The Prediction of Heart Failure based on Four Machine Learning Algorithms

Ming Zeng *

Eller College of Management, University of Arizona, Arizona, United States

* Corresponding author email: zengming@arizona.edu

Abstract. The main purpose of this study is to observe the correlations between clinical features (input) with heart disease (output) in the practice of machine learning and modeling. Primarily, this research conducts a data exploration of the attributes and the output to observe the relationships between the attributes and heart disease. The experiment method covers the Decision Tree, K-Nearest Neighbor, Support Vector Machine and Extreme Gradient Boosting algorithms after the data exploration. This research concludes that exercise-induced angina and ST depression are two factors that are highly related to heart disease. Older males should take more care of these features because they are more likely to have heart disease. Moreover, based on the output of four algorithms, the SVM is the best method for predicting the probability of the occurrence of heart disease since SVM output the highest values on the accuracy, recall rate, and f1-score.

Keywords: Heart Disease; Decision Tree (DT); K-Nearest Neighbor (KNN); Support Vector Machine (SVM); Extreme Gradient Boosting (XGB).

1. Introduction

Heart disease is defined as a general phrase that includes diverse kinds of heart problems, and all kinds of heart disease can be classified under the cardiovascular disease category [1]. In the United States, heart disease is the most frequent cause of death for both men and women, as well as people of most ethnic groups. According to the data from the Centers for Disease Control and Prevention, around 697,000 people died due to cardiovascular disease in 2020 in the US [2]. Therefore, it has great significance to conduct research on heart disease in clinical medicine.

Moreover, there is abundant research related to the clinical heart disease field for many decades. Mohan et al conducted a total of 10 machine learning techniques such as Logistic Regression, Linear Model, Naïve Bayes, and HRFLM for a similar dataset. Their research shows that the HRFLM approach, which is used to associate Random Forest with Linear Method, has the highest accuracy [3]. Furthermore, Fitriyani et al have implemented 7 machine learning techniques for the same dataset, and they concluded that the prediction model HDPM, which was developed by integrating DBSCAN, SMOTE-ENN, and XGB-based MLA, has the most efficiency for predicting heart disease [4]. Garg et al have also performed K-Nearest Neighbor, Random Forest methods based on the same dataset, the K-Nearest Neighbor reaches a higher accuracy with 86.885% [5]. To achieve better performance, more machine learning methods (Decision Tree, K-Nearest Neighbor, Support Vector Machine, Extreme Gradient Boosting) with higher accuracies are conducted in this paper, which is used to investigate the correlation between the input clinical features with the output attribute for the sake of providing suggestions to reduce the likelihood of heart disease.

2. Material and Methods

2.1 Dataset

This research is based on the heart disease prediction dataset provided by an author in Kaggle, the dataset consists of 5 datasets that are not combined before but are available for public use independently [6]. The 5 datasets are combined into one dataset that includes 11 input clinical features and one output class with a total of 918 entries. The descriptions and the types of all attributes are presented in Tab 1, the research in this paper is based on these 12 attributes. Specifically, the attributes

apart from Age, RestingBP, Cholesterol, FastingBS, Oldpeak, and HeartDisease are all nominal types, which means non-numeric.

Table 1. Dataset information

No.	Attribute	Description	Type
1	Age	Patient's age [Numeric value]	Numeric
2	Sex	Patient's gender [M: Male; F: Female]	Binary
3	ChestPainType	Type of chest pain [ATA: Atypical Angina; NAP: Non-Anginal Pain; ASY: Asymptomatic; TA: Typical Angina]	Nominal
4	RestingBP	Resting blood pressure [mm Hg]	Numeric
...	....	.....	....
12	HeartDisease	Output attribute [1: heart disease; 0: Normal]	Binary

2.2 Decision Tree

In this paper, the research chooses the Decision Tree algorithm as one of the prediction methods. The DT algorithm is one kind of supervised learning in the machine learning field. A decision tree is a hierarchical instance of knowledge relationships that include multiple nodes and links, and the algorithm is a tree-based technique that started with the root, any path after the root is illustrated by a data separating order down to a Boolean outcome is concluded at the leaf node [7]. The structure of the DT is shown in Fig. 1.

