Prediction of heart attack based on ensemble learning

Xiaotong Zhang

School of Mechanical, Electrical & Information Engineering, Shandong University, Jinan, 264209, China

202100800167@mail.sdu.edu.cn

Abstract. Heart disease has always been a high proportion of diseases worldwide, and it is also one of the leading causes of death worldwide. The prevention and detection of heart disease is still a key issue. To enhance patient outcomes in the real world and add to the expanding body of knowledge in heart disease research, this work aims to develop an integrated learning model to predict heart disease using machine learning techniques. The main research object of this paper is the heart disease data set. This paper's primary approach is to use 10 distinct machine-learning algorithms to construct a model for predicting heart attacks. And then, score and compare the performance of the models with some fundamental metrics. According to the result of the performance, the best three are selected, and a better-ensembled model is built through the ensemble learning of the voting method with Multilayer Perceptron Classifier (MLPC), the Support Vector Machine (SVM), and the Naive Bayes. This may be likely to help with real-life diagnoses in the future.

Keywords: Heart-attack prediction, Machine learning, Classification, Ensemble learning.

1. Introduction

According to the latest World Health Statistics report, non-communicable chronic diseases (NCDs) are responsible for the highest burden of disease worldwide, with heart disease accounting for the largest share. However, traditional medical methods make it difficult to accurately analyze and diagnose the disease. Machine learning (ML) is currently widely applied in many different domains to help humans solve a variety of complex challenges, such as healthcare, finance, environment, marketing, security, and industry. The ML methods are characterized by the ability to examine a great deal of data and discover the correlations, providing explanations for analysis. Moreover, ML can help improve the reliability, performance, predictability, and accuracy of the diagnosis of many diseases.

A variety of classification algorithms learn from the collected data and can make correct diagnoses and prediction probability of the diseases, combined with the analysis of experts in related diseases, to achieve better medical diagnosis results and improve the prevention and diagnosis of diseases.

The researchers developed a heart disease prediction system using three data mining classification modeling techniques, including Naive Bayes, Neural Networks, and Decision Trees [1]. In this research paper, the Naive Bayes algorithm, Decision List, and KNN are used to improve the classification accuracy of the data set for diagnosis [2]. The paper focused on the use of different algorithms and the combination of some target attributes to make predictions of heart disease [3]. The researchers added two additional attributes obesity and smoking, using three algorithms - Decision Trees, Naive Bayes, and Neural Networks - to predict and compare them [4].

This paper begins with a simple exploratory data analysis (EDA) of the heart disease dataset to better understand the data. Then ten kinds of ML classification algorithms are used to classify and predict the disease, and the performance of each algorithm in dealing with the problem is compared. According to the comparison results, three algorithms are selected to integrate the model using a voting mechanism, and an integrated classifier is established. This paper hopes to obtain a classifier with high accuracy through these operations, which can provide some really useful help for doctors' diagnoses in practice.

2. Data

2.1. Exploratory data analysis

Understanding the structure and properties of our data at this stage of the analysis is essential for us to be able to make wise judgments about the next steps of the research.

The data information is as follows:

Table 1. Features of dataset.

No	Attribute	Description
1	age	Age of the person
2	sex	Gender of the person
3	ChestPainType	Chest pain type (1-4)
4	RestingBP	Resting blood pressure
5	chol	cholesterol in mg/dl fetched via BMI sensor
6	FastingBS	Fasting blood sugar>120 (1=true,0=false)
7	RestingECG	Resting electrocardiographic results (0-2)
8	MaxHR	Maximum heart rate
9	ExerciseAngina	Exercise include angina(1=yes,0=no)
10	old peak	Previous peak
11	Slope	Slope
12	MajorVessels	Number of major vessels (0-3)
13	StressTest	That rate

According to the data type, data can be divided into continuous data and categorical data.

①Continuous data: [age, RestingBP, chol, FastingBS, RestingECG, MaxHR, oldpeak];

②Categorical data: [sex, ChestPainType, ExerciseAngina, Slope, MajorVessels, StressTest].

2.2. Categorical data

For categorized data, use a count plot to count the numbers of different values and the count plot according to the output.

