards childbirth, comes with a multitude of health risks, potentially culminating in maternal mortality. Pregnant women frequently die of problems owing to poor knowledge about maternal health care during and after pregnancy [1]. Despite substantial research in this domain, studies often fall short of providing robust classification frameworks that are vital for assessing and mitigating these risks. For instance, it classified the risk of death for pregnant women but did not incorporate a dynamic clustering-based approach, which can potentially enhance predictive accuracy [2]. Alongside time, the integration of clustering methods such as K-Means with sophisticated algorithms like Random Forest, as explored in the works [3], presents an opportunity to develop more nuanced maternal health risk evaluation models.

The importance of predictive analytics in maternal health cannot be overstated [4]. The exploration of maternal health risks through computational models has been a focal area of research in recent years. The challenge that this study is trying to solve is how to create an appropriate classification of the optimal factors that cause maternal health risks, which are often comparable [5]. Our research is motivated by the need to bridge these gaps, particularly by harnessing the synergistic potential of unsupervised and supervised learning methods to improve classification accuracy. This research addresses a critical literature shortage concerning the optimal combination of clustering algorithms and Random Forests. By introducing the Davies-Bouldin Index (DBI) into the clustering process, we propose a novel approach that not only categorizes maternal health risks with high precision but also provides a clearer delineation of risk categories, thereby contributing significantly to the field of predictive health analytics. According to the research motivation, this research objective is to generate classification rules for maternal health risk using optimal clusters through a combination of clustering and classification algorithms.

Several theoretical studies have addressed the necessity of thoroughly classifying the factors impacting maternal health risk. Classification must be based on several dominant factors [6]. A body of research, including that by [7], has shown that classifications are needed to evaluate the danger to the fetus of pregnant mothers with a history of heart disease. More advanced approaches have incorporated machine learning algorithms to enhance the classification and prediction of health risks. For instance, the Decision Tree and BilTCN frameworks have been used to estimate risk in pregnant women [4], and Artificial Neural Networks (ANN) have been combined with Random Forests to yield a model accuracy of 95% [6]. However, these studies typically apply algorithms in isolation or simple combinations, not fully leveraging the complexity inherent in maternal health data. Unsupervised learning, especially clustering, plays a significant role in understanding and categorizing health risks. The study of [8] demonstrated that unsupervised learning could uncover hidden patterns in healthcare data, but they stopped short of combining these insights with supervised learning to improve prediction models. This study aims to fill this void by using unsupervised learning not as an end but to enhance the supervised learning process, providing a comprehensive risk classification system. In another study, the combination model that involved the K-Means algorithm was merged with the Random Forest method to estimate the gas content of coalbed methane reservoirs [9], but has not explored identifying the optimal cluster. Previous research has shown that combining unsupervised and supervised learning algorithms in numerous disciplines still promises advancement.

Recent advancements in clustering methods have shown considerable promise in health data analysis. Previous studies highlight that clustering, particularly K-Means, can illuminate patterns within complex health datasets [10]. Nonetheless, the challenge lies in cluster validation, a crucial yet frequently overlooked step. The clustering of health data to inform dominant factors revelation of outbreak's cause, as seen in studies [11]. Yet, these studies rarely utilize clustering quality indices like the Davies-Bouldin Index (DBI) that can validate the coherence of clusters, which is critical for the accuracy of subsequent classifications. Similarly, investigations into the effects of individual health conditions, such as diabetes, on maternal and neonatal outcomes [12], provide valuable insights but do not extend to the multifaceted risk categorization necessary for comprehensive maternal health analysis. By integrating DBI as a measure of cluster validity, our research contributes a methodological improvement that enhances the reliability of cluster-based risk stratification in maternal health datasets.

The Random Forest algorithm has been lauded for its performance in high-dimensional health data, where its ability to handle numerous input variables without overfitting is particularly valuable [13]. Yet, there remains a considerable gap in the literature regarding optimizing Random Forest parameters in conjunction with clustering outputs. Notably, the Random Forest algorithm has emerged as a robust classifier when dealing with high-dimensional data [14]. Its application to maternal health data presents promising results; however, there is less research integrating the Random Forest with an optimal cluster approach using DBI [15] Our research leverages DBI-optimized clusters as input for Random Forest, hypothesizing that this synergy will provide a robust model for classifying maternal health risks. This integration is hypothesized to yield superior classification performance due to the improved cluster quality, ultimately leading to the development of more accurate predictive models.