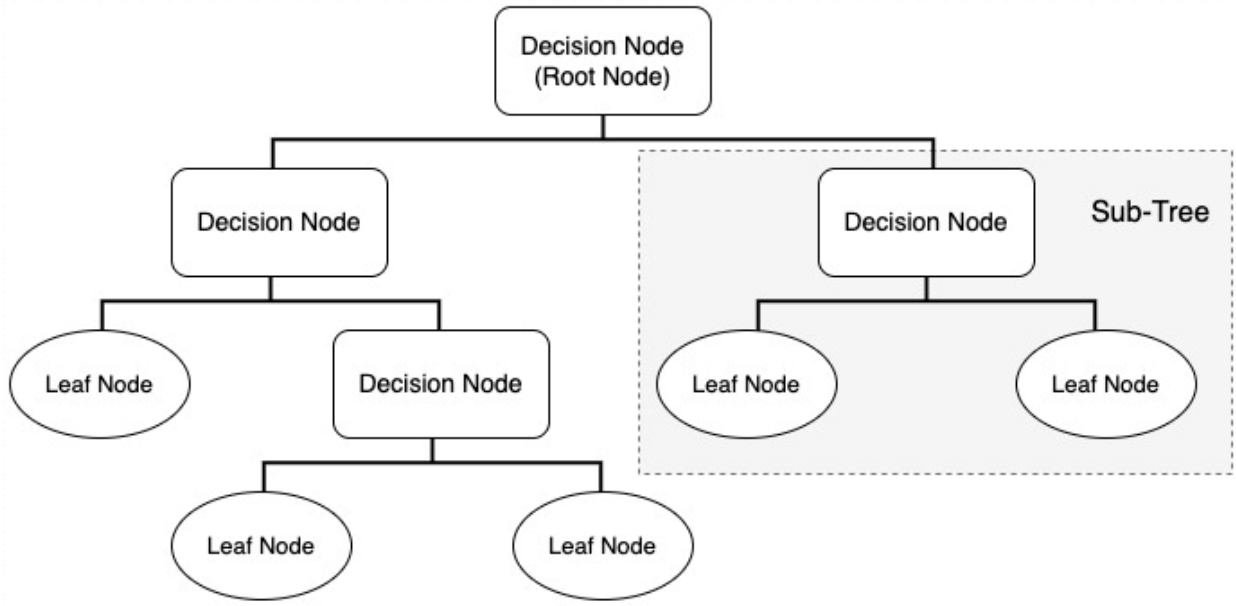

Fig 1. Decision Tree algorithm structure

2.3 K-Nearest Neighbor (KNN)

The KNN technique is a non-parametric, supervised learning classification method that can help resolve regression and classification problems. There are two primary research directions in the K-Nearest Neighbor algorithm, one is to create an appropriate K value, and the other one is the distance metric for determining KNN [8]. The algorithm clusters each individual data point to develop classifications or forecasts of a dataset based on the proximity of the data. Equation 1 shows the Euclidean distance equation, which is the most commonly used equation in the KNN for determining the distance metrics. In this research, the KNN also be implemented for finding the optimal model.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(yi - xi)^2} \quad (1)$$

2.4 Support Vector Machine (SVM)

The subsequent algorithm that is used for the research in this paper is the SVM, which is kind of the most famous supervised learning algorithms technique that can be used for handling both classification and regression problems. The SVM technique was originally invented by Vladimir N. Vapnik for linear models in 1963, and later was broadened to non-linear training data in 1995 [9]. The SVM algorithm produces predictions according to the function:

$$y(x;w) = \sum_{i=1}^{N} w_i K(x,x_i) + w_0 \quad (2)$$

Where $K(x,x_i)$ is a kernel function [10]. The SVM technique solves both linear and non-linear by converting the initial training data into a greater dimension through the non-linear mapping method. The algorithm then searches for a linear optimal disconnected hyperplane within the new dimension, which allows data from two clusters to be divided by the hyperplane with a fitting non-linear mapping to an appropriately high dimension [11].

