IT Journal Research and Development (ITJRD)
Vol.9, No.2, March 2025, E-ISSN : 2528-4053 | P-ISSN : 2528-4061
DOI : 10.25299/itjrd.2025.17941

Predicting Heart Disease Using Machine Learning: An Evaluation of Logistic Regression, Random Forest, SVM, and KNN Models on the UCI Heart Disease Dataset

Nurliana Nasution[1], Feldiansyah Bakri Nasution[2], Mhd Arief Hasan[3]
Informatic Engineering Study Program, Universitas Lancang Kuning [1,2,3]
nurliana@unilak.ac.id[1], feldiansyah@unilak.ac.id[2], m.arif@unilak.ac.id[3]

Article Info

Article history:

Received Jul 4, 2024
Revised Aug 21, 2024
Accepted Apr 24, 2025

Keyword:

Heart Disease Prediction
Machine Learning
Random Forest
SVM
UCI Dataset

ABSTRACT

This study evaluates the performance of three machine learning models—Random Forest, Support Vector Machine (SVM), and Logistic Regression—in predicting heart disease using the "Heart Disease UCI" dataset from Kaggle. The models were assessed based on accuracy, precision, recall, and F1-score, both with and without feature selection techniques such as Chi-Square and Mutual Information. Without feature selection, Random Forest achieved the highest performance with an accuracy of 89.7%, followed by SVM with 87.0% and Logistic Regression with 84.2%. Using Mutual Information for feature selection, Random Forest achieved an accuracy of 85.3%, SVM 87.0%, and Logistic Regression 82.6%. With Chi-Square feature selection, Random Forest and Logistic Regression both showed an accuracy of 83.2%, while SVM achieved 82.6%. The results indicate that Random Forest consistently performs well across different scenarios, making it a robust choice for heart disease prediction. Feature selection did not significantly enhance model performance, suggesting that the initial features in the dataset are already highly relevant. These findings highlight the potential of machine learning, especially Random Forest, in aiding the clinical diagnosis of heart disease. Further research is needed to validate these models on larger, more diverse datasets and to explore advanced feature selection techniques for improved model performance.

Corresponding Author:

Nurliana Nasution
Informatic Engineering Study Program
Univeritas Lancang Kuning
Rumbai, Pekanbaru, Indonesia
Email: nurliananst@unilak.ac.id

1. INTRODUCTION

Heart disease is one of the leading causes of death worldwide. According to the World Health Organization (WHO), millions of people die each year from cardiovascular diseases, including coronary heart disease and stroke. Early and accurate diagnosis of heart disease is crucial for reducing mortality rates and improving patients' quality of life. However, diagnosing heart

disease is often complex and requires various expensive and invasive medical tests. Therefore, there is an urgent need for more effective and efficient diagnostic methods.

In the current digital era, machine learning has emerged as a powerful tool for analyzing medical data and diagnosing disease. This technology enables the processing and analysis of large amounts of data quickly and accurately. Machine learning algorithms can uncover patterns and correlations that are not immediately apparent through traditional analysis. However, the main challenge in applying machine learning to heart disease diagnosis is the selection of relevant and optimal features. Irrelevant or excessive features can reduce the accuracy of the predictive model and increase computational complexity[1], [2][3].

This study presents a comprehensive analysis of the application of various feature selection techniques in machine learning models for heart disease prediction. The uniqueness of this research lies in the direct comparison between models with and without feature selection using Chi-Square and Mutual Information methods[4], [5]. The study aims to determine the most efficient feature selection technique by comparing these methods. This approach provides deep insights into how feature selection can affect model performance and helps choose the most relevant features for heart disease diagnosis.

This study aims to evaluate the performance of four machine learning algorithms—Logistic Regression, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—in predicting heart disease. The evaluation is conducted with and without feature selection, using two different feature selection methods. By doing this, the study seeks to understand how feature selection impacts model accuracy and complexity. Thus, this study aims to identify the most effective combination of algorithms and feature selection methods for heart disease diagnosis.

