Indonesian Journal of Data and Science

Volume 4 Issue 2
ISSN 2715-9930
https://doi.org/10.56705/ijodas.v4i2.70

Research Article

Comparison of K-Nearest Neighbor and Decision Tree Methods using Principal Component Analysis Technique in Heart Disease Classification

Al Danny Rian Wibisono [1], Syahrul Hidayat [2], Humam Maulana Tsubasanofa Ramadhan [3], Eva Yulia Puspaningrum [4]

[1]Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, Indonesia, 20081010010@student.upnjatim.ac.id
[2]Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, Indonesia, 20081010076@student.upnjatim.ac.id
[3]Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, Indonesia, 20081010084@student.upnjatim.ac.id
[4]Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, Indonesia, evapuspaningrum.if@upnjatim.ac.id

Correspondence should be addressed to Al Danny Rian Wibisono; 20081010010@student.upnjatim.ac.id

Received 10 June 2023; Accepted 18 June 2023; Published 31 July 2023

© Authors 2023. CC BY-NC 4.0 (non-commercial use with attribution, indicate changes).
License: https://creativecommons.org/licenses/by-nc/4.0/ — Published by Indonesian Journal of Data and Science.

Abstract:

Heart disease has become a global health issue that can threaten anyone, regardless of age. Numerous research efforts have been made to develop classification methods that can aid in diagnosing heart disease. In this study, we compared two classification methods, namely K-Nearest Neighbor (KNN) and Decision Tree, by applying Principal Component Analysis (PCA) technique to the heart disease classification. The dataset used contains relevant clinical attributes. After analyzing the dataset and performing data preprocessing, we applied PCA to reduce the dataset's dimensions. PCA models with KNN and Decision Tree were implemented and evaluated using performance metrics such as Confusion Matrix, F1 Score, and Accuracy. The analysis results showed that the PCA model with Decision Tree outperformed the PCA model with KNN in terms of accuracy. The Decision Tree model successfully classified all data correctly, while KNN had some misclassifications. This research recommends using the PCA model with Decision Tree for heart disease classification with the best performance. However, further research with larger datasets is needed for a deeper understanding.

Keywords: K-Nearest Neighbor, Decision Tree, Principal Component Analysis, Penyakit Jantung, Analisis.
Dataset link:

1. Introduction

The heart is a vital organ essential for living beings, especially humans, as it functions to pump blood throughout the body to ensure proper blood circulation every day. However, the heart also poses a threat to humans when it experiences disorders or diseases. According to data from the World Health Organization (WHO), heart disease is the leading cause of death worldwide, with approximately 17.9 million deaths annually. Therefore, early and accurate diagnosis of heart disease is crucial to provide appropriate treatment and reduce the risk of more serious complications.

In the field of medicine, the use of classification methods becomes an effective approach to diagnose heart disease based on various clinical attributes. Two commonly used classification methods in medical data analysis are K-Nearest Neighbor (KNN) and Decision Tree. KNN is a method based on the concept of nearest neighbors, where a data point is classified based on the majority class of its nearest neighbors. On the other hand, Decision Tree is a method that uses a decision tree to represent decision rules based on the features of the data.

One of the challenges in medical data analysis is the high dimensionality of attributes. Numerous attributes can impact the performance and interpretation of classification results. Therefore, an effective dimension reduction technique, such as Principal Component Analysis (PCA), is needed to reduce data complexity while preserving relevant information. The use of PCA in heart disease classification can optimize the performance of the classification model, identify important features, and facilitate overall data analysis.

However, there is still a lack of research comparing the performance of KNN and Decision Tree using PCA technique in heart disease classification. Hence, this study aims to compare both methods in classifying heart disease using PCA technique. It is expected that this research will provide a deeper understanding of the effectiveness and advantages of each method in heart disease classification and the contribution of PCA technique in improving classification performance.

