JOURNAL MEDICAL INFORMATICS TECHNOLOGY
medinftech.org ISSN 2988-7003 (online)

Optimization of The Machine Learning Approach using Optuna in Heart Disease Prediction

Wan Ahmad Gazali Kodri[1] *and Sri Hadianti*[2*]

[1,2]*Universitas Nusa Mandiri, Indonesia*

ARTICLE HISTORY
Received: 06 September 23
Final Revision: 10 September 23
Accepted: 11 September 23
Online Publication: 30 September 23

KEYWORDS
Machine Learning, Heart Disease, Aproach, Predicting, Optuna

CORRESPONDING AUTHOR
sri.shv@nusamandiri.ac.id

DOI
10.37034/medinftech.v1i3.15

A B S T R A C T

Heart disease prediction is a critical area in healthcare, as early identification and accurate assessment of cardiovascular risks can lead to improved patient outcomes. This study explores the application of machine learning techniques for predicting heart disease. Various data attributes, including medical history, clinical measurements, and lifestyle factors, are utilized to develop predictive models. A comprehensive analysis of different machine learning algorithms is conducted to determine their efficacy in classification tasks. The dataset used for experimentation is sourced from a diverse patient population, enhancing the generalizability of the findings. Through rigorous evaluation and validation, the study aims to identify the most suitable machine learning approach for effectively predicting heart disease. The results highlight the potential of machine learning as a valuable tool in assisting healthcare professionals in making informed decisions and providing personalized care to individuals at risk of heart disease.

1. Introduction

Cardiovascular diseases (CVDs) including coronary heart disease (heart attack), stroke, and heart failure are a major burden of disease globally [1]. According to the World Health Organization (WHO), CVD including Heart Disease (HD) is responsible for 31% of total deaths worldwide [2]. HD occurs when the heart is unable to provide enough blood throughout the body. This can be affected by high blood pressure, diabetes, coronary heart disease, and other heart problems or disorders [3].

The human body consists of several tissues. These tissues need oxygen and nutrients to work properly. The heart is the main organ that supplies blood to all parts of the body using the circulatory system. Through this system, it supplies nutrients and oxygen to the tissues. If there is a problem that causes the heart to not function properly, the circulatory system will experience a blockage and will cause heart failure [1], [4]. There are many forms of heart disease; However, cardiovascular disease (CVD) is the most lethal [2]. CVD is one of the diseases that causes the most deaths worldwide [3]. More than 31% of global deaths occurdue to heart failure. By 2030, it is predicted that there will be more than 22 million deaths due to heart problems [5]. The American Heart Association says that more than 121.5 million adults suffer from heart disease [6]. Several factors that cause heart disease, lack of exercise, smoking, drinking alcohol, poor lifestyle, eating junk food, etc., are the main factors for heart disease [7].

Doctors and medical professionals use angiography to treat heart disease, but there are some drawbacks associated with this method, including requiring human assistance, so it will take a lot of time to produce results, and because humans are operators, there is a high probability of getting wrong results, and the most importantly, this procedure is very expensive; everyone can't afford it. Therefore it is necessary to identify cardiovascular disease so that patients can take the necessary precautions to prevent a severe heart attack. In this study, disease identification was carried out using a machine learning approach with several methods, namely Random Forest (RF), Logistic Regression (LR), and k-Nearest Neighbor (KNN) models which were optimized using optuna.

2. Research Method

The overall research methodology is described in Figure 1, which starts from Exploratory Data Analysis (EDA), dataset preprocessing, upsampling to balance target variables, modeling using Random Forest (RF), Logistic Regression (LR), k-Nearest Neighbor (KNN), and evaluation to determine the best model in Figure 1.