The benefits of this research are that it provides practical guidance for medical professionals and researchers in using machine learning for heart disease diagnosis. By understanding the impact of feature selection on model performance, this study can help develop more efficient and accurate diagnostic systems. The insights gained from this research can also aid in the early detection and treatment of heart disease, ultimately saving lives. Additionally, the results of this research can serve as a foundation for further studies in medical diagnosis using machine learning.

2. RESEARCH METHOD

This study aims to evaluate the performance of various machine learning algorithms in predicting heart disease, both with and without feature selection. The dataset used in this research was obtained from Kaggle and contains information about patients, including attributes such as age, gender, blood pressure, cholesterol levels, and several other indicators relevant to diagnosing heart disease.

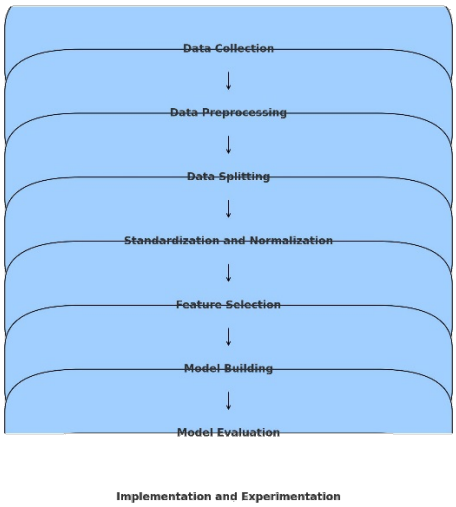


Figure 1. Research Method

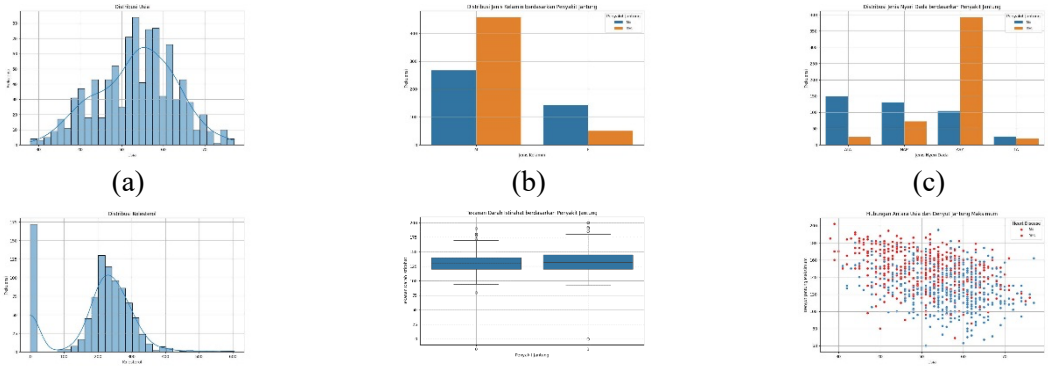
2.1 Data Gathering

The dataset used in this research is the "Heart Disease UCI" available on Kaggle. This dataset consists of 14 features encompassing various risk factors for heart disease and one target column indicating whether the patient has heart disease. These features include various clinical and historical attributes relevant to the diagnosis of heart disease. Below are the details of the features in the dataset:

Table 1. Dataset Features

No	Feature Name	Description	Data Type	Category
1	Age	Age of the patient	Numeric	
2	Sex	Sex of the patient (1 = male, 0 = female)	Categorical	(1 = male, 0 = female)
3	ChestPainType	Type of chest pain experienced by the patient (TA = typical angina, ATA = atypical angina, NAP = non-anginal pain, ASY = asymptomatic)	Categorical	(TA, ATA, NAP, ASY)
4	RestingBP	Resting blood pressure (mm Hg)	Numeric	
5	Cholesterol	Serum cholesterol level (mg/dl)	Numeric	
6	FastingBS	Fasting blood sugar (> 120 mg/dl, 1 = true, 0 = false)	Categorical	(1 = true, 0 = false)
7	RestingECG	Resting electrocardiographic results (Normal, ST = having ST-T wave abnormality, LVH = showing probable or definite left ventricular hypertrophy by Estes' criteria)	Categorical	(Normal, ST, LVH)
8	MaxHR	Maximum heart rate achieved	Numeric	
9	ExerciseAngina	Exercise-induced angina (1 = yes, 0 = no)	Categorical	(1 = yes, 0 = no)
10	Oldpeak	ST depression induced by exercise relative to rest	Numeric	
11	ST_Slope	Peak exercise ST segment slope (Up, Flat, Down)	Categorical	(Up, Flat, Down)
12	NumVessels	Number of major vessels (0-3) colored by fluoroscopy	Numeric	(0-3)
13	Thal	Thalassemia (Normal, Fixed Defect, Reversible Defect)	Categorical	(Normal, Fixed Defect, Reversible Defect)
14	heart disease	Whether the patient has heart disease (1 = yes, 0 = no)	Categorical	(1 = yes, 0 = no)