2. Method

The general stages of this research include problem identification, data collection, system design, coding, and drawing conclusions. The detailed research stages are as follows:

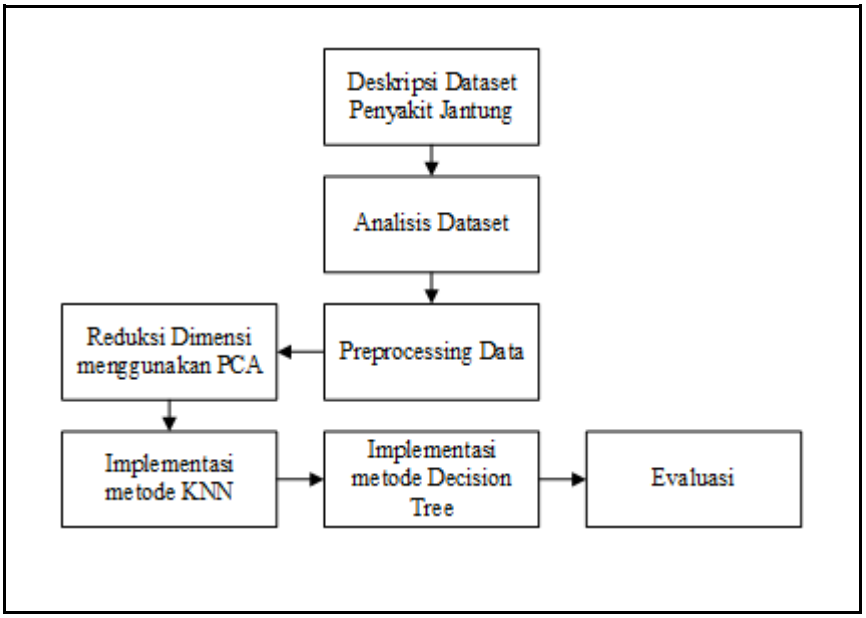

Figure1. Tahap Desain Penelitian Sistem yang dilakukan

Based on **Figure 1**, the stages of the research system design are as follows:

1. The first stage involves preparing the segmented heart disease dataset. The dataset is sourced from Kaggle and contains 1025 samples with 14 attributes/features. The explanation of the attributes/features used is as follows:
   - age: contains age data
   - sex: contains gender data
   - cp: contains brain paralysis data
   - trestbps: contains blood pressure data
   - chol: contains cholesterol data
   - fbs: contains blood sugar data
   - restecg: contains electrocardiogram data
   - thalach: contains maximum heart rate data
   - exang: contains chest pain data
   - oldpeak: contains ST segment depression data
   - slope: contains ST segment slope data
   - ca: contains cancer data
   - thal: contains heart status data
   - target: contains the target/variable to be predicted

2. The second stage involves analyzing the heart disease dataset, including the structure analysis, handling missing values, and obtaining statistical summaries of the dataset. The heart disease dataset consists of 1025 data samples without any missing values. The statistical summary of the heart disease dataset provides descriptive information about the numerical columns in the dataset. It gives an initial understanding of the data distribution and variation of values in each column. Some statistics included in the descriptive summary are as follows:
   - count: the number of non-null data in the column.
   - mean: the average value of data in the column.
   - std: the standard deviation, indicating the spread of data from the mean value.
   - min: the smallest value in the column.
   - quartiles (25%, 50%, 75%): values that divide the data into four equally sized groups. The 50th percentile (50%) represents the middle value or median.
   - max: the largest value in the column.
3. The third stage involves data preprocessing to prepare the dataset for classification. In this stage, the dataset is split into features (X) and variables (Y) and undergoes preprocessing using StandardScaler. The formula for preprocessing using StandardScaler is as follows:
   - Calculating the mean (average) of each feature in the dataset:
     mean = sum(x) / n
   - Calculating the standard deviation of each feature in the dataset:
     std_deviation = sqrt(sum((x - mean)^2) / n)
   - Standardizing each value in the features using the formula:
     x_scaled = (x - mean) / std_deviation

   In the formula above:

   - x: the original value of the feature in the dataset.
   - mean: the mean of the feature in the dataset.
   - std_deviation: the standard deviation of the feature in the dataset.
   - x_scaled: the standardized value of the feature in the dataset.

   By performing preprocessing using StandardScaler, the features in the dataset will have a mean of 0 and a standard deviation of 1. This helps improve the performance of the classification model, especially if the method used is sensitive to data scaling.