The combination of supervised and unsupervised learning methods has the potential to create a more nuanced understanding of data [16]. In maternal health, this approach has yet to be fully exploited. Previous research has not resolved some gaps, namely, failure to enhance classification algorithm performance by using optimized cluster findings and failure to identify cluster results, including maternal health risk levels properly. The difference between this research and the previous one is that this research combines the clustering capabilities of K-Means optimal with the classification strength of Random Forest, a synthesis scarcely represented in the existing literature. The optimal cluster is obtained from the DBI process. We propose a model that identifies clusters within maternal health data and utilizes these clusters to build a highly accurate classification system, thereby setting a new precedent in classification analytics. This research aims to generate classification rules for maternal health risk using optimal clusters. The optimal cluster is obtained from the process carried out by the validity cluster.

2. RESEARCH METHOD

This research employs a mixed-methods approach that synergistically combines the K-Means clustering algorithm with the Random Forest classifier, supplemented by the Davies-Bouldin Index (DBI) for cluster validation. The objective is to enhance the predictive accuracy of maternal health risk classification. To achieve this, mixed methods are arranged as follows in Figure 1. Every step in Figure 1 describes an innovative stage that was constructed to achieve the research purpose with the following explanation:

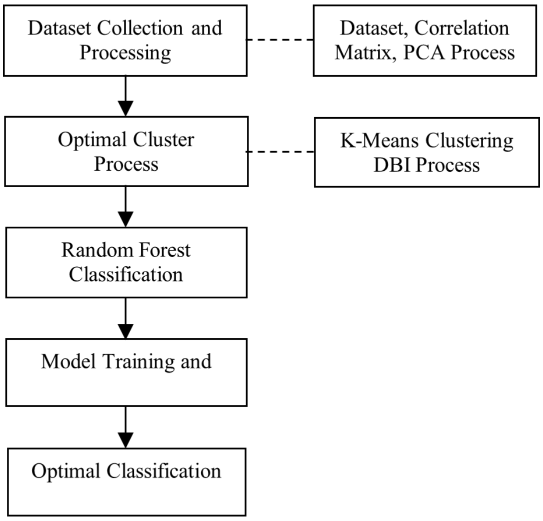

Dataset Collection and Processing
Dataset, Correlation Matrix, PCA Process
Optimal Cluster Process
K-Means Clustering DBI Process
Random Forest Classification
Model Training and
Optimal Classification

Figure 1. The research stages

2.1. Data Collection and Processing

The dataset, comprising various maternal health risk factors, was sourced from a public repository on Kaggle (2022). The dataset to be processed contains 1,456 entries. The columns include various health measurements and a risk assessment level. The data has six attributes, as shown in Table 1, and has three categories of risk level, namely low risk (LR), mid risk (MR), and high risk (HR). The data used is based on the six attributes shown in Table 1. The data has been sorted to prevent data inaccuracies. Table 2 contains the data applied in this research.

Table 1. The Attributes

No	Attributes	Information	The Value of Attributes
1	Age	The age of individuals	Ranging from 10 to 70 years
2	Systolic BP (Systolic Blood Pressure)	The pressure in the arteries when the heart beats	ranging from 70 to 160 mmHg
3	Diastolic BP (Diastolic Blood Pressure)	The pressure in the arteries when the heart rests between beats	ranging from 49 to 100 mmHg
4	BS (Blood Sugar)	The concentration of glucose in the blood	ranging from 6.0 to 19.0 mg/dL
5	Body Temp (Body Temperature)	The internal temperature of the body	mostly clustered around 98.0 to 103.0 °F
6	Heart Rate	The number of heartbeats per minute	from 7 to 90 beats per minute

Table 2. The Initial Dataset

No	Age	Systolic BP	Diastolic BP	Blood Sugar	Body Temp	Heart Rate	Risk Level
1	25	130	80	15	98	86	HR
2	35	140	90	13	98	70	HR
3	29	90	70	8	100	80	HR
4	30	140	85	7	98	70	HR
5	35	120	60	6.1	98	76	LR
6	23	140	80	7.01	98	70	HR
7	23	130	70	7.01	98	78	MR
	...	...	...	...	...	...	...
1456	16	120	75	7.9	98	7	LR

Knowing how each attribute relates to each other is necessary to achieve the best classification model. This research will examine the attributes in Table 2 for their closeness. The correlation method and Principal Component Analysis (PCA) are utilized to examine the impact of attributes on the classification results. The correlation coefficient is a statistical measure often used in studies to show an association between variables [17]. Principal Component Analysis (PCA) was utilized for feature reduction, simplifying the high-dimensional dataset while preserving essential information critical for subsequent analysis [18]. It achieves this by compressing the data into fewer components, which may be considered feature analysis [19]. PCA also supports the preprocessing stages of data processing using Euclidean distance [20]. The reduction process utilized the same data to test the proposed algorithm.