To understand the "Heart Disease UCI" dataset more deeply, exploratory data analysis (EDA) was performed through data visualization. This visualization helps identify patterns, trends, and anomalies that are not visible in tabular data form. Additionally, data visualization allows us to see the distribution of features, the relationships between features, and the influence of certain features on the target variable (heart disease).



(a) (b) (c)

(d) (e) (f)

Figure 2. Plot Data

The following are some of the charts used in this analysis:

- Age Distribution: A histogram showing the age distribution of patients in the dataset. It helps understand the age range most affected by heart disease.
- Sex Distribution: A bar chart showing the number of male and female patients. It helps see the difference in the prevalence of heart disease between genders.
- Chest Pain Type Distribution: A bar chart showing the number of patients based on the type of chest pain experienced (TA, ATA, NAP, ASY). It helps understand the most common type of chest pain in patients with heart disease.
- Cholesterol Distribution: A histogram showing the distribution of cholesterol levels in patients' blood. It helps identify cholesterol levels at high risk for heart disease.
- Resting Blood Pressure by Heart Disease: A box plot showing the distribution of resting blood pressure in patients based on whether they have heart disease. It helps understand the relationship between blood pressure and heart disease.
- Relationship Between Age and Maximum Heart Rate: A scatter plot showing the relationship between age and the maximum heart rate patients achieve. It helps see patterns or trends between these two features.

A correlation heatmap was used to understand the relationship between all the features in the dataset. This chart shows the correlation between the features, helping to identify features with strong relationships with the target variable or other features. A high correlation between certain features can provide insights into feature redundancy or relevance in the context of heart disease prediction. The correlation heatmap helps determine which features should be more closely examined or removed to improve model performance.

Figure 3. Heatmap of Correlations Between Variables

2.2 Data Preprocessing

The first step in data analysis is data preprocessing, which includes:

1. Handling Missing Values: Ensuring no missing values in the dataset. If there are, the missing values will be filled using appropriate methods (e.g., mean or median).
2. Converting Categorical Data to Numeric: Categorical columns such as sex (Sex), chest pain type (ChestPainType), resting ECG results (RestingECG), and others are converted to numeric format using label encoding.
3. Data Splitting: The dataset is split into training data (80%) and testing data (20%) to evaluate model performance. This split uses stratified sampling to balance the target class proportions between training and testing sets.
4. Standardization and Normalization: For some feature selection methods, like Chi-Square, data needs to be normalized to the range [0, 1] using MinMaxScaler. For other algorithms, features are standardized using StandardScaler to ensure all features are on the same scale.

2.3 Feature Selection

Feature selection is a crucial step in machine learning modeling aimed at choosing a subset of all available features so that the built model can achieve optimal performance. This process helps reduce model complexity, improve accuracy, and reduce the risk of overfitting. In this study, two different feature selection techniques are used to determine which features are most relevant in predicting heart disease:

This study uses two different feature selection techniques:

1. Chi-Square (χ^2) Test: Used to measure the dependency between features and the target class[4], [6]. The basic formula for chi-square is:

$$X^2 = \sum \frac{(0_i - E_i)^2}{E_i} \quad (1)$$

Where ($0_i$is the observed frequency and $E_i$is the expected frequency. This technique is particularly useful for categorical features as it can identify features with significant correlations with the target variable.