4. The fourth stage involves applying dimensionality reduction technique, namely Principal Component Analysis (PCA), to reduce the dimensions of the heart disease dataset. PCA is used to identify the principal components that contribute the most to data variance. Manually, PCA has the following procedural formulas and steps:
   - PCA aids in the normalization process, where each data variable is transformed to have a mean of zero and variance of one, using the formula:

     $$X' = \frac{x - \bar{x}}{\sigma}$$

     In the formula above:
     x' = normalized variable value
     x = original variable value
     $\bar{x}$ = mean of the variable
     σ = standard deviation of the variable

   - PCA calculates the covariance matrix with the formula
     - $Cov\ (X_i - X_j) = \frac{1}{n-1} \Sigma_{k=1}^{n} (X_{ki} - \bar{x}_i)\ (X_{kj} - \bar{x}_j)$

     dimana:
     $Cov\ (X_i - X_j)$ = kovarians $(X_i - X_j)$
     n = number of samples

X_{ki} = value of variable X_i in sample ke-k
$\bar{x}_i$ = mean value of variable X_i
$\bar{x}_j$ = mean value of variable X_j

- Next, PCA computes the eigenvalues and eigenvectors with the equation:

$$Cv = \lambda v$$

In the formula above:

C = Matriks Kovarians
v = Vektor Eigen
λ = Nilai Eigen

- Finally, PCA selects the principal components

5. The fifth stage involves classifying heart disease using the K-Nearest Neighbor (KNN) method. In this stage, the KNN model will be trained using the training data that has gone through the dimensionality reduction using PCA, and then used to make predictions on the testing data. Manually, KNN has the following procedural formulas and steps:
   - Menghitung jarak

$$d(x,y) = \sqrt{\Sigma_{i=1}^{n}\left(x_i - y_i\right)^2}$$

dimana :
d(x,y) = jarak antara data baru x dan data latih y
n = jumlah fitur atau atribut dalam data
x_i = nilai atribut ke-i dari data baru x
y_i = nilai atribut ke-i dari data baru y

   - Menentukan tetangga terdekat

   - Mengklasifikasikan data baru

6. The sixth stage involves classifying heart disease using the Decision Tree method. In this stage, the Decision Tree model will be trained using the training data that has gone through dimensionality reduction using PCA, and then used to make predictions on the testing data. Decision tree also has useful procedures and manual steps as follows:
   - Mengukur keberagaman (*Impurity Measure)*
     Pengukuran keberagaman memiliki 2 jenis yakni
     - Entropy
       A measurement of uncertainty in the data.

$$Entropy\ (D) = -\Sigma_{i=1}^{c} p_i log_2(p_i)$$

dimana :
Entropy (D) = entropy of the data set D
c = number of classes in the data
p_i = proportion of the number of samples belonging to class i to the total samples

     - Gini Index
       A measurement of the level of classification error if one object is randomly chosen.

$$Gini\ (D) = 1 - \Sigma_{i=1}^{c}\ p_i^2$$

where:
Gini (D) = Gini index of the data set D

c = number of classes in the data
p_i = proportion of the number of samples belonging to class i to the total samples

- Selecting the best feature (Best Split)
  The selection of the best feature has 2 types, namely:
  - Information Gain
    Measures the reduction of entropy achieved by splitting data based on a particular feature.

$$IG(D,F) = Entropy\ (D) - \Sigma_{v \epsilon Values(F)} \frac{|D_v|}{|D|} \cdot Entropy\ (D_v)$$

    where:
    IG(D,F) = information gain of the data set D by dividing it based on feature F
    Entropy (D) = entropy of the data set D
    $Values(F)$ = set of possible values of feature F
    D_v = subset of data D where feature F has value υ
    $|D_v|$ dan $|D|$ = number of samples in D_υ and D, respectively

  - Gini Gain
    Measures the reduction of Gini index achieved by splitting data based on a particular feature

$$GG(D,F) = Gini\ (D) - \Sigma_{v \epsilon Values(F)} \frac{|D_v|}{|D|} \cdot Gini\ (D_v)$$

    where: :
    GG(D,F) = *gini gain of the data set D by dividing it based on feature F*
    Gini (D) = *initial Gini index of the data set D*
    $Values(F)$ = set of possible values of feature F
    D_v = subset of data D where feature F has value υ
    $|D_v|$ dan $|D|$ = number of samples in D_υ and D, respectively

7. The seventh stage involves evaluating the performance of the KNN and Decision Tree models on the reduced heart disease classification data. Evaluation is done using evaluation metrics such as confusion matrix, f1-score, and accuracy. The goal of this stage is to compare the performance of the two methods and evaluate how effective they are in classifying heart disease based on the processed and reduced dataset using PCA.

