Predicting Heart Failure with ML Algorithms

Sai Chandan Reddy Koncha
Department of Computer Science
Virginia Tech
Falls Church, VA
saichandanreddyk@vt.edu

Sakshi Mhatre
Department of Computer Science
Virginia Tech
Falls Church, VA
sakshimm@vt.edu

Aniruddha Hore
Department of Computer Science
Virginia Tech
Falls Church, VA
aniruddhah@vt.edu

Abstract

Heart disease is one of the causes of high mortality in the world today and the diagnosis of heart disease mostly depends on the analysis of clinical and pathological data. Analyzing the data is complex and it should be meticulous, there is a significant amount of interest among professionals and researchers regarding the efficiency and accuracy in predicting heart disease. In this project, we develop a heart disease prediction system that can assist medical professionals in predicting heart disease based on the clinical data of patients. We have seen various machine learning techniques used for predicting heart disease. In this project, we evaluate the performance of multiple machine learning models classifying heart disease based on the clinical features present in the dataset.

1 Motivation

Heart failure is one of the leading causes of death worldwide. Predicting heart failure and its fatal symptoms is crucial for timely interventions and prevention strategies. Machine learning algorithms have shown promise in predicting heart failure by analyzing and predicting the outcomes of diseases. However, the existing models have a low accuracy rate, ranging from 52% to 67.7%. Therefore, there is a need to improve the accuracy rate of predicting heart failure using machine learning algorithms.

2 Methodology

We used the Heart Disease dataset available on Kaggle, which contains over 1000 observations and 14 attributes. The dataset includes attributes such as age, sex, chest pain type, blood pressure, serum cholesterol levels, and other health conditions that may or may not affect the heart. Our first step was data cleaning and preprocessing, where we removed unwanted features and filled in missing values using imputation techniques. We then applied feature selection methods to find informative attributes of the dataset that have a higher impact rate on the result. We used univariate feature selection, which selects features based on univariate statistical tests, and recursive feature elimination, which recursively removes the least important features until the optimal subset of features is achieved.

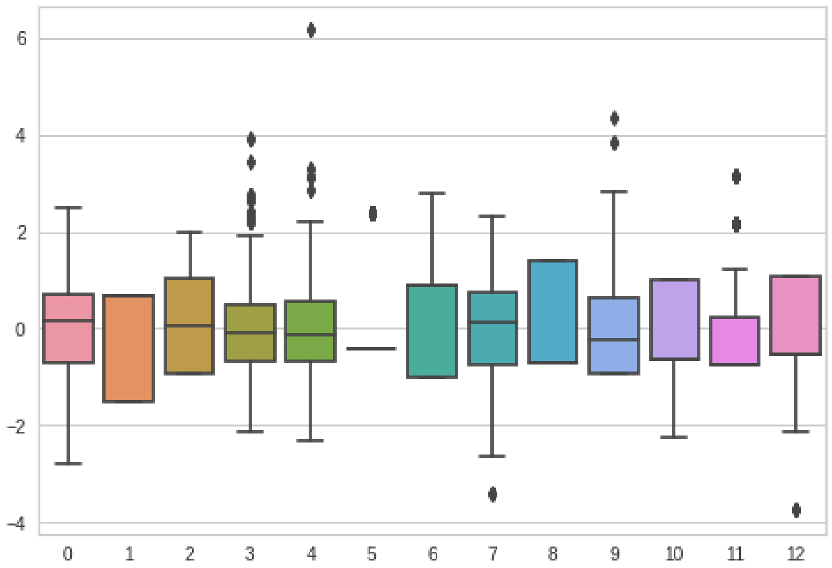

Figure 1: Box plot

2.1 Data Preprocessing

We checked for null and missing values once the dataset was loaded. This resulted in showing us that there is none of the sort. Now that we have determined that there are no null/missing values, we can proceed with the dataset as it is for further preprocessing. Then checking for correlation between the features, we can see that it is a strong correlation, proving the strength of the linear relationship between two variables and amongst each other. To further standardize the dataset, we converted integer values to float for the models to comprehend easily. Finally, the dataset was standardized to scale it to appropriate sizing and maintain consistency. By plotting a boxplot of the dataset, we can see the outliers present in the dataset 1. Although it is usually advisable to remove these outliers, we will not be because these are 'true outliers.' They are actual responses from people, and removing them would distort the study's and dataset's authenticity.

2.2 Training methods

2.2.1 Naive Bayes

Since naive bayes is generally used for models with binary and multiclass features performs well in cases of categorical input variables compared to numerical variables. It is useful for making predictions and forecasting data based on historical results. Due to its nature of being able to handle both continuous and discrete data, it is highly scalable with the number of predictors and data points. Also it is fast and can be used to make real-time predictions. And as for outliers, it is not sensitive to irrelevant features. These features make Naive Bayes a suitable model to use and implement when it comes to data with values and features regarding people and their health.

2.2.2 Random Forest

It builds trees on dofferent samples and teales their majority vote for classification and average on case regression. One of the most prominent features of Random Forest is that it can handle dataset containing continuous variables as in the case for regression, and categorical values as classification as well. Mostly it performs better for classification and regression tasks. Due to its diversity of features while making a tree, not all features/attributes are consnidered and hence each tree is different. It is also immune to the curse of dimensionality and is extremely stable as a result is based on majority voting/ averaging.

