Integrating Multiple Machine Learning Models to Predict Heart Failure Risk

Integrasi Model *Multiple Machine Learning* untuk Memprediksi Resiko Gagal Jantung

Tuahta Hasiholan Pinem[1], Yan Rianto[2]

[1,2] Ilmu Komputer, Universitas Nusa Mandiri Margonda, Indonesia

[1*]14220020@nusamandiri.ac.id, [2] yan.yrt@nusamandiri.ac.id

Informasi Artikel
Received: December 2023
Revised: January 2024
Accepted: January 2024
Published: June 2024

Abstract

The research aims to create and evaluate machine learning models for the prognosis of heart failure based on patient medical information. Various predictive models have been created employing algorithms like logistic regression, decision trees, random forests, K-nearest neighbors, naive Bayes, support vector machines (SVMs), neural networks, and ensemble voting classifiers. The dataset utilized comprises diverse clinical characteristics from patients diagnosed with heart failure. The data underwent division into training and testing sets in an 80:20 ratio. Metrics including accuracy, Cross Validation Score, and ROC_AUC Score score were used to assess the models' performance. The findings reveal that the Voting Classifier, amalgamating the Logistic Regression and Support Vector Classifier models, demonstrated superior performance with an accuracy of 88.04%, a cross-validation score of 91.01%, and a ROC_AUC score of 88.00%. Further scrutiny suggested that blood pressure and cholesterol levels serve as substantial indicators of heart failure. This study presents a notable advancement in the utilization of machine learning models for heart failure prediction by scrutinizing diverse algorithms and pinpointing the most pertinent clinical characteristics. These outcomes hint at the potential for the development of machine learning-driven clinical tools to facilitate early detection and enhance medical interventions.

Keywords: Machine Learning Models; Voting Classifier Algorithm; Feature Binning
Kata kunci: Model Machine Learning, Algoritma Voting Classifier, Pengelompokan Fitur

Abstrak

Penelitian bertujuan mengembangkan dan mengevaluasi model machine learning untuk memprediksi gagal jantung berdasarkan data medis pasien telah dilakukan. Model prediksi dibangun menggunakan algoritma *Logistic Regression, Decision Tree, Random Forest, K-Neares Neighbor, Naive Bayes, Support Vector Machine (SVM), Neural Network* dan *Voting Classifier*. Dataset yang digunakan mencakup berbagai fitur klinis dari pasien yang didiagnosis dengan gagal jantung. Data telah dibagi menjadi set pelatihan dan pengujian dengan rasio 80:20. Evaluasi model menggunakan metrik akurasi, Cross Validation Score, dan ROC_AUC Score untuk menilai kinerja masing-masing model. Hasil menunjukkan bahwa model Voting Classifier yang menggabungkan model Logistic Regression dan Support Vector Classifier menghasilkan kinerja terbaik dengan nilai akurasi sebesar 88.04%, Cross Validation Score sebesar91.01%, dan ROC_AUC Score sebesar 88.00%. Analisis lebih lanjut mengindikasikan bahwa fitur tekanan darah dan kadar kolesterol adalah prediktor yang signifikan untuk gagal jantung. Penelitian ini memberikan kontribusi signifikan dalam aplikasi model machine learning untuk memprediksi gagal jantung dengan mengevaluasi berbagai algoritma dan mengidentifikasi fitur klinis yang paling relevan. Hasil ini menunjukkan bahwa pengembangan alat klinis berbasis machine learning yang dapat mendukung deteksi dini dan intervensi medis yang lebih efektif.

1. Introduction

Heart failure (HF) poses a significant global health burden, with increasing prevalence associated with factors such as an aging population, improved survival rates from cardiovascular diseases, and lifestyle changes [1]. Frailty and high BMI notably heighten the risk of heart failure, with frail and prefrail individuals showing significantly higher risks, particularly when combined with obesity [2]. The early stages of heart failure in elderly patients elevate the risk of cardiovascular events such as hospitalization, ischemic heart disease, stroke, and all-cause mortality, with non-cardiovascular hospitalizations also being common [3]. Lifestyle factors significantly impact the risk of heart failure, with an unfavorable lifestyle increasing the risk by 2.90 times compared to a favorable lifestyle, regardless of metabolic or genetic risk status [4]

Traditionally, heart failure diagnosis relies on a combination of clinical evaluations, including characteristic symptoms and physical examination findings, to determine the presence and type of heart failure. The diagnostic process involves confirming the presence of heart failure, identifying the underlying cardiac dysfunction, and determining the etiology of the dysfunction. These methods are crucial in diagnosing heart failure based on symptoms such as dyspnea, orthopnea, and systemic edema, as well as objective evidence of cardiac dysfunction such as left ventricular systolic dysfunction (LVSD) [5]. However, traditional approaches have limitations in sensitivity and specificity, often necessitating additional tests for confirmation. Emerging technologies offer more precise and efficient methods for diagnosing heart failure, potentially revolutionizing its management in the future.

Machine learning techniques have been widely employed to predict heart failure risk. Various methods have demonstrated promising results in early detection and risk prediction of heart failure, with models like the HarT deep learning model showing significant improvements in predicting heart failure trajectories in patients with congenital heart disease [6]. The application of machine learning in heart failure research continues to evolve, focusing on enhancing diagnosis, prognosis, classification, and precision treatment for heart failure patients [7]. Studies have employed supervised learning approaches such as logistic regression, decision trees, Random Forest, support vector machines, K-Nearest Neighbors, and Naive Bayes to develop predictive models for early detection of heart failure, with Random Forest demonstrating the best performance. These studies highlight the effectiveness of machine learning in predicting heart failure risk and improving patient outcomes.