2.2. Optimal Cluster by K-Means Clustering and DBI

Perform data processing with K-Means clustering and DBI to obtain optimal clusters with some k tests. In this research, the number of k-tests was 23. This method divides data into clusters such that data with comparable features are clustered together, while data with distinct features is placed in various clusters [21]. To measure data similarity, use the Euclidean distance formulas. The Euclidean distance (d) of two data cases (x1, x2) is defined as the square root of the sum of squared differences as below [10].

$$d(i,j) = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \ldots + (x_{ki} - x_{kj})^2} \tag{1}$$

Where $d(i,j)$ is the i data distance to j cluster centroid, x_{ki} is the i data on the k data attribute and x_{kj} is the j data on the k data attribute. In the current research, K-Means clustering was utilized to generate maternal pregnancy health risk data groups based on Age, Systolic Blood Pressure (Systolic BP), Diastolic Blood Pressure (Diastolic BP), Blood Sugar (BS), Body Temperature, and Heart Rate. This data set is then optimized so that the resulting classification rules have an established relationship. The DBI was a matrix used to evaluate clustering algorithms. This research used DBI because of its ability to check whether or not a certain number of clusters was correctly divided using the K-Means clustering algorithm [22]. DBI also produces the greatest results for the K-Means method with the Euclidean Distance measuring approach [23]. DBI is an internal evaluation technique that assesses cluster evaluation in a grouping manner based on cohesion and separation values. It is written as equation 2 [24]. Where σ_i is the average distance of all points in cluster i to centroid c_i, σ_j is the average distance of all points in cluster j to centroid c_j, $d(c_i, c_j)$ is the distance between centroids c_i and c_j , and K is the total number of clusters.

$$DBI = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \tag{2}$$

2.3. Random Forest Classification

Categorize the data to be classified using the Random Forest Algorithm. The Random Forest approach mixes the output of numerous decision trees to provide a homogenous and constructive classification result [25]. As an Ensemble model, a Random Forest algorithm may create decision trees and apply its rules to generate final conclusions [26]. The more trees employed, the better the decision accuracy increases. Random Forest classification is determined based on the vote results and the trees produced [27, 28]. Random Forest algorithm starts from: a) Select N Random Records: from the original dataset, select N records randomly with replacement; b) Choose Number of Features: determine the number of features m to consider at each split in a decision tree; c) Build Decision Trees. For each bootstrap sample, grow a decision tree, and at each node, randomly select m features from the total features, determine the best split using the selected m features based on an objective metric like Gini impurity or entropy in the case of classification, and split the node into child nodes using the best split. Continue this process recursively until a stopping criterion is met (e.g., maximum depth of the tree, minimum number of samples per leaf, no further improvement in the impurity measure). Each tree is grown to the largest extent possible; d) Repeat the Process: repeat steps 1 to 3 for K times to create K decision trees in the forest; e) Aggregation (voting or averaging). Classification: For a new record, make each of the K decision trees in the forest predict the class label and choose the classification with the most votes over all the trees as the final prediction (majority voting). Regression: for regression tasks, average the numerical outputs of all trees to obtain the final prediction; and f) Output the Result: the aggregated output from step 5 is considered the predicted result for the input record.

2.4. Model Training and Validation

The Random Forest model was trained on the clustered data, with the number of trees set based on preliminary tests to optimize accuracy and prevent overfitting. Model performance was evaluated using standard metrics, including accuracy, precision, recall, and F1 score, derived from the confusion matrix with the following equation 3. The matrix utilized can be adjusted to meet accuracy requirements. The classification rule with the best accuracy is the optimal classification rule [21, 22]. The accuracy metric displays the overall number of correct predictions made by the predictive model out of all predictions. TPR (True Positive Rate) = positive predictive model is true, FPR (False Positive Rate) = positive predictive model is false, TNR (True Negative Rate) = negative predictive model is true, FNR (False Negative Rate) = negative predictive model is false.

$$Accuracy = \frac{TPR + TNR}{TPR + FPR + TNR + FNR} \tag{3}$$

3. RESULT AND ANALYSIS

3.1. Results

Data on maternal health risk was used to obtain optimal classification results. Referring to Figure 1, The process towards results begins with measuring the correlation coefficient. Statistics criteria analysis is displayed with Python, as shown in Figure 2. Figure 2 shows the correlation between each attribute. The Age row and column show a moderate positive correlation with both Systolic BP and Diastolic BP (around 0.43 and 0.41, respectively), suggesting that as age increases, blood pressure tends to rise as well. Age also has a moderate positive correlation with BS (Blood Sugar), around 0.47, indicating that higher blood sugar levels tend to occur with increasing age. Systolic BP and Diastoli c BP have a very strong positive correlation of 0.77, meaning they tend to increase or decrease together, which is expected since they are both measures of blood pressure. Systolic BP, Diastolic BP, and BS show similar levels of positive correlation with one another, ranging from 0.40 to 0.41, indicating a moderate tendency to rise together. Negative correlations are present between Body Temp and other variables such as Age (-0.27), Systolic BP (-0.29), and Diastolic BP (-0.26), suggesting that higher body temperatures are less common as these other variables increase. Heart Rate shows very low or negligible correlations with most parameters, indicating it does not strongly associate with age, blood pressure, blood sugar, or body temperature in this dataset.