2. Mutual Information: Mutual Information Measures the dependency between two variables. The mutual information between two variables X and Y can be expressed as[7], [8][9]:

$$I\ (X;Y) = \sum p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (2)$$

where $p(x,y)$ is the joint distribution of X and Y, and $p(x)$ and $p(y)$ are the marginal distributions of X and Y. This technique is used for categorical and numerical features as it can measure the mutual information between features and the target variable, regardless of their distribution types.

2.4 Model Building

Model building is a critical stage in machine learning research where various algorithms make predictions based on processed data. This study uses four different machine learning algorithms to predict heart disease. Below is a detailed explanation of each algorithm:

1. **Logistic Regression:** This algorithm is used to model the probability of a binary event occurring. It calculates the target variable's logit (log odds) as a linear function of input features[10]. Logistic Regression is very useful for binary classification, such as determining whether a patient has heart disease. The basic formula for logistic regression is:

$$X^2 = \sum \frac{(0_i - E_i)^2}{E_i} \quad (3)$$

where Oi is the observed frequency, and Ei is the expected frequency. This technique is particularly useful for categorical features as it can identify features with significant correlations with the target variable.

2. **Random Forest:** A Random Forest is an ensemble algorithm with many decision trees. This algorithm works by building several decision trees during training and producing a class as the mode of the classes (classification) or the average prediction (regression) of the individual trees. Random Forest increases accuracy and reduces overfitting by combining the results of many different trees.
3. **Support Vector Machine (SVM)**: SVM is an algorithm that seeks the optimal hyperplane that separates classes in the feature space. This hyperplane is chosen to maximize the margin between the two different classes. SVM is very effective in high-dimensional spaces and cases where the number of dimensions exceeds the number of samples. SVM also uses the kernel trick to make more complex predictions by mapping data to a higher-dimensional space[11]. The basic formula for the hyperplane in SVM is:

$$\mathrm{w} \cdot \mathrm{X} + \mathrm{b} = 0 \quad (4)$$

where w is the weight vector, X is the input features, and b is the bias.

4. **K-Nearest Neighbors (KNN):** KNN is an algorithm that classifies data points based on the majority class of their k-nearest neighbors. This algorithm is very simple and intuitive, and it works by calculating the distance (usually Euclidean) between the unknown data point and all data points in the training set, then taking the majority class of the k-nearest neighbors to determine the class of the unknown data point[12]. The basic formula for the Euclidean distance between two points A and B is:

$$d(A,B) = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2 + \ldots (A_n - B_n)^n} \quad (5)$$

KNN is very useful for classification when the boundaries between classes are not linear and require a model that can adapt well to new data.

These four algorithms were chosen because each has unique advantages that can help in predicting heart disease with high accuracy[13], [14], [15]. This study evaluates the performance of each algorithm with and without feature selection to determine the best combination that can be used for heart disease diagnosis.

2.5 Model Evaluation

Model evaluation is an important stage in machine learning research to ensure that the built model performs well and can be relied upon to make predictions. The performance of each model is evaluated using the following metrics:

1. **Accuracy**: Measures the proportion of correct predictions out of the total predictions. Accuracy is calculated using the formula:

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \quad (6)$$

Here, TP is true positive, TN is true negative, FP is false positive, and FN is false negative. Accuracy provides a general overview of model performance but does not always reflect good performance on imbalanced datasets.
2. **Confusion Matrix**: A matrix that shows the number of correct and incorrect predictions for each class. The confusion matrix provides more detailed information on how the model makes predictions by displaying the number of TP, TN, FP, and FN. By examining the confusion matrix, we can understand where the model makes errors and which class predictions are more accurate.
3. **Precision, Recall, and F1-Score:** Used to provide deeper insights into model performance, especially on imbalanced datasets:
    - **Precision:** Measures the proportion of positive predictions that are actually positive. Precision is calculated using the formula::

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

High precision indicates that the model has few incorrect positive predictions (false positives).

- Recall: Measures the proportion of actual positives that are detected by the model. Recall is calculated using the formula:

$$Recall = \frac{TP}{TP + FN}$$

High recall indicates that the model successfully captures most of the actual positive data.