Classification and prediction models can aid the medical field by demonstrating how to efficiently utilize medical data. Research conducted by Fahd Saleh Alotaibi aimed to improve heart failure prediction accuracy using the UCI heart disease dataset. Various machine learning approaches were employed to understand the data and predict the likelihood of heart failure in the medical database. The study results indicated that the most effective models for detecting heart failure were SVM and Decision Tree, with accuracy rates of 92.30% and 93.19%, respectively [8].

Several previous studies on heart failure detection have utilized various machine learning techniques to predict heart disease using vocal sounds recorded during patient admission and discharge while pronouncing five Korean vowels ('a/e/i/o/u') for 3 seconds. Low-level audio features were extracted for classification. Mel-Spectrograms were then extracted and used as input features for deep learning models. Two types of deep learning-based classification models, convolutional neural networks and Transformers were adapted for analysis. The best-performing model, DenseNet201, achieved a classification accuracy of 85.11%. This accuracy increased to 92.76% with ViT-16-large after incorporating additional features of heart failure. Adding low-level audio features improved the classification task accuracy by approximately 2% on DenseNet201. These results propose the clinical potential of voice as a biomarker for early-stage ADHF detection [9].

In order to improve the accuracy of predictions and offer a thorough comparative examination, this research utilizes seven distinct machine learning algorithms: Logistic Regression, Support Vector Classifier (SVC), Decision Tree Classifier, Random Forest Classifier, K-Neighbors Classifier, Naive Bayes, and Neural Network. Additionally, a composite model incorporating a Voting Classifier is implemented, a technique not previously explored in the literature.

2. Method

The investigation was carried out by employing the Python programming language for code development. Furthermore, the researchers made use of various libraries offered by Python for the purpose of machine learning. The code was executed on Google's cloud services, leveraging Google’s graphics processors that accelerate the machine learning training process. The research workflow began with sourcing the dataset from the UCI Machine Learning Repository [10]. After acquiring the data, the dataset was divided into two types: training data used in the model-building process and test data used for model evaluation. The next step involved feature selection to reduce the dataset without diminishing the essential values used in classification. This feature selection also speeds up the classification process. The subsequent step was classification using several algorithms and evaluation of the resulting models. The complete research workflow can be seen in **Figure 1**.

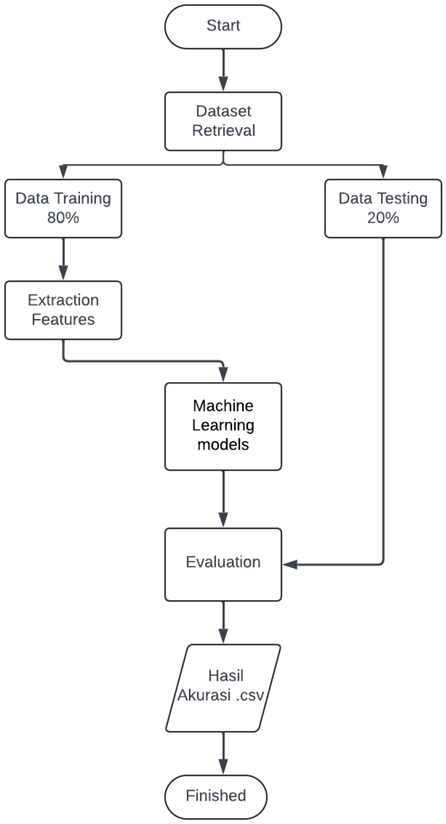


Figure 1. Flow Model

2.1. Dataset

The dataset for this research is sourced from the University of the Government College, Fedesoriano, and is freely accessible on the UCI Machine Learning Repository page. The dataset used consists of 918 records and 12 attributes, which are used in the measurements: Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope, and HeartDisease. The complete dataset along with attribute descriptions can be seen in **Table 1**.

Table 1. Dataset Descriptions

No	Attribute	Information
1	Age	Patient age [years]
2	Gender	Patient gender [M: Male, F: Female]
3	ChestPainType	type of chest pain [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Angina Pain, ASY: Asymptomatic]
4	RestingBP	RestingBP [mmHg]
5	Cholesterol	Serum cholesterol [mm/dl]

No	Attribute	Information
6	FastingBS	fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7	RestingECG	Resting electrocardiogram results [Normal: Normal, ST: has ST-T wave abnormalities (T wave inversion and/or ST elevation or depression > 0.05 mV), LVH: indicates possible or definite left ventricular hypertrophy based on Estes criteria]
8	MaxHR	Maximum heart rate achieved [Numerical value between 60 and 202]
9	ExerciseAngina	Angina caused by physical exercise [Y: Yes, N: No]
10	Oldpeak	oldpeak = ST [Numerical value measured in depression]
11	ST_Slope	Peak exercise ST segment slope [Up: uphill, Flat: flat, Down: downhill]
12	Heart Disease	Output class [1: heart disease, 0: Normal]

The dataset analyzed in this study provides detailed information on patients' health conditions relevant to heart disease. The data, consisting of several important attributes, include age (Age), gender (Sex), type of chest pain (ChestPainType), RestingBP (RestingBP), total cholesterol level (Cholesterol), fasting blood sugar (FastingBS), resting electrocardiogram results (RestingECG), maximum heart rate (MaxHR) achieved during exercise, presence of exercise-induced angina (ExerciseAngina), ST segment depression (Oldpeak), slope of the peak exercise ST segment (ST_Slope), and the heart disease diagnosis status (HeartDisease).