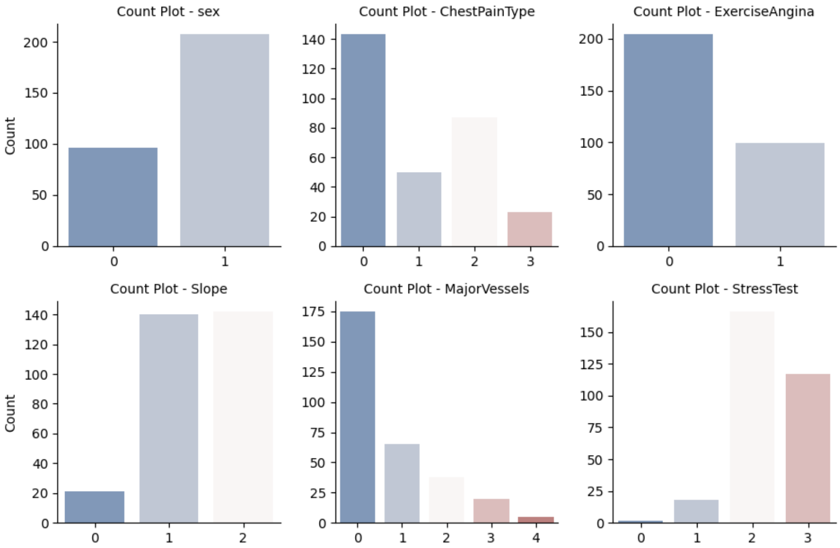


Figure 1. Count Plot of Categorical Data.

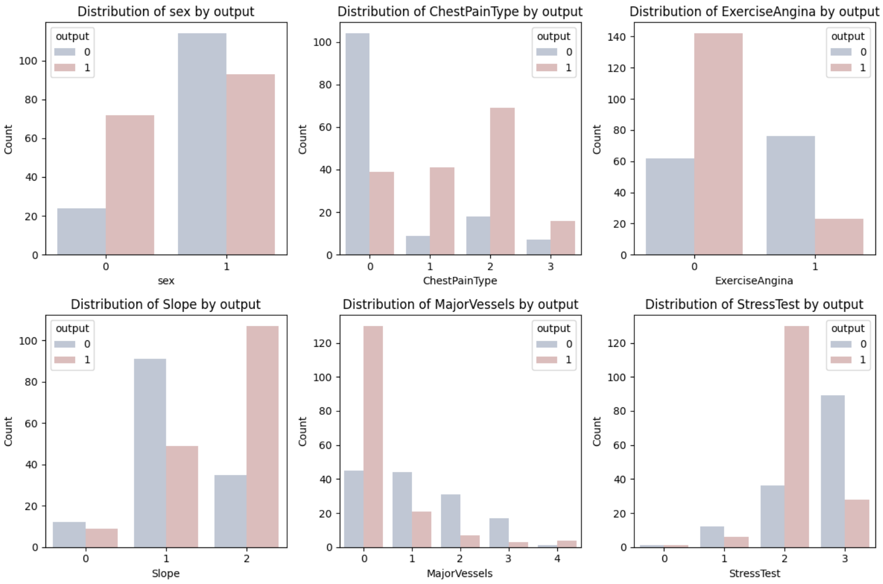


Figure 2. Distribution of Categorical Data by Output.

2.3. Continuous data

For continuous data, histogram statistics are performed for each feature.

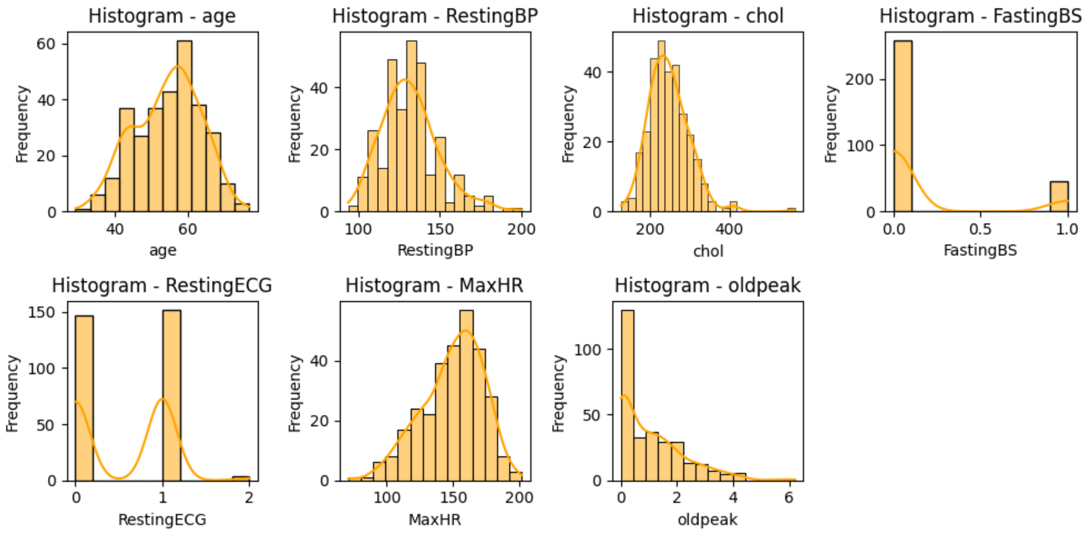


Figure 3. Histogram of Continuous Data.