- F1-Score: It is the harmonic mean of precision and recall, providing a measure of the balance between them. The F1-Score is calculated using the formula:

$$F1 - Score = 2\,X \frac{Precision\ x\ Recall}{Precision + Recall}$$

A high F1-Score indicates that the model has a good balance between precision and recall, which is very important for imbalanced datasets.

The use of these metrics allows for a comprehensive assessment of the performance of machine learning models. This ensures that researchers can confirm that the model not only has high accuracy but is also capable of making reliable predictions under various dataset conditions. This evaluation helps in selecting the best model that can be used for a more accurate and reliable heart disease diagnosis.

2.6 Implementation and Experimentation

This study implements the above models using Python and the sci-kit-learn library. Experiments were conducted with and without feature selection to understand its impact on model performance. The results of these experiments were then analyzed and compared to determine the best combination of algorithms and feature selection techniques.

By this method, the study aims to identify the best machine learning model for heart disease prediction and understand the impact of feature selection techniques on model performance. The results of this research are expected to contribute to the development of more accurate and efficient diagnostic systems for heart disease.

3. RESULTS AND ANALYSIS

The results of the experiments conducted on the "Heart Disease UCI" dataset using four machine learning models—Random Forest, Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbors (KNN)—are presented and analyzed in this section. Each model's performance was evaluated with and without feature selection techniques such as Chi-Square and Mutual Information.

3.1 Without Feature Selection:

Random Forest demonstrated the highest performance with an accuracy of 89.7%. The precision, recall, and F1-score were 0.86, 0.90, and 0.88, respectively, indicating that the model can correctly identify both positive and negative cases of heart disease. SVM followed with an accuracy of 87.0%, precision of 0.84, recall of 0.84, and an F1-score of 0.84. Logistic Regression achieved an accuracy of 84.2%, with precision, recall, and F1-score being 0.77, 0.88, and 0.82, respectively. KNN showed an accuracy of 84.8%, with precision, recall, and F1-score being 0.79, 0.87, and 0.83, respectively:

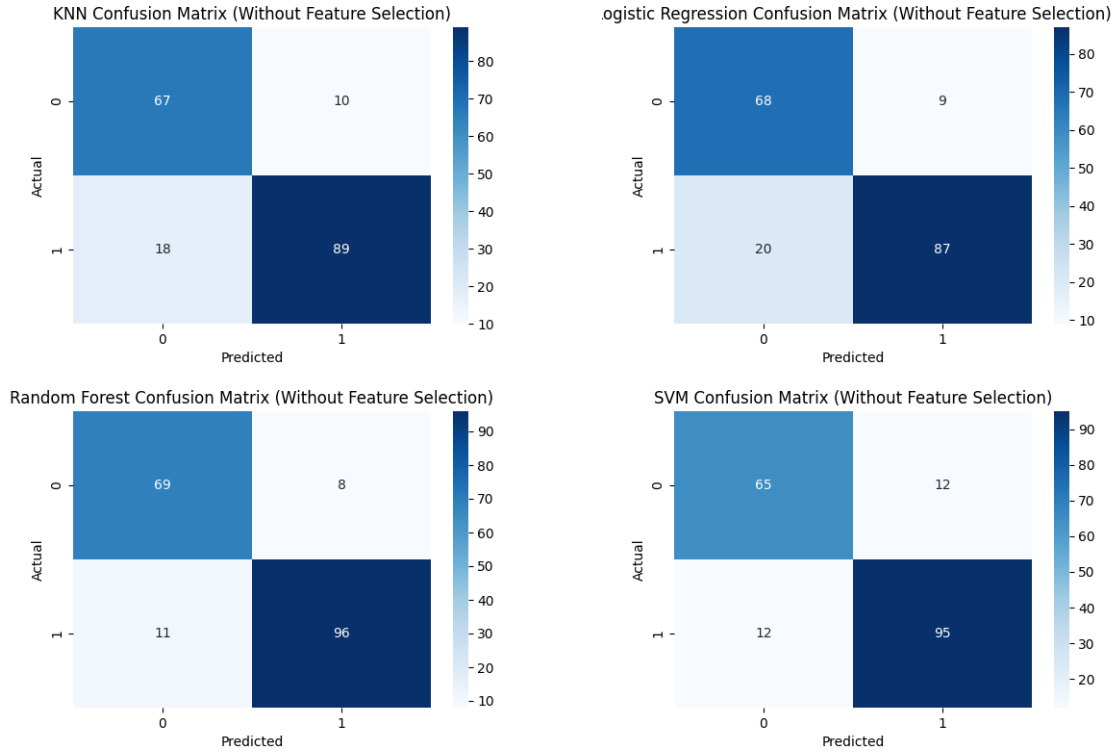
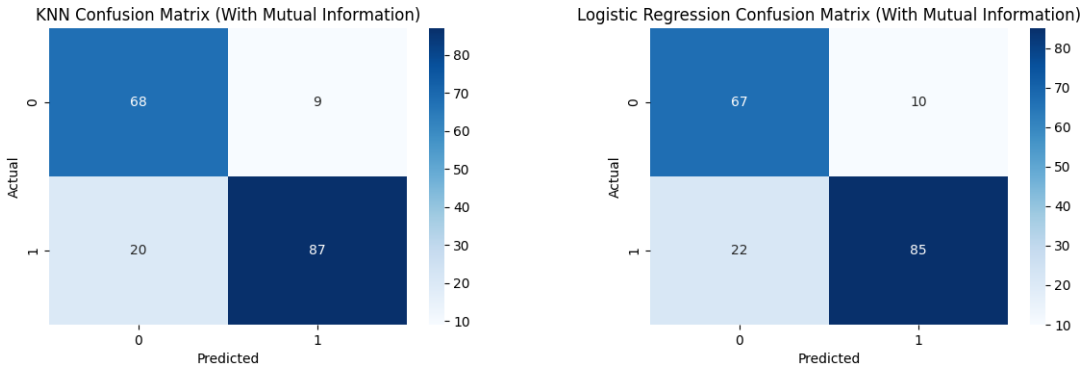