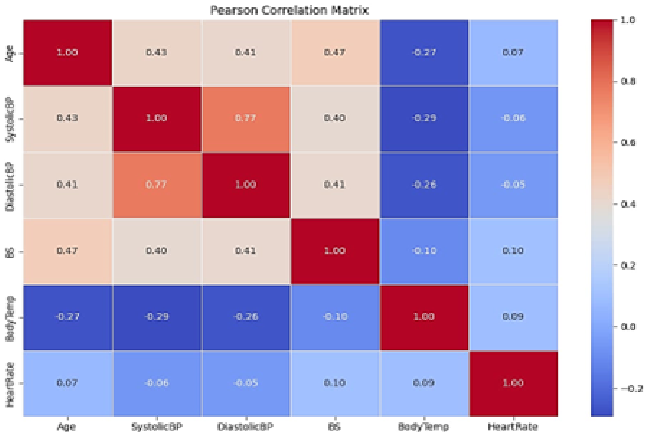
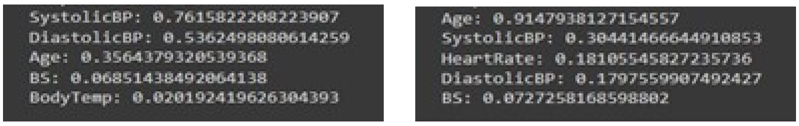


Figure 2. Correlation matrix

Initial preprocessing involved cleaning the data for missing values, outliers, and normalization. Following preprocessing, PCA was utilized to reduce the feature of maternal health data. The reduction process utilized the same data to test the proposed algorithm. The PCA results for Maternal Health attributes are processed by Python and shown in Figure 3. PCA minimizes attributes with low values. Figure 3 depicts the results of PCA processing and explains that component 1 decreases the Heart Rate, which has the lowest value among the six attributes utilized, whereas Component 2 reduces the Body Temp. This suggests that the Heart Rate and Body Temp attributes have the smallest impact on processing. In line with the correlation results, both of these attributes have a weak or very weak correlation, allowing them to influence the performance of other attributes more.



```
SystolicBP: 0.7615822208223907
DiastolicBP: 0.5362498080614259
Age: 0.3564379320539368
BS: 0.06851438492064138
BodyTemp: 0.020192419626304393
```

(a) Principal Component 1

```
Age: 0.9147938127154557
SystolicBP: 0.30441466644910853
HeartRate: 0.18105545827235736
DiastolicBP: 0.1797559907492427
BS: 0.07272581685988802
```

(b) Principal Component 2

Figure 3. PCA values of maternal health dataset

Figure 4 is a visualization of PCA that has the highest values. Component 1 displays the highest values for Systolic BP, Diastolic BP, and Age. The attributes with the greatest values in component 2 are Age, Systolic BP, and Heart Rate. The correlation matrix and PCA analysis support the idea of improving the accuracy of the chosen classification model for grouping maternal health attributes. To accomplish the research purposes, the maternal health risk data was processed initially with and without clustering the data. The first experiment was to process data without clustering. The Random Forest algorithm was used to classify data directly. Classification training used random tree numbers of 3, 10, and 22. The process used a Random Forest algorithm. A confusion matrix was utilized to determine the accuracy of each classification that refers to equation 3. All processes were carried out using Python. The following displays 3, 10, and 22 tree confusion matrix.

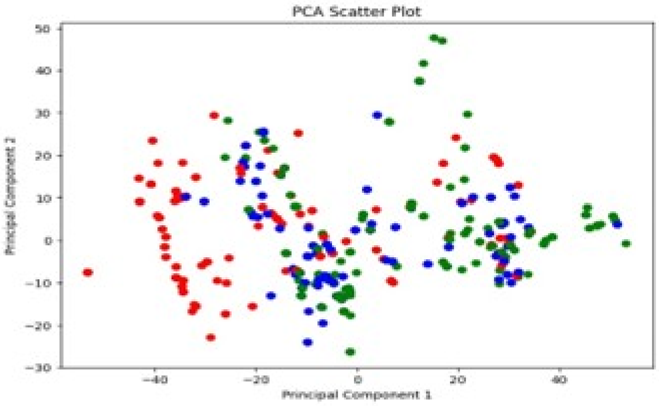


Figure 4. PCA visualization

According to Table 3, the classification rules have high accuracy and the average accuracy of 87.3%. This indicates that the classification rules developed are appropriate. Next, clustering algorithms are used to analyze the data using the K-Means method and the Euclidean distance calculation. The process refers to equation 1. The data grouping attempts to get solid group members to build classification rules while keeping the categories defined in the maternal health risk data, including LR, MR, and HR. This research took the risk of grouping findings with an infinite number of groups since the data employed does not have processing constraints after reduction, as in previous studies [9]. Creating classification rules has proven to be a challenge in itself. The proposed cluster optimization approach allows the data utilized for classification to be specific to the data produced by optimal clusters. DBI makes identifying the best cluster members easy, and the calculation refers to Equation 2. All processes were using Python.

