JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 1, Desember 2024, hlm. 9 – 16 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i1.3498
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

HEART DISEASE RISK PREDICTION: EVALUATING MACHINE LEARNING ALGORITHMS WITH FEATURE REDUCTION USING LDA

Nurliana Nasution[1*], Feldiansyah Nasution[1], Mhd Arief Hasan[1]
[1]Informatic Enginering Study Program, Universitas Lancang Kuning
email: *nurliananst@unilak.ac.id

Abstract: Heart disease is one of the leading causes of death worldwide, making early detection and accurate diagnosis crucial for reducing mortality rates and improving patient outcomes. This study aims to evaluate the effectiveness of four machine learning algorithms—Logistic Regression, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—in predicting heart disease, with a focus on enhancing model performance using Linear Discriminant Analysis (LDA) for feature reduction. Among the models, SVM achieved the highest accuracy at 84.24%, followed by Logistic Regression at 83.70%. Although Random Forest and KNN showed lower accuracies, all models benefited from LDA's dimensionality reduction. This study suggests that SVM, combined with LDA, offers an optimal solution for early and accurate heart disease prediction in the healthcare industry.

Keywords: feature reduction; heart disease; linear discriminant analysis (LDA); machine learning; SVM

Abstrak: Penyakit jantung merupakan salah satu penyebab utama kematian di seluruh dunia, sehingga deteksi dini dan diagnosis yang akurat sangat penting untuk menurunkan angka kematian dan meningkatkan hasil pengobatan pasien. Penelitian ini bertujuan untuk mengevaluasi efektivitas empat algoritma pembelajaran mesin—Regresi Logistik, Random Forest, Support Vector Machine (SVM), dan K-Nearest Neighbors (KNN)—dalam memprediksi penyakit jantung, dengan fokus pada peningkatan kinerja model menggunakan Analisis Diskriminan Linear (LDA) untuk reduksi fitur. Di antara model yang diuji, SVM mencapai akurasi tertinggi sebesar 84,24%, diikuti oleh Regresi Logistik dengan 83,70%. Meskipun Random Forest dan KNN menunjukkan akurasi yang lebih rendah, semua model memperoleh manfaat dari reduksi dimensi yang diberikan oleh LDA. Studi ini menunjukkan bahwa SVM yang dikombinasikan dengan LDA merupakan solusi optimal untuk prediksi penyakit jantung secara dini dan akurat dalam industri kesehatan.

Kata kunci: linear discriminant analysis (LDA); machine learning; penyakit jantung; reduksi fitur; SVM.

INTRODUCTION

Heart infection is one of the driving causes of passing around the world. Concurring to the World Wellbeing Organization (WHO), millions of individuals kick the bucket each year due to cardiovascular illnesses, counting coronary

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 1, Desember 2024, hlm. 9 – 16 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i1.3498
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

heart illness and stroke[1]. Early and accurate diagnosis of this condition is essential to reduce mortality rates and improve patient's quality of life. However, diagnosing heart disease is often challenging due to its complex nature, requiring a series of expensive and invasive medical tests. In this context, there is an urgent need for more efficient and accurate solutions to support the diagnostic process.

With the progression of techn ology, machine learning has appeared to have significant potential in helping the diagnosis of heart illness by analyzing quiet information[2], [3], [4]. Numerous past things have utilized different machine learning calculations, such as Calculated Relapse, Irregular Woodland, Bolster Vector Machine (SVM), and K-Nearest Neighbors (KNN), to anticipate the chance of heart infection. Be that as it may, one of the biggest challenges confronted in actualizing machine learning is the choice of relevant highlights, as unessential highlights can diminish demonstrate execution and lead to overfitting [5], [6], [7].

LDA is known for maximizing the partition between classes, but its application as a include choice strategy in heart illness forecast has not been broadly talked about [8], [9]. This creates a gap in research, particularly in comparing the performance of machine learning models with and without the use of LDA. This study's uniqueness lies in applying LDA for dimensionality reduc tion in heart disease prediction. It evaluates the effectiveness of LDA as a feature selection method and compares the performance of four machine learning algorithms—Logistic Regression, Random Forest, SVM, and KNN—to identify the best combination for improving prediction accuracy.

METHOD

This study utilizes machine learning to predict heart disease based on the Heart Disease UCI dataset from Kaggle, starting with data collection, prepro cessing, and dimensionality reduction using Linear Discriminant Analysis (LDA). Four machine learning algorithms—Logistic Regression, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—were evaluated using accuracy, precision, recall, and F1-Score, with and without LDA application.