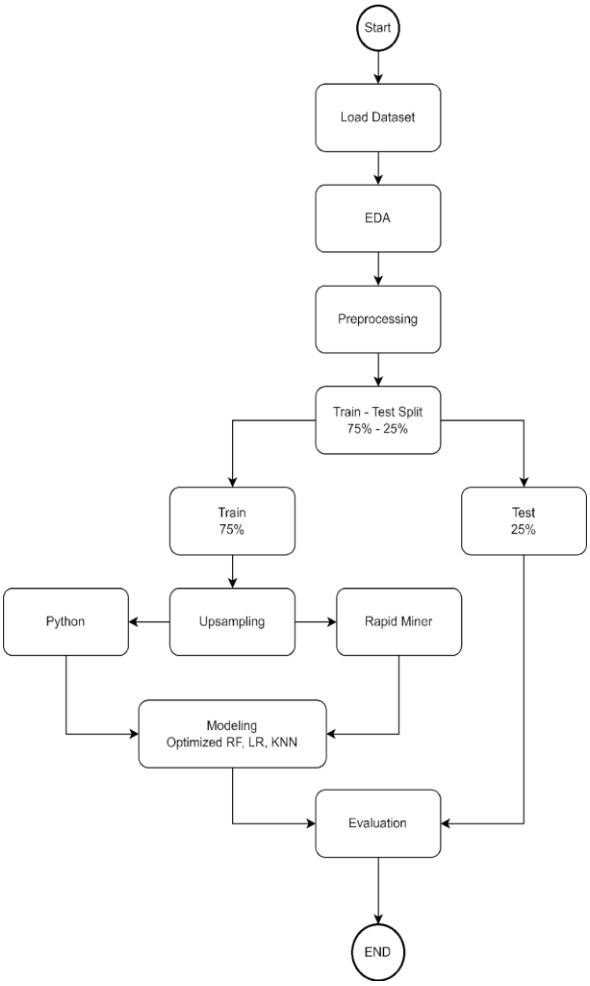

Figure 1. Research Methodology

2.1. Data Set

Researchers used the Cleveland dataset, which is a heart disease dataset collected by Andras Janosi, M.D. (Hungarian Institute of Cardiology. Budapest), William Steinbrunn, M.D. (University Hospital, Zurich, Switzerland), Matthias Pfisterer, M.D. (University Hospital, Basel, Switzerland) and Robert Detrano, M.D. Ph. D. (V.A. Medical Center, Long Beach and Cleveland Clinic Foundation). This dataset consists of 14 features (Table 1) and 606 observations. The main task of this dataset is to predict whether a patient has heart disease or not based on the attributes given. The goal of our experiments is to diagnose and find insights from this data set that can help in understanding heart disease.

2.2. Exploratory Data Analysis

We performed Exploratory Data Analysis (EDA) on the dataset to see data dimensions, data distribution, feature significance by Chi2 test and T test, correlation test using Pearson and multicollinearity test using variance inflation factor (VIF) in Table 1.

Table 1. Cleveland Dataset

No	Fitur	Deskripsi
1	age	Age of the patient in years
2	sex	Male/Female
3	cp	chest pain type ([typical angina, atypical angina, non-anginal, asymptomatic])
4	trestbps	resting blood pressure (resting blood pressure (in mm Hg on admission to the hospital))
5	chol	(serum cholesterol in mg/dl)
6	fbs	(if fasting blood sugar > 120 mg/dl)
7	restecg	resting electrocardiographic results ([normal, stt abnormality, lv hypertrophy])
8	thalach	maximum heart rate achieved
9	exang	exercise-induced angina (True/ False)
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	the slope of the peak exercise ST segment
12	ca	number of major vessels (0-3) colored by fluoroscopy
13	thal	[normal; fixed defect; reversible defect]
14	target	the predicted attribute

2.3. Preprocessing

Based on EDA, we preprocess the data by: Discarding duplicate observations (304 observations), Discarding outlier observations (13 observations) using z-score, Implement feature transformation. We apply feature transformation with the min-max data normalization method. Normalization is needed because some ML models try to search for and discover patterns in datasets by comparing attributes and data ranges. Differences in data scale will cause problems if used to train ML models. Simply put, scale differences between features will result in a weak ML model. To ensure each feature has the same scale, we apply min-max data normalization. Min-max normalization is a linear data transformation method, where the minimum value of the data becomes 0 and the maximum value of the data becomes 1. This method is applied to each feature. Each feature is normalized using this Equation (1).

$$x' = x - \max(x).\ x - \min(x) \qquad (1)$$

Where x' is new value of each entry, x is attribute data value, $\max(x)$ is absolute maximum value of A, and $\min(x)$ is absolute minimum value of A.