Table 3. Pembagian data untuk Training dan Testing

Number of Tree	Accuracy
3	86%
10	88%
22	88%
Average Accuracy	87.3%

The training method employed k-tests of more than 20. This is because the optimal cluster resulting from k-test 20 is always in the second k-test with an accuracy of less than 65%. This outcome is deemed inadequate. A better optimal cluster is formed only at k-tests = 23, starting from k-tests = 2 to k-tests = 24. The clustering process refers to equation 1 and determining optimal clusters refers to Equation 2. The visualization of DBI value for all k tests is shown in Figure 4. According to Figure 4, the lowest DBI is found in k-test = 22 with a value of 0.8543108991651. Hence, this cluster is chosen as the most optimal cluster. All k-test DBI value is performed in Table 4—the visualization of cluster data distribution at k-tests = 22 in Figure 5.

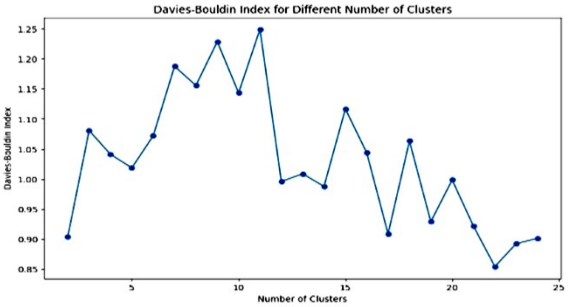


Figure 5. The DBI value of each cluster in k tests = 22

Table 4. DBI Score for All K-test

k- Tests	DBI Score	k-Tests	DBI Score
2	0.904043671	14	0.987769666
3	1.080791516	15	1.116691961
4	1.041063112	16	1.04365708
5	1.018679919	17	0.908618581
6	1.071878163	18	1.063340366
7	1.187666987	19	0.928981922
8	1.155703875	20	0.998524411
9	1.228709122	21	0.921398445
10	1.143374677	22	**0.854310899**
11	1.249293439	23	0.8926579
12	0.995972945	24	0.901199243
13	1.008693767		

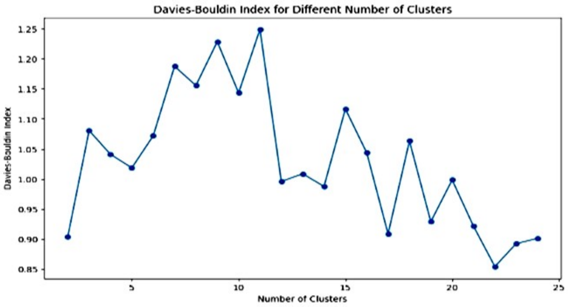


Figure 6. Optimal cluster visualization in k-tests = 22

In Figure 6, the data distribution indicates the number of individuals from each cluster in the optimal cluster. Scattered cluster members exhibit data that have several attributes, making a cohesive group. Based on the members of this cluster, classification is performed using the Random Forest algorithm to generate classification rules. Each group obtains aggregate rules by using three random trees. The aggregate results become classification rules obtained from each group, and their accuracy is calculated using a confusion matrix. Table 5 shows the classification accuracy of data processing in each group in the optimal cluster. The process refers to equation 3.

Apart from the rule accuracy values, Table 5 also shows the categories of maternal health risk data contained in each group. The categories refer to the three trees formed by the Random Forest algorithm, and the complete category with the highest accuracy is found in the 20th cluster with the tree form. The grouping by the K-Means algorithm has divided the data according to similar attributes. Hence, the data no longer contains all categories of maternal health data. Cluster 20 has a 97% accuracy rating based on the accuracy value of the rules created by the Random Forest algorithm and comprises all three categories of maternal health risk data. The aggregate is used to establish the rules. All three trees comply with the identical rules, resulting in 97% accuracy. The accuracy of a proposed algorithm increased to 9.7% compared to the average accuracy of classification rules that did not apply the clustering and optimization process (Table 3).

