Transactions on Materials, Biotechnology and Life Sciences
ISSN: 2960-1932, eISSN: 2960-2319 | Volume 3
BLSME 2024

Heart disease prediction utilizing machine learning techniques

Litian Chen *

Department of Statistics, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519087, China

* Corresponding Author Email: q030005005@mail.uic.edu.cn

Abstract. Heart disease is an international public health issue and a significant health risk for many people. The World Health Organization (WHO) reported that heart disease has been identified as one of the primary causes of death. Owing to its rapid development and application in many areas, Machine Learning has become an effective technique for predicting heart disease. Machine learning, along with large-scale medical data and advanced algorithms, can assist healthcare professionals in accurately predicting the risk of heart disease, thus providing early intervention and treatment for patients. This research paper uses the "heart_2020_cleaned.csv" dataset, containing 319795 instances and 18 attributes, of which 70% of instances were randomly selected for the training set and 30% for testing. Applying machine learning algorithms in data mining such as Decision Tree (DT), LightGBM, Random Forest (RF) and Logistic Regression (LR) to forecast heart disease. Before constructing models, data cleaning, feature selection and hyperparameter tuning processes were done, aiming to explore the potential patterns among data. Comparative Analysis was conducted on the external test set to compare the prediction performance of different models at the same level. The result reported that the highest accuracy achieved with LightGBM was 76.9%, followed by Logistic Regression and Random Forest, with Decision Tree being the worst.

Keywords: Heart disease prediction; Machine learning algorithms; Comparative analysis.

1. Introduction

1.1. Heart disease

Globally, Cardiovascular Disease (CVD) is the main cause of death. Until now, the number of CVD deaths has been steadily increasing. In 2019, a total of 9.6 million men and 8.9 million women passed away due to CVD, accounting for one -third of all global deaths.

The most widespread kind of CVD is coronary artery disease. Coronary artery disease occurs when the coronary artery is unable to transport enough oxygen-rich blood to the myocardium. Then a waxy material called plaque accumulates in the artery, which causes it to narrow. As time goes by, arteries that have become narrowed can lead to angina. Myocardial infarction, commonly known as a heart disease, is the consequence of a severe decrease or stoppage of blood flow to the heart [1].

1.2. Various risk factors

There are a variety of risk factors contributing to the probability of CVD, mainly including behavioral risk factors, potential factors and other factors.

Behavioral risk factors such as smoking, drinking alcohol in excess, lack of physical activity, an unhealthy diet, being overweight or obese, high cholesterol, high blood pressure and high blood sugar are all important for heart disease. Studies have shown that by implementing strategies to manage these risk factors, people can improve their cardiovascular health and reduce the risk of developing heart diseases and associated complications. Other factors include age, gender, and family history [2,3].

1.3. Machine learning (ML)

In the digital era, ML techniques have been widely adopted in various areas. Especially in the medical area, the collection and sharing of large-scale medical data have promoted the continual improvement of ML algorithms and provided effective assistance for disease prediction and diagnosis.

2. Literature review

Cardiovascular disease (CVD) is a major health issue that has been gaining attention worldwide due to its high prevalence. As people become more aware of their health, the importance of identifying and diagnosing heart disease has become increasingly important.

Machine learning algorithms have been successfully implemented in the prediction of heart disease. At present, in the area of heart disease prediction, the majority of studies evaluate the prediction performance of different machine learning models on the same benchmark. Some authors accept that tree or tree-based classifiers are more effective than traditional methods. The study in [5] achieved a higher accuracy of 99.5% with a Decision Tree using the "Heart Attack Prediction dataset of UCI" that includes 303 instances. The study in [6] achieved an accuracy of 99% with a Random Forest under the same dataset in a short time. However, some papers obtain different results on the prediction of heart disease. For instance, in the comparison between Decision Tree, Random Forest and Logistic Regression, the study in [7] shown that the latter displayed the best accuracy, at 84.3%.

Rather than concentrating on comparing machine learning algorithms to achieve better prediction performance, some researchers may prefer to identify effective features for predicting heart disease from a comprehensive set of features including personal information, physiological indicators, pathological features, health behavior and drug use. M Rizwan, S Arshad, et al. conducted a feature importance analysis and determined that the number of major arteries that are blocked was the most significant feature on heart disease predictions, followed by the maximum heart rate of the patient [8]. B Patil and D Kumaraswamy have conducted research to explore the correlations between features in predicting heart disease. They found that cholesterol, blood pressure, blood sugar and abnormal ECG were significant [9].

In this paper, the goal is to optimize the prediction performance, while taking into account feature selection and model comparison. At the same time, more reliable results can be obtained by increasing the sample size in the dataset that includes various features.

3. Methods

3.1. Data preparation

3.1.1. Data source

This study is a cross-sectional study. The "heart_2020_cleaned.csv" dataset we used is from Kaggle (https://www.kaggle.com/datasets/luyezhang/heart-2020-cleaned).

3.1.2. Data exploring

The dataset totally includes 319,795 instances and 18 attributes including target (HeartDisease). The descriptive summaries of the dataset are shown in Table 1:

Table 1. Attributes of the heart disease dataset

Feature	Levels	Mean±SD or N (%)
HeartDisease	No	292422 (91.40%)
	Yes	27373 (8.60%)
BMI	Mean ± SD	28.30 ± 6.40
Smoking	No	187887 (58.80%)
	Yes	131908 (41.20%)
AlcoholDrinking	No	298018 (93.20%)
	Yes	21777 (6.80%)
Stroke	No	307726 (96.20%)
	Yes	12069 (3.80%)
PhysicalHealth		3.40 ± 8.00
MentalHealth		3.90 ± 8.00
DiffWalking	No	275385 (86.10%)
	Yes	44410 (13.90%)
Sex	Female	167805 (52.50%)
	Male	151990 (47.50%)
AgeCategory	18-24	21064 (6.60%)
	25-29	16955 (5.30%)
	30-34	18753 (5.90%)
	35-39	20550 (6.40%)
	40-44	21006 (6.60%)
	45-49	21791 (6.80%)
	50-54	25382 (7.90%)
	55-59	29757 (9.30%)
	60-64	33686 (10.50%)
	65-69	34151 (10.70%)
	70-74	31065 (9.70%)
	75-79	21482 (6.70%)
	80 or older	24153 (7.60%)
Race	American Indian/Alaskan Native	5202 (1.60%)
	Asian	8068 (2.50%)
	Black	22939 (7.20%)
	Hispanic	27446 (8.60%)
	Other	10928 (3.40%)
	White	245212 (76.70%)
Diabetic	No	269653 (84.30%)
	No, borderline diabetes	6781 (2.10%)
	Yes	40802 (12.80%)
	Yes (during pregnancy)	2559 (0.80%)
PhysicalActivity	No	71838 (22.50%)
	Yes	247957 (77.50%)
GenHealth	Excellent	66842 (20.90%)
	Fair	34677 (10.80%)
	Good	93129 (29.10%)
	Poor	11289 (3.50%)
	Very good	113858 (35.60%)
SleepTime		7.10 ± 1.40
Asthma	No	276923 (86.60%)
	Yes	42872 (13.40%)
KidneyDisease	No	308016 (96.30%)
	Yes	11779 (3.70%)
SkinCancer	No	289976 (90.70%)
	Yes	29819 (9.30%)

3.1.3. Operating environment

This study is based on the R software program (version 4.3.1). RStudio provides an interactive and user-friendly graphical interface in an integrated development environment, allowing users to freely edit, debug and use various convenient functions.

R language has the ability to perform complex data analysis. The Tidyverse-package (version 2.0.0) is an integrated package for data processing and visualization, aiming to offer users a succinct and stylish programming experience. It can also help users to get familiar with the steps and fundamentals of data processing and improve the efficiency of their work. The Mlr3verse-package (version 0.2.8) provides a unified interface for a range of R language-based machine learning methods packages, with the goal of constructing a comprehensive machine learning workflow framework. Machine learning users can benefit from utilizing this process, which covers everything from data pre-processing to model evaluation, thus increasing the efficiency and quality of their models.

3.1.4. Data pre-processing

The “heart_2020_cleaned.csv” dataset totally includes 319,795 cases whose 27,373 are heart disease patients and 292,422 are normal individuals. The extreme imbalanced samples in the target often lead to unavoidable bias in model prediction. The model is likely to assign higher predictive probabilities to the category with a larger sample size, potentially leading to a high false positive rate. Randomly undersampling is a common method to solve the unbalanced problem. Since the number of cases in this study is sufficient, this method's shortcoming, which may lead to information loss, can be partially ignored. Finally, 50,000 completely balanced cases were selected and randomly divided in a ratio of 70% and 30%, for the training set and the test set respectively. Figure 1 explains the sequential chart of our proposed model. The sequential chart is shown in Figure 1.

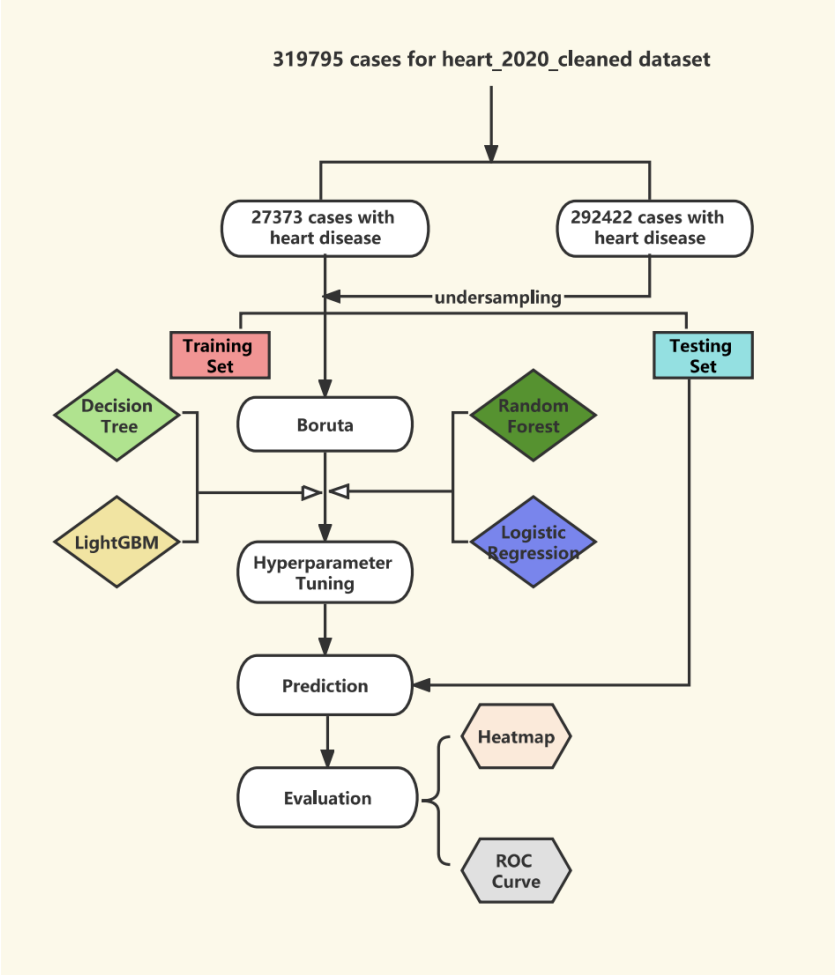

Figure 1. Sequential chart for the system

3.2. Feature selection approaches

Feature selection is a vital step that should be completed before constructing a model. Feature selection not only reduces the dimension of the dataset, thus decreasing model complexity, but it can also improve model prediction performance.

The Boruta algorithm introduces all the features in the dataset and then performs a top-down search by comparing the feature importance of the original variables with the shadow variables to eliminate irrelevant features [10]. The mean decrease accuracy is utilized to obtain the feature importance [11].