Apply categorical encoding. We apply categorical encoding to numeric features that have <= 5 unique records. We consider these features to be categorical features. Figure 4 exemplifies the transformation of numerical features into categorical features using categorical encoding.

2.4. Train – Test Split

By applying stratified random sampling, we divided this dataset into two parts, namely training data (75%) and test data (25%). The stratified division aims to maintain class proportions on the dependent variable

(55% for class 1 and 44% for class 0). Test data is not included in model training with the aim of avoiding overfitting and increasing fairness at the model evaluation stage so that the evaluation results truly illustrate the feasibility of the model

2.5. Upsampling

We apply the upsampling/oversampling method using SMOTE to balance the distribution of the dependent variable. SMOTE is an oversampling method that generates new synthetic samples using interpolation to balance the number in the dependent class. SMOTE will create synthetic samples based on the proximity of the samples to the minority class. Synthetic samples are generated by taking the difference between the feature vectors to be enhanced and the closest observation. This difference is then multiplied by a random number between 0 and 1 and added to the feature vector to be enhanced. This approach forces the decision region of the minority class to become more common. The newly generated synthetic minority class, x_{new}, is located between the observations x_i and x_k.

2.6. Tools

In building the prediction model, we used two tools, namely Python with Python version 3.10.12 and RapidMiner with version 10.1.003. Both of them processed the same data.

2.7. Modeling

We used the Random Forest (RF), Logistic Regression (LR), and k-Nearest Neighbors (KNN) models, optimized using Optuna in Python and Grid Search Optimization in RapidMiner (Figure 5), to predict heart disease. In both Python and RapidMiner, each model was trained using the 10-fold cross-validation method. Training with the 10-fold cross-validation method randomly divides the training data into 10 parts while maintaining the same proportion (stratified) for each class in each training and testing part.

Using OPTUNA for search is an efficient and beneficial approach considering the search speed and the improvement in model accuracy. OPTUNA is responsible for finding the best combination among the available hyperparameters. This step is called hyperparameter optimization [8].

RF is a type of classification algorithm that consists of multiple Decision Trees (DT), analogous to how a forest has many trees.

Deep DT can cause a problem known as overfitting during the training stage with the training dataset, resulting in significant changes in classification outcomes for small differences in test samples. Various DTs, which are part of the RF, are trained with different parts of the training dataset [9]. Input values must be sent along with each DT in the forest to identify new samples. Each DT then uses a specific part of the input values and returns its result as a classification output. The forest then selects the output with the highest "votes" (for categorical segmentation output) or the sum of all trees in the forest (for numerical segmentation output). Since the results from multiple DTs are considered by the RF, the variation caused by one DT for similar datasets will be reduced [10].

The LR model describes and estimates the relationship between one binary dependent variable, also known as the outcome variable, and one or more independent variables, also known as covariates or explanatory variables. The LR model has a strong interpretation. It is used to analyze retrospective data, including case-control studies, as well as to create prediction algorithms. LR is commonly used to solve two-class classification problems [11].

KNN is a generalization algorithm for the nearest neighbor rule. Its inductive bias is the class label of the k-nearest samples with the label closest to the test sample. The nearest neighbor rule can be described as a simple class determination, where the test sample is assigned the class of the nearest sample. If the training set and the distance metric remain unchanged, the decision outcome of the nearest neighbor rule will be uniquely determined for each test instance [12] in Figure 2.