2.4. *Data correlation*

Finally, to investigate possible connections between numerical variables and assist in identifying important dependencies and interactions in our dataset, we shall do a correlation analysis. This research will yield important insights into the possible interactions between these variables and how they affect the risk of heart disease as a whole.

A half correlation matrix of all features is drawn, that is, a heat map, which can convey the correlation between the two features. The association is the larger the value. In particular, the output line shows the degree of influence of each feature on the results. It is not difficult to find that the most important influencing features are the type of chest pain, maximum heart rate, and slope. Therefore, this can serve as a reminder to pay more attention to these major influencing factors in medical diagnosis.

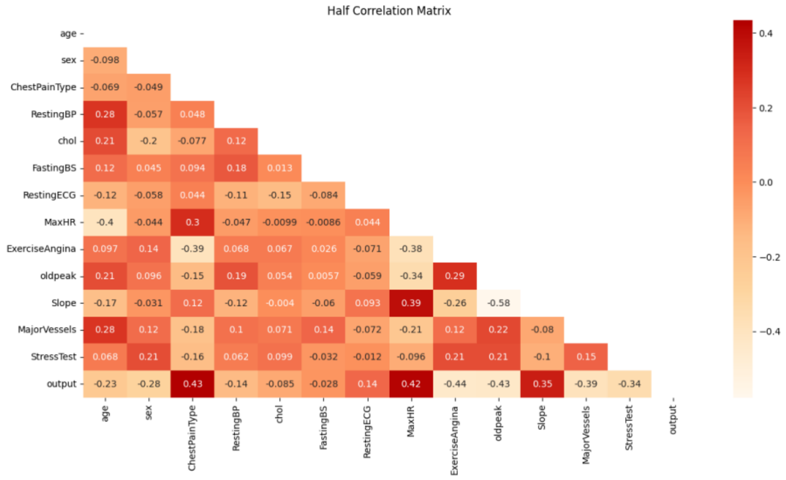


Figure 4. Half Correlation Matrix.

2.5. *Data preprocessing*

The technique before applying machine learning algorithms to data is called data preprocessing [5]. On the one hand, data preprocessing can improve data quality, and on the other hand, data can be better adapted to specific processing methods or tools.

The majority of real-world data is noisy, may have missing values, or be in unsuitable forms that prevent it from being directly utilized in machine learning models. To clean up the data and prepare it for machine learning models, a technique known as data preprocessing is required. This increases the accuracy as well as the efficacy of machine learning models.

For different datasets, the data preprocessing steps may be slightly different and often need to be determined according to the specific situation of the dataset and the subsequent use of processing tools or methods.

For the dataset of this paper, the first step to be carried out is to check the missing values. If there are missing values, it is easy to lead to insufficient extracted information and errors in the conclusion [6]. Because the dataset used in this paper is not complex and complete, and we have done some work in the previous data analysis, the second step is to distinguish between independent variables and dependent variables, that is, features and targets. This is so that the classification algorithm used later knows the learning purpose and finds the relationship between features and the target. The next thing to do is divide the dataset. There are usually three ways to divide datasets, namely the hold-out method, cross-validation (CV) method, and bootstrapping method. The stratified k-fold CV is chosen in this paper, which belongs to the type of CV, and can maintain the proportion of each category in the original data at each fold. Compared with the simplest hold-out method, CV can reduce contingency through multiple partitioning. Simultaneously, following successive segmentation and training, the model may come into contact with additional data, enhancing its capacity for generalization. In a k-fold CV, the choice of k value is very important. According to a large number of experimental data, the deviation is smaller when k=5 or k=10 [7], so this paper chooses k=10 for CV. The last step before the formal use of ML algorithms is feature scaling. The value of features will affect the weight obtained in the model, which is likely to lead to a decrease in prediction accuracy. The best solution is to scale all the features to a unified interval. In this paper, the MinMaxScaler is chosen, then the value of the feature is controlled between 0 and 1.

3. Classifier Model (Classification Algorithm)

In this paper, ten classification models are selected from many machine learning models, which are: Logistic Regression, Decision Trees, Random Forest, SVM (RBF), Naive Bayes, K-Nearest Neighbors, Gradient Boosting, XGBoost, Bagging, Multilayer Perceptron Classifier. The primary objective of the study is to learn the given data using the aforementioned classifiers, determine the correlation between features and targets, and forecast the outcomes.