2.5 Extreme Gradient Boosting (XGB)

The principle of the XGB machine learning method is to develop a powerful learning model by combining a series of weak models in order to enhance the performance of machine learning and outputs better prediction outcomes than single models [12]. Specifically, the XGB algorithm establishes a cluster of classification with regression trees. Then, within each step of the algorithm, the point that is larger than the threshold with the maximum gain is determined to be a splitting point, which is then separated for obtaining a new tree thereafter. Further, for forecasting the score of a sample, the algorithm estimates all individual tree features of the sample to obtain a score, then sums the scores to get the final predicted value of the sample [13]. The XGB method is used in this research aiming to make comparisons with other models and determine the potential most fitting predictive models for the dataset.

3. Material and Methods

The source of the dataset and the machine learning methods have been discussed in the former section. The next section demonstrates the whole experiment process and the result of the experiment.

3.1 Exploring Data

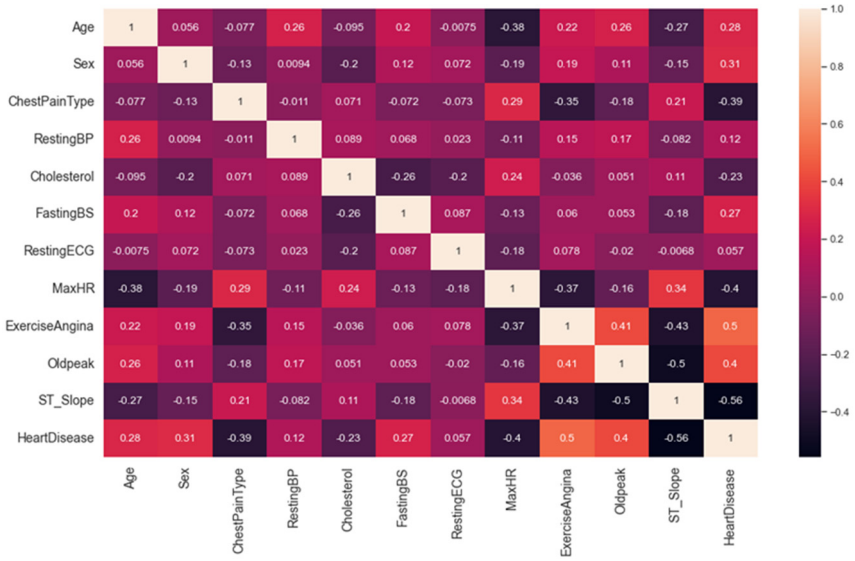

Fig 2. Heatmap for 12 attributes

Since there are 5 features that are non-numeric, these 5 features need to be transformed into numeric values so that the research can be advanced. Subsequently, detect the outliers within the dataset according to the calculation of the quartile range. Only one outlier is detected and dropped from the dataset, and the dataset now contains 917 observations. After dropping the outlier, the data is ready to be plotted in the heatmap, and the correlations among all attributes have been indicated in Fig. 2. The angina and the ST depression induced by exercise are highly related to the motivation for disease. Additionally, according to the statistics from American Heart Association (AHA), 52.9% were men and 47.1% were women among the 5.1 million cases of heart disease. Furthermore, the estimation of the mean year of the first heart disease was 65.6 years for men and 72.0 years for women in the United States. In brief, the older male generally has a higher risk of having heart disease compared to older female, moreover, the exercise-induced angina and ST depression can be treated as a forewarning of heart disease.