Figure 4. Without Feature Selection

3. 2 With Mutual Information Feature Selection

Using Mutual Information for feature selection, the Random Forest model achieved an accuracy of 85.3%, with precision of 0.80, recall of 0.87, and an F1-score of 0.83. SVM maintained a high performance with an accuracy of 87.0%, precision of 0.84, recall of 0.86, and an F1-score of 0.85. Logistic Regression showed a slight decrease in performance with an accuracy of 82.6%, precision of 0.75, recall of 0.87, and an F1-score of 0.81. KNN had an accuracy of 84.2%, with precision, recall, and F1-score being 0.77, 0.88, and 0.82, respectively.



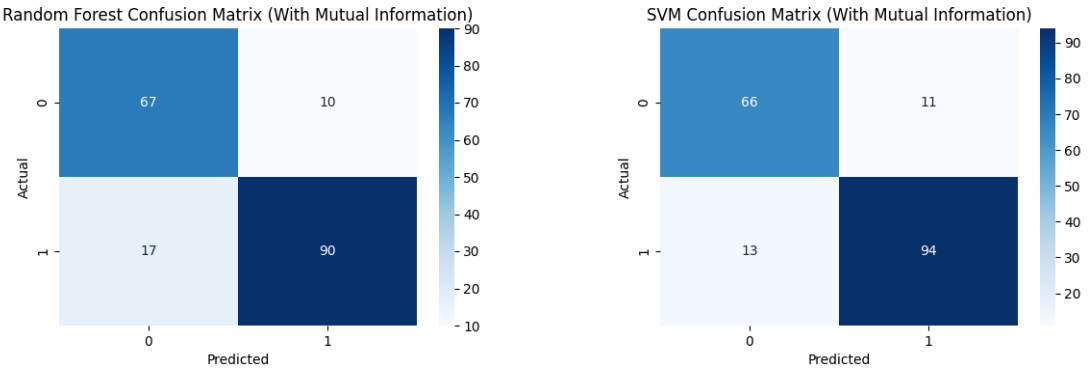


Figure 5. With Mutual Information Feature Selection

3. 3 With Chi-Square Feature Selection

When Using Chi-Square for feature selection, Random Forest and Logistic Regression both showed an accuracy of 83.2%. For Random Forest, the precision was 0.76, the recall was 0.87, and F1-score was 0.81. For Logistic Regression, the precision was 0.76, the recall was 0.87, and F1-score was 0.81. SVM had an accuracy of 82.6%, precision of 0.78, recall of 0.81, and an F1-score of 0.79. KNN showed an accuracy of 81.5%, with precision, recall, and F1-score being 0.75, 0.84, and 0.79, respectively.