In machine learning, an algorithm may be specifically divided into several sub-algorithms based on important influencing factors. Therefore, some of these algorithms need to be explained briefly:

The choice of kernel function is crucial for building a support vector machine with high performance as the algorithm's performance mostly depends on it. There are now three different types of kernel functions: radial basis function (RBF), polynomial kernel, and linear kernel. Until now, there is no general rule about which kernel should be used. However, according to the research, RBF core is the better choice in practical applications [8].

In the KNN algorithm, the k value is a very important factor. A small value of k indicates a high model complexity, decreased training error, weakened generalization ability, and ease of overfitting; a large value of k indicates a low model complexity, increased training error, somewhat improved generalization ability, and ease of underfitting. During the experiment, this paper draws the learning curve with the k value (1-40) as the X-axis and the model score as the Y-axis and calculates the error rate of each k value. These are visualized as follows:

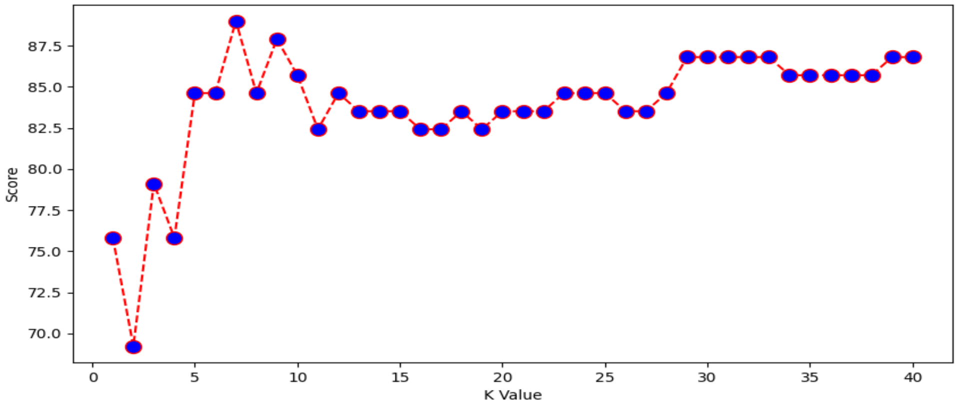


Figure 5. The Learning Curve by k value.

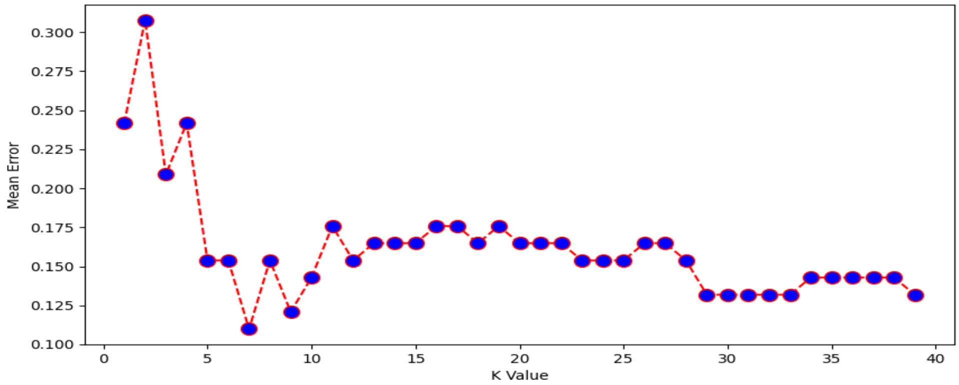


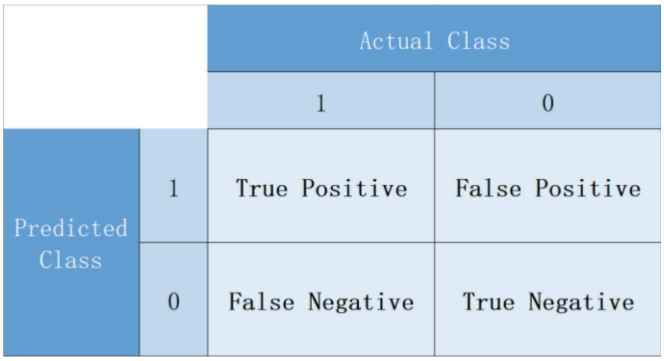
Figure 6. The Error Rate by k value.