3.3. Hyperparameter tuning

All machine learning models have a set of default parameters. Tuning hyperparameters involves manually adjusting combinations of parameters. The hyperparameter search space chosen determines the complexity of the model, the training process and the prediction performance. However, it is a challenging work to obtain the optimal hyperparameter values. There is no universal definition for the "optimal" hyperparameter, as the results of different tasks, models and performance measures used in tuning can produce entirely dissimilar outcomes. In addition, identifying the ideal hyperparameter may take a lot of time. To achieve a balance between performance and speed, various novel hyperparameter tuning techniques have been developed and utilized, including Bayesian Optimization, Heuristic algorithm and Black-box Optimization [12].

Due to the fewer number of parameters to be adjusted, a Random Search technique is proposed with accuracy as the performance measure. Random Search is a popular technique for hyperparameter tuning, which involves randomly sampling the search space to generate and evaluate different combinations of hyperparameter.

3.4. Algorithms used

3.4.1. Decision Tree (DT)

DT algorithm involves creating a tree-like structure graph where features are divided based on a specific criterion. The root node acts as the beginning point of the tree, representing the entire dataset. Decision nodes located within the tree, are utilized to select features. Leaf nodes found at the bottom of the tree represents the final outcomes. Every node is connected with branches. With the help of recursive branch conditions, the tree is developed downwards. The DT algorithm's clearly structure makes it highly interpretable [13].

DT algorithm includes a variety of models. CART is currently one of the most popular frameworks. CART utilizes a greedy approach to select partition features and cut-off points that minimize the Gini Index [10]. The steps of CART include:

(1) For each feature in the dataset, select the feature with the lowest Gini Index as the root node. The Gini index of the root node can be obtained by the formula (1) and (2) as follows:

$$GINI(t) = 1 - \sum_j [p(t)]^2 \tag{1}$$

$$GINI_{split} = \sum_{t=1}^{k} \frac{n_t}{n} GINI\ (t) \tag{2}$$

(2) For each feature, traverse all of its possible values of each feature with the lowest Gini Index and determine the cut-off point as the decision node. The Gini index of the decision nodecan be obtained by the formula (3) and (4) as follows:

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2 \tag{3}$$

$$GINI_{split} = \sum_{t=1}^{k} \frac{n_t}{n} GINI\ (t) \quad (4)$$

(3) Repeat step 2 until it meets the termination condition.

The depth of a tree represents its complexity, which is controlled by the decision nodes. Obviously, as the tree becomes increasingly complex, the model is more likely to provide a higher accuracy for the training data. Generally, trees with moderate complexity (affected by the dataset size) are more preferred as they provide a balance between fitting and generalization, focusing on obtaining better prediction performance on unknown data. To achieve this, pre-processing or pruning are implemented. Pre-processing is a technique used in the formation of a decision tree to prevent it from growing too large. Pruning is commonly seen as a post-processing technique. After a tree has been created, some nodes and branches in the tree may be omitted in order to make it smaller.

3.4.2. LightGBM

Boosting is an effective method of combining multiple weak learners to construct a more powerful one that can successfully decrease the bias. In recent years, boosting-based ensemble algorithms have become increasingly popular from Adaboost, GBM, to XGBoost and the currently fashionable LightGBM.

LightGBM, which keeps a balance between prediction performance and training speed, is an efficient algorithm to process large-scale data. LightGBM algorithm is based on the gradient boosting technique, wherein the residuals are used as the loss function to iteratively refine the learner.

LightGBM algorithm utilizes some special techniques [14], including:

(1) EFB (Exclusive Feature Bundling): For unordered factors, the traditional approach involves one-hot coding, where each level of a factor is transformed into a binary feature. The potential risks of this process include dimensional explosion, data sparsity and an increasing complexity of the model, resulting in a higher cost of training and prediction. EFB technique is able to merge similar features, while preserving the information of the original features and reducing sparsity.

(2) Histogram Optimization: For tree-based models, when selecting features and cut-off points, a common approach is to traverse the feature values and compute the objective function such as the Gini Index. Histogram Optimization minimize the computational complexity and time consumption of global search by dividing features into a set of bins.

(3) Leaf-Wise: A technique for constructing tree structures. Unlike the more common approach that splits all nodes at the same level simultaneously, Leaf-Wise only splits the nodes that optimize the objective function. This is good for speeding up the process of modeling. As it may have a deeper vertical structure in this way, it is necessary to set a reasonable maximum depth for the boosting tree to avoid overfitting.

3.4.3. Random Forest (RF)

Bagging is another ensemble learning method that creates lots of learners using the bootstrap technique and combines their results to make predictions based on the majority, which can effectively decrease the variance.

Ideally, each learner is independent, so:

$$Var\ \left(\frac{\sum f(\boldsymbol{x};D)}{n}\right) = \frac{Var\ (f(\boldsymbol{x};D))}{n} \quad (5)$$

As the quantity of learners is more than 1, so:

$$Var\ \left(\frac{\sum f(\boldsymbol{x};D)}{n}\right) < Var\ (\sum f(\boldsymbol{x};D)) \quad (6)$$

RF algorithm is an application of the bagging method in the DT framework. Due to the risk of overfitting in a tree, it is always beneficial to combine multiple trees to obtain more convincing results. The randomness of the RF algorithm mainly consists of two parts: the features and the instances that are randomly selected for each tree, both of which are obtained through the bootstrap technique according to the pre-determined ratio [15].

3.4.4. Logistic Regression (LR)

LR algorithm is a popular model in machine learning due to its computational efficiency and interpretability [15].

The assumptions of the LR algorithm are: The target variable follows a Bernoulli distribution, and that the probability of the log transformation has a linear or nearly linear relationship with the linear combinations of the independent variables. Then the equation of logistic regression is defined as:

$$logit(p) = log\ \frac{p(y=1|x)}{p(y=0|x)} = w^T x + b \tag{7}$$

Actually, the logistic function is utilized as the link function in the LR algorithm to map the results of linear regression into a range between 0 and 1, thus obtaining the prediction probabilities. The logistic function is:

$$f = \frac{1}{1+e^{-w^T x+b}} \tag{8}$$

LR models can be constructed using a feature or several features. It is important to note that after performing multiple LR, it is necessary to assess the correlation (multicollinearity) of each variable using the Generalized Variance Inflation Factor (GVIF). Commonly, when GVIF is greater than 5, it indicates the presence of multicollinearity in the given variable, which can lead to an unstable parameter estimation and an unclear interpretation of the model. To address this, regularization and feature selection techniques can be utilized.