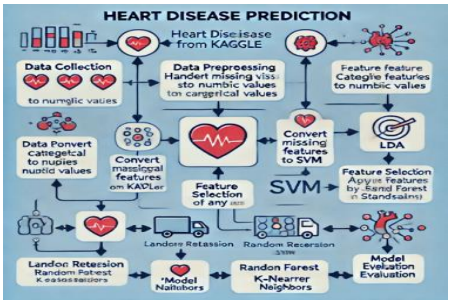

Image 1. Research Method

Table 1. Dataset Features

No	Nama Fitur	Tipe Data
1	Age	Numeric
2	Sex	Categorical
3	ChestPainType	Categorical
4	RestingBP	Numeric
5	Cholesterol	Numeric
6	FastingBS	Categorical
7	RestingECG	Categorical
8	MaxHR	Numeric
9	ExerciseAngina	Categorical
10	Oldpeak	Numeric
11	ST_Slope	Categorical
12	NumVessels	Numeric
13	Thal	Categorical
14	HeartDisease	Categorical

Data Collection

The dataset used in this study is the publicly available Heart Disease UCI dataset from Kaggle, containing patient

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 1, Desember 2024, hlm. 9 – 16 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i1.3498
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

data with various risk factors related to heart disease. It includes 14 features representing clinical properties such as age, sex, blood pressure, cholesterol levels, and electrocardiogram results, and one target column indicating the presence or absence of heart disease. This dataset was selected for its broad applicability in modeling heart disease risks and frequent use in machine learning research.

Data Preprocessing

Data preprocessing is a crucial step to ensure the information used in machine learning models is in an optimal format. Key steps include cleaning and preparing the data before applying predictive algorithms.:

Handling Missing Values:

The dataset used in this research has no missing values, so no imputation was necessary. However, if missing values were present, mean or median imputation could be applied using the formula:

$$x_{new} = \frac{1}{N}\sum_{i=1}^{n} x_i \quad (1)$$

where x_{new} is the replacement value (mean or median) and n is the number of samples with available values.

Conversion of Categorical Data to Numeric

Categorical features such as Sex, ChestPainType, RestingECG, Exercise Angina, and ST_Slope were converted into numeric form using Label Encoding techniques. For example, the Sex feature was encoded as 1 for males and 0 for females. This label encoding allows machine learning algorithms to process categorical features.

Conversion Examples:

$$Sex = \begin{cases} 1 & \text{If Man} \\ 2 & \text{If Woman} \end{cases} \quad (2)$$

Feature Normalization:

Normalization is essential for scaling features, especially for algorithms like K-Nearest Neighbors (KNN) and Support Vector Machine (SVM), which are sensitive to feature scaling. StandardScaler was used in this study, transforming the data to have a mean (μ) of 0 and a standard deviation (σ) of 1 using the formula::

$$z = \frac{x - \mu}{\sigma} \quad (3)$$

where z is the normalized include esteem, x is the initial esteem, μ is cruel, and σ is the standard deviation of the highlight.

Linear Discriminant Analysis (LDA)

After normalization, Linear Discriminant Analysis (LDA) was applied to perform dimensionality reduction and feature selection. LDA projects the original features onto a new space with components that maximize the separation between the two target classes. LDA uses the following formula:

$$w = S\frac{-1}{w}(m_1 - m_2) \quad (5)$$

where w is the weight vector, S_w is the within-class scatter matrix, and m_1 and m_2 are the means of the two classes. LDA is used to reduce data complexity and improve model accuracy by focusing on the most relevant features.

Machine Learning Algorithms

This investigation utilizes four fundamental machine learning calculations to anticipate heart maladies: Calculated Relapse, Irregular Timberland, Bolster Vector Machine (SVM), and K-Nearest Neighbors (KNN). Each calculation contains a distinctive approach to handling information and making fore-

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 1, Desember 2024, hlm. 9 – 16 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i1.3498
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

casts, as depicted underneath.

Logistic Regression

Calculated Relapse is an algorithm utilized to demonstrate the likelihood of a twofold result (e.g., having or not having heart illness). This calculation calculates the likelihood of an occasion based on the logit work (logarithm of the chances), which is communicated as:

$$P(y = 1 \mid x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}} (6)$$

Where $P(y = 1 \mid x)$ is the probability that the patient has heart disease, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of each feature $x_1, x_2, \ldots, x_n$ and e is the exponential constant. Logistic Regression is used because it is effective for binary classification, such as predicting whether a patient has heart disease or not.