The age column records the patients' age in years at the time of data collection, while the gender column indicates the distribution between male ("M") and female ("F"). Chest pain type is labeled as Typical Angina (ATA), Non-anginal Pain (NAP), and Asymptomatic (ASY), whereas RestingBP and total cholesterol level are measured in standard units of mm Hg and mg/dL, respectively. Fasting blood sugar is categorized into two values: "0" for less than 120 mg/dL and "1" for 120 mg/dL or more.

Resting electrocardiogram results reflect the normal state or the presence of ST-T abnormalities. Maximum heart rate is recorded in bpm, while the presence of exercise-induced angina is indicated by "Y" for yes and "N" for no. ST segment depression and the slope of the ST segment at peak exercise provide additional information regarding the heart's response to physical activity.

The heart disease diagnosis column provides binary labels: "0" for patients without heart disease and "1" for those with heart disease. This dataset has the potential to support the development of machine learning models to predict the risk of heart disease based on recorded health parameters, enabling early diagnosis and more effective management of patients' heart conditions.

2.2. Dataset Division

Splitting the dataset involves dividing the obtained dataset into two types: training data and testing data. The training process aims to create a machine learning model, which is then used to classify the testing data. When developing a machine learning model, the training data is utilized for model construction, while the testing data from the overall dataset is used for evaluating the completed model. This separation is necessary to assess the classification

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v20i2. 13006

performance accurately. The data used for testing is kept separate from the training data to ensure that the model genuinely learns from new data during testing. The distribution of the dataset size used is 8:2, with 80% allocated for training data and 20% for testing data.

2.3. Data Classification

Classification is an analytical process that aims to categorize data within a dataset based on its type. The built classification model can be used to determine the class of labeled data. This model, known as a classifier, enables the identification of various classes within the dataset. In this study, the classification model is used to group the dataset into two main polarities: polarity 1 and polarity 0 .

The classification process is performed by applying several algorithms, including Logistic Regression, Support Vector Classifier (SVC), Decision Tree Classifier, Random Forest Classifier, K-Neighbors Classifier, Naive Bayes, and Neural Network. After obtaining the classification results, the authors compare the performance of these various algorithms to identify the most effective classification method. Additionally, the combination of models using a Voting Classifier is considered to enhance classification accuracy.

2.4. Feature Extraction

Feature extraction is the process of reducing the features in a dataset. This reduction is useful for refining data, speeding up computation time, and improving classification performance. The optimal feature selection process involves identifying the closest relationship between the features and the classification target. One approach used for feature extraction is binning, where data can be grouped into specific intervals, which helps in simplifying features and enhancing the interpretability of the classification model. This approach can reduce noise in the data and facilitate the analysis of the relationship between features and the target.

2.5. Machine Learning Models

Machine learning (ML) models play a crucial role in predicting cardiac outcomes, such as heart failure (HF) prognosis and heart disease status. Various ML models, including Random Forest, Gradient Boosting, Support Vector Machines, Neural Networks, Logistic Regression, Decision Trees, K-Nearest Neighbors (KNN), Gaussian Naive Bayes, among others, have been employed in numerous studies to predict HF readmission and mortality . These models take into account various predictive variables, such as age, sex, activity level, BMI, and medical history, to enhance accuracy in identifying individuals at high risk for heart disease. Studies have demonstrated that ML models can outperform traditional methods in predicting heart disease, with the integration of ML algorithms into the diagnosis and treatment of heart disease significantly improving patient outcomes and overall health.

2.5.1. Decision Tree Classifier

The Decision Tree classifier is a machine learning algorithm that has been extensively studied in various contexts. A decision tree is a predictive model used in machine learning to classify or predict a value based on decisions made at each node of the tree [11], as shown in **Figure 2**.

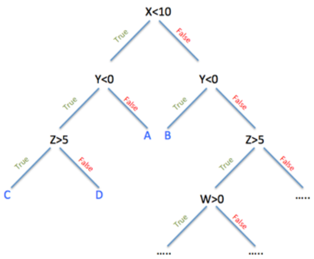


Figure 2. Decision Tree Classifier Stuctur

Figure 2 illustrates that a decision tree starts with an initial question or decision, such as "X < 10" in the example shown. Based on the answer to this condition (yes or no), the tree branches to the left (if "Yes") or to the right (if "No"). Each node in the tree represents a decision that evaluates a specific condition. Each decision leads to a more specific branch; in the example, it moves from "Y < 0" to "Z > 5" if the result is "No." This process further divides the data and narrows down the possible classifications or predictions, and so on.

2.5.2. Random Forest Classifier

Random Forest consists of multiple decision trees that work collectively to produce more accurate predictions. The algorithm's ability to handle high-dimensional features and effectively classify various types of data has been demonstrated in these diverse applications. Random Forest has shown a high level of accuracy in distinguishing distributions within stochastic systems [11]. The structure of the Random Forest is shown in Figure 3.