It is not difficult to find that when k=7, the score is the highest and the error rate is the smallest, so k=7 is chosen for subsequent learning.

4. Comparison

When evaluating model performance, many metrics can be used to evaluate the performance of a model. The following metrics are used in this paper: accuracy, precision, recall F1 score, and AUC score.

In machine learning, the confusion matrix is an important visualization tool, so the following discussed evaluation metrics are mostly based on the confusion matrix. For a data set, we can separate positive and negative according to the actual situation, and combine the predicted positive and negative of the model results to generate a confusion matrix, as shown in Figure 7.



		Actual Class	
		1	0
Predicted Class	1	True Positive	False Positive
	0	False Negative	True Negative

Figure 7. Confusion Matrix.

4.1. Accuracy

The ratio of the number of samples with a valid prediction to the total number of samples participating in the prediction is used to calculate the accuracy of the model's prediction of the samples in the dataset.:

$$Accuracy = \frac{\text{the number of samples with correct prediction}}{\text{the total number of samples participating in the prediction}}$$

It also can be described as:

$$Accuracy = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Negative} + \text{False Negative}}$$

Accuracy is a very commonly used evaluation index, but high accuracy does not always mean that the classification algorithm is good. Especially when the samples of each category are unevenly distributed, even if the accuracy rate is 99%, it is meaningless. Therefore, it is not comprehensive to evaluate a model only by its accuracy.

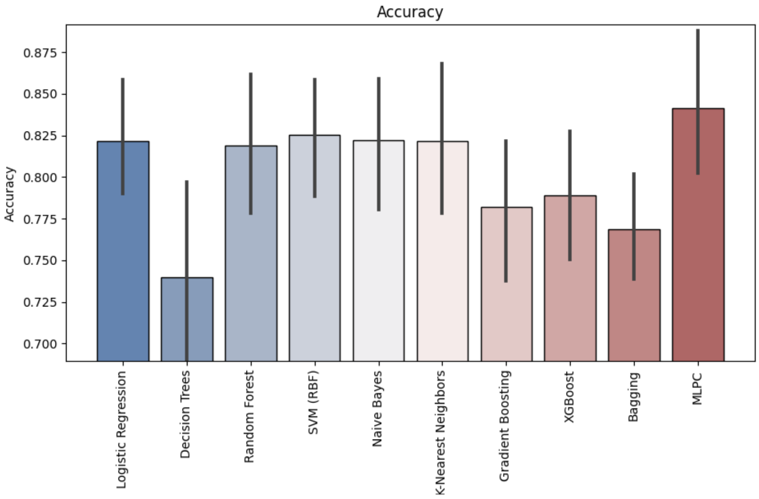


Figure 8. Bar plot of Accuracy.

Among the methods selected, all the accuracy rates are above 70%, but the overall difference is not large, and the highest is Multilayer Perceptron Classifier (MLPC).

4.2. Precision

Precision refers to the percentage of samples that are positive cases out of all samples that are predicted to be positive cases, that is:

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Although they have similar attributes, accuracy, and precision are very separate ideas. Accuracy is the total prediction's accuracy taking into account both positive and negative samples, whereas precision is the prediction's accuracy as it relates to the outcomes of the positive sample. According to precision, we can know the ability of the model to predict correctly.

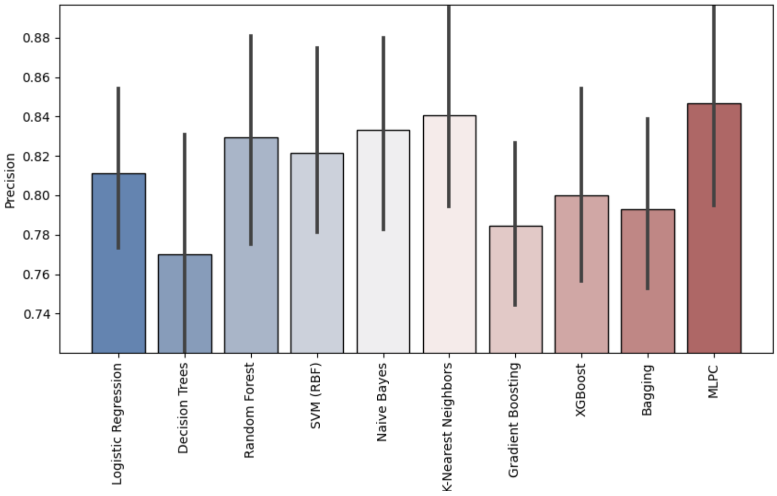


Figure 9. Bar plot of Precision.