Table 5. Accuracy Classification Training with Clustering Process

Cluster	Number of Members	Accuracy	Categorization
0	79	88%	HR, LR, MR
1	106	64%	HR, LR, MR
2	221	87%	LR, MR
3	60	100%	HR
4	170	85%	HR, LR, MR
5	60	83%	HR, LR, MR
6	30	100%	LR, MR
7	44	100%	LR, MR
8	28	83%	HR, LR, MR
9	55	73%	MR, HR
10	64	92%	HR, LR, MR
11	52	64%	HR, LR, MR
12	86	89%	HR, LR, MR
13	4	100%	LR
14	19	75%	LR, MR
15	53	91%	HR, LR, MR
16	44	100%	HR
17	77	100%	LR, MR
18	60	92%	MR, HR
19	53	73%	HR, LR, MR
20	**72**	**97%**	**HR, LR, MR**
21	19	100%	LR, MR

3.2. Analysis

This research's findings can be regarded from several perspectives. First, the cluster's data processing makes it easier to build high-accuracy classification methods. Table 5 shows that the data group created by the optimal cluster has more than 60% accuracy. Second, reliable conclusions must incorporate the completeness of maternal health risk data categories, which it turns out do not exist in all data groups. Table 5 shows seven clusters with an accuracy of up to 100%. Clusters 3 and 16, for example, are 100% accurate. Clusters with the same member attributes are particularly good for K-Means algorithm clusters. Only members of clusters 3 and 16 are classified as HR. Similarly, cluster 13 contains members that are in the LR category. Other clusters are divided into two categories, and this approach is incompatible with classifications that need many objective attributes. The primary aim of this classification is to keep data categories, beginning with the clustering process and progressing to the establishment of classification rules. This is why clusters with only one or two categories are not chosen, even if they are 100% accurate. Third, attribute reduction via the PCA procedure generates optimal classification results. Heart Rate and Body Temp were included in the computation despite having the lowest values, implying that these two variables had no meaningful effect on cluster formation. However, using both variables in the distance calculation determines the number of cluster members with the same features (closest distance). This means that it influences the DBI calculation of each cluster's members, resulting in properly classified clusters. Of course, the results would be different if the two PCA-reduced attributes were removed. Our research shows that the results align with [9]. This can be seen from the classification process, which combined the K-Means algorithm with Random Forest. We employed optimal K-Means results to increase Random Forest performance, which was not employed by previous research. The findings of this study showed that the concept was successful in boosting the algorithm's performance even though attributes that were thought to be unimportant got involved in the process.

4. CONCLUSION

The study findings reveal an increase in the accuracy of the rules created after combining clustering algorithms, DBI, and Random Forest. This study's findings highlight that classification rules are built from data with close similarities. This shows that theoretical implications have been satisfactorily verified. The optimal data selection can increase the classification algorithm's performance. The correlation between attributes examined using PCA can influence the effectiveness of both unsupervised and supervised learning systems. Attributes with low PCA values may not impact categorization outcomes. However, it is advisable to consider employing these attributes in the testing process.

This study highlights how the proposed approach can accurately identify each classification level, offering confidence in the impact of attributes on maternal health risks. The research may still be expanded on the amount of data, the clustering method, and distance measurements employed. Other cluster evaluation techniques must consider the adequacy of the clustering algorithm's distance measurements. Many experiments are required to demonstrate that combining unsupervised with supervised algorithms yields classification rules with optimum accuracy. Hence, the findings of this research are expected to provide new information to improve the performance of machine learning algorithms.

5. ACKNOWLEDGEMENTS

Universitas Gunadarma and Institut Bisnis dan Teknologi Pelita Indonesia partially support this work. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

6. DECLARATIONS

AUTHOR CONTIBUTION

Prihandoko: Data investigation, methodology, supervision, writing review and editing. Deny Jollyta: Conceptualization, Original Draft. Gusrianty: Validation and interpretation. Muhammad Siddik: Validation and editing. Johan: Editing.

FUNDING STATEMENT

This research received no specific grant from any funding agency.

COMPETING INTEREST

The authors declare no conflict of interest

REFERENCES

[1] P. Sahithi, S. Amulya, A. S. Gajapathi, S. V. V. Raju, and S. V. Murthy, "Deep Learning Based Risk Level Prediction Model For Maternal Mortality," vol. 13, no. 4, pp. 70–80, 2023, https://doi.org/10.9790/9622-13047080.

[2] M. Y. Al-Hindi, T. A. Al Sayari, R. Al Solami, A. K. AL Baiti, J. A. Alnemri, I. M. Mirza, A. Alattas, and Y. A. Faden, "Association of Antenatal Risk Score With Maternal and Neonatal Mortality and Morbidity," *Cureus*, vol. 12, no. 12, pp. 1–8, 2020, https://doi.org/10.7759/cureus.12230.