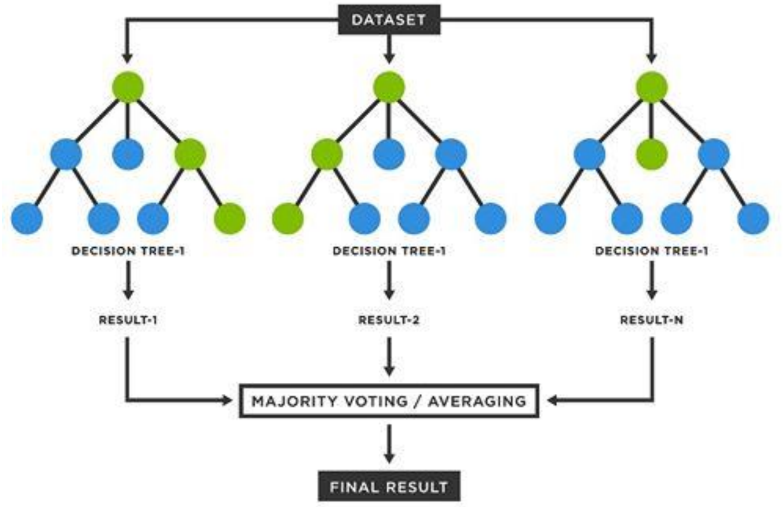


Figure 3. Random Forest Classifier Stuctur

The figure shows different decision trees, each built using a different subset of data. The results from each decision tree are combined through a process called voting (for classification) or averaging (for regression). The final result of the Random Forest is determined based on the majority vote from these trees.

2.5.3. K-Nearest Neighbor

K-Nearest Neighbor (KNN) is an algorithm widely used in various fields such as data analysis, disease prediction, and graph prediction. It involves determining the K closest points in a dataset to a specific data point based on similarity measures. This method can be enhanced with techniques such as multi-layer locality-sensitive hashing to improve search efficiency and accuracy [12].

$$dis(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2} \qquad (1)$$

x_1 and x_2 these are the two points in n-dimensional space. x_{1i} and x_{2i} these are the i-th coordinates of the points $x_1$and x_2, respectively. n are the dimension of the space.

2.5.4. Naive Bayes

Naive Bayes is a popular document classification model due to its simplicity and efficiency [13]. Although traditionally used for clustering and classification, recent advancements have highlighted its potential for general probabilistic learning and inference, providing accuracy and learning times comparable to context-independent Bayesian networks, but with significantly faster inference speeds [14]. The Naive Bayes formula is shown in equation (2).

$$P(A|B) = \frac{P(A\cap B)}{P(B)} = \frac{P(A)\cdot P(B|A)}{P(B)} \qquad (2)$$

P(A|B) are the probability of event A occurring given that event B has occurred. *P(A∩B)* are the probability that both events A and B occur simultaneously. P(A) are the prior probability of event A. *P(B/A)*: The probability of event B occurring given that event A has occurred. *P(B)* are the probability of event B occurring.

2.5.5. Logistic Regression

Logistic regression is a powerful statistical method widely used in medical research to analyze the impact of independent variables on binary outcomes. It allows for measuring the unique influence of each variable by identifying the strongest linear combinations of factors associated with the observed outcome. This versatile method accommodates both continuous and categorical variables [12].

$$f(x) = \frac{1}{1+e^{-x}} \qquad (3)$$

f(x) are sigmoid function, when *x* approaches negative infinity (−∞), the value of e^{-x} becomes very large, causing *f(x)* to approach 0, indicating that the probability of an event is almost nonexistent. Conversely, when *x* approaches positive infinity (+∞), the value of e^{-x} approaches 0, so *f(x)* approaches 1, indicating that the probability of the event is almost certain to occur.

2.5.6. Support Vector Machine (SVM)

Support Vector Machines (SVM) is a robust and accurate algorithm in data mining, with applications in both classification and regression [15]. SVM identifies the hyperplane that optimally separates labeled data points into classes, supports vectors at the margin, and can be used for nonlinear smoothing in communication systems. In some cases, it has outperformed neural networks, as illustrated in **Figure 4**.

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v20i2. 13006

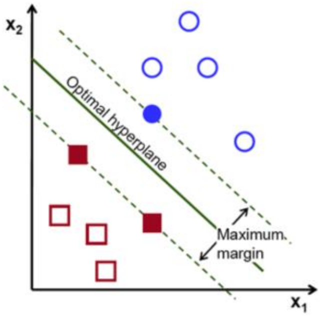


Figure 4. Hyperplane Support Vector Machine (SVM)

The thick green line in the middle is the hyperplane or optimal separating line that divides the two different data classes. In this case, the data consists of red squares (first class) and blue circles (second class). The points located near the separating line (optimal hyperplane) and lying on the dashed lines are the support vectors. The distance between the dashed green lines on both sides of the optimal hyperplane is the maximum margin.

2.5.7. Neural Network

Neural Networks (NN) are a fundamental machine learning method inspired by biological neural systems, consisting of interconnected neurons that process data to solve various problems [16]. While traditional NN use simple activation functions such as the sigmoid function, recent advancements propose the use of additive Gaussian process regression to generate individually optimal activation functions for each neuron, enhancing computational efficiency and expressive power. The neural network learning process is illustrated in **Figure 5**.

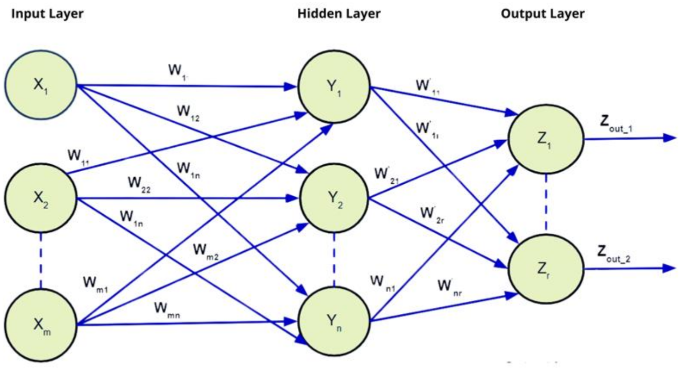


Figure 5. Neural Network Learning Process

In the figure, there is an input layer consisting of input neurons (X_1, X_2, *…*, X_m). This layer receives raw data that will be processed by the neural network. Next, there is a hidden layer in the middle, consisting of hidden neurons (Y_1, Y_2, *…*, Y_n). Each neuron in this layer is connected to the input and output neurons through weights (*W*). Finally, there is an output layer, consisting of output neurons (Z_1, Z_2), which represents the output or prediction of the network for the given data.