Random Forest

Arbitrary Timberland is a gathering calculation composed of different choice trees. This calculation works by building a few choice trees autonomously amid the preparing handle and after that combining the comes about of these trees (through voting or averaging) to create a last forecast [10], [11]. The basic formula for a decision tree in Random Forest is:

$$\text{Gini} = 1 - \sum_{i=1}^{n} P_i^2 \quad (7)$$

where the Gini Index measures the impurity at a given node in the tree, and P_i is the probability of an observation falling into a specific class[12], [13]. Random Forest uses the average result from multiple decision trees to reduce the risk of overfitting and improve accuracy.

Support Vector Machine (SVM)

SVM is an algorithm that searches for the optimal hyperplane that separates two classes in the feature space[10], [14], [15]. This algorithm works by maximizing the margin between the data points of the two classes. The basic formula for the hyperplane in SVM is:

$$w^T x + b = 0 \quad (8)$$

After normalization, Linear Discriminant Analysis (LDA) is applied for dimensionality reduction and feature selection. LDA projects the original features into a new space with components that maximize the separation between the two target classes. LDA uses the following formula:.

K-Nearest Neighbors (KNN):

KNN is an instance-based calculation that works by classifying unused information focuses based on the lion's share lesson of their k closest neighbors. The calculation calculates the separation between the unused information point and all the focuses within the preparing dataset, at that point decides the course based on the larger part course of the k closest neighbors [16]. The formula used to calculate the Euclidean distance between two points $A(x_1, y_1)$ and $B(x_2, y_2)$ is:

$$d(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (9)$$

After calculating the distance, the algorithm identifies the k nearest neighbors, and the most frequent class among those neighbors becomes the prediction for the new data point.

RESULT AND DISCUSSION

This chapter presents the results of applying machine learning algorithms to predict heart disease, including data visualization, preprocessing, feature selection with LDA, and classification using four models. The models' performance is evaluated using accuracy, precision, recall, F1-score, and confusion matrix, with comparisons made between

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 1, Desember 2024, hlm. 9 – 16 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i1.3498
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

models with and without LDA to assess its impact on accuracy.

Data Distribution Visualization

This section presents the visualization of key features from the dataset to understand their distributions and detect potential patterns.

Age, Resting Blood Pressure, Cholesterol, Exercise-Induced Angina, and Heart Disease Correlations

The age distribution (Image 2) shows that most patients are between 50 and 60 years old, with fewer patients under 40 or over 70, indicating that middle-aged and older individuals are more likely to be represented in this dataset, which aligns with typical heart disease risk factors.

The resting blood pressure distribution (Image 3) shows that most patients have resting blood pressure values between 120 and 140 mm Hg, peaking around 130 mm Hg. A small portion of patients falls outside this range, with a few cases below 100 mm Hg or above 160 mm Hg. Similarly, the cholesterol distribution (Image 4) indicates that most patients have cholesterol levels between 200 and 300 mg/dl, though a notable number of zero values suggest missing data, with some outliers exceeding 400 mg/dl. The exercise-induced angina distribution (Image 5) shows that the majority of patients do not experience exercise-induced chest pain, while a smaller group does, making this a significant factor in diagnosing heart disease.

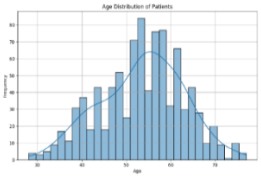

Image 2. Age Distribution

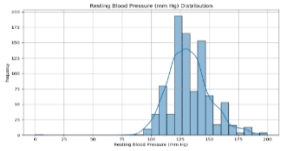

Image 3. Resting Blood Pressure Distribution (RestingBP)

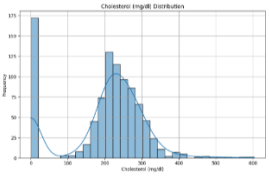

Image 4. Cholesterol Levels (Cholesterol):

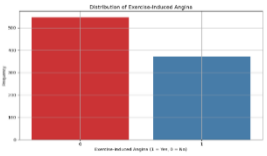

Image 5. Exercise-induced angina (Exercise Angina)

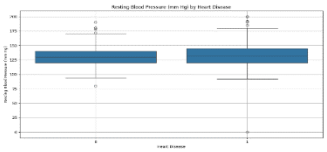

Image 6. Resting Blood Pressure Analysis by Heart Disease Status

Image 7. Scatter Plot Relationship Between Age and Maximum Heart Rate (MaxHR)

.

