Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13198
e-ISSN : 2541-2019
p-ISSN : 2541-044X

GridSearch and Data Splitting for Effectiveness Heart Disease Classification

Rusyda Tsaniya Eka Putri[1)*], Junta Zeniarja[2)], Sri Winarno[3)], Ailsa Nurina Cahyani[4)], Ahmad Alaik Maulani[5)]
[1,2,3,4,5)] Dian Nuswantoro University Semarang, Indonesia
[1)]111202013008@mhs.dinus.ac.id, [2)]junta@dsn.dinus.ac.id, [3)]sri.winarno@dsn.dinus.ac.id,
[4)]111202012636@mhs.dinus.ac.id, [5)]111202013154@mhs.dinus.ac.id

Submitted : Nov 25, 2023 | **Accepted** : Nov 28, 2023 | **Published** : Jan 1, 2024

Abstract: Cardiovascular disease (CVD) is a major global health issue that affects death rates significantly. This research aims to improve the early detection and diagnosis of cardiovascular illness by utilizing machine learning methods, particularly classification algorithms. According to estimates from the World Health Organization (WHO), cardiovascular disease (CVD) caused 17.9 million deaths globally in 2019, or 32% of all fatalities. The treatment and prognosis of cardiovascular illness are greatly improved by early detection and diagnosis. Classification, in particular, machine learning, has become a prominent tool for solving problems connected to heart disease. The main objective of this project is to assess how well Grid Search and various data-sharing methods classify cardiac disease. SVM, Random Forest Classifier, Logistic Regression, Naïve Bayes, Decision Tree Classifier, KNN, and XGBoost Classifier are just a few machine learning methods. The UCI heart disease dataset, which contains information from 303 heart disease patients and 165 healthy participants, is used for the evaluation. Performance parameters like recall, accuracy, precision, and F1 score are considered to evaluate the algorithms' efficacy. The investigation's expected outcomes are intended to increase doctors' ability to diagnose cardiac disease more accurately. Moreover, these results may aid in creating more complex classification models for diagnosing cardiac conditions.

Keywords: Cardiovascular Disease; Machine Learning; Classification Algorithms; Grid Search; Data-Sharing Techniques

INTRODUCTION

Cardiovascular disease (CVD) is a globally recognized and highly hazardous illness. WHO estimates that 17.9 million deaths worldwide in 2019 were related to cardiovascular disease. This figure represents 32% of all fatalities globally. Because the coronary arteries provide blood to the heart, coronary artery constriction and blockage are two of the most common causes of cardiovascular disease(Ayon et al., 2022). By 2040, estimates indicate that over 30 million deaths worldwide would result from cardiovascular disease. Cardiovascular disease can be caused by several risk factors, including smoking, gender differences, poor diet, inactivity, diabetes, high blood pressure, obesity, and past family history(Bays et al., 2022).

Early identification and diagnosis are critical to enhance cardiovascular disease treatment and prognosis. Modern civilization is developing at a quick pace, which is contributing to an increase in the incidence of cardiovascular diseases and a younger patient base(Shi et al., 2022). Thus, by emphasizing lifestyle and proper treatment, rigorous screening and proactive risk factor management can be used to lower the incidence of cardiovascular disease. Classification is one of the most popular uses of machine learning about heart disease.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i1.13198 p-ISSN : 2541-044X

Artificial intelligence (AI) was first used in 1959 by American artificial intelligence pioneer Arthur Lee Samuel, who presented a chess program displaying AI capabilities. Since then, machine learning applications have been utilized in various industries, including healthcare, financial services, robotics, autonomous vehicles, facial and voice recognition, and more(Tougui et al., 2020). Machine learning (ML) is a set of math tools that help machines make intelligent choices when given much information. This helps scientists solve big problems that are hard for machines to solve on their own(Hagan et al., 2021).

This study compares the efficacy of multiple data-sharing strategies and Grid Search for heart disease classification. This study makes use of multiple machine-learning algorithms. This study will employ three distinct strategies for data division: 90:10, 80:20, and 70:30. SVM, Random Forest Classifier, Logistic Regression, Naïve Bayes, Decision Tree Classifier, KNN, and XGBoost Classifier are some of the techniques employed. The UCI heart disease dataset includes data from 303 patients with heart disease and 165 healthy volunteers. The SMOTE method will be utilized to balance the data classes. To assess its efficacy, we will examine several variables: recall, accuracy, precision, and F1 score. The findings of this study should increase the precision with which medical professionals identify cardiac problems. Furthermore, these results might aid in creating more intricate classification models for diagnosing cardiac conditions.

LITERATURE REVIEW

This study investigated the classification of heart disease using different algorithms and the UCI Machine Learning dataset. The study applied three different feature selection techniques, which were chi-square, ANOVA, and mutual information(Biswas et al., 2023) for predicting heart disease; six machine learning models were utilized, namely logistic regression, support vector machine, random forest, naïve Bayes, K-nearest neighbor, and decision tree. The Random Forest algorithm exhibited the highest accuracy of 94.51%, while logistic regression achieved a respectable accuracy of 93.41%. The Support Vector Machine had an accuracy of 75.82%, while the other models demonstrated accuracies ranging from 84.61% to 92.31%.