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v20i2. 13006

2.5.8. Voting Classifier

This classifier leverages the concept of ensemble learning by combining predictions from multiple base classifiers to make a final decision, as shown in **Figure 6**. The Voting Classifier approach has shown promising results in enhancing prediction accuracy and robustness across various applications by harnessing the diversity of multiple classifiers [17].

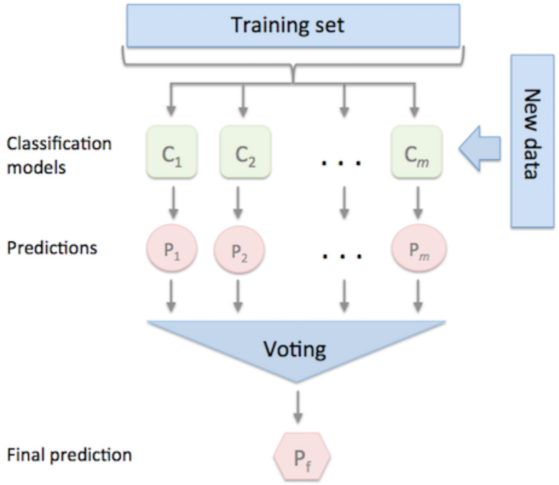


Figure 6. Voting Classifier Structure

The training set at the top is used to train several different classification models, represented in the figure as C_1, C_2, …, C_m. Each trained classification model will make predictions (P_1, P_2, …, P_m), and a voting stage is conducted to determine the final prediction (P_f).

2.5.9. Evaluation

Evaluating machine learning models is crucial to ensure their effectiveness and reliability. Several metrics are commonly used to assess the performance of these models, including Accuracy, Cross Validation Score, and ROC_AUC Score. Each of these metrics offers unique insights into the model's behavior and performance in different scenarios. [18].

Accuracy

Accuracy is one of the best and most broadly utilized assessment measurements. It is characterized as the proportion of correctly anticipated occasions to the entire number of occurrences. Exactness measures how regularly the show adjusts expectations. Whereas precision gives a clear measure of how frequently the demonstration is adjusted, it may not be the finest metric for imbalanced datasets. In cases where one course altogether dwarfs the others, a demonstrate may accomplish tall precision by just foreseeing the larger part course [19]. The calculations for Accuracy are given in equations (4).

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{4}$$

Cross Validation Score

Cross Approval Score may be a strong metric utilized to gauge how well a demonstration performs on unused information. It includes isolating the dataset into different subsets, preparing the demonstration on a few subsets testing it on the rest, and rehashing this preparation a few times. The foremost common strategy is k-fold cross-validation, where the information is part into k subsets. The demonstration is prepared and approved k times, with each subset serving as the approval set once and the remaining subsets as the preparing information. The cross-validation score is as a rule the normal of these approval scores, advertising a comprehensive appraisal of the model's execution without overfitting to any specific subset of information [20]. The formula for k-fold cross-validation is given in equation (5).

$$Cross\ Validation\ Score = \frac{1}{k}\sum_{i}^{k} Validation\ Score_i \tag{5}$$

k, hhis denotes the number of folds or splits in the cross-validation process. For example, in *k*-fold cross-validation, the dataset is divided into *k* equal parts. The model is trained *k* times, each time using a different fold as the validation set and the remaining folds as the training set.

ROC_AUC Score (Receiver Operating Characteristic - Area Under the Curve)

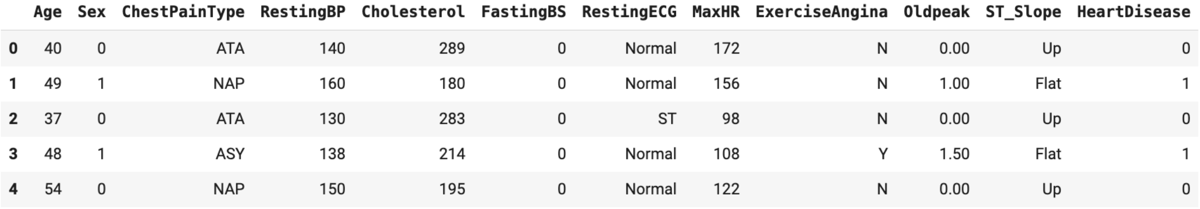
ROC_AUC Score is another capable metric, particularly for parallel classification issues. The ROC bend plots the genuine positive rate (review) against the untrue positive rate, outlining the trade-off between affectability and specificity. The AUC (Zone Beneath the Bend) evaluates the general capacity of the show to separate between positive and negative classes. A demonstrate with a ROC_AUC score of 1.0 demonstrates culminated classification, whereas a score of 0.5 proposes no discriminative control, comparable to arbitrary speculating [19]. The ROC_AUC score is calculated as the area under the ROC curve is given in equations (6).

$$ROC_{AUC} = \int_{0}^{1} ROC(x)dx \tag{6}$$

While accuracy provides a quick snapshot of a model’s performance, cross-validation score and ROC_AUC score offer deeper insights, especially in cases of imbalanced datasets or varying decision thresholds. Using a combination of these metrics ensures a comprehensive evaluation, guiding the selection and improvement of machine learning models.