Resting Blood Pressure by Heart Disease Status and Relationship Between Age and MaxHR

The resting blood pressure analysis by heart disease status (Image 6) reveals that both patients with and without heart disease have similar median blood pressure levels. However, there is greater variability among patients with heart disease, as well as more outliers among those without heart disease, suggesting potential correlations between blood pressure variability and heart disease risk. In the relationship between age and maximum heart rate (MaxHR) (Image 7), a clear trend emerges: patients without heart disease generally have higher MaxHR values across age groups, while those with heart disease show lower MaxHR.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 1, Desember 2024, hlm. 9 – 16 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i1.3498
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Confusion Matrix Analysis After LDA

The confusion matrix shows the Logistic Regression model's performance after applying LDA, with 86 true positives and 68 true negatives accurately classified. However, the model produced 9 false positives and 21 false negatives, missing some heart disease cases. Despite its overall strong performance, the 21 false negatives highlight the need for further refinement to improve accuracy in detecting heart disease. Reducing false negatives is particularly crucial in medical diagnosis, as missed cases could lead to severe consequences for patients. Enhancing the model's sensitivity through additional feature selection or algorithm tuning could help mitigate this issue.

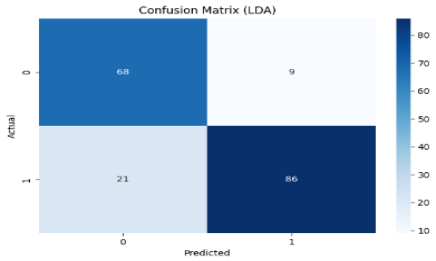

Image 8. Confusion Matrix Analysis After LDA

Classification Results Using Linear Discriminant Analysis (LDA)

After applying Linear Discrimi nant Analysis (LDA), the performance of four machine learning algorithms—Logistic Regression, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—was assessed. Logistic Regression achieved an accuracy of 83.70%, with balanced precision for both class 0 (0.76) and class 1 (0.91), making it effective for heart disease classification. Random Forest, with an accuracy of 77.72%, struggled with class 1 recall (0.72), indicating difficulties in identifying heart disease cases. KNN performed well in class 1 precision (0.89) but had limitations in distinguishing heart disease cases, achieving an accuracy of 81.52%.

Support Vector Machine (SVM) emerged as the top performer, with the highest accuracy of 84.24% and balanced precision and recall for both classes. Both SVM and Logistic Regression demons trated consistent reliability, whereas Random Forest lagged in both precision and recall, and KNN showed particular weakness in class 0 precision. Overall, SVM and Logistic Regression, when combined with LDA, proved to be the best models for heart disease classification in this dataset. These findings highlight the effectiveness of LDA in improving model performance for predicting heart disease.

Table 2. Classification Results

Model	Accuracy	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)
Logistic Regression	83.70%	0.76	0.91	0.88	0.80
Random Forest	77.72%	0.69	0.88	0.86	0.72
Support Vector Machine (SVM)	84.24%	0.78	0.90	0.87	0.82
K-Nearest Neighbors	81.52%	0.74	0.89	0.87	0.78

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 1, Desember 2024, hlm. 9 – 16 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i1.3498
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