In another study(Dissanayake & Johar, 2021), ten techniques were used to select significant features from the Cleveland Heart Disease dataset. These techniques included ANOVA, chi-square, mutual information, Relief F, inverse feature selection, full feature selection, feature elimination regression, LASSO regression, and Ridge regression with Grid Search CV. In addition, six classification methods were employed, including Naïve Bayes, Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines. Ultimately, the most accurate algorithm was the Decision Tree, which achieved an accuracy value of 88.52%.

To predict heart illness, the study(Asif et al., 2023) utilized ensemble learning methods, including random forest, extra tree classifier, XGBoost, and CatBoost. Two hyperparameter optimization techniques, namely GridSearchCV and Randomized Search CV, were employed to optimize the model performance. Kaggle data was used to split the dataset into an 80:20 ratio. The suggested model achieved an accuracy of 98.15% in detecting both the presence and absence of heart disease. Based on the recall value measure, the proposed algorithm provided a score of 98.09%.

Studies have compared model performance with and without Grid Search CV (G. N. Ahmad et al., 2022). Several machine learning techniques, including XGBoost Classifier, KNN, SVM, and Logistic Regression, were applied using the UCI Kaggle dataset. According to the results, XGBoost used Grid Search CV to obtain the highest accuracy of 99.03%. The LR and SVM algorithms came in second and third with an identical accuracy rating of 88.41%. The XGB model's optimal parameters are n-estimation, maximum depth, and learning rate. The optimal parameters for the LR method are the solver and learning rate (C). The optimal parameters for the Support Vector Machine are the learning rate (C), the gamma, and the rbf kernel.

Machine learning techniques like Random Forest, Decision Tree Classifier, Multilayer Perceptron, and XGBoost were utilized in a related work (Bhatt et al., 2023). In addition, this work uses the GridSearchCV method for hyperparameter tuning and a public dataset consisting of 70,000 Kaggle examples. After training on an 80:20 data split, the model will yield the following accuracy: Decision Tree yields 86.37% accuracy with cross-validation and 86.53% accuracy without; XGBoost yields

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i1.13198 p-ISSN : 2541-044X

86.87% accuracy with cross-validation and 87.02% accuracy without; Random Forest yields 87.05% accuracy with cross-validation and 86.92% accuracy without; and MLP yields 87.28% accuracy with cross-validation and 86.94% accuracy without. The AUC values of all suggested algorithms fall between 0.94 and 0.95.

This study by(Chandrasekhar & Peddakrishna, 2023) proposes to apply six machine learning algorithms, including Random Forest, KNN, Naïve Bayes, Gradient Boosting, and AdaBoost Classifier, to improve the accuracy of heart disease identification. Cleveland and IEEE data port datasets will be used to assess and train the model. Five-fold cross-validation and GridSearchCV will be used to optimize the models. AdaBoost achieves 90% accuracy with the IEEE Dataport dataset, whereas Logistic Regression yields the most outstanding performance with 90.16% accuracy using the Cleveland dataset.

METHOD

Research Flow

Five primary methodologies were used in this research: data processing, data preparation, model training, model optimization, and model evaluation. A few actions must be taken in phases, including data processing and preparation, before moving on to data training, which can be seen in Figure 1:

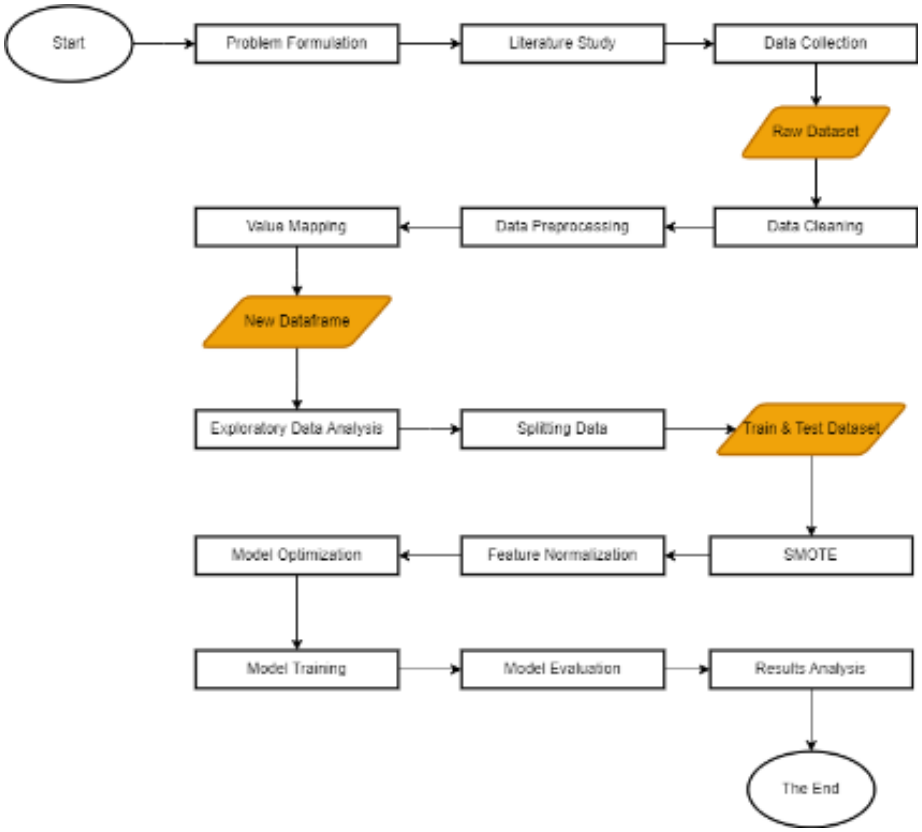

Figure 1 Research Stages

Data Collecting

The study employs the 'Heart Disease' public dataset from the UCI Machine Learning Repository (Janosi, 1988). Although there are 76 features in this database, only 14 attributes are mentioned in the published studies, as seen in Table 1. Certain factors or indicators in this dataset can be utilized to identify patients. Specifically, the Cleveland database is the only machine learning experts have used thus far. The UCI Heart Disease target dataset values are (0, 1, 2, 3, 4). In this case, 0, 1, 2, 3, and 4 denote the absence of cardiovascular disease and excellent health, respectively. However, the target to be used will be restricted to (0,1) for this study. The levels (1, 2, 3, 4) will be lowered to 1(Louridi et al., 2021).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13198
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 1 Heart Disease UCI Repository

Attributes	Type
age	Integer
sex	Categorical
cp	Categorical
tretsbps	Integer
chol	Integer
fbs	Categorical
restecg	Categorical
thalach	Integer
exang	Categorical
oldpeak	Integer
slope	Categorical
ca	Integer
thal	Categorical
num	Integer

Data Preparation

The process of preparing raw data that can be used for additional processing and analysis is known as data preparation. The primary tasks of preparing the raw data for machine learning (ML) algorithms are gathering, cleaning, and labeling it. Next, the data must be explored and visualized. A key component of data mining is optimal data preparation. Finding patterns or links in vast volumes of data involves data mining, which includes gathering and analyzing historical data(Pamungkas et al., 2019).

Data cleansing is crucial when analyzing data to classify cardiac disease using machine learning algorithms. Data cleaning involves processing erroneous values or errors, deleting duplicate or missing data, and formatting inconsistent data. This is similar to how mean/mode imputation is used in the journal(Apriyanto Alhamad, 2019) to manage missing value data. Value mapping aims to expedite classification by reducing the prior target class values of 0, 1, 2, 3, and 4 to 0 and 1(A. A. Ahmad & Polat, 2023).

Next, univariate analysis and multivariate analysis are the two steps of EDA that are conducted. When dataset items that only have one relevant entity are subjected to univariate analysis. In contrast, multivariate analysis uses more than two characteristics to find the association between the features in the data. An approach called SMOTE (Synthetic Minority Oversampling Technique) addresses the class imbalance in a set of data(Elreedy & Atiya, 2019).

By generating convex combinations of neighboring instances, the Min Max Scaler approach generates additional examples of the minority class, balancing the data set(Chauhan & Singh, 2022). Changing the feature values within a specific range is the goal of feature normalization, as follows in the formula:

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Splitting Data

Data decomposition is primarily used to give an unbiased assessment of the model's performance and avoid overfitting, in which the model predicts seen data exceptionally well but performs poorly when applied to fresh data(Handayani et al., 2021). Although data can generally be categorized as training, test, and validation data, an approach was taken in this study to separate data into training and test categories. This study uses Three data division ratios: 90:10, 80:20, and 70:30. These three variants yield varying degrees of model accuracy.

Model Optimization

This stage aims to improve the model's ability to predict or generalize new data. In model optimization, grid search(Prabu et al., 2022) is a parameter search technique that entails attempting every

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13198
e-ISSN : 2541-2019
p-ISSN : 2541-044X

possible combination of parameter values from a preset set of values. In grid search, the model is trained and tested on every possible combination of parameter values to evaluate each combination.

Model Training

Methods like SVM, Random Forest, Naïve Bayes, LR, Decision Trees, and K-nearest neighbors are commonly employed to address classification problems in supervised learning. These approaches employ prior theoretical knowledge to develop a model of the data by estimating the function's outcome using a set of training data(Taranto-Vera et al., 2021). The training or learning on a target feature to map each collection of properties (features) to one of the many available class labels is known as classification.