3.2 Experiment Design

For attaining higher accuracy, the data needs to be standardized first. After standardization, the data is separated into four variables ('x_train', 'x_test', 'y_train', 'y_test') for predicting and testing purposes while setting 70% of the data for prediction and 30% for testing. Once all the preprocessing steps have been completed, the four variables are ready to be applied to the selected machine learning algorithms for the experiment. The first method is the Decision Tree algorithm. Setting the number of features to consider when the algorithm is looking for the optimal split (max_features) as 8 while other parameters are set as default. Next, to join the KNN algorithm, simply import the library and import the four variables directly to the function, and set the tuning parameter (K) as 5 since it produces the highest accuracy. Moreover, for the third algorithm SVM, set all the parameters in default and apply the trained and testing variables to the SVM functions. Lastly, import the functions in the XGB library and apply the four variables to the functions without changing the functions' settings. Then, for each machine learning algorithm used above, print out the classification result as well as plotting the prediction outcome into a confusion matrix.

3.3 Evaluation Criteria

In this research, the evaluation criteria of the model are based on the accuracy, recall rate and f1-score in each model's classification report. Primarily, each model produces its own confusion matrix, and the three indexes can be calculated according to the values in the confusion matrix. The confusion matrix's sample diagram is shown in Tab 2.

Table 2. Confusion matrix

		Predicted Label	
		0	1
Actual Label	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

The equations for calculating three primary indexes through the confusion matrix have been shown below.

$$Accuracy = \frac{TN+TP}{TN+FP+TP+FN} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

$$F1-score = \frac{Precision \times Recall \times 2}{Precision+Recall} \tag{5}$$

The precision rate can be expressed as $Precision = \frac{TP}{TP+FP}$ for calculating the f1-score.

3.4 Experiment Result

The DT, KNN, SVM, and XGB algorithms are implemented in this research for forecasting the possibility of the occurrence of heart disease based on 11 clinical features. The four models are evaluated according to the accuracy, recall rate and f1-score. The four algorithms' confusion matrixes are shown in Fig 3 in sequence.

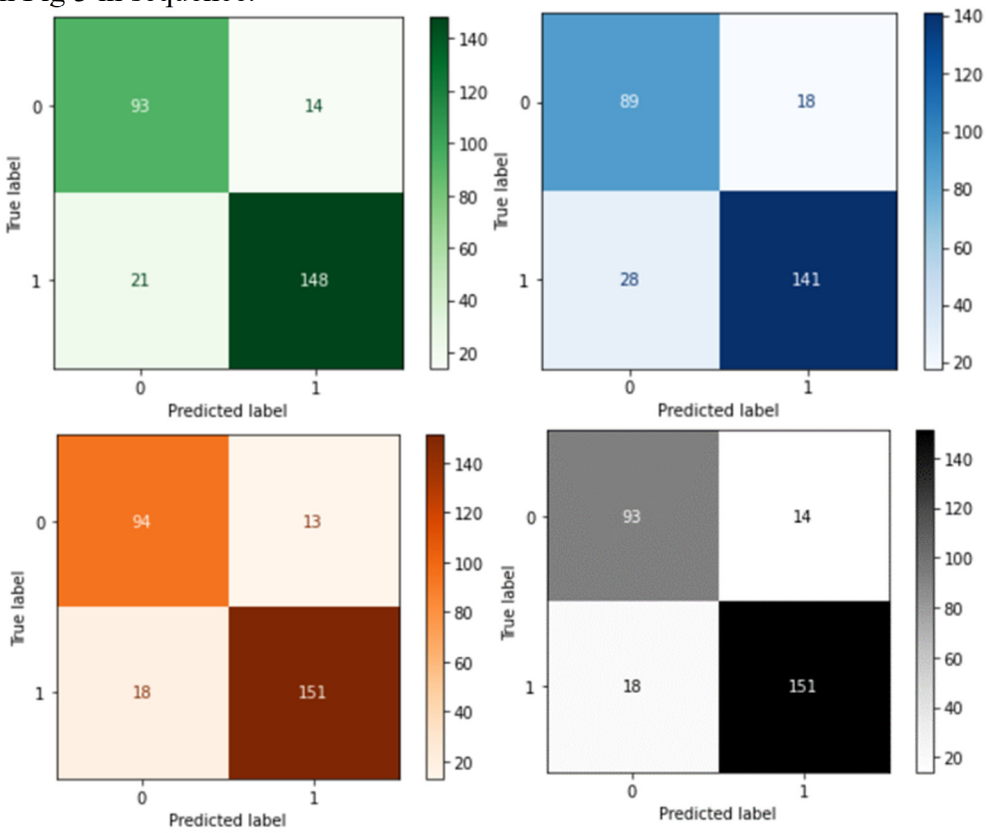

Fig 3. Confusion Matrix

The accuracy, recall rate and f1-score are calculated through the equations listed above. The output of each model performance has been calculated (round to 3 decimal places) and shown in Tab 3.