3. Results and Discussion

This study was conducted using heart failure medical record data from the UCI Machine Learning Repository. The feature selection process employed binning, converting the value of Sex into binary numbers: 'M' was changed to 0 and 'F' to 1. The results of the binning process are shown in the **Figure 7**.



	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	0	ATA	140	289	0	Normal	172	N	0.00	Up	0
1	49	1	NAP	160	180	0	Normal	156	N	1.00	Flat	1
2	37	0	ATA	130	283	0	ST	98	N	0.00	Up	0
3	48	1	ASY	138	214	0	Normal	108	Y	1.50	Flat	1
4	54	0	NAP	150	195	0	Normal	122	N	0.00	Up	0

Figure 7. Dataset

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v20i2. 13006

Converting the "Sex" feature into binary values, where "0" represents male (M) and "1" represents female (F), is useful in analysis because it makes the data compatible with modeling algorithms that require numerical features, such as logistic regression and decision trees, thereby enhancing computational efficiency in model training and prediction. Additionally, this simple and easily interpretable binary representation facilitates understanding the relationship between the "Sex" feature and the target variable. By avoiding issues that may arise from categorical features, this conversion ensures that algorithms can process the data effectively. The binary standardization also creates consistency within the dataset, which is essential for statistical analysis techniques and calculations such as mean and standard deviation. Overall, binning the "Sex" feature into a binary format not only simplifies data analysis and modeling but can also improve the accuracy and performance of classification models.

Furthermore, the researcher provides a visual representation of the distribution of heart disease cases within a dataset. Each graph offers a different perspective on the prevalence of heart disease, as shown in the figure.

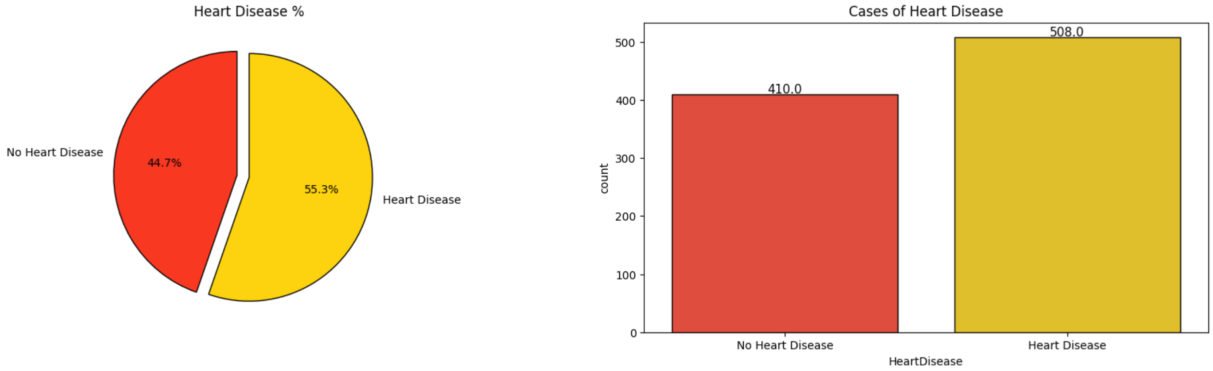


Figure 8. Pie Chart and Bar Char Heart Disease

The pie chart illustrates the percentage of individuals with and without heart disease. It shows that 55.3% of individuals have heart disease, depicted in yellow, while 44.7% do not have heart disease, shown in red. This indicates that the majority of the population in this dataset is affected by heart disease. The Figure 8 displays the absolute number of individuals with and without heart disease. Here, we see that 508 individuals have heart disease (yellow bar) compared to 410 individuals who do not have heart disease (red bar). The significant difference between these bars reinforces the information provided by the pie chart, highlighting the higher prevalence of heart disease.

Figure 9 is a correlation matrix illustrating the relationships between various variables in a dataset related to heart disease.

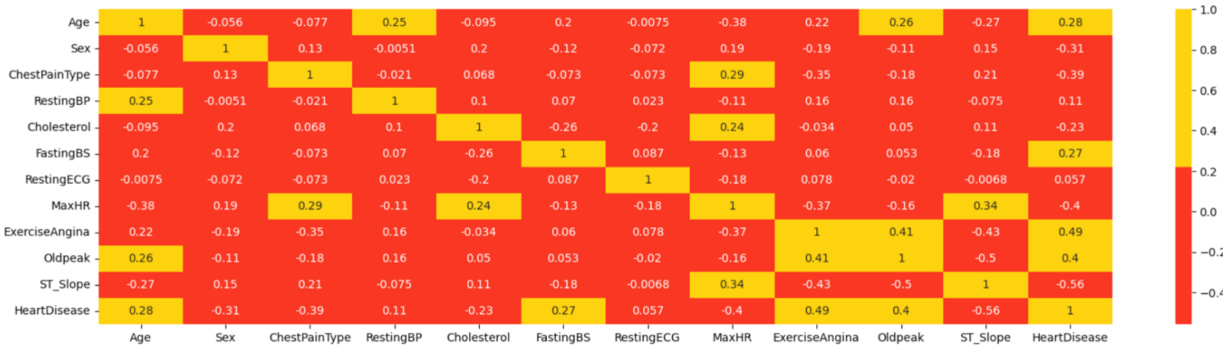


Figure 9. Heatmap

This correlation map uses a color scheme to indicate the degree of correlation between these variables, with yellow indicating a strong positive correlation and red indicating a strong negative correlation.