[3] J. Lopes, T. Guimaraes, and M. F. Santos, "Identifying Diabetic Patient Profile Through Machine Learning-Based Clustering Analysis," in *Procedia Computer Science*, vol. 220. Elsevier B.V., 2023, pp. 862–867, https://doi.org/10.1016/j.procs.2023.03.116.

[4] A. Raza, H. U. R. Siddiqui, K. Munir, M. Almutairi, F. Rustam, and I. Ashraf, "Ensemble learning-based feature engineering to analyze maternal health during pregnancy and health risk prediction," *PLoS ONE*, vol. 17, no. 11, pp. 1–29, 2022, https://doi.org/10.1371/journal.pone.0276525.

[5] M. N. Islam, S. N. Mustafina, T. Mahmud, and N. I. Khan, "Machine learning to predict pregnancy outcomes: a systematic

review, synthesizing framework and future research agenda," *BMC Pregnancy and Childbirth*, vol. 22, no. 1, pp. 1–19, 2022, https://doi.org/10.1186/s12884-022-04594-2.

[6] T. O. Togunwa, A. O. Babatunde, and K. U. R. Abdullah, "Deep hybrid model for maternal health risk classification in pregnancy: synergy of ANN and random forest," *Frontiers in Artificial Intelligence*, vol. 6, no. July, pp. 1–11, 2023, https://doi.org/10.3389/frai.2023.1213436.

[7] G. J. Paul, S. A. Princy, S. Anju, S. Anita, M. C. Mary, G. Gnanavelu, K. Kanmani, M. Meena, M. Nandakumaran, S. Ramya, G. Ravishankar, G. Shaanthi, S. Shoba, V. Sangareddi, S. Vijaya, Gomathy, Geetha, U. Rani, N. Tamil Selvi, Sarala, B. Tamil Selvi, Prema Elizabeth, Nalina, Priyadarsene, Kasthuri, Sadhana, Sindhumathy, Sudarshini, Nazreeen, Devika, Shoba Sivakumar, C. Umarani, R. Priya, Kaleeswari, Suganya, R. M. Shunmugam, P. Ganapathy, M. Chandran, S. Nagarajan, M. Ganesan, A. M. Angappamudali, N. Jeyabalan, B. P. Palani, Saravanababu, K. Srinivasan, E. M. Elangovan, N. P. Mohandoss, E. Chandrasekaran, R. R. Duraipandian, P. K. Gorijavaram, T. Kunjjitham, Ravindran, Dharmarajan, T. Kaliyamurthy, J. Sreeram, A. Seeralan, Mangalabharathi, B. Mariappan, C. Manimaran, and E. J. Kumar, "Pregnancy outcomes in women with heart disease: the Madras Medical College Pregnancy And Cardiac (M-PAC) Registry from India," *European Heart Journal*, vol. 44, no. 17, pp. 1530–1540, 2023, https://doi.org/10.1093/eurheartj/ehad003.

[8] A. A. Sinha and S. Rajendran, "A novel two-phase location analytics model for determining operating station locations of emerging air taxi services," *Decision Analytics Journal*, vol. 2, no. June 2021, p. 100013, 2022, https://doi.org/10.1016/j.dajour.2021.100013.

[9] J. Yu, L. Zhu, R. Qin, Z. Zhang, L. Li, and T. Huang, "Combining k-means clustering and random forest to evaluate the gas content of coalbed bed methane reservoirs," *Geofluids*, vol. 2021, no. -, pp. 1–8, 2021, https://doi.org/10.1155/2021/9321565.

[10] A. Ultsch and J. Lötsch, "Euclidean distance-optimized data transformation for cluster analysis in biomedical data (EDOtrans)," *BMC Bioinformatics*, vol. 23, no. 1, pp. 1–18, 2022, https://doi.org/10.1186/s12859-022-04769-w.

[11] W. Ramdhan, O. S. Sitompul, E. B. Nababan, and Sawaluddin, "A Framework for Dominant Factors Revelation of the Outbreak's Cause," in *2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics, DATABIA 2021 - Proceedings*. IEEE, 2021, pp. 52–57, https://doi.org/10.1109/DATABIA53375.2021.9649732.

[12] K. Rodolaki, V. Pergialiotis, N. Iakovidou, T. Boutsikou, Z. Iliodromiti, and C. Kanaka-Gantenbein, "The impact of maternal diabetes on the future health and neurodevelopment of the offspring: a review of the evidence," *Frontiers in Endocrinology*, vol. 14, no. July, pp. 1–19, 2023, https://doi.org/10.3389/fendo.2023.1125628.

[13] W. Li, "Optimization and Application of Random Forest Algorithm for Applied Mathematics Specialty," *Security and Communication Networks*, vol. 2022, no. -, pp. 1–9, 2022, https://doi.org/10.1155/2022/1131994.