Looking at the comparison chart, it is clear that precision and accuracy differ greatly, but the best is still the Multilayer Perceptron Classifier (MLPC).

4.3. Recall

The ratio of samples correctly classified as positive by the model to the total number of positive samples is known as recall. So recall is used to determine whether all the actual positive examples in the sample have been found. It is also an indicator of the sensitivity of a model, known as the 'True Positive Rate'(TPR), namely:

$$Recall = TPR = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

From this, we can know the ability of the model to predict the overall positives.

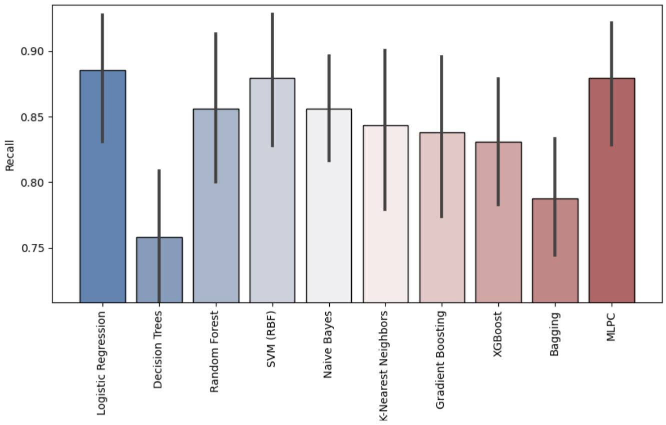


Figure 10. Bar plot of Recall.

For recall rates, we can see that logistic regression performs better, closely followed by MLPC and SVM.

4.4. F1-score

F1-Score is a measure combining both precision and recall. Calculating precision and recall is only to calculate the characteristics of a certain classification. In general, precision and recall are contradictory measures. To better characterize the performance measure of the learner in precision rate and recall rate, F1 scores are introduced. It is therefore frequently referred to as the harmonic mean of the two. Simply put, the harmonic mean is an alternative method of calculating an "average" of numbers. It is generally accepted that this method works better for ratios (such as recall and precision) than the standard arithmetic mean. The formula used for the F1-score in this case is:

$$F1\ score = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1-Score ranges from 0 to 1, with 1 representing the model's best output and 0 representing the model's worst output.

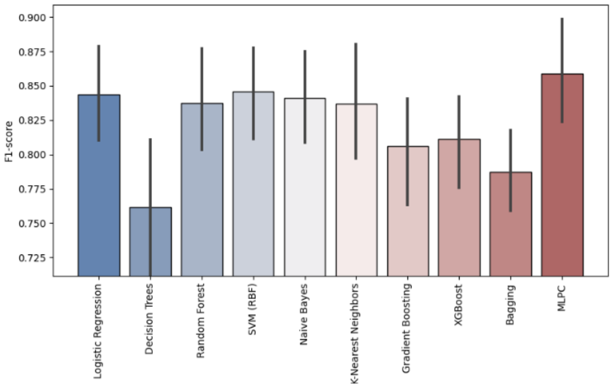


Figure 11. Bar chart of F1-Score.

For MLPC, which does well in precision and recall, it's no surprise that it also has the best F1 score. More than half of the models have an F1 score of more than 0.8, indicating good performance.

4.5. AUC score

The AUC is the area under curve (ROC curve), which has a value between 0.5 and 1. The AUC value can be used to measure the performance of the classifier, by estimating the predictive accuracy of classifying models derived from presence and absence data [9]. An AUC of 1 would indicate a flawless classifier, whereas an AUC of 0.5 would indicate a random estimate. In general, if the AUC is greater than 0.8, we can consider the performance of the classifier to be good.

Talking about the Receiver Operator Characteristic(ROC) Curve, its horizontal coordinate is ‘False Positive Rate’ (FPR) which is regarded as specificity and its vertical coordinate is TPR (sensitivity) mentioned in the recall part. The formula for FPR is as follows:

$$FPR = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}}$$

The ROC curve is obtained by taking the threshold value between 0 and 1 to different TPR and FPR values. For the result, we want to get the largest possible TPR and the smallest possible FPR, that is, the largest possible AUC value. ROC curve and AUC score are very important in the study of medical diagnosis.