Model Evaluation

Model evaluation aims to determine the accuracy of a machine learning model's predictions. A confusion matrix is a table that summarizes a model's performance on a test data set. The number of TP, TN, FP, and FN that the model produced on the test data is displayed in the matrix. We must ascertain the performance measure values to assess the confusion matrix's performance. Four categories comprise the values of the performance measure: recall, accuracy, precision, and F1 score(Xu et al., 2020). The formula for performance metrics is as follows:

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (2)$$

$$Precision = \frac{(TP)}{(TP+FP)} \quad (3)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (4)$$

$$F-1\ Score = 2\frac{(Recall \times Precision)}{(Recall+Precision)} \quad (5)$$

RESULT

Value Mapping

The outcomes of data cleansing, value mapping, EDA, altering column names and specific values, feature normalization, and SMOTE during the data preparation step. Separating the target data (y) for prediction and the data (X) for model training is the next step after the data has been cleaned. Given that y contains five categories—0, 1, 2, 3, and 4. As a result of this value mapping process, the initial 5-point objective for heart disease detection is now only at 2. Next, the reduced data is added to the new data frame. Following value mapping, the next stage generates a new data frame and then sees the target the technique has achieved. As seen in Figure 2, the number for category 0, or not having heart disease, is 160, while the number for category 1, or having heart disease, is 137:

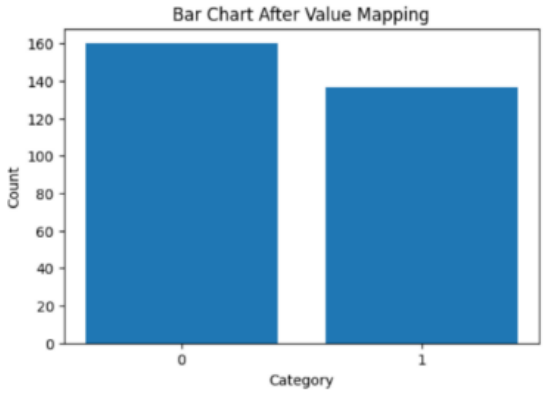

Figure 2 Visualization after Value Mapping

To display the percentage of class data used in this study, EDA is run at this point. Multivariate analysis was carried out to determine the correlation between the variables using Pearson's correlation, and the results were shown on a heatmap. The variables are not connected when the value is around 0.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13198
e-ISSN : 2541-2019
p-ISSN : 2541-044X

If the value is near 1, it is said to have a positive correlation; if it is near -1, it is said to have a negative correlation. Figures 3 and 4 below display the findings of the univariate and multivariate EDA analysis of heart disease in the target data.
:

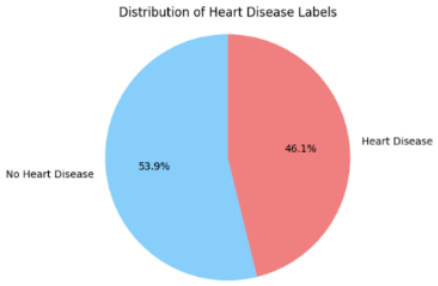

Figure 3 Comparison of Class Labels

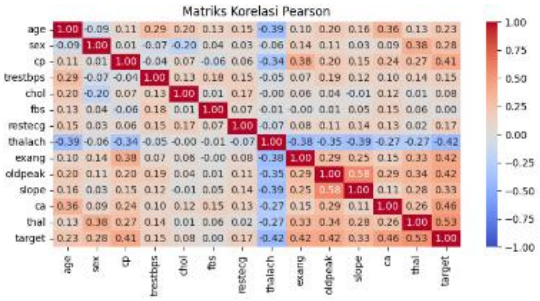

Figure 4 Correlation Between Variable

According to the EDA results, just one feature—FBS (Fasting Blood Sugar)—has a low correlation (around 0) to the label since the dataset contains uneven data. Two characteristics, 'thal' (thalassemia) and 'ca' (number of major vessels), have a strong association (almost 1). As per the findings, 53.9% of the target data has the label 'No Heart Disease,' whereas 46.1% has the label 'Heart Disease'. It is clear from this analysis that the dataset has target values that are out of balance. The SMOTE approach leads to a rise in the proportion of minority class representations. As previously mentioned, there are 160 and 137 label classes, which is an imbalance. To have a balanced one, this SMOTE technique helps raise the value of label class 1, which comes from 137 to 160. When feature normalization is performed with Min Max Scaler, the data is placed into a predetermined range between 0 and 1. This procedure entails deducting each characteristic's minimum value and dividing the outcome by the range of each characteristic's maximum and minimum values.

Splitting Data

Training and testing data will be divided based on specific percentages, with 80% used for training, 20% for testing, 70% for training, 30% for testing, 90% for training, and 10% for testing. This allocation ensures a distinct separation between training and test data. The outcomes of the data division are displayed in Figure 5–7:

```
shape X_train:  (267, 13)
shape X_test:  (30, 13)
shape y_train:  (267,)
shape y_test:  (30,)
```

Figure 5 Splitting Data with 90:10

```
shape X_train:  (237, 13)
shape X_test:  (60, 13)
shape y_train:  (237,)
shape y_test:  (60,)
```

Figure 6 Splitting Data with 80:20

```
shape X_train:  (207, 13)
shape X_test:  (90, 13)
shape y_train:  (207,)
shape y_test:  (90,)
```

Figure 7 Splitting Data with 70:30

Analysis Model Optimization

The Grid Search Technique is used to hyperparameter tune each of the seven machine learning algorithms for model optimization. Train-validation datasets are used for hyperparameter tuning. This optimization also impacts the variation in data sharing. It should be noticed that when using Grid Search, the Gaussian Naïve Bayes algorithm does not have a parameter value. Tables 2 through 4 display the optimal hyperparameter values obtained with Grid Search:

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13198
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 2 Grid Parameters with Test Size 0.1

No	Hyperparameter	Value	Algorithms
1	'max_depth'	10	Random Forest Classifier
2	'min_samples_leaf'	1	
3	'min_samples_split'	5	
4	'n_estimators'	100	
5	'C'	0.1	Support Vector Machine
6	'gamma'	'scale'	
7	'C'	0.1	Logistic Regression
8	'penalty'	'I2'	
9	'max_depth'	3	Decision Tree
10	'min_samples_leaf'	3	
11	'n_neighbors'	9	K – Nearest Neighbor
12	'weights'	'uniform'	
13	'learning_rate'	0.1	XGBoost Classifier
14	'max_depth'	7	
15	'n_estimators'	500	

Table 3 Grid Parameters with Test Size 0.2

No	Hyperparameter	Value	Algorithms
1	'max_depth'	10	Random Forest Classifier
2	'min_samples_leaf'	2	
3	'min_samples_split'	2	
4	'n_estimators'	200	
5	'C'	1	Support Vector Machine
6	'gamma'	'scale'	
7	'C'	1.0	Logistic Regression
8	'penalty'	'I2'	
9	'max_depth'	3	Decision Tree
10	'min_samples_leaf'	5	
11	'n_neighbors'	9	K - Nearest Neighbor
12	'weights'	'uniform'	
13	'learning_rate'	0.01	XGBoost Classifier
14	'max_depth'	7	
15	'n_estimators'	1000	

Table 4 Grid Parameters with Test Size 0.3

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13198
e-ISSN : 2541-2019
p-ISSN : 2541-044X

No	Hyperparameter	Value	Algorithms
1	'max_depth'	5	Random Forest Classifier
2	'min_samples_leaf'	4	
3	'min_samples_split'	5	
4	'n_estimators'	200	
5	'C'	10	Support Vector Machine
6	'gamma'	'scale'	
7	'C'	0.1	Logistic Regression
8	'penalty'	'I2'	
9	'max_depth'	3	Decision Tree
10	'min_samples_leaf'	5	
11	'n_neighbors'	9	K – Nearest Neighbor
12	'weights'	'uniform'	
13	'learning_rate'	0.01	XGBoost Classifier
14	'max_depth'	3	
15	'n_estimators'	1000	

Analysis Model Evaluation

The confusion matrix and three additional assessment metrics—accuracy, recall, and area under the curve (AUC)—are employed in testing this model. Recall that AUC and the confusion matrix are essential for measuring model performance. The findings of the model performance evaluation will be presented in line with the stages as intended, from beginning to end. For every data division, various findings are obtained using the evaluation metrics. As stated in Table 5, the assessment metrics show the following outcomes with a data division of 90:10 and a total of 30 support values (19 for healthy patients and 11 for patients with heart disease):

Table 5 Evaluation Metrics on Test Size 0.1

No	Algorithms	Evaluation Metrics	Score (0)	Score (1)
1	Random Forest Classifier	Precision	0.94	0.83
		Recall	0.89	0.91
		F1-score	0.92	0.87
		Akurasi	0.90	0.90
		AUC	0.96	0.96
2	Support Vector Machine	Precision	0.94	0.83
		Recall	0.89	0.91
		F1-score	0.92	0.87
		Akurasi	0.90	0.90
		AUC	0.97	0.97
4	Naïve Bayes	Precision	0.89	0.82
		Recall	0.89	0.82
		F1-score	0.89	0.82
		Akurasi	0.87	0.87
		AUC	0.96	0.96
5	Logistic Regression	Precision	0.94	0.83
		Recall	0.89	0.91
		F1-score	0.92	0.87
		Akurasi	0.90	0.90
		AUC	0.97	0.97
6	Decision Tree	Precision	0.94	0.77

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13198
e-ISSN : 2541-2019
p-ISSN : 2541-044X

		Recall	0.84	0.91
		F1-score	0.89	0.83
		Akurasi	0.87	0.87
		AUC	0.92	0.92
7	K- Nearest Neighbor	Precision	0.89	0.82
		Recall	0.89	0.82
		F1-score	0.89	0.82
		Akurasi	0.87	0.87
		AUC	0.95	0.95
8	XGBoost Classifier	Precision	0.89	0.75
		Recall	0.84	0.82
		F1-score	0.86	0.78
		Akurasi	0.83	0.83
		AUC	0.86	0.86

Table 5 demonstrates that the algorithms with the highest score values, 0.97, are SVM and Logistic Regression. Furthermore, the accuracy of both models, including Random Forest's, demonstrates its peak performance of 0.90. On the other hand, the best recall value for label class 1 is 0.91 with Random Forest, SVM, LR, and Decision Tree. Sixty support values—36 for healthy patients and 24 for patients with heart disease—were obtained from the 80:20 data split. The outcomes of the assessment metrics in Table 6 are as follows:

Table 6 Evaluation Metrics on Test Size 0.2

No	Algorithms	Evaluation Metrics	Score (0)	Score (1)
1	Random Forest Classifier	Precision	0.92	0.88
		Recall	0.92	0.88
		F1-score	0.92	0.88
		Akurasi	0.90	0.90
		AUC	0.94	0.94
2	Support Vector Machine	Precision	0.89	0.87
		Recall	0.92	0.83
		F1-score	0.90	0.85
		Akurasi	0.88	0.88
		AUC	0.94	0.94
4	Naïve Bayes	Precision	0.90	0.95
		Recall	0.97	0.83
		F1-score	0.93	0.89
		Akurasi	0.92	0.92
		AUC	0.95	0.95
5	Logistic Regression	Precision	0.89	0.83
		Recall	0.89	0.83
		F1-score	0.89	0.83
		Akurasi	0.87	0.87
		AUC	0.94	0.94
6	Decision Tree	Precision	0.88	0.71
		Recall	0.78	0.83
		F1-score	0.82	0.77
		Akurasi	0.80	0.80
		AUC	0.90	0.90
7	K- Nearest Neighbor	Precision	0.89	0.83

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13198
e-ISSN : 2541-2019
p-ISSN : 2541-044X

		Recall	0.89	0.83
		F1-score	0.89	0.83
		Akurasi	0.87	0.87
		AUC	0.95	0.95
8	XGBoost Classifier	Precision	0.88	0.74
		Recall	0.81	0.83
		F1-score	0.84	0.78
		Akurasi	0.82	0.82
		AUC	0.91	0.91

Table 6 shows that the Naïve Bayes and KNN algorithms yield the most excellent AUC value of 0.95. Random Forest and Naïve Bayes yield the best model accuracy. The support value in the 70:30 data division is 90 (49 for patients who are healthy and 41 for those who have cardiac disease). The evaluation metrics' outcomes are displayed in Table 7:

Table 7 Evaluation Metrics on Test Size 0.3

No	Algorithms	Evaluation Metrics	Score (0)	Score (1)
1	Random Forest Classifier	Precision	0.85	0.87
		Recall	0.90	0.80
		F1-score	0.87	0.84
		Akurasi	0.86	0.86
		AUC	0.95	0.95
2	Support Vector Machine	Precision	0.83	0.89
		Recall	0.92	0.78
		F1-score	0.87	0.83
		Akurasi	0.86	0.86
		AUC	0.93	0.93
4	Naïve Bayes	Precision	0.88	0.92
		Recall	0.94	0.85
		F1-score	0.91	0.89
		Akurasi	0.90	0.90
		AUC	0.96	0.96
5	Logistic Regression	Precision	0.88	0.92
		Recall	0.94	0.85
		F1-score	0.91	0.89
		Akurasi	0.90	0.90
		AUC	0.95	0.95
6	Decision Tree	Precision	0.76	0.74
		Recall	0.80	0.71
		F1-score	0.78	0.72
		Akurasi	0.76	0.76
		AUC	0.86	0.86
7	K- Nearest Neighbor	Precision	0.90	0.92
		Recall	0.94	0.88
		F1-score	0.92	0.90
		Akurasi	0.91	0.91
		AUC	0.94	0.94
8	XGBoost Classifier	Precision	0.84	0.82
		Recall	0.86	0.80
		F1-score	0.85	0.81
		Akurasi	0.83	0.83
		AUC	0.92	0.92

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13198
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 7 reveals that the Naïve Bayes method outperforms other methods with a remarkable AUC value of 0.96. In contrast, the K-nearest neighbor method boasts the highest accuracy rate of 0.91. To evaluate the model, a confusion matrix illustrates the number of correct and incorrect predictions. This matrix representation effectively demonstrates the classification model's efficiency. It is important to note that the confusion matrix findings may vary depending on the data distribution. The TN, TP, FN, and FP values produced in the 90:10 data division are as follows in Table 8:

Table 8 Confusion Matrix Results on Test Size 0.1

No	Algorithms	Evaluation Metrics	Score
1	Random Forest Classifier	True Negative	17
		True Positive	9
		False Negative	2
		False Positive	2
2	Support Vector Machine	True Negative	17
		True Positive	10
		False Negative	1
		False Positive	2
3	Naïve Bayes	True Negative	17
		True Positive	9
		False Negative	2
		False Positive	2
4	Logistic Regression	True Negative	17
		True Positive	10
		False Negative	1
		False Positive	2
5	Decision Tree	True Negative	16
		True Positive	10
		False Negative	1
		False Positive	3
6	K- Nearest Neighbor	True Negative	17
		True Positive	9
		False Negative	2
		False Positive	2
7	XGBoost Classifier	True Negative	16
		True Positive	9
		False Negative	2
		False Positive	3

Table 8 shows that there are ten positive samples that the majority of models accurately anticipate. There are seventeen negative samples that the majority of models accurately anticipate. There are few falsely anticipated positive and negative samples in the 90:10 data split. Thus, it can be verified that this data division method works well for heart disease prediction. It does not, however, rule out the possibility that this kind of data division also yields a relatively tiny support value. We will next go over the evaluation metric's value using an 80:20 data divide, as shown in Table 9:

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13198
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 9 Confusion Matrix Results on Test Size 0.2