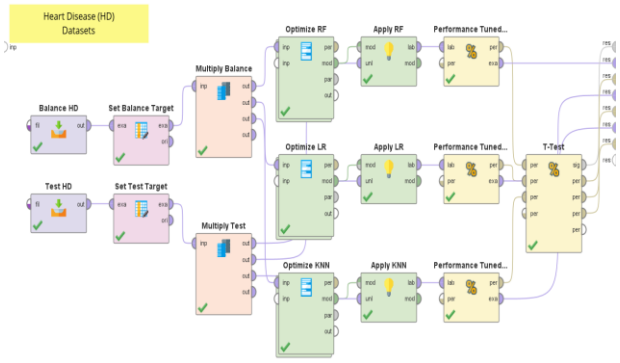

Figure 2. Modeling in Rapid Miner

2.8. Evaluation

In evaluating the model built, we use measurements of accuracy, precision, sensitivity, specificity, F-measure, g-mean, MCC and AUC. In a binary class classification task, there are two possible outcome classes: True (1) and False (0). The results of the correct and incorrect class predictions are depicted in the Confusion matrix in Table 2.

Tabel 2. Confusion Matrix

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

a. Accuracy

In evaluating prediction models for classification cases, the accuracy metric is the most commonly used metric. However, for predictions of class unequal classification cases, accuracy metrics can be misleading due to prediction bias towards the majority class. Therefore, other metrics are needed that are more useful in evaluating the reliability of the model. In classification, accuracy is defined as the ratio of the total number of correct predictions compared to the total number of instances, which can be described in the following Equation (2).

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

b. Precision

Precision or positive predictive value (PPV) is the ratio of correct predictions in the positive class compared to the overall positive predicted results. In general, a model with high precision is selected if the model user expects TP to occur and does not really expect FP to occur. The following is the equation for the Precision metric in the following Equation (3).

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

c. Sensitivity

Sensitivity can also be called recall, hit rate, or true positive rate (TPR), this metric describes the performance of a classification model in predicting positive classes. The high sensitivity value reflects that the classification model is reliable in predicting the positive class. Sensitivity is described in the following Equation (4).

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

d. Area Under Curve (AUC)

AUC is also called ROC-AUC. ROC stands for Receiver Operating Characteristics and AUC is the Area under the TPR vs. ROC Curve. False Positive Rate (FPR). A good model has a high TPR, and a low FPR. AUC has a value range from 0.5 to 1. The higher the AUC value, the better the model performance.

3. Result and Discussion

Model training using Python shows that the LR model is the best model compared to the RF and KNN models, while in Rapid Miner, the RF model is the best model compared to the LR and KNN models. Overall, this research produces the best RF model trained on Python with optimization using Optuna based on the accuracy (93.15%) and F-Measure (93.83%) metrics, however this model is not good enough in predicting TN which can be seen from the specificity metric in the RF rapid miner model (95%) outperforms this model (Table 6). In addition, the significance test shows that the fbs (p-value 0.761) and chol (p-value 0.158) features are features that are not significant in predicting HD. Feature selection using the significance test method increases performance on the optimized KNN model, but conversely decreases performance on the optimized RF model in Table 3 and Table 4.

Table 3. Model Performance Comparison (%)

Index	Python			Rapid Miner		
	Optuna-RF	Optuna-LR	Optuna-KNN	GS-RF	GS-LR	GS-KNN
Accuracy	93.15	84.93	80.82	86.30	83.56	83.56
Precision	92.68	83.72	80.95	92.59	86.21	80.00
Sensitivity/Recall	95.00	90.00	85.00	75.76	75.76	84.85
Specificity	90.91	78.79	75.76	95.00	90.00	82.50
F-Measure	93.83	86.75	82.93	83.33	80.65	82.35
Roc_Auc	92.95	84.39	80.38	94.00	94.50	95.20