Comparing a univariate model and a multivariate model is meaningful, as the latter considers the cumulative effect of multiple features on prediction. Comparison can also be utilized to assess the model's robustness. If two results are similar, the model has enough reliability in interpretation.

3.5. Performance measure

Performance measures are utilized to assess the prediction results on various aspects. Not only are they used for the final model evaluation, but they also play an important role in feature selection and hyperparameter tuning to make improvements to the model. For a binary classification problem, classifiers obtain two kinds of outcomes: positive or negative. The selection of appropriate performance measures is vital.

3.5.1. Confusion matrix

For a binary classification problem, the confusion matrix presents a 2x2 table that compares the actual labels with the predictions, which is shown in Table 2 [7].

Table 2. Confusion matrix for the system

Prediction	Reference	
	Positives	Negatives
Positives	True Positives (TP)	False Positives (FP)
Negatives	False Negatives (FN)	True Negatives (TN)

According to the confusion matrix, some performance measures can be calculated using the equations as follows:

$$Sensitivity\ (\%) = Recall\ (\%) = \frac{TP}{TP+FN} \tag{9}$$

$$Specificity\ (\%) = \frac{TN}{TN+FP} \tag{10}$$

$$Accuracy\ (\%) = \frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

$$Precision\ (\%) = \frac{TP}{TP+FP} \tag{12}$$

But in some cases, it is difficult for the traditional performance measures above to make fair judgements. For example, when the target is unbalanced, the model is likely to make predictions for the majority, making it difficult to assess the model's performance.

To address these issues, advanced performance measures such as the Matthews Correlation Coefficient (MCC) and F1-Score should be used. The equations for these measures are as follows:

$$MCC = \frac{(TP\times TN - FP\times FN)}{\sqrt{[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]}} \tag{13}$$

$$F1-Score = \frac{2\times(Precision\times Recall)}{Precision+Recall} \tag{14}$$

3.5.2. Receiver Operating Characteristic (ROC) Curve

ROC Curves provide sensitivity and specificity of the model at various classification thresholds [16]. AUROC is equivalent to the probability that sensitivity is higher than specificity, that is:

$$AUROC = Prob(Sensitivity > 1 - Specificity) \tag{15}$$

The AUROC ranges from 0 to 1; the higher the value, the better the predictive performance. In general, when AUC is equal to 0.5, the model has no ability to classify cases; when AUC is higher than 0.8, the model has excellent classification ability; when AUC is equal to 1, there exists at least one classification threshold that allows the model to perfectly classify all the cases.

4. Results

4.1. Feature selection

The results of feature selection using Boruta algorithm are shown in Figure 2. The study initially provided a total of 17 features and retained all of them. AgeCategory has a much greater impact on prediction than other features.

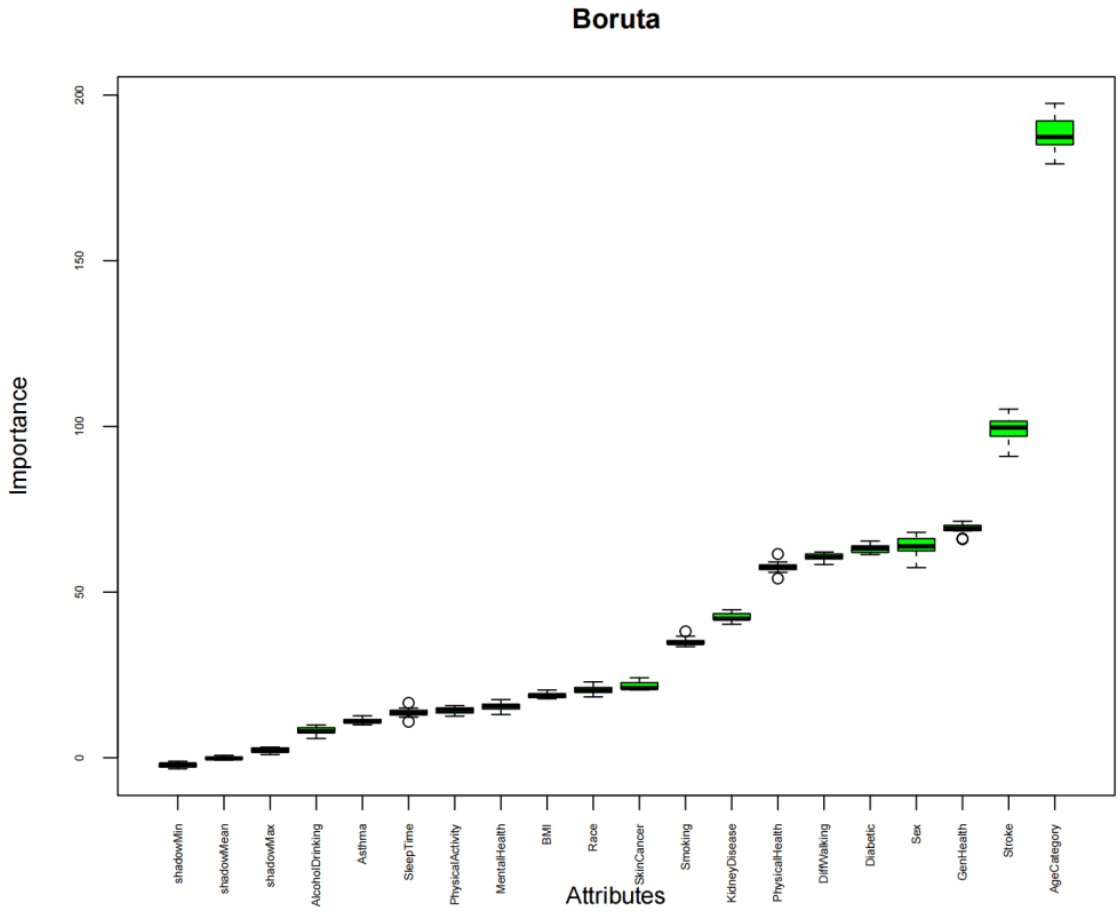

Figure 2. Boxplots of the feature selection results using the Boruta algorithm

4.2. Machine learning models

4.2.1. Tree

The final tree of DT algorithm is shown in Figure 3. AgeCategory is the root node located at the top of the tree, indicating that age is the most significant feature impacting the tree predictions. As the root node, AgeCategory is split into two branches: 18-54 and 55-80 or older, implying that those in the older age groups are more likely to suffer from heart disease. GenHealth is the second most significant feature, showing that sometimes people's self-assessment can be linked to whether they have heart disease.