Data Mining

Data Mining is a process of extracting or filtering valuable information from a large set of data through various processes to gain insights from the data. Data mining utilizes various methods such as classification, clustering, association, regression, prediction, forecasting, sequencing, and descriptive techniques, depending on the specific case being addressed.

K-Nearest Neighbor (KNN)

K-Nearest Neighbor is a machine learning algorithm used for classification or regression on a dataset. The KNN algorithm can be used on datasets with numerical or categorical values. The working principle of the KNN algorithm is to find the K nearest neighbors of a new data point to be predicted and then choose the majority class of those K nearest neighbors as the prediction for the new data point. If used for regression, the KNN algorithm predicts the numerical value of a new data point by taking the average of the K nearest neighbors' values.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a data analysis technique used in statistics and machine learning to reduce the dimensions of a dataset. The goal of PCA is to reduce the complexity of a dataset by projecting it into a lower-dimensional space while preserving the relevant information of the dataset. PCA looks for linear combinations of variables in the dataset that have high variance. These combinations are called principal

components and are used to represent the dataset in a lower-dimensional space. The first principal component captures the highest variance in the dataset, and subsequent principal components capture the variance that is not present in the previous principal components.

Decision Tree

Decision Tree is a machine learning algorithm that uses a tree-like structure to perform classification and regression. This algorithm solves problems by depicting decision rules in the form of a tree, where each branch represents a decision based on specific conditions. Decision tree works by dividing the dataset based on the most significant attribute to separate the data into different classes. Each branch in the tree represents a condition or rule that the data must satisfy. The results of each branch are predictions or decisions made by the algorithm.

3. Results and Discussion

Performance Analysis of K-Nearest Neighbor (KNN)

In the code that has been implemented, the K-Nearest Neighbors (KNN) method is applied for heart disease classification. From the analysis conducted, the following information is obtained:

1. The normalized dataset is divided into two sets, training and testing, with a proportion of 80:20 (80% data for training, 20% data for testing).
2. Classification is performed using the KNN method, and the following evaluation results are obtained:

a. Confusion Matrix:

- There are 71 data correctly classified as non-heart disease (true negative).
- There are 31 data falsely classified as non-heart disease (false positive).
- There are 9 data falsely classified as heart disease (false negative).
- There are 94 data correctly classified as heart disease (true positive).

b. F1 Score :

- The F1 score of 0.82 indicates a relatively good performance of the model in predicting heart disease classification. A high F1 score indicates a good balance between the model's precision and recall.

c. Accuracy :

- The accuracy of 80.49% indicates the percentage of successful classification of data. Although this accuracy is relatively high, it is essential to note that other evaluation results such as Confusion Matrix also provide important information about misclassifications.

Performance Analysis of Decision Tree

In the code that has been implemented, the Decision Tree method is applied for heart disease classification. From the analysis conducted, the following information is obtained:

1. The normalized dataset is divided into two sets, training and testing, with a proportion of 80:20 (80% data for training, 20% data for testing).
2. Classification is performed using the Decision Tree method, and the following evaluation results are obtained.

a. Confusion Matrix:

- There are 102 data correctly classified as non-heart disease (true negative).
- There are 0 data falsely classified as non-heart disease (false positive).
- There are 3 data falsely classified as heart disease (false negative).
- There are 100 data correctly classified as heart disease (true positive).

b. F1 Score:

- The F1 score of 0.99 indicates a relatively good performance of the model in predicting heart disease classification. A high F1 score indicates a good balance between the model's precision and recall.

c. Accuracy:

- The accuracy of 98.54% indicates the percentage of successful classification of data. Although this accuracy is relatively high, it is essential to note that other evaluation results such as Confusion Matrix also provide important information about misclassifications.