Figure 9 presents a correlation matrix depicting the relationships between variables related to heart disease. The analyzed variables include demographic and clinical factors, such as Age, Sex, Chest Pain Type (ChestPainType), RestingBP (RestingBP), Cholesterol, Fasting Blood Sugar (FastingBS), Resting Electrocardiogram (RestingECG), Maximum Heart Rate (MaxHR), Exercise-Induced Angina (ExerciseAngina), ST segment depression after exercise (Oldpeak), ST segment slope during stress tests (ST_Slope), and the presence of heart disease (HeartDisease).

Based on this correlation matrix, several variables show a strong correlation with HeartDisease, such as ST_Slope with a negative correlation (-0.56), Oldpeak with a positive correlation (0.4), and ExerciseAngina with a positive correlation (0.49). A negative correlation suggests that an increase in variables like ST_Slope is associated with a decreased risk of heart disease, whereas a positive correlation, as seen with ExerciseAngina, indicates that more frequent angina during exercise is associated with a higher likelihood of heart disease. Additionally, MaxHR also has a negative correlation (-0.49) with HeartDisease, indicating that a lower maximum heart rate is associated with an increased risk of heart disease.

Overall, variables such as ST_Slope, ExerciseAngina, and MaxHR appear to play significant roles in predicting the presence of heart disease, demonstrating meaningful relationships in this context. The correlations presented offer important insights into the influence of clinical variables on heart disease risk, which can serve as a reference for developing predictive models or risk evaluation frameworks.

Then the selected features are used as input for several classification algorithms. The classification results from the research conducted based on equations (4), (5), and (6) are shown in **Table 2**.

Table 2. Models Clasification Result

Algoritma	**Accuracy**	**Cross Validation Score**	**ROC_AUC Score**

Logistic Regression	87.50	91.12	87.43
Support Vector Classifier	87.50	90.53	87.43
Decision Tree Classifier	84.78	89.09	84.62
Random Forest Classifier	83.70	92.93	83.50
KNeighbors Classifier	81.52	89.34	81.36
Naïve bayes	85.87	91.36	85.75
Neural Network	80.98	91.38	80.90
Voting Classifier (SVC & Logistic Regression)	88.04	91.01	88.00

The table provides a comparison of the performance of various machine learning algorithms used for classifying heart failure data based on three key metrics: Accuracy, Cross Validation Score, and ROC_AUC Score. The tested algorithms include Logistic Regression, Support Vector Classifier (SVC), Decision Tree Classifier, Random Forest Classifier, KNeighbors Classifier, Naïve Bayes, Neural Network, and Voting Classifier (a combination of SVC and Logistic Regression).

The cross-validation method was used to improve the accuracy of model evaluation. The configuration used was n_splits=10, n_repeats=3, and random_state=1. Data was divided into 10 folds in each cross-validation iteration, where the model was trained on 9 folds and tested on 1 fold. This process was repeated 10 times so that each fold alternates as test data. Next, with n_repeats=3, the entire cross-validation process was repeated three times with different data splits. This repetition aimed to reduce variability in evaluation results that might arise from a single data split, resulting in more stable and consistent model performance values. Setting random_state=1 makes the data split randomization process reproducible with the same results in each repetition, ensuring consistent and repeatable evaluation.

Logistic Regression and Support Vector Classifier (SVC) show the same performance in terms of Accuracy with a value of 87.50 and an ROC_AUC Score of 87.43. However, Logistic Regression has a slightly higher Cross Validation Score of 91.12 compared to SVC's value of 90.53. The Decision Tree Classifier shows an Accuracy of 84.78, a Cross Validation Score of 89.09, and an ROC_AUC Score of 84.62. The Random Forest Classifier has the highest Cross Validation Score of 92.93 but has lower Accuracy and ROC_AUC Scores of 83.70 and 83.50, respectively, compared to some other algorithms.

The KNeighbors Classifier has an Accuracy of 81.52, a Cross Validation Score of 89.34, and an ROC_AUC Score of 81.36. Naïve Bayes shows an Accuracy of 85.87, a Cross Validation Score of 91.36, and an ROC_AUC Score of 85.75. The Neural Network has the lowest Accuracy of 80.98 but high Cross Validation Score of 91.38 and an ROC_AUC Score of 80.90.

The Voting Classifier, which combines SVC and Logistic Regression, shows the best overall performance with an Accuracy of 88.04, a Cross Validation Score of 91.01, and an ROC_AUC Score of 88.00. Based on these results, the Voting Classifier is the most suitable algorithm for the heart failure classification problem in this study. The results provide valuable insights for selecting the most effective algorithm for specific classification tasks, assisting researchers and practitioners in making better decisions based on relevant performance metrics.

The results of this study provide a broader comparison of machine learning methods compared to previous research. In previous studies, a maximum of 5 methods were used, whereas this study includes 7 methods, offering more comprehensive information on the most accurate methods for predicting heart failure risk.

4. Conclusions and Suggestions

Based on the results of the research described earlier, it can be concluded that the combined Voting Classifier algorithm, which integrates SVC and Logistic Regression, performs exceptionally well for classifying heart failure mortality with an accuracy of 88.01%. Additionally, the incorporation of feature binning methods can enhance the effectiveness of the existing dataset without losing its important values. Furthermore, the findings of this study can serve as a reference for future researchers in both medical and non-medical classification analyses. The author recommends that future research explore the use of deep learning algorithms to achieve even better results.