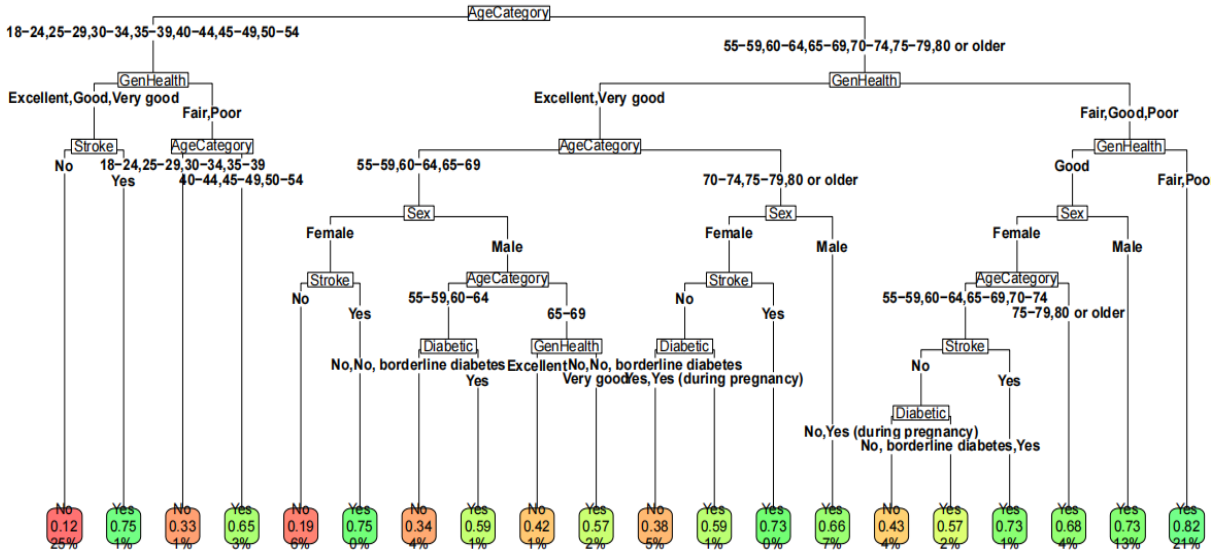

Figure 3. The Final structure of the Decision Tree after tuning hyperparameters

4.2.2. LR

The results of LR are shown in Table 5. A univariate and a multivariate LR models were constructed, wherein the latter used all 17 features. In the multivariate model, all features except for PhysicalActivity shown a significant influence on predictions (significant if p value < 0.05). The small GVIF of each feature indicated that there was no multicollinearity. In addition, the univariate and the multivariate LR models showed similar results, indicating that the models were robust.

Table 3. Result summaries of the Logistic Regression models

	Describe	Univariate Models			Multivariate Model			
Characteristic	N = 38,322[1]	OR[2]	95%- CI[2]	p value	OR[23]	95%- CI[2]	p value	GVIF[2]
HeartDisease	19,132 (50%)							
BMI	28 (24, 32)	1.03	1.02, 1.03	<0.001	1.01***	1.00, 1.01	<0.001	1.2
Smoking	18,768 (49%)							1.1
No		—	—		—	—		
Yes		2.14	2.06, 2.23	<0.001	1.45***	1.38, 1.53	<0.001	
AlcoholDrinking	2,098 (5.5%)							1.0
No		—	—		—	—		
Yes		0.57	0.52, 0.62	<0.001	0.84**	0.75, 0.94	0.002	
Stroke	3,566 (9.3%)							1.0
No		—	—		—	—		
Yes		7.08	6.44, 7.81	<0.001	3.40***	3.06, 3.79	<0.001	
PhysicalHealth	0 (0, 5)	1.05	1.05, 1.06	<0.001	1.00*	1.00, 1.01	0.014	1.7
MentalHealth	0 (0, 3)	1.01	1.01, 1.01	<0.001	1.01***	1.00, 1.01	<0.001	1.2
DiffWalking	9,283 (24%)							1.4
No		—	—		—	—		
Yes		4.20	3.99, 4.43	<0.001	1.24***	1.16, 1.33	<0.001	
Sex								1.1
Female	18,164 (47%)	—	—		—	—		
Male	20,158 (53%)	1.70	1.64, 1.77	<0.001	2.13***	2.02, 2.25	<0.001	
AgeCategory								1.4
18-24	1,426 (3.7%)	—	—		—	—		
25-29	1,205 (3.1%)	1.31	0.98, 1.77	0.071	1.10	0.81, 1.49	0.556	
30-34	1,381 (3.6%)	1.84	1.41, 2.42	<0.001	1.50**	1.13, 1.99	0.005	
35-39	1,595 (4.2%)	2.32	1.80, 3.01	<0.001	1.65***	1.27, 2.17	<0.001	
40-44	1,699 (4.4%)	3.51	2.76, 4.50	<0.001	2.29***	1.78, 2.97	<0.001	
45-49	1,783 (4.7%)	5.73	4.55, 7.30	<0.001	3.36***	2.63, 4.33	<0.001	