Performance Analysis of K-Nearest Neighbor (KNN) with PCA

In the code that has been implemented, PCA (Principal Component Analysis) is implemented, followed by the application of the K-Nearest Neighbors (KNN) method for heart disease classification. From the analysis conducted, the following information is obtained:

1. The normalized dataset, after dimensionality reduction using PCA, results in 2 principal components that represent linear combinations of the original features in the heart disease dataset. 'X_pca' represents the representation of the dataset that has been reduced in dimensions.
2. The 'X_pca' feature dataset is divided into two sets, training and testing, with a proportion of 80:20 (80% data for training, 20% data for testing).
3. Classification is performed using the KNN method, and the following evaluation results are obtained.

a. Confusion Matrix:

- There are 75 data correctly classified as non-heart disease (true negative).
- There are 27 data falsely classified as non-heart disease (false positive).
- There are 16 data falsely classified as heart disease (false negative).
- There are 87 data correctly classified as heart disease (true positive).

b. F1 Score:

- The F1 score of 0.80 indicates a relatively good performance of the model in predicting heart disease classification. A high F1 score indicates a good balance between the model's precision and recall.

c. Accuracy:

- The accuracy of 79.02% indicates the percentage of successful classification of data. Although this accuracy is relatively high, it is essential to note that other evaluation results such as Confusion Matrix also provide important information about misclassifications.

Performance Analysis of Decision Tree with PCA

In the code that has been implemented, PCA (Principal Component Analysis) is implemented, followed by the application of the Decision Tree method for heart disease classification. From the analysis conducted.

1. The normalization of the dataset followed by dimensionality reduction using PCA resulted in 2 principal components representing linear combinations of the original features in the heart disease dataset. 'X_pca' represents the reduced dimensionality representation of the dataset.
2. The 'X_pca' feature dataset is divided into two sets, training and testing, with a proportion of 80:20 (80% data for training, 20% data for testing).
3. Classification is performed using the K-Nearest Neighbors (KNN) method, and the following evaluation results are obtained.

a. Confusion Matrix:

- There are 102 data correctly classified as non-heart disease (true negative).
- There are 0 data falsely classified as non-heart disease (false positive).
- There are 0 data falsely classified as heart disease (false negative).
- There are 103 data correctly classified as heart disease (true positive).

b. F1 Score:

- Obtained an F1 score of 1.0. This score indicates excellent performance in classifying data into the correct classes.

c. Accuracy:

- Achieved an accuracy level of 100%. This indicates that the KNN model can correctly classify all testing data.

Comparison of KNN & Decision Tree Performance

Table 1. Comparison of Classification Model Performance

	KNN	**Decission Tree**	**PCA & KNN**	**PCA & Decision Tree**
Confusion Matrix	[[71 31] [9 94]]	[[102 0] [3 100]]	[[75 27] [16 87]]	[[102 0] [0 103]]

	KNN	Decission Tree	PCA & KNN	PCA & Decision Tree
F1 Score	0.82	0.99	0.80	1.0
Accuracy	80.49%	98.54%	79.02%	100.0%

K-Nearest Neighbor (KNN):

- **Confusion Matrix:** The matrix shows that the KNN model made some errors in classifying the testing data. There are 31 data from the negative class falsely classified as positive, and 9 data from the positive class falsely classified as negative.
- **F1 Score:** The obtained F1 score is 0.82, indicating that the KNN model has a good balance between precision and recall in predicting the target classes, but not as high as the Decision Tree model.
- **Accuracy:** The accuracy of the KNN model reaches 80.49%, meaning that the model successfully classifies most of the testing data correctly, but slightly lower compared to the Decision Tree model.

Decision Tree:

- **Confusion Matrix:** The matrix shows that the Decision Tree model made some errors in classifying the testing data. There are 3 data from the positive class falsely classified as negative.
- **F1 Score:** The obtained F1 score is 98.5365, indicating that the Decision Tree model has a good balance between precision and recall in predicting the target classes.
- **Accuracy:** The accuracy of the Decision Tree model reaches 98.54%, meaning that the model successfully classifies most of the testing data correctly.

