Machine Learning Model to Predict Heart Failure Disease

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. [1]

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. [2]

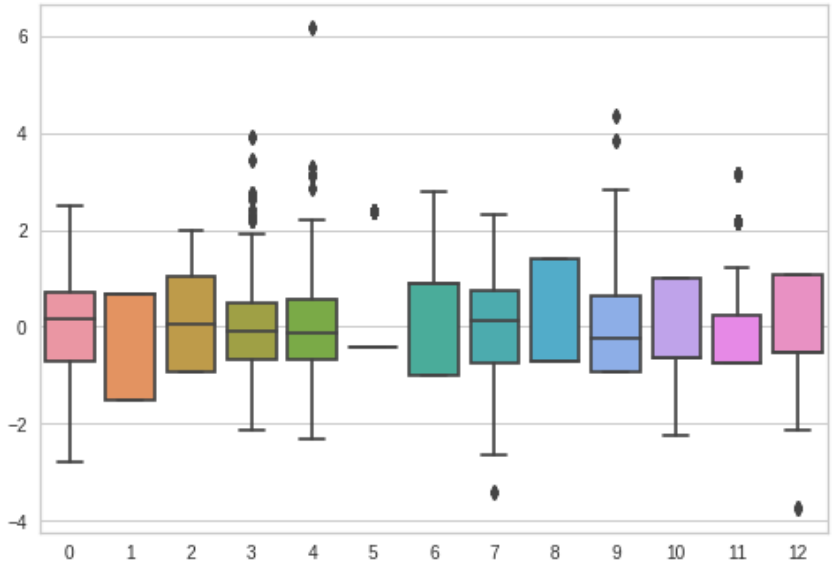

Figure 1: Box plot

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.[3]

2.2 Training methods

2.2.1 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. [4]

2.2.2 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. [5]

2.2.3 Random Forest

It builds trees on different 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. [6]

2.2.4 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.[7]

2.2.5 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. [8]

3 Implementation

3.0.1 Support Vector Machine

Even if there is a distinct separation between the two classes, there are many possible hyper planes that can be chosen. Being more effective in high dimensional spaces it gave the highest accuracy rate with 91.707% with a precision of 86.735%

accuracy	91.707
precision	86.735
recall	95.506
f1 score	90.909

3.0.2 Naive Bayes

Being a model which is easy to implement and handles both continuous and discrete data. Although with being highly scalable with a large number of predictors and data points, it is also assumes all the predictors are independent which rarely is the case. Yet, the accuracy with Naive Bayes was 81.463% and precision rate of 72.448% [9]

accuracy	81.463
precision	72.448
recall	86.585
f1 score	78.888

3.0.3 Random Forest

Since this model reduces overfitting in decision trees and aids to improve the accuracy, it is extremely pliable for both regression and classification problems. But with complex data, it is unable to explicitly describe the relationship within data. To increase accuracy, the model requires more trees which can cause the model to slow down substantially. With this dataset, the model had an accuracy of 83.415% and a precision rate of 73.469%