	Describe	Univariate Models			Multivariate Model			
Characteristic	N = 38,322[1]	OR[2]	95%- CI[2]	p value	OR[23]	95%- CI[2]	p value	GVIF[2]
50-54	2,562 (6.7%)	8.82	7.07, 11.1	<0.001	4.95***	3.91, 6.32	<0.001	
55-59	3,286 (8.6%)	12.8	10.3, 16.2	<0.001	6.56***	5.22, 8.35	<0.001	
60-64	4,336 (11%)	17.2	13.9, 21.6	<0.001	8.68***	6.92, 11.0	<0.001	
65-69	4,945 (13%)	21.3	17.2, 26.7	<0.001	11.1***	8.86, 14.1	<0.001	
70-74	5,116 (13%)	28.2	22.7, 35.3	<0.001	14.4***	11.5, 18.3	<0.001	
75-79	3,947 (10%)	36.6	29.4, 46.0	<0.001	18.6***	14.8, 23.7	<0.001	
80 or older	5,041 (13%)	44.7	36.0, 56.1	<0.001	25.7***	20.5, 32.8	<0.001	
Race								1.1
American Indian/Alaskan Native	714 (1.9%)	—	—		—	—		
Asian	658 (1.7%)	0.26	0.21, 0.33	<0.001	0.58***	0.43, 0.77	<0.001	
Black	2,550 (6.7%)	0.64	0.54, 0.76	<0.001	0.70***	0.57, 0.86	<0.001	
Hispanic	2,736 (7.1%)	0.44	0.37, 0.52	<0.001	0.74**	0.60, 0.92	0.006	
Other	1,274 (3.3%)	0.74	0.61, 0.89	0.001	0.90	0.71, 1.13	0.370	
White	30,390 (79%)	0.81	0.70, 0.94	0.006	0.85	0.71, 1.03	0.097	
Diabetic								1.1
No	28,734 (75%)	—	—		—	—		
No, borderline diabetes	924 (2.4%)	2.02	1.77, 2.31	<0.001	1.23**	1.06, 1.44	0.008	
Yes	8,421 (22%)	3.93	3.72, 4.15	<0.001	1.65***	1.54, 1.76	<0.001	
Yes (during pregnancy)	243 (0.6%)	0.63	0.47, 0.82	<0.001	1.15	0.82, 1.59	0.422	
PhysicalActivity	27,318 (71%)							1.2
No		—	—		—	—		
Yes		0.49	0.47, 0.51	<0.001	1.01	0.95, 1.07	0.785	
GenHealth								1.8
Excellent	5,240 (14%)	—	—		—	—		
Fair	6,764 (18%)	11.1	10.2, 12.1	<0.001	4.75***	4.27, 5.28	<0.001	
Good	12,189 (32%)	4.81	4.46, 5.20	<0.001	2.72***	2.49, 2.98	<0.001	
Poor	3,236 (8.4%)	20.9	18.6, 23.4	<0.001	6.36***	5.47, 7.40	<0.001	

	Describe	Univariate Models			Multivariate Model			
Characteristic	N = 38,322[1]	OR[2]	95%- CI[2]	p value	OR[23]	95%- CI[2]	p value	GVIF[2]
Very good	10,893 (28%)	2.08	1.93, 2.25	<0.001	1.51***	1.39, 1.65	<0.001	
SleepTime	7.00 (6.00, 8.00)	1.02	1.01, 1.03	0.001	0.98*	0.96, 1.00	0.012	1.1
Asthma	5,947 (16%)							1.1
No		—	—		—	—		
Yes		1.47	1.39, 1.56	<0.001	1.41***	1.31, 1.51	<0.001	
KidneyDisease	2,992 (7.8%)							1.0
No		—	—		—	—		
Yes		4.95	4.50, 5.44	<0.001	1.86***	1.68, 2.08	<0.001	
SkinCancer	5,095 (13%)							1.1
No		—	—		—	—		
Yes		2.45	2.30, 2.61	<0.001	1.17***	1.08, 1.26	<0.001	
[1]n (%)								
[2]OR = Odds Ratio, CI = Confidence Interval, GVIF = Generalized Variance Inflation Factor								
[3]*p < 0.05; **p < 0.01; ***p < 0.001								

4.3. Performance comparison

The comparative results of external test set are shown in Figure 4. The Heatmap demonstrates that LightGBM is the better choice for heart disease prediction, achieving an accuracy of 76.9%, while LR is slightly behind at 76.8%. All of the models have significantly lower sensitivity than specificity, meaning that they were more likely to overestimate the probability of having heart disease.

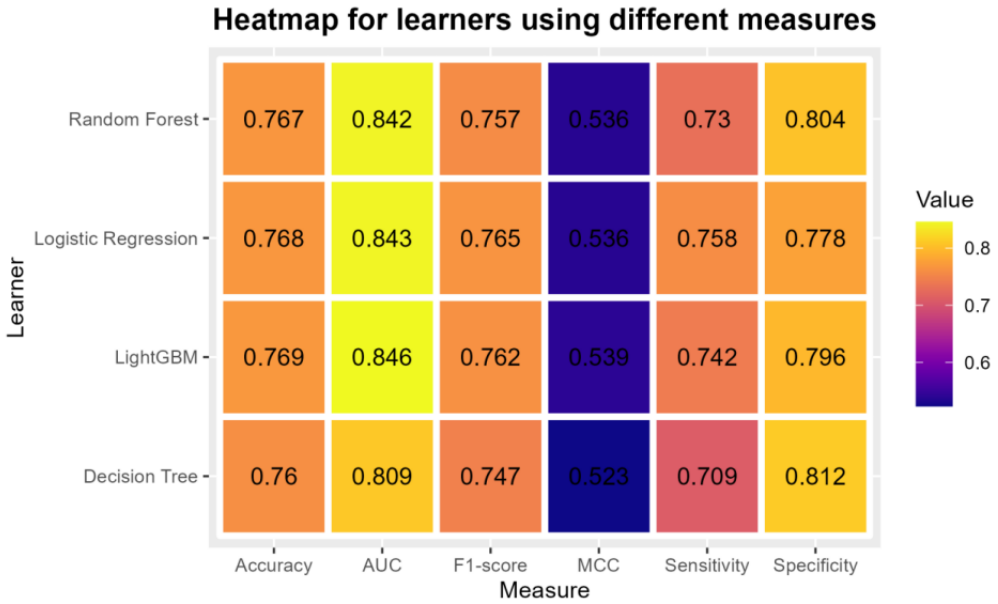

Figure 4. Comparative heatmap using prediction performance measures obtained from machine learning algorithms

Receiver Operating Characteristic (ROC) Curves are shown in Figure 5. Based on it, the values of AUROC on the following three models were similar and ideal (AUROC of 84.3% for LR, AUROC of 84.2% for RF, AUROC of 84.6% for LightGBM) . The DT obtains a lower AUROC of 80.9%.