2.2.3 Logistic Regression

Logistic regression is used to obtain odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to multiple linear regression, with the exception that the response variable is binomial. The result is the impact of each variable on the odds ratio of the observed event of interest. Being one of the most efficient algorithms, it supports when the different outcomes or distinctions are represented by linearly separable data. Which is the type of data we are working with. And since the outputs have a decent probabilistic outcome, they can be regularized to avoid overfitting.

2.2.4 Decision Tree

Decision trees are excellent tools for helping you to choose between several courses of action. They provide a highly effective structure within which you can lay out options and investigate the possible outcomes of choosing those options. It can be used for both classification and regression problems: Decision trees can be used to predict both continuous and discrete values i.e. they work well in both regression and classification tasks.

2.2.5 Support Vector Machine

SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane. Since it can handle both classification and regression on linear and non-linear data, it is ideally suited for data in the medical field.

3 Implementation

We have implemented Gaussian Naive Bayes algorithm present in sklearn package and got an accuracy of the 81.46% and precision of 72.44%

accuracy	81.46
precision	72.44
recall	86.58
f1 score	78.88

4 Evaluation

The dataset is the Heart Disease Dataset available on kaggle which has over 1000 observations and 14 attributes. These attributes range from gender, age to health diseases that may or may not affect the heart.The problem here is a binary classification and output of each model is either 1 or 0 class 1 means that patient has a chance of heart failure and class 0 means that patient has a healthy heart Evaluation would be done by comparing the performance of multiple models with the help of a confusion matrix and calculating the accuracy, precision and recall. The model with the highest accuracy would be selected as the final one, which can further be used by healthcare professionals for their use.

5 Conclusion

In conclusion, machine learning algorithms can be effective in predicting heart failure. Our preliminary results show that the Naive Bayes algorithm gave the accuracy of 81% .The results also indicate that feature selection plays a critical role in improving the performance of machine learning models. However, further investigation is needed to validate the results and determine if the difference in accuracy rates is statistically significant. This project has the potential to improve healthcare outcomes by providing early prediction of heart failure, which can lead to timely interventions and prevention strategies. Future work can involve applying other machine learning algorithms, to further improve the accuracy of heart failure prediction. Additionally, more data can be collected to increase

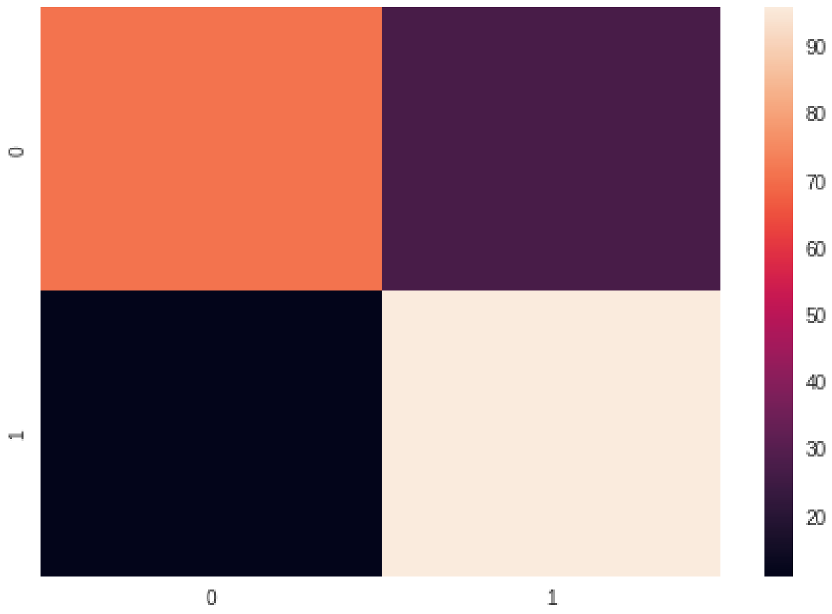

Figure 2: Heatmap for the confusion matrix.

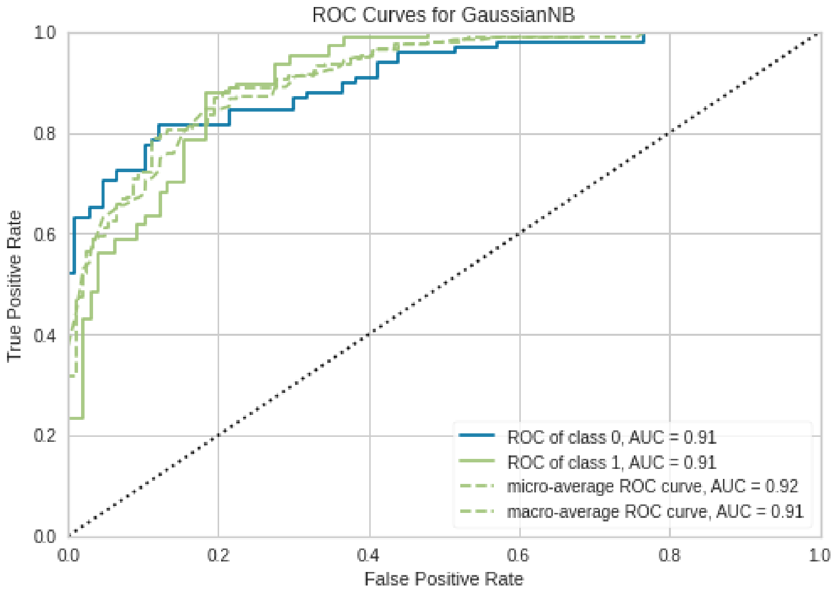

Figure 3: ROC curve

the sample size and diversity of the dataset, which can improve the generalizability of the machine learning models.