References

[1] B. Shahim, C. J. Kapelios, G. Savarese and L. H. Lund, "Global Public Health Burden of Heart Failure: An Updated Review," *Cardiac Failur Review*, vol. 11, no. Doi https://doi.org/10.15420/cfr.2023.05, 27 Juli 2023.

[2] B. Tajik, A. Voutilainen, R. Sankaranarayanan, A. Lyytinen, J. Kauhanen, G. Y. Lip, T.-P. Tuomainen and M. Isanejad, "Frailty alone and interactively with obesity predicts heart failure: Kuopio Ischaemic Heart Disease Risk Factor Study," *ESC Heart Failure*, vol. 10, pp. 2354-2361, 10 May 2023.

[3] S. Parveen, B. Zareini, A. Arulmurugananthavadivel, C. Kistorp, J. Faber, L. Køber, C. Hassager, T. B. Sørensen, C. Andersson, D. Zahir, K. Iversen, E. Wolsk, G. Gislason, F. Gaborit and M. Schou, "Association between early detected heart failure stages and future cardiovascular and non-cardiovascular events in the elderly (Copenhagen Heart Failure Risk Study)," *BMC Gereatrics*, vol. 22 (230), pp. 1-10, 2022.

[4] Z. Zhu, F.-R. Li, Y. Jia, Y. Li, D. Gu, J. Chen, H. Tian, J. Yang, H.-H. Yang, L.-H. Chen, K. Zhang, P. Yang, L. Sun, M. Shi, Y. Zhang, L.-Q. Qin and G.-C. Chen, "Association of Lifestyle With Incidence of Heart Failure According to Metabolic and Genetic Risk Status: A Population-Based Prospective Study," *Circulation: Heart Failure*, vol. 15 (9), pp. 851-859, September 2022.

[5] C. Fonseca, "Diagnosis of heart failure in primary care," *Heart Fail Rev*, vol. 11 (2), pp. 95-107, Juni 2006.

[6] H. Moroz, Y. Li and A. Marelli, "hART: Deep Learning-Informed Lifespan Heart Failure Risk Trajectories," *medRxiv preprint*, 5 September 2023.

[7] D. Yu, S. Yang, R. Wang, K. Wang, W. Han, H. Wu, W. Wang and X. Wang, "Machine Learning in Heart Failure Research: A Bibliometric Analysis from 2003 to 2023," *Research Square*, pp. 1-55, Juni 2023.

[8] F. S. Alotaibi, "Implementation of Machine Learning Model to Predict Heart Failure Disease," *(IJACSA) International Journal of Advanced Computer Science and Applications,* vol. 10 (6), pp. 261-268, 2019.

[9] J. Lee, G. Kim, I. Ham, K. Ko, S. Park, Y.-J. Choi, D. O. Kang, J. Y. Choi, E. J. Park, S. Lee, S. Y. Roh, D.-I. Lee, J. O. Na, C. U. Choi, J. W. Kim, S.-W. Rha, C. G. Park, E. J. Kim and H. Ko, "Voice as a Biomarker to Detect Acute Decompensated Heart Failure: Pilot Study for the Analysis of Voice Using Deep Learning Models," *medRxiv,* pp. 1-44, 12 September 2023.

[10] F. "Heart Failure Prediction Dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction. [Accessed Juni 2024].

[11] M. Sulewski and A. Ozka, "Application of The Forest Classifier Methode for Desription of Movements of an Oscillator Forced by a Stochastic Series of Impulses," *Journal of Theoretical and Applied Mechanics,* vol. 61 (4), pp. 819-831, 2023.

[12] J. C. Stoltzfus, "Logistic Regression: A Brief Primer," *Academic Amergency Medicine,* vol. 18, pp. 1099-1104, 2011.

[13] L. Nguyen, "Tutorial on Support Vector Machine," *Applied and Computational Mathematics,* Vols. 6 (4-1), pp. 1-15, 2017.

[14] S. Manzhos and M. Ihara, "Neural network with optimal neuron activation functions based on additive Gaussian process regression," *arXiv,* vol. 2, pp. 1-24, 19 Januari 2023.

[15] V. D. Cong and T. T. Hiep, "Support vector machine-based object classification for robot arm system," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 13 (5), pp. 5047-5053, Oktober 2023.

[16] M. Thorat, S. Pandit and S. Balote, "Artificial Neural Network: A brief study," *Asian Journal of Convergence in Technology,* vol. VIII, no. III, pp. 12-16, 2022.

[17] J.-h. Kim, "Ensemble Approach for Predicting the Diagnosis of Osteoarthritis Using Song Voting Classifier," *medRxiv,* pp. 1-22, 28 Januari 2023.

[18] M. O. Adjei, J. B. H. Acquah, T. Frimpong and G. A. Salaam, "A systematic review of prediction accuracy as an evaluation measure for determining machine learning model performance in healthcare systems," *medRxiv preprint,* no. Doi: https://doi.org/10.1101/2023.06.01.23290837, pp. 1-23, 4 Juni 2023.

[19] C. M. Bishop, "Pattern Recognition and Machine Learning," in *Information Science and Statistics*, New York, Springer New York, 2006, pp. XX, 778.

[20] A. C. Müller and S. Guido, Introduction To Machine Learning With Python: A Guide for Data Scientists, Sebastopol, CA: O'Reilly Media, Inc., 2017.