No	Algorithms	Evaluation Metrics	Score
1	Random Forest Classifier	True Negative	33
		True Positive	21
		False Negative	3
		False Positive	3
2	Support Vector Machine	True Negative	33
		True Positive	20
		False Negative	4
		False Positive	3
3	Naïve Bayes	True Negative	35
		True Positive	20
		False Negative	4
		False Positive	1
4	Logistic Regression	True Negative	32
		True Positive	20
		False Negative	4
		False Positive	4
5	Decision Tree	True Negative	28
		True Positive	20
		False Negative	4
		False Positive	8
6	K- Nearest Neighbor	True Negative	32
		True Positive	20
		False Negative	4
		False Positive	4
7	XGBoost Classifier	True Negative	29
		True Positive	20
		False Negative	4
		False Positive	7

According to Table 9's data, a maximum of 21 positive samples are accurately predicted. The maximum number of successfully predicted negative samples is 35. Compared to other methods, the technique still generates many positive samples incorrectly projected on the Decision Tree for the 80:20 data division. The evaluation metric's outcome using the confusion matrix on the 70:30 data split is as follows and is shown in Table 10:

Table 10 Confusion Matrix on Test Size 0.3

No	Algorithms	Evaluation Metrics	Score
1	Random Forest Classifier	True Negative	44
		True Positive	33
		False Negative	8
		False Positive	5
2	Support Vector Machine	True Negative	45
		True Positive	32
		False Negative	9
		False Positive	4
3	Naïve Bayes	True Negative	46
		True Positive	35

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i1.13198 p-ISSN : 2541-044X

		False Negative	6
		False Positive	3
4	Logistic Regression	True Negative	46
		True Positive	35
		False Negative	6
		False Positive	3
5	Decision Tree	True Negative	39
		True Positive	29
		False Negative	12
		False Positive	10
6	K- Nearest Neighbor	True Negative	46
		True Positive	36
		False Negative	5
		False Positive	3
7	XGBoost Classifier	True Negative	42
		True Positive	33
		False Negative	8
		False Positive	7

Accurately dividing data into training and testing sets is crucial in assessing the effectiveness of a predictive model. Table 10 (70:30) shows that the K-nearest neighbor (KNN) algorithm has achieved the highest score of 36 in predicting heart disease, outperforming other algorithms. Hence, it can be inferred that KNN is the most effective algorithm for predicting heart disease among the algorithms compared. The resulting support value in this data division typically tends to be higher than previously. Furthermore, the model with the confusion matrix is said to provide higher True Positive and True Negative values when Grid Search optimization is used.

DISCUSSIONS

The model's performance will be compared to earlier studies in this session. Grid Search yields the most outstanding results in this study with a 90:10 data division. Regarding precision, recall, f1-score, accuracy, and AUC, the evaluation metrics used to assess the performance of various machine learning algorithms, the Random Forest, SVM, and Logistic Regression algorithms, have demonstrated recall values of 91% and accuracy values of 90%. These results indicate that the algorithms above are highly effective and efficient for the tasks they were designed for. This indicates that the model can accurately predict heart disease using various data division techniques. Similar studies with varying accuracy and recall values were proposed by(Asif et al., 2023) (Bhatt et al., 2023) (Chandrasekhar & Peddakrishna, 2023). According to research, the Extra Tree Classifier algorithm yielded a recall value of 98.72%, and the Cat Boost and Random Forest algorithms produced an accuracy of 98.61%. The research used Grid Search CV to achieve a recall value of 84.85% and an accuracy of 87.02% using the XGBoost algorithm. Subsequent investigation yielded results of 95% for the Random Forest Classifier algorithm recall and 90% for the accuracy of Logistic Regression.

CONCLUSION

The performance of classification models improves significantly when the optimal parameters for machine learning approaches are found using GridSearch. The training and testing data sets must be divided appropriately to increase model accuracy. Of all the algorithms examined, Random Forest Classifier, SVM, and Logistic Regression yielded the highest accuracy of 90% in the 90:10 data split approach. At an 80:20 data split, Naïve Bayes achieved 92% accuracy without GridSearch, and Random Forest Classifier came in second with 90% accuracy. On a 70:30 approach, KNN obtained 91% accuracy, while Naïve Bayes and Logistic Regression achieved 90% accuracy. Given the nature of the research on the classification of heart disease, the recall value is quite significant. A 90:10 data split using GridSearch produced a recall value of 91%. This research brought significant fresh data to developing categorization models for heart disease. By increasing diagnostic accuracy, these models can assist healthcare providers in diagnosing patients with cardiac disease more accurately and swiftly.

*name of corresponding author

REFERENCES

Ahmad, A. A., & Polat, H. (2023). Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm. *Diagnostics*, *13*(14), 2392. https://doi.org/10.3390/diagnostics13142392

Ahmad, G. N., Fatima, H., Shafiullah, Salah Saidi, A., & Imdadullah. (2022). Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques with and Without GridSearchCV. *IEEE Access*, *10*, 80151–80173. https://doi.org/10.1109/ACCESS.2022.3165792

Apriyanto Alhamad, A. I. S. A. , B. S. , S. T. (2019). Prediksi Penyakit Jantung Menggunakan Metode-Metode Machine Learning Berbasis Ensemble – Weighted Vote. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)*, *5*(3), 352–360.