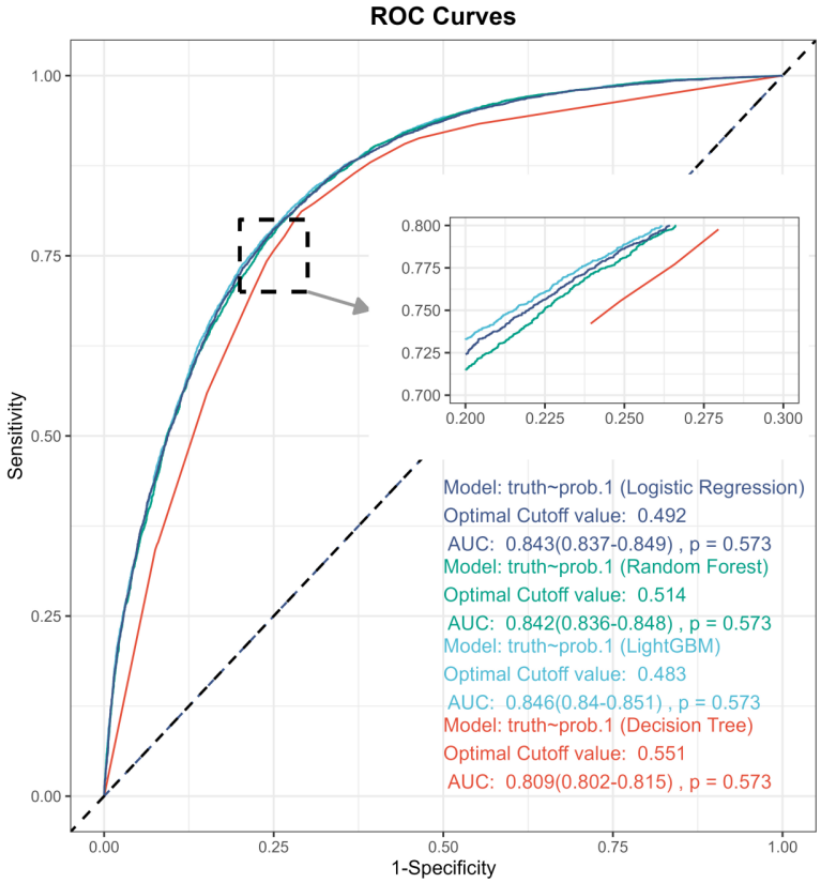

Figure 5. ROC Curve obtained from ML algorithms

5. Discussion

Heart disease is one of the most frequent causes of mortality globally. As lifestyles have changed and unhealthy habits have become more prevalent in recent years, the rate of heart disease is steadily increasing, making heart disease prediction play an increasingly significant role in modern life.

The purpose is to utilize supervised classification algorithms such as DT, LightGBM, RF and LR to predict whether an individual has heart disease or not. Several other algorithms including Naïve Bayes (NB), WARM, KNN, XGB, SVM have been observed in similar studies. For different studies, comparison of accuracy of heart disease prediction with different data sources and algorithms is given in Table 4.

It is likely that the differences arising from the papers above are due to the fact that a different data source was used. Heart disease prediction is a complex process, as it takes into account various features and the selection of different subjects, both of which can significantly influence the model's prediction performance.

Table 4. Comparison of results with other papers

Reference	Dataset	Model	Accuracy
[5]	-	NB	80.444%
		WARM	58.552%
		DT	99.5%
[6]	Heart Attack Analysis & Prediction Dataset from Kaggle	NB	88.2%
		KNN	90.8%
		DT	80.3%
		RF	86.8%
[7]	Heart Attack Prediction dataset of UCI	LR	81.9%
		DT	81.4%
		RF	77%
		SVM	66.2%
[8]	Heart Attack Analysis & Prediction Dataset from Kaggle	DT	86.8%
		KNN	90.1%
		LR	88.5%
		RF	88.5%
		XGB	86.8%
		SVM	70.4%
Current Study	"heart_2020_cleaned.csv" dataset	DT	76%
		LightGBM	76.9%
		RF	76.7%
		LR	76.8%

Aiming to obtain classification accuracy as high as possible on heart disease prediction, LR, RF, LightGBM and DT algorithms were applied. LR, RF and LightGBM shown the best and similar results, with LightGBM having the highest accuracy of 76.9% on a 30% external test set. One of the most important features we found is AgeCategory, meaning that elderly people should pay more attention to the prevention of heart disease. Our research's limitation is the inadequate use of information due to the undersampling method used to deal with imbalanced target. Future work could apply more advanced imbalanced techniques to improve information utilization and obtain more accurate predictions.

Appendix A

Table 5. Results of hyperparameter tuning on ML algorithms

Learner	Hyperparameter	Before Tuning (Default)	Tuning Range	Logscale	After Tuning
DT	minsplit	20	{50, 51, ..., 100}		54
	cp	0.01	[0.001, 0.1]	√	0.001
	maxdepth	30	{3, 4, ..., 7}		7
	minbucket	NONE	{5, 6, ..., 10}		5
LightGBM	num_iterations	100	[50, 1000]	√	173
	learning_rate	0.1	[0.001, 0.1]	√	0.053
	max_depth	NONE	{3, 4, ..., 12}		3
	feature_fraction	1	[0.4, 1]		0.466
	lambda_l1	0	[0.1, 10]	√	1.903
	lambda_l2	0	[0.1, 10]	√	1.278
	boosting	"GBDT"	NONE		"GBDT"
RF	min.node.size	NONE	{5, 6, ..., 50}	√	34
	mtry	NONE	{3, 4, ..., 14)		3
	num.trees	500	(100, 101, ..., 1000)	√	745

References

[1] What is a heart attack? American Heart Association. https://www.heart.org/en/health-topics/heart-attack/about-heart-attacks. Accessed March 29, 2022.