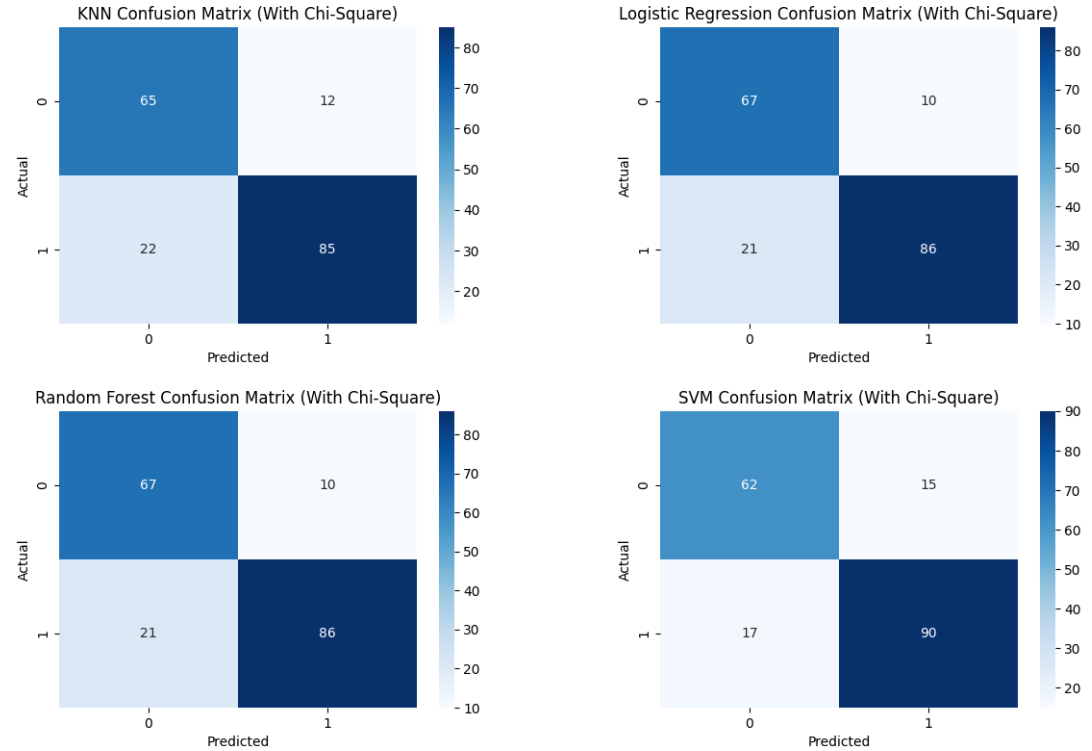


Figure 5. With Chi-Square Feature Selection

3. 4 Discussion of Findings

The results indicate that Random Forest consistently performs well across different scenarios, making it a robust choice for heart disease prediction. SVM also shows strong performance, particularly without feature selection and with Mutual Information. Logistic Regression and KNN exhibit reasonable performance but are slightly less effective compared to Random Forest and

SVM. Feature selection techniques such as Chi-Square and Mutual Information do not significantly enhance model performance, suggesting that the initial features in the dataset are already highly relevant for predicting heart disease.

In this study, feature selection did not provide significant benefits in improving model performance because the "Heart Disease UCI" dataset already contains features that are highly relevant and informative for heart disease prediction. These features adequately represent the influencing variables, making feature selection techniques like Chi-Square and Mutual Information unable to enhance model performance further. Additionally, these techniques have limitations in capturing more complex correlations between features and the target variable.

Furthermore, models like Random Forest can naturally handle overfitting without the need for feature selection, as it uses many decision trees, each relying on subsets of features. With a limited number of features (14 features) and features that are already relevant, feature selection becomes less necessary. The results of this study indicate that feature selection techniques do not significantly improve model performance for datasets with sufficiently representative features.