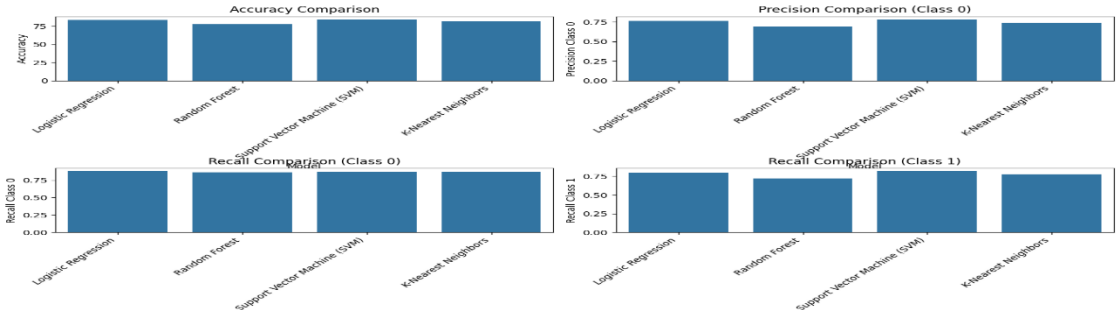

Image 9. Classification Results

CONCLUSION

In this study, four machine learning algorithms—Logistic Regression, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—were applied to predict heart disease using Linear Discriminant Analysis (LDA) for feature reduction. SVM achieved the highest accuracy at 84.24%, followed by Logistic Regression at 83.70%, while Random Forest and KNN performed slightly lower. The application of LDA improved model performance by simplifying features and retaining key information, with SVM and Logistic Regression emerging as the optimal models for heart disease prediction in this dataset.

ACKNOWLEDGMENTS

Gratitude is extended to the Faculty of Computer Science at Universitas Lancang Kuning for the financial support provided for this research through the 2023/2024 research funding program. This study would not have been possible without their encouragement and assistance. Appreciation is also given to all colleagues and individuals who contributed to the success of this research.

BIBLIOGRAPHY

[1] J. Heo, J. G. Yoon, H. Park, Y. D. Kim, H. S. Nam, and J. H. Heo, "Machine learning–based model for prediction of outcomes in acute stroke," *Stroke*, vol. 50, no. 5, pp. 1263–1265, 2019.

[2] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Comput. Sci.*, vol. 1, no. 6, p. 345, 2020, doi: 10.1007/s42979-020-00365-y.

[3] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart disease prediction using hybrid machine learning model," in *2021 6th international conference on inventive computation technologies (ICICT)*, 2021, pp. 1329–1333.

[4] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE access*, vol. 7, pp. 81542–81554, 2019.

[5] W. E. Pratiwi *et al.*, "Classification of Orange Fruit Using Convolutional Neural Network, Support Vector Machine, K-Nearest Neighbor and Naive Bayes Methods Based on Color Analysis," in *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, 2023, pp. 484–488. doi: 10.1109/ICCoSITE57641.2023.10127775.

[6] M. Sitompul, M. A. Hasan, and M. Devega, "Forecasting Simcard Demand Using Linear Regression Method," *J. Res. Dev.*, vol. 8, no. 1, 2023, doi: 10.25299/itjrd.2022.12202.

[7] N. Nasution, D. Setiawan, and M. A. Hasan, "PKM Sosialasi Aplikasi Pengelolaan Laboratorium Pertanian Fakultas Pertanian Universitas Lancang Kuning," 2021.

[8] P. M. and T. T. B. Xanthopoulos Petrosand Pardalos, "Linear Discriminant Analysis," in *Robust Data Mining*, New York, NY: Springer New York, 2013, pp. 27–33. doi: 10.1007/978-1-4419-9878-1_4.

[9] C. H. Park and H. Park, "A comparison of generalized linear discriminant analysis algorithms," *Pattern Recognit.*, vol. 41, no. 3, pp. 1083–1097, 2008, doi: https://doi.org/10.1016/j.patcog.2007.07.022.

[10] V. V Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques : A survey Heart disease prediction using machine learning techniques : a survey," no. August, 2019, doi: 10.14419/ijet.v7i2.8.10557.

[11] G. Biau and E. Scornet, "A Random Forest Guided Tour," Nov. 2015, [Online]. Available: http://arxiv.org/abs/1511.05741

[12] P. Palimkar, R. N. Shaw, and A. Ghosh, "Machine learning technique to prognosis diabetes disease: Random forest classifier approach," in *Advanced computing and intelligent technologies: proceedings of ICACIT 2021*, 2022, pp. 219–244.

[13] W. Xu, J. Zhang, Q. Zhang, and X. Wei, "Risk prediction of type II diabetes based on random forest model," in *2017 third international conference on advances in electrical, electronics, information, communi cation and bio-informatics (AEE ICB)*, 2017, pp. 382–386.

[14] G. Parthiban, "Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients," vol. 3, no. 7, pp. 25–30, 2012.

[15] T. Joachims, "SVMLight: Support Vector Machine," 2018. [Online]. Available: https://www.researchgate.net/publication/243763293

[16] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2888, pp. 986–996, 2003, doi: 10.1007/978-3-540-39964-3_62.