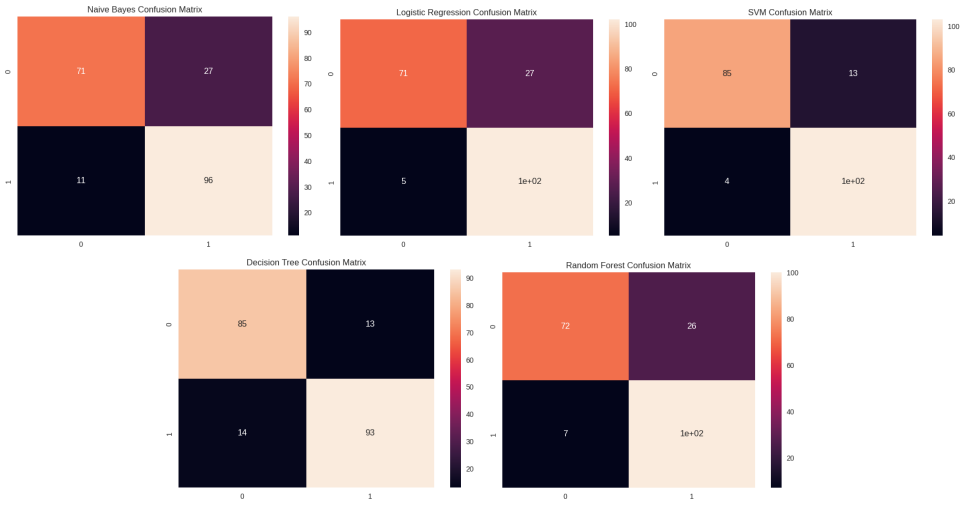

Figure 2: Confusion Matrix Heatmap

accuracy	83.415
precision	73.469
recall	90.000
f1 score	80.899

3.0.4 Logistic Regression

The outputs with Logistic Regression have a nice probabilistic interpretation, and the algorithm can be regularized to avoid the problem of overfitting, yet it tends to under-perform when there are multiple non-linear decision boundaries. This is mainly due to the fact that it assumes linearity between the predicted and the predictor variables. Logistic Regression gave an accuracy of 84.390% and a precision rate of 72.449%

accuracy	84.390
precision	72.449
recall	93.421
f1 score	81.609

3.0.5 Decision Tree

Decision Trees are very simple to understand and interpret. The ability to handle both numerical and categorical data along with multi-output problems makes it the second best model to work with. But with even a slight variation in data there needs to be a completely new tree generated and this can sometimes create extremely complex trees that do not generalize the data properly. With a accuracy of 86.829% and precision rate of 86.735% it has potential to improve.

accuracy	86.829
precision	86.735
recall	85.859
f1 score	86.294

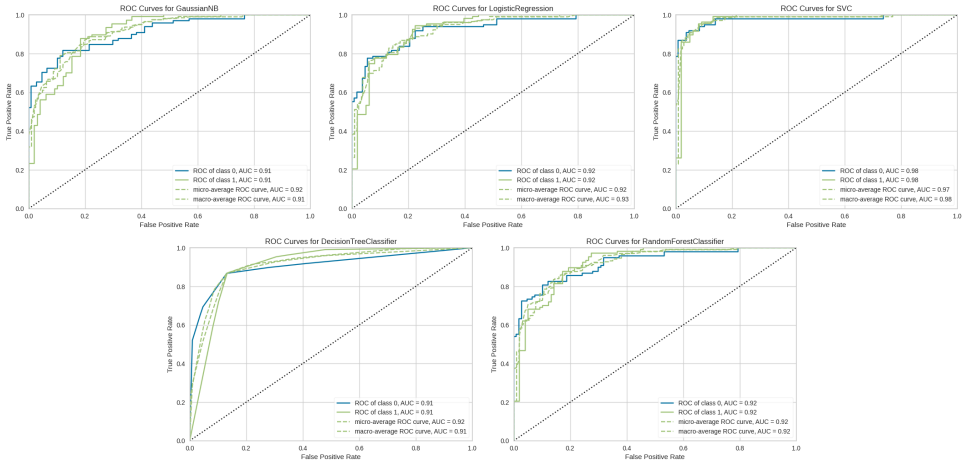

Figure 3: ROC Curve

4 Results

Model	Accuracy
Support Vector Machine	91.707
Decision Tree	86.829
Logistic Regression	84.390
Random Forest	83.415
Naive Bayes	81.463

The table above displays the model used and the accuracy achieved by each of them respectively. Support Vector Machine gave the highest accuracy due to its excellent generalization capacity and robust nature towards outliers which are present in this data, since it is a 'real dataset'. By using this model for prediction, it is definitely designed to creative a positive impact in the healthcare domain.

5 Conclusion

In conclusion, machine learning algorithms can be effective in predicting heart failure. Our results show that Support Vector Machine gave the highest accuracy .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 the sample size and diversity of the dataset, which can improve the generalizability of the machine learning models.[10]

6 Contributions

Name	Contribution
Sai Chandan Reddy K	Worked on Data Collection and Preprocessing,Helped in analysing models
Sakshi Mhatre	Worked on Model training(SVM,Decision Tree,Logistic Regression) and model metrics
Aniruddha Hore	Worked on Model training(Random Forest,Naive Bayes)

References

[1] Ahmad, T., Lund, L. H., Rao, P., Ghosh, R., Warier, P., Vaccaro, B., & Desai, A. S. (2018). Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. Journal of the American Heart Association, 7(8), e008081.

[2] Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies. MIT Press.

[3] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., ... & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American journal of cardiology, 64(5), 304-310.

[4] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[5] Zhang, H. (2004). The optimality of naive Bayes. Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, 562-567.

[6] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[7] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

[8] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

[9] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

[10] Johnson, K. W., Torres Soto, J., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., ... & Dudley, J. T. (2018). Artificial intelligence in cardiology. Journal of the American College of Cardiology, 71(23), 2668-2679.