[2] Gregory A. Roth, George A. Mensah, Catherine O. Johnson, et al. Global burden of cardiovascular diseases and risk factors, 1990-2019: Update from the GBD 2019 Study. J Am Coll Cardiol. Dec 09, 2020.

[3] Gregory A. Roth, George A. Mensah, Valentin fuster the global burden of cardiovascular diseases and risks: a compass for global action. J Am Coll Cardiol. Dec 09, 2020.

[4] Heart attack. National heart, lung, and blood institute. https: //www.nhlbi.nih.gov/health/heart-attack/causes. Accessed March 29, 2022.

[5] J. N, D. P, M. E, R. Santhosh, R. Reshma and D. Selvapandian, "Heart attack prediction using machine learning," 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2022, pp. 854-860, doi: 10.1109/ICIRCA54612. 2022. 9985736.

[6] V. Sharma, S. Yadav and M. Gupta, "heart disease prediction using machine learning techniques," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020, pp. 177 - 181, doi: 10.1109/ICACCCN51052.2020.9362842.

[7] J. S. Rose, P. Malin Bruntha, S. Selvadass, R. M. V, B. C. Mary M and M. J. D, "Heart attack prediction using machine learning techniques," 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2023, pp. 210-213, doi: 10. 1109/ICACCS57279. 2023. 10113045.

[8] M. Rizwan, S. Arshad, H. Aijaz, R. A. Khan and M. Z. U. Haque, "Heart attack prediction using machine learning approach," 2022 Third International Conference on Latest trends in Electrical Engineering and Computing Technologies (INTELLECT), Karachi, Pakistan, 2022, pp. 1-8, doi: 10.1109/INTELLECT55495. 2022. 9969395.

[9] S. B. Patil and D. Kumaraswamy, "Extraction of significant patterns from heart disease warehouses for heart attack prediction", IJCSNS International Journal of Computer Science and Network Security, vol. 9, no. 2, pp. 228 - 235, 2009.

[10] R. Tang and X. Zhang, "CART Decision Tree combined with Boruta feature selection for medical data classification," 2020 5th IEEE International Conference on Big Data Analytics (ICBDA), Xiamen, China, 2020, pp. 80 - 84, doi: 10.1109/ICBDA49040. 2020. 9101199.

[11] A. Behnamian, K. Millard, S. N. Banks, L. White, M. Richardson and J. Pasher, "A systematic approach for variable selection with random forests: achieving stable variable importance values," in IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 11, pp. 1988 - 1992, Nov. 2017, doi: 10. 1109/LGRS. 2017. 2745049.

[12] F. Arden and C. Safitri, "Hyperparameter tuning algorithm comparison with machine learning algorithms," 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, 2022, pp. 183 - 188, doi: 10.1109/ICITISEE57756. 2022. 10057630.

[13] Kappelhof N, Ramos LA, Kappelhof M, van Os HJA, Chalos V, van Kranendonk KR, Kruyt ND, Roos YBWEM, van Zwam WH, van der Schaaf IC, van Walderveen MAA, Wermer MJH, van Oostenbrugge RJ, Lingsma H, Dippel D, Majoie CBLM, Marquering HA. Evolutionary algorithms and decision trees for predicting poor outcome after endovascular treatment for acute ischemic stroke. Comput Biol Med. 2021 Jun; 133: 104414. doi: 10.1016/j.compbiomed.2021. 104414. Epub 2021 Apr 21. PMID: 33962154.

[14] Z. Wang, H. Ren, R. Lu and L. Huang, "Stacking based LightGBM-CatBoost-RandomForest algorithm and its application in big data modeling," 2022 4th International Conference on Data-driven Optimization of Complex Systems (DOCS), Chengdu, China, 2022, pp. 1-6, doi: 10.1109/DOCS55193. 2022. 9967714.

[15] S. Naveen, S. K. Ravindran, S. G and S. N. Ameen, "Effective heart disease prediction framework using Random Forest and Logistic regression," 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN), Vellore, India, 2023, pp. 1-6, doi: 10.1109/ViTECoN58111. 2023. 10157078.

[16] P. Adeodato and S. Melo, "Kolmogorov-Smirnov and ROC curve metrics for binary classification performance assessment are equivalent," 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 2022, pp. 1194 - 1199, doi: 10. 1109/ICPR56361. 2022. 9956449.

[17] P. Pongthanoo and W. Songpan, "Feature selection and reduction based on SMOTE and information gain for sentiment mining," 2020 5th International Conference on Computer and Communication Systems (ICCCS), Shanghai, China, 2020, pp. 109 - 114, doi: 10. 1109/ICCCS49078. 2020. 9118467.

[18] T. Manvitha and K. S. Rekha, "Improved accuracy for prediction of leaf wetness using Logistic Regression algorithm compared with Decision Tree algorithm," 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2023, pp. 1 - 5, doi: 10.1109/ICONSTEM56934. 2023. 10142550.

[19] Rethinking drinking: alcohol and your health. National Institute on Alcohol Abuse and Alcoholism. https: //www.niaaa.nih.gov/niaaa-publications-order-form#pub-1. Accessed March 29, 2022.

[20] Health Education & Content Services (Patient Education). Sex and heart disease. Mayo Clinic; 2011.

[21] 2020-2025 Dietary Guidelines for Americans. U.S. Department of Health and Human Services and U.S. Department of Agriculture. https: //www.dietaryguidelines.gov. Accessed March 29, 2022.

[22] coronary artery disease. Mayo Clinic. https: //www.mayoclinic.org/diseases-conditions/coronary-artery-disease/symptoms-causes/syc-2 0350613. Accessed March 29, 2022.