Table 4. Modeling Evaluation Using Feature Selection Significance Test Results (%)

Index	Python					
	Optuna-RF	Optuna-LR	Optuna-KNN	Sel-Optuna-RF	Sel-Optuna-LR	Sel-Optuna-KNN
Accuracy	93.15	84.93	80.82	90.41	84.93	82.19
Precision	92.68	83.72	80.95	90.24	82.22	84.62
sensitivity/recall	95.00	90.00	85.00	92.50	92.50	82.50
specificity	90.91	78.79	75.76	87.88	75.76	81.82
F-Measure	93.83	86.75	82.93	91.36	87.06	83.54
roc_auc	92.95	84.39	80.38	90.19	84.13	82.16

The results of measuring feature importance in the best model (Optuna-RF) describe that cp with a value of 0 is the best predictor of HD, followed by ca, thal with a value of 2, oldpeak, thalach, and age in Figure 3.

Fitur	Skor Importance
cp_0.0	0.24323
ca	0.17686
thal_2.0	0.14736
oldpeak	0.09741
thalach	0.06579
age	0.06127
chol	0.04116
slope_1.0	0.03300
trestbps	0.02997
thal_3.0	0.02697
exang	0.02571
sex	0.01910
slope_2.0	0.01141
restecg_1.0	0.00793
restecg_0.0	0.00731
cp_2.0	0.00255
fbs	0.00162
cp_3.0	0.00130
cp_1.0	0.00005
thal_1.0	0.00000
slope_0.0	0.00000
restecg_2.0	0.00000

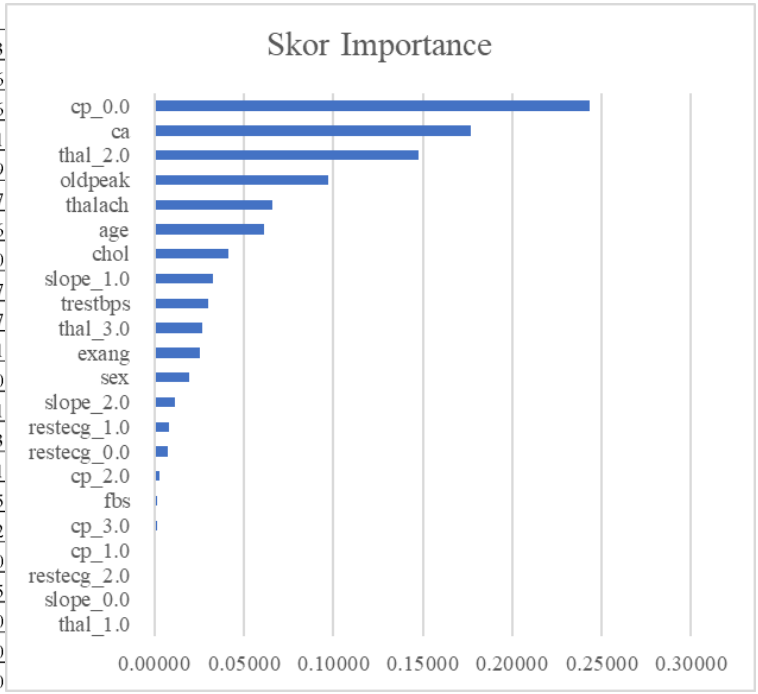

Figure 3. Feature Importance of the Optuna-RF model

The results of the EDA found that: There were 304 observations of duplication of data, no null data (NULL Value) in the observations, the dataset had an imbalanced class (imbalance class) with a distribution of 138 class 0 and 164 class 1, 44 observations were outliers if detected using the IQR method, 13 observations are outliers when detected using the z-score method in Figure 4.