PCA & K-Nearest Neighbor (KNN):

- **Confusion Matrix:** The matrix shows that the implementation of PCA on the KNN model made some errors in classifying the testing data. There are 27 data from the negative class falsely classified as positive, and 16 data from the positive class falsely classified as negative..
- **F1 Score:** The obtained F1 score is 0.8018, indicating that the KNN model with PCA has a good balance between precision and recall in predicting the target classes, but not as high as the Decision Tree model.
- **Accuracy:** The accuracy of the KNN model with PCA reaches 79.02%, meaning that the model successfully classifies most of the testing data correctly, but slightly lower compared to the Decision Tree model.

PCA & Decision Tree:

- **Confusion Matrix:** The matrix shows that the implementation of PCA on the Decision Tree model perfectly classifies all testing data correctly, with no errors in classification.
- **F1 Score:** The obtained F1 score is 1.0, indicating that the Decision Tree model with PCA has an excellent balance between precision and recall in predicting the target classes.
- **Accuracy:** The accuracy of the Decision Tree model with PCA reaches 100%, meaning that the model successfully classifies all testing data correctly.

4. Conclusion

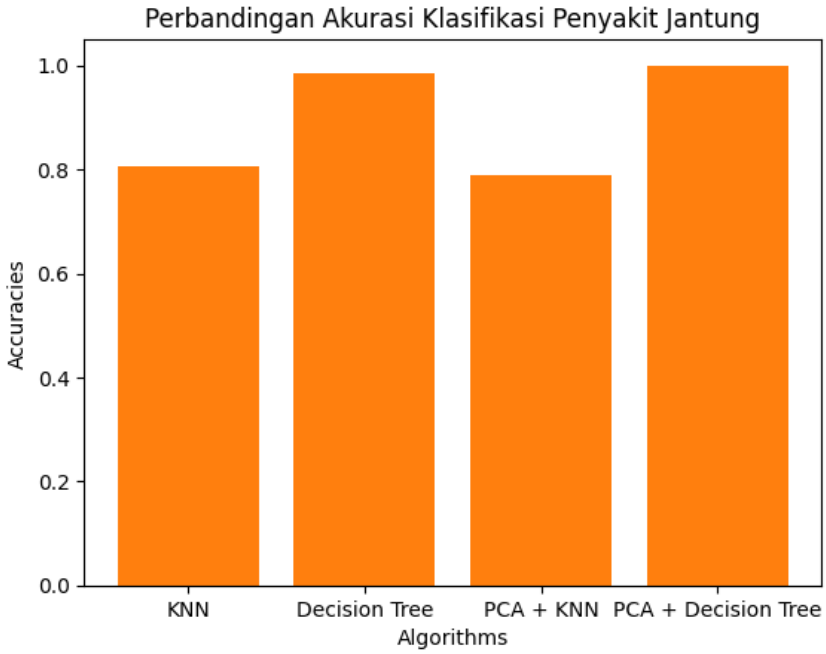

Figure 2. Comparison of Classification Model Accuracy Graph

Based on the analysis of the performance comparison between the KNN and Decision Tree models on the testing of the heart disease dataset, it can be concluded that the Decision Tree model provides better performance with high F1 scores and higher accuracy. The implementation of PCA on the Decision Tree model resulted in excellent performance, achieving 100% accuracy and an F1 score of 1.0. However, the implementation of PCA on the KNN model still yields lower performance compared to the Decision Tree model. It can be said that the dimension reduction of PCA can help improve accuracy in the Decision Tree model, but it is less effective in the KNN model, as it actually reduces accuracy. Therefore, if the main goal is to predict heart disease classification with high accuracy, then the Decision Tree model with or without PCA is a better choice. However, it is essential to note that model performance may vary depending on the dataset and parameters used.

Acknowledgments

The author expresses heartfelt gratitude to the Supervisor of the Data Mining course, Mrs. Eva Yulia Puspaningrum, S.Kom, M.Kom, who provided various suggestions and inputs in completing this article.

References

[1] F. C. Anggoro, "Penerapan Metode K-Nearest Neighbour Untuk Menganalisis Investasi Budidaya Lobster Air Tawar Berbasis Web," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. Vol. 3 No. 1, no. 106, pp. 104-109, 2019.

[2] N. M. Pane, M. S. S. Umam and F. N. Fauziah, "Perancang Sistem Pakar Diagnosis Kerusakan Perangkat Keras Menggunakan Pohon Keputusan," *Jurnal METHODIKA*, vol. Vol. 6 No. 2, no. 30, pp. 29-33, 2020.