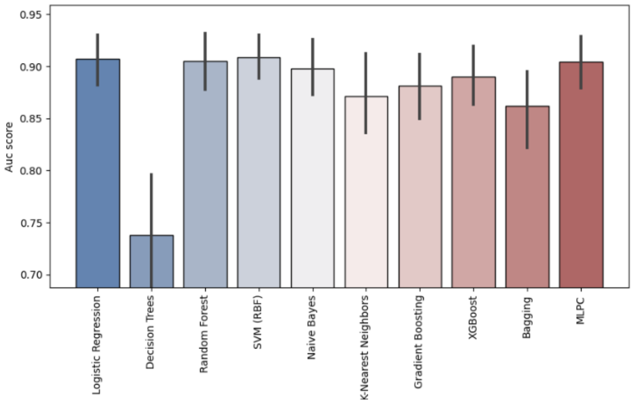


Figure 12. Bar plot of Auc-score.

In addition to the decision trees algorithm, the rest of the algorithms have relatively good performance.

5. Ensemble Learning

5.1. Selection

After 10-fold cross-validation of each model, the scores in the following table are obtained.

Table 2. Scores for each model.

	Auc score	Accuracy	Precision	Recall	F1-score
Model					
MLPC	0.904408	0.841613	0.846776	0.879412	0.859009
SVM(RBF)	0.908813	0.825269	0.821320	0.879779	0.845583
Naive Bayes	0.898154	0.822043	0.833144	0.856250	0.841278
KNN	0.871463	0.821828	0.840527	0.843750	0.837183
Logistic Regression	0.907078	0.821720	0.811352	0.885294	0.843846
Random Forest	0.904963	0.818710	0.829542	0.855882	0.837654
XGBoost	0.889928	0.789032	0.800093	0.830882	0.811341
Gradient Boosting	0.881436	0.782151	0.784463	0.838235	0.806108
Bagging	0.861690	0.768602	0.793080	0.787868	0.787516
Decision Trees	0.737744	0.738570	0.770044	0.758456	0.761588

The paper chose the three that performed the best, making them particularly promising candidates for further improvement through voting to create new models. These models are the Multilayer Perceptron Classifier (MLPC), the Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel, and the Naive Bayes. Let’s explore each one:

MLPC: This is a classifier based on a feedforward neural network. The neural network model uses a positive transmission mechanism and a layer or layers of hidden nodes are contained between the input layer and the output layer. The intermediate node uses the Sigmoid (logistic) function, and the output node uses the Softmax function. The number of nodes in the output layer indicates how many classes the classifier has. The backpropagation (BP) algorithm is used in the MLPC learning process.

SVM (RBF): Finding the ideal hyperplane to divide samples of several classes is the aim of SVM. It is widely used in classification and regression problems. In contrast to logistic regression and neural networks, it offers a more effective and transparent method for learning intricate nonlinear equations [10].

Naive Bayes: The basic idea of the Naive Bayes algorithm is based on the probabilistic relationship between the features and target in the training data, and the classification prediction is made by calculating the posterior probability, which has strong interpretability [11]. For the multi-class classification problem, it can maintain good performance even when the number of classes is large.

5.2. Voting Classifier

Voting is a combination strategy for classification problems in ensemble learning. This ensemble learning model, which lowers variance by integrating different models, increases the model's resilience (i.e., how well the algorithm tolerates changes in data). It operates on the concept of the minority obeying the majority. The voting method's prediction effect should ideally outperform all of the base models. There are two types of voting procedures: hard voting and soft voting. The hard voting chosen in this paper will output the class label directly, and the predicted result will be the class with the most voting results.

For this integrated model, the training used in this paper is the same as the previous method for single classifiers, which is still stratified with 10-fold cross-validation.

5.3. Results

The training results of the hard voting integrated model in this paper are as follows:

Table 3. Scores of Voting Classifier.

	Accuracy	Precision	Recall	F1-score
Model				
Voting Classifier	0.835161	0.834919	0.879779	0.853554

For this ensemble learning model, the maximum and minimum values of the four main evaluation indicators in the 10-fold cross-validation are as follows:

Table 4. Minimum and Maximum of Metrics.

	Accuracy	Precision	Recall	F1-score
Min	0.74	0.75	0.71	0.75
Max	0.94	1.0	0.94	0.94

In the final stage, to maximize the amount of information available to our model, the entire data set is utilized for final training. The final score of the integrated learning model is 0.874587454587878746.

6. Conclusion

The study gained an understanding of the data distribution and looked at the fundamental statistical characteristics of the data throughout the data analysis step. This involved displaying the distribution of different characteristics and the target variable, as well as comprehending the many sorts of variables and looking for missing values. This first stage is essential to every data science project since it enables us to find patterns, correlations, trends, and anomalies in the data. We performed a correlation analysis to understand the relationships between different numerical features. We were able to determine which characteristics were positively or negatively linked with the goal variable, "output," and with each other by using a heatmap display of the correlation matrix.