[14] M. Jiang, J. Wang, L. Hu, and Z. He, "Random forest clustering for discrete sequences," *Pattern Recognition Letters*, vol. 174, no. September, pp. 145–151, 2023, https://doi.org/10.1016/j.patrec.2023.09.001.

[15] M. Savargiv, B. Masoumi, and M. R. Keyvanpour, "A new random forest algorithm based on learning automata," *Computational Intelligence and Neuroscience*, vol. 2021, no. -, pp. 1–19, 2021, https://doi.org/10.1155/2021/5572781.

[16] S. Kumar, P. Kaur, and A. Gosain, "A Comprehensive Survey on Ensemble Methods," in *2022 IEEE 7th International conference for Convergence in Technology, I2CT 2022*, no. April, 2022, pp. 1–8, https://doi.org/10.1109/I2CT54291.2022.9825269.

[17] R. J. Janse, T. Hoekstra, K. J. Jager, C. Zoccali, G. Tripepi, F. W. Dekker, and M. Van Diepen, "Conducting correlation analysis: Important limitations and pitfalls," *Clinical Kidney Journal*, vol. 14, no. 11, pp. 2332–2337, 2021, https://doi.org/10.1093/ckj/sfab085.

[18] A. Nobi, K. H. Tuhin, and J. W. Lee, "Application of principal component analysis on temporal evolution of COVID-19," *PLoS ONE*, vol. 16, no. 12 December, pp. 1–12, 2021, https://doi.org/10.1371/journal.pone.0260899.

[19] S. P and K. Pothuganti, "Overview on Principal Component Analysis Algorithm in Machine Learning," *@International Research Journal of Modernization in Engineering*, vol. 02, no. 10, pp. 241–246, 2020.

[20] M. Greenacre, P. J. Groenen, T. Hastie, A. I. D'Enza, A. Markos, and E. Tuzhilina, "Principal component analysis," *Nature Reviews Methods Primers*, vol. 2, no. 1, pp. 1–24, 2022, https://doi.org/10.1038/s43586-022-00184-w.

[21] G. J. Oyewole and G. A. Thopil, *Data clustering: application and trends*. Springer Netherlands, 2023, vol. 56, no. 7, https://doi.org/10.1007/s10462-022-10325-y.

[22] K. A. Abbas, A. Gharavi, N. A. Hindi, M. Hassan, H. Y. Alhosin, J. Gholinezhad, H. Ghoochaninejad, H. Barati, J. Buick, P. Yousefi, R. Alasmar, and S. Al-Saegh, "Unsupervised machine learning technique for classifying production zones in unconventional reservoirs," *International Journal of Intelligent Networks*, vol. 4, no. October 2022, pp. 29–37, 2023, https://doi.org/10.1016/j.ijin.2022.11.007.

[23] R. Buaton and S. Solikhun, "The Application of Numerical Measure Variations in K-Means Clustering for Grouping Data," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 23, no. 1, pp. 103–112, 2023, https://doi.org/10.30812/matrik.v23i1.3269.

[24] F. Ros, R. Riad, and S. Guillaume, "PDBI: A partitioning Davies-Bouldin index for clustering evaluation," *Neurocomputing*, vol. 528, no. -, pp. 178–199, 2023, https://doi.org/10.1016/j.neucom.2023.01.043.

[25] B. Zagajewski, M. Kluczek, E. Raczko, A. Njegovec, A. Dabija, and M. Kycko, "Comparison of random forest, support vector machines, and neural networks for post-disaster forest species mapping of the krkonoše/karkonosze transboundary biosphere reserve," *Remote Sensing*, vol. 13, no. 2581, pp. 1–23, 2021, https://doi.org/10.3390/rs13132581.

[26] M. Aria, C. Cuccurullo, and A. Gnasso, "A comparison among interpretative proposals for Random Forests," *Machine Learning with Applications*, vol. 6, no. April, p. 100094, 2021, https://doi.org/10.1016/j.mlwa.2021.100094.

[27] A. D. Purwanto, K. Wikantika, A. Deliar, and S. Darmawan, "Decision Tree and Random Forest Classification Algorithms for Mangrove Forest Mapping in Sembilang National Park, Indonesia," *Remote Sensing*, vol. 15, no. 16, pp. 1–31, 2023, https://doi.org/10.3390/rs15010016.

[28] T. G. Pratama, R. Hartanto, and N. A. Setiawan, "Machine learning algorithm for improving performance on 3 AQ-screening classification," *Communications in Science and Technology*, vol. 4, no. 2, pp. 44–49, 2019, https://doi.org/10.21924/cst.4.2.2019.118.

[This page intentionally left blank.]