[3] A. W. Alwi and A. Sauddin, "Pengembangan Media Pembelajaran Analisis Komponen Utama Berbasis Web Menggunakan Shiny R," *jurnal Matematika dan Statistika serta Aplikasinya*, vol. Vol.10 No. 2, no. 64 , pp. 63-72, 2022.

[4] R. Kosasih, "Kombinasi Metode Isomap Dan Knn Pada Image Processing Untuk Pengenalan Wajah," *CESS (Journal of Computer Engineering System and Science)*, vol. Vol. 5 No. 2, pp. 166-170, 2020.

[5] D. Cahyanti, A. Rahmayanti and S. A. Husniar, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indonesian Journal of Data and Science*, Vols. Vol 1, No 2, pp. 39-43, 2020.

[6] D. G. Pradana, M. L. Alghifari, M. . F. Juna and S. D. Palaguna, "Klasifikasi Penyakit Jantung Menggunakan Metode Artificial Neural Network," *Indonesian Journal of Data and Science (IJODAS)*, Vols. Vol 3, No 2, pp. 55-60, 2022.

[7] D. P. Utomo, P. Sirait and R. Yunis, "Reduksi Atribut Pada Dataset Penyakit Jantung dan Klasifikasi Menggunakan Algoritma C5.0," *Jurnal Media Informatika Budidarma*, Vols. Volume 4, Nomor 4, pp. 994-1006, 2020.

[8] Derisma, "Perbandingan Kinerja Algoritma untuk Prediksi Penyakit Jantung dengan Teknik Data Mining," *Journal of Applied Informatics and Computing (JAIC)*, Vols. Vol.4, No.1, pp. 84-88, 2020.

[9] H. Azis, P. F. Fattah and I. P. Putri, "Performa Klasifikasi K-NN dan Cross-validation pada Data Pasien Pengidap Penyakit Jantung," *ILKOM Jurnal Ilmiah*, vol. Vol. 12 No. 2, pp. 81-86, 2020.

[10] L. Andiani, S. and D. P. Rini, "Analisis Penyakit Jantung Menggunakan Metode KNN Dan Random Forest," *Annual Research Seminar (ARS) 2019 Fakultas Ilmu Komputer UNSRI*, vol. Vol.5 No.1, pp. 165-169, 2019.

[11] F. Handayani, K. S. Kusuma, H. L. Asbudi, R. . G. Purnasiwi, R. Kusuma, A. Sunyoto and W. M. Pradnya, "Komparasi Support Vector Machine, Logistic Regression Dan Artificial Neural Network dalam Prediksi Penyakit Jantung," *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, vol. Vol. 7 No. 3, pp. 329-334, 2021.

[12] Sahar, "Analisis Perbandingan Metode K-Nearest Neighbor dan Naïve Bayes Classiffier pada Data Set Penyakit Jantung," *Indonesian Journal of Data and Science (IJODAS)*, Vols. Vol 1, No 3, pp. 79-86, 2020.

[13] N. M. Sunariadi, S. N. Fadilah and D. C. R. Novitasari, "Analisis Resiko Kanker Serviks Menggunakan PCA-ANFIS Berdasarkan Historical Medical Record," *JURNAL MEDIA INFORMATIKA BUDIDARM*, vol. 6, pp. 1349-1355, 2022.

[14] A. Islamiyati, S. Sahriman and S. Oktoni, "Studi Longitudinal Pada Analisis Data Gula Darah Pasien Diabetes melalui Principal Component Analysis," *Jambura Journal Of Mathematics*, vol. 4, pp. 41-49, 2022.

[15] S. Mutmainah, "Penanganan Imbalance Data Pada Klasifikasi Kemungkinan Penyakit Stroke," *Jurnal SNATi*, vol. 1, pp. 10-16, 2021.

[16] A. L. Unihehu and I. Suharjo, "Klasifikasi Jenis Ikan Berbasis Jaringan Saraf Tiruan Menggunakan Algoritma Principal Component Analysis (PCA)," *Jurnal Ilmiah Ilmu Komputer*, vol. 7, pp. 27-32, 2021.