Asif, D., Bibi, M., Arif, M. S., & Mukheimer, A. (2023). Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization. *Algorithms*, *16*(6). https://doi.org/10.3390/a16060308

Ayon, S. I., Islam, M. M., & Hossain, M. R. (2022). Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques. *IETE Journal of Research*, *68*(4), 2488–2507. https://doi.org/10.1080/03772063.2020.1713916

Bays, H. E., Kulkarni, A., German, C., Satish, P., Iluyomade, A., Dudum, R., Thakkar, A., Rifai, M. Al, Mehta, A., Thobani, A., Al-Saiegh, Y., Nelson, A. J., Sheth, S., & Toth, P. P. (2022). Ten things to know about ten cardiovascular disease risk factors – 2022. In *American Journal of Preventive Cardiology* (Vol. 10). Elsevier B.V. https://doi.org/10.1016/j.ajpc.2022.100342

Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms*, *16*(2). https://doi.org/10.3390/a16020088

Biswas, N., Ali, M. M., Rahaman, M. A., Islam, M., Mia, M. R., Azam, S., Ahmed, K., Bui, F. M., Al-Zahrani, F. A., & Moni, M. A. (2023). Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques. *BioMed Research International*, *2023*. https://doi.org/10.1155/2023/6864343

Chandrasekhar, N., & Peddakrishna, S. (2023). Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization. *Processes*, *11*(4). https://doi.org/10.3390/pr11041210

Chauhan, N. K., & Singh, K. (2022). Performance Assessment of Machine Learning Classifiers Using Selective Feature Approaches for Cervical Cancer Detection. *Wireless Personal Communications*, *124*(3), 2335–2366. https://doi.org/10.1007/s11277-022-09467-7

Dissanayake, K., & Johar, M. G. M. (2021). Comparative study on heart disease prediction using feature selection techniques on classification algorithms. *Applied Computational Intelligence and Soft Computing*, *2021*. https://doi.org/10.1155/2021/5581806

Elreedy, D., & Atiya, A. F. (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. *Information Sciences*, *505*, 32–64. https://doi.org/10.1016/j.ins.2019.07.070

Hagan, R., Gillan, C. J., & Mallett, F. (2021). Comparison of machine learning methods for the classification of cardiovascular disease. *Informatics in Medicine Unlocked*, *24*, 100606. https://doi.org/10.1016/j.imu.2021.100606

Handayani, F., Sari Kusuma, K., Leoni Asbudi, H., Guines Purnasiwi, R., Kusuma, R., Sunyoto, A., & Mega Pradnya, W. (2021). Komparasi Support Vector Machine, Logistic Regression Dan Artificial Neural Network dalam Prediksi Penyakit Jantung. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)* , *7*(3), 329–334.

Janosi, A. S. W. P. M. & D. R. (1988). Heart Disease. In *UCI Machine Learning Repository*. https://archive.ics.uci.edu/dataset/45/heart+disease

Louridi, N., Douzi, S., & El Ouahidi, B. (2021). Machine learning-based identification of patients with a cardiovascular defect. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00524-9

Pamungkas, F. S., Prasetya, B. D., & Kharisudin, I. (2019). Perbandingan Metode Klasifikasi Supervised Learning pada Data Bank Customers Menggunakan Python. *PRISMA, Prosiding Seminar Nasional Matematika*, *3*, 689–694. https://journal.unnes.ac.id/sju/index.php/prisma/

Prabu, S., Thiyaneswaran, B., Sujatha, M., Nalini, C., & Rajkumar, S. (2022). Grid Search for Predicting Coronary Heart Disease by Tuning Hyper-Parameters. *Computer Systems Science and Engineering*, *43*(2), 737–749. https://doi.org/10.32604/csse.2022.022739

*name of corresponding author

Shi, Y., Zhang, H., Huang, S., Yin, L., Wang, F., Luo, P., & Huang, H. (2022). Epigenetic regulation in cardiovascular disease: mechanisms and advances in clinical trials. In *Signal Transduction and Targeted Therapy* (Vol. 7, Issue 1). Springer Nature. https://doi.org/10.1038/s41392-022-01055-2

Taranto-Vera, G., Galindo-Villardón, P., Merchán-Sánchez-Jara, J., Salazar-Pozo, J., Moreno-Salazar, A., & Salazar-Villalva, V. (2021). Algorithms and software for data mining and machine learning: a critical comparative view from a systematic review of the literature. *Journal of Supercomputing*, *77*(10), 11481–11513. https://doi.org/10.1007/s11227-021-03708-5

Tougui, I., Jilbab, A., & Mhamdi, J. El. (2020). Heart disease classification using data mining tools and machine learning techniques. *Health and Technology*, *10*, 1137–1144. https://doi.org/10.1007/s12553-020-00438-1/Published

Xu, J., Zhang, Y., & Miao, D. (2020). Three-way confusion matrix for classification: A measure driven view. *Information Sciences*, *507*, 772–794. https://doi.org/10.1016/j.ins.2019.06.064

*name of corresponding author