4. CONCLUSION

This study evaluated the performance of four machine learning models—Random Forest, Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbors (KNN)—in predicting heart disease using the "Heart Disease UCI" dataset from Kaggle. The results show that Random Forest consistently outperforms the other models, achieving the highest accuracy and a good balance between precision, recall, and F1-score. SVM also showed strong performance, while Logistic Regression and KNN performed slightly lower. Feature selection techniques such as Chi-Square and Mutual Information did not significantly enhance model performance, suggesting that the initial features in the dataset were already relevant.

The results are presented clearly, with each model's performance described before and after feature selection. The findings highlight the robustness of Random Forest, which maintains superior performance across all tested conditions. Although feature selection did not significantly improve performance, the data supports this conclusion, showing that the dataset's features were already optimal. The study suggests that further research should explore larger, more diverse datasets and investigate advanced feature selection techniques to improve prediction accuracy and efficiency.

ACKNOWLEDGEMENTS

We want to extend our heartfelt gratitude to the Faculty of Computer Science, Universitas Lancang Kuning, for their generous funding and support of this 2023/2024 research budget research project.

REFERENCES

[1] A. Tucker, Z. Wang, Y. Rotalinti, and P. Myles, “Generating high-fidelity synthetic patient data for assessing machine learning healthcare software,” *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–13, 2020.

[2] H. Habehh and S. Gohel, “Machine learning in healthcare,” *Curr. Genomics*, vol. 22, no. 4, p. 291, 2021.

[3] D. S. Char, M. D. Abràmoff, and C. Feudtner, “Identifying ethical considerations for machine learning healthcare applications,” *Am. J. Bioeth.*, vol. 20, no. 11, pp. 7–17, 2020.

[4] C. Shen, S. Panda, and J. T. Vogelstein, “The chi-square test of distance correlation,” *J. Comput. Graph. Stat.*, vol. 31, no. 1, pp. 254–262, 2022.

[5] N. S. Turhan, “Karl Pearson’s Chi-Square Tests.,” *Educ. Res. Rev.*, vol. 16, no. 9, pp. 575–580, 2020.

[6] Z. Wang, W. Chen, S. Gu, Y. Wang, and J. Wang, “Evaluation of trunk borer infestation duration using MOS E-nose combined with different feature extraction methods and GS-SVM,” *Comput. Electron. Agric.*, vol. 170, no. December 2019, p. 105293, 2020, doi: 10.1016/j.compag.2020.105293.

[7] M. Zhou, K. Yan, J. Huang, Z. Yang, X. Fu, and F. Zhao, "Mutual information-driven pan-sharpening," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1798–1808.
[8] H. Zhou, X. Wang, and R. Zhu, "Feature selection based on mutual information with correlation coefficient," *Appl. Intell.*, vol. 52, no. 5, pp. 5457–5474, 2022.
[9] J. D. Morgenstern *et al.*, "Predicting population health with machine learning: a scoping review," *BMJ Open*, vol. 10, no. 10, p. e037860, 2020.
[10] G. Biau and E. Scornet, "A Random Forest Guided Tour," Nov. 2015, [Online]. Available: http://arxiv.org/abs/1511.05741
[11] P.-H. Chen, C.-J. Lin, and B. Schölkopf, "A Tutorial on ν-Support Vector Machines."
[12] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, "Learning k for kNN Classification," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, Jan. 2017, doi: 10.1145/2990508.
[13] N. Nasution, F. Feldiansyah, A. Zamsuri, and M. A. Hasan, "Synthetic Minority Oversampling Technique for Efforts to Improve Imbalanced Data in Classification of Lettuce Plant Diseases," *J. Teknol. DAN OPEN SOURCE*, pp. 31–40, Feb. 2023, doi: 10.36378/jtos.v6i1.2883.
[14] 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.
[15] N. Nasution, F. B. Nasution, E. Erlin, and M. A. Hasan, "Evaluation Study of the Chi-Square Method for Feature Selection in Stroke Prediction with Random Forest Regression," EAI, 2024. doi: 10.4108/eai.30-10-2023.2343096.