Ten distinct machine-learning models were used to predict heart disease during the machine-learning prediction phase, and their effectiveness was assessed using metrics. Multilayer Perceptron Classifier (MLPC), the Support Vector Machine (SVM) (RBF), and the Naive Bayes emerged as the models with better performance. The ensemble learning model combined with the hard voting method further improves the performance of prediction.

In this experiment, we can see that the performance of the decision tree algorithm is very average for all performance evaluation indicators. A decision tree is a simple and easy-to-understand algorithm in machine learning, but its accuracy is not as good as other algorithms, and the instability of the tree leads to its disadvantage is easy to overfit. So for decision tree algorithms, the biggest challenge is to determine when to stop growing the tree [12]. The most common stopping strategies are setting the depth of the decision tree or requiring a minimum number of nodes [13]. However these methods are computationally expensive, and using them overrides the simplicity of the decision tree itself. A novel hyperparameter-free approach to decision tree construction to avoid overfitting was proposed by Leiva et al.[12]. In the future, we will use it in this study to maximize the performance of the decision tree, and then improve the voting classifier of ensemble learning in this paper.

It would be advantageous to carry out more studies in real-world applications, such as adjusting hyperparameters and using strategies to validate the outcomes. We hope that the study in this paper will inspire more researchers to conduct further research in this direction, and ultimately, related medical fields will benefit from this direction.

References

[1] Palaniappan, Sellappan, and Rafiah Awang. 2008. “Intelligent Heart Disease Prediction System Using Data Mining Techniques.” In 2008 IEEE/ACS International Conference on Computer Systems and Applications. doi:10.1109/aiccsa.2008.4493524.

[2] Rajkumar, Asha, and G.Sophia Reena. 2010. “Diagnosis of Heaer Disease Using Datamining Algorithm.” Global Journal of Computer Science and Technology, Global Journal of Computer Science and Technology, September.

[3] Soni, Jyoti, Ujma Ansari, Dipesh Sharma, and Sunita Soni. 2011. “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction.” International Journal of Computer Applications, April, 43–48. doi:10.5120/2237-2860.

[4] S. Dangare, Chaitrali, and Sulbha S. Apte. 2012. “Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques.” International Journal of Computer Applications, July, 44–48. doi:10.5120/7228-0076.

[5] García, Salvador, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez, and Francisco Herrera. 2016. “Big Data Preprocessing: Methods and Prospects.” Big Data Analytics, December. doi:10.1186/s41044-016-0014-0.

[6] Wang, Hai, and Shouhong Wang. 2010. “Mining Incomplete Survey Data through Classification.” Knowledge and Information Systems, August, 221–33. doi:10.1007/s10115-009-0245-8.

[7] Rodriguez, J.D., A. Perez, and J.A. Lozano. 2010. “Sensitivity Analysis of K-Fold Cross Validation in Prediction Error Estimation.” IEEE Transactions on Pattern Analysis and Machine Intelligence, March, 569–75. doi:10.1109/tpami.2009.187.

[8] A. Patel and D. S. Chouhan, "SVM kernel functions for classification," 2013 International Conference on Advances in Technology and Engineering (ICATE), Mumbai, India, 2013, pp. 1-9, doi: 10.1109/ICAdTE.2013.6524743.

[9] Lobo, Jorge M., Alberto Jiménez-Valverde, and Raimundo Real. 2008. “AUC: A Misleading Measure of the Performance of Predictive Distribution Models.” Global Ecology and Biogeography, March, 145–51. doi:10.1111/j.1466-8238.2007.00358.x.

[10] Mountrakis, Giorgos, Jungho Im, and Caesar Ogole. 2011. “Support Vector Machines in Remote Sensing: A Review.” ISPRS Journal of Photogrammetry and Remote Sensing, May, 247–59. doi:10.1016/j.isprsjprs.2010.11.001.

[11] Friedman, Nir, Dan Geiger, and Moises Goldszmidt. 1997. “Bayesian Network Classifiers.” Machine Learning, January, 131–63. doi:10.1023/a:1007465528199.

[12] Leiva, RafaelGarcia, AntonioFernández Anta, Vincenzo Mancuso, and Paolo Casari. 2019. “A Novel Hyperparameter-Free Approach to Decision Tree Construction That Avoids Overfitting by Design.” arXiv: Learning,arXiv: Learning, June.

[13] Rokach, Lior, and Oded Maimon. 2007. “Data Mining with Decision Trees: Theory and Applications.” Series in Machine Perception and Artificial Intelligence, Series in Machine Perception and Artificial Intelligence, December.