Table 3. Performance result of each model

Models	Accuracy	Recall	F1-score
DT	0.833	0.834	0.860
KNN	0.873	0.876	0.894
SVM	0.888	0.893	0.907
XGB	0.884	0.893	0.904

According to the table shown, the SVM algorithm has the best performance among other models since the SVM algorithm produces the highest values on all three standards. Therefore, the SVM algorithm is the most appropriate method for this dataset for predicting the probability of heart disease.

4. Conclusion

Overall, this study observes the heart disease dataset by discovering the correlation among the clinical features with the output class as well as applying DT, KNN, SVM, XGB methods to forecast the probability of heart disease. The four methods are evaluated according to their accuracy, recall and f1-score to determine the best model for this dataset. The dataset provided by a Kaggle author is a combination of 5 independent datasets, which includes a total of 918 patient records. The heatmap has proven that the angina (ExerciseAngina) and ST depression (Oldpeak) induced by exercise should be paid more attention to since these two features are highly related to the occurrence of heart disease. Furthermore, for the performance of four models, as the SVM method outputs the highest accuracy (88.8%), recall rate (89.3%) and f1-score (90.7%), the SVM is the best machine learning algorithm for predicting the probability of heart disease. This dataset contains a total of 918 instances with 11 clinical features, but more features and patient examples are required for further analysis. In the future, more machine learning methods will be used for analyzing the factors that would result in heart disease. This research provides conducive suggestions for the clinical medicine field for knowing the features of heart disease and shows examples for applying machine learning skills for analysis.

References

[1] “Heart Disease.” MedlinePlus, U.S. National Library of Medicine, <https:// medlineplus. gov/ heartdiseases. html> (2022, September 2).

[2] “Heart Disease Facts.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, <https://www.cdc.gov/heartdisease/facts.htm> (2022, July 15).

[3] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." IEEE access 7 (2019): 81542-81554.

[4] Fitriyani, Norma Latif, et al. "HDPM: an effective heart disease prediction model for a clinical decision support system." IEEE Access 8 (2020): 133034-133050.

[5] Garg, Apurv, Bhartendu Sharma, and Rijwan Khan. "Heart disease prediction using machine learning techniques." IOP Conference Series: Materials Science and Engineering. Vol. 1022. No. 1. IOP Publishing, 2021.

[6] fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved (2022, August 23) from <https:// www.kaggle.com/fedesoriano/heart-failure-prediction> (2021, September 22).

[7] Members, Writing Group, et al. "Executive Summary: Heart Disease and Stroke Statistics--2016 Update: A Report from the American Heart Association." Circulation 133.4 (2016): 447-454.

[8] Virani, Salim S., et al. "heart disease and stroke statistics--2021 update: a report from the American Heart Association." Circulation 143.8 (2021): e254-e743.

[9] Otchere, Daniel Asante, et al. "Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models." Journal of Petroleum Science and Engineering 200 (2021): 108182.

[10] Tipping, Michael E. "Sparse Bayesian Learning and the Relevance Vector Machine." Journal of Machine Learning Research 1.3 (2001): 211-44. Web.

[11] Vijayarani, S., and S. Dhayanand. "Liver disease prediction using SVM and Naïve Bayes algorithms." International Journal of Science, Engineering and Technology Research (IJSETR) 4.4 (2015): 816-8201.

[12] Li, Hua, et al. "XGBoost model and its application to personal credit evaluation." IEEE Intelligent Systems 35.3 (2020): 52-61.

[13] HUANG, Jingwei, et al. "An XGB-based runtime prediction algorithm for cloud workflow tasks."