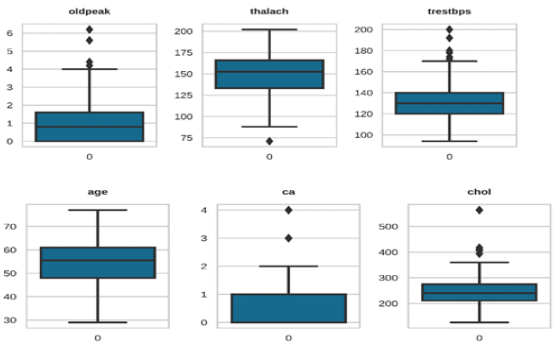

Figure 4. Boxplot for each numerical feature

The results of the correlation test show that there are no features that have a high correlation (<0.75) with the target in Figure 5.

Figure 5. Correlation Test Results

4. Conclusion

From the research results, it can be concluded that by using Optuna, the Machine Learning model can be optimized more efficiently and produce more accurate predictions in identifying heart disease. In this study, the author successfully improved the accuracy of heart disease predictions using optimization techniques provided by Optuna. As a result, the prediction of heart disease can be enhanced. This research has important implications in the field of health, where early detection of heart disease can assist in more effective diagnosis and treatment.

References

[1] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," BMC Med Inform Decis Mak, vol. 20, no. 1, Feb. 2020, doi: 10.1186/s12911-020-1023-5.

[2] World Health Organization, "Cardiovascular diseases (CVDs)," https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[3] L. and B. I. National Heart, "What Is Heart Failure, 2022," https://www.nhlbi.nih.gov/health/heart-failure.

[4] Stuart J. Pocock et al., "Predicting survival in heart failure: a risk scorebased on 39 372 patients from 30 studies," Eur Heart J, vol. 34, no. 19, pp. 1391–1392, May 2013, doi: 10.1093/eurheartj/ehs363.

[5] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, and D. S. Lee, "Using methods from the data-mining and machine-learning literature for disease classification and prediction: A case study examining classification of heart failure subtypes," J Clin Epidemiol, vol. 66, no. 4, pp. 398–407, Apr. 2013, doi: 10.1016/j.jclinepi.2012.11.008.

[6] S. A. Hunt et al., "2009 Focused Update Incorporated Into the ACC/AHA 2005 Guidelines for the Diagnosis and Management of Heart Failure in Adults. A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines Developed in Collaboration With the International Society for Heart and Lung Transplantation," Journal of the American College of Cardiology, vol. 53, no. 15. Elsevier USA, Apr. 14, 2009. doi: 10.1016/j.jacc.2008.11.013.

[7] D. S. Lee et al., "Relation of disease pathogenesis and risk factors to heart failure with preserved or reduced ejection fraction: Insights from the framingham heart study of the national heart, lung, and blood institute," Circulation, vol. 119, no. 24, pp. 3070–3077, Jun. 2009, doi: 10.1161/CIRCULATIONAHA.108.815944.

[8] P. Srinivas and R. Katarya, "44 hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost," Biomed Signal Process Control, vol. 73, Mar. 2022, doi: 10.1016/j.bspc.2021.103456.

[9] A. Nugroho and H. Suhartanto, "17_1 Hyper-Parameter Tuning based on Random Search for DenseNet Optimization," in 7th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2020 - Proceedings, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 96–99. doi: 10.1109/ICITACEE50144.2020.9239164.

[10] M.M. Raihan et al., "Chronic renal disease prediction using clinical data and different ML techniques," in 2nd International Informatics and Software Engineering Conference, IISEC, 2021.

[11] E. C. Zabor, C. A. Reddy, R. D. Tendulkar, and S. Patil, "Logistic Regression in Clinical Studies," Int J Radiat Oncol Biol Phys, vol. 112, no. 2, pp. 271–277, Feb. 2022, doi: 10.1016/j.ijrobp.2021.08.007.

[12] W. Xing and Y. Bei, "Medical Health Big Data Classification Based on KNN Classification Algorithm," IEEE Access, vol. 8, pp. 28808–28819, 2020, doi: 10.1109/ACCESS.2019.2955754..