# Predicting and Analyzing Heart Disease Using Machine Learning Techniques

Kenneth Lam
November 2022

## 1 Introduction

According to the American Heart Association, in 2020 approximately 19.1 million deaths were attributed to cardiovascular disease globally. The most common causes are coronary artery disease and heart attacks, which can come unexpectedly in some people and complicate their health. Predicting heart disease in its early stages can help patients avoid fatal events and seek preventative care when needed. However, there are a number of factors that come into play when looking at the causes and significant indicators of heart disease. There have been a number of studies that show how smoking, alcohol, physical health, and mental health can affect the rate of heart disease progression in at-risk patients. Currently, healthcare professionals can provide insight into a person's cardiovascular health through many of the available medical technologies available these days. The main problem most of the population has to deal with is the high prices associated with getting medical checks, consulting a specialized doctor, getting blood work done, and all the different assessments that can be made to evaluate heart health. Getting the right attention and treatment as early as possible can be expensive and complicated, but there have been numerous studies that provide possible risk factors that increase the likelihood of heart disease in patients. With machine learning, this concept can be further explored to determine a person's chances of having a heart disease given medical history and background through different algorithms to create an accurate predictive model that assesses a person's health and risks.

This project aims to identify the key predictors of heart disease and predict the chances of heart disease using machine learning models. logistic regression, naive Bayes, trees, LDA, deep learning, etc.). We intend to experiment with different prediction models and verify which are the most accurate in evaluating heart disease risk factors. The data comes from the UCI Machine Learning Repository, which consists of over 1,000 observations sampled from Cleveland, Hungary, Switzerland, and LongBeach capturing a large portion of the population. The traits included in the data set are age, sex, type of chest pain, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise-induced angina, ST-induced depression induced by exercise relative to rest, the slope of peak exercise ST segment, number of major vessels colored by fluoroscopy, and thallium defect. and target(whether the patient has heart disease or not). A journal called Beyond Established and Novel Risk Factors from the American Heart Association has identified indicators of heart disease that go along with the predictors used in the dataset and the goal of this project is to replicate some of the results produced in the paper.

## 2 Related Works

Our paper focuses on the roles the indicators of heart disease play in the correct diagnosis of patients with possible heart disease, as well as the accuracy with which

we are able to predict without having too many false positives/negatives. Heart disease prediction is a problem that has been worked on for the last 20 years in the medical field with increasingly accurate algorithms being created to improve upon previous works. In our research, we found other works that have done similar processes to ours and found that the most helpful measurements used to predict heart disease come from electrocardiogram readings (ECG), as cited in [2]. In [3], the researchers fit a variety of machine learning models, getting very similar results that were around 89-93% accurate in predicting cardiovascular disease. We look to improve upon these results and try to understand more about what predictors have an important role in predicting the outlook of patients.

The data we have chosen has been used before in previous papers as well. By focusing less on the medical interpretation of our findings, we hope to not only be able to replicate results found in papers like [1] and [3], but hopefully build upon them slightly more. Nowadays, machine learning models are used vastly in the medical field for cardiovascular disease prediction, as it allows doctors to validate their diagnosis and help improve the accuracy with which they can identify potential at-risk patients.

## 3 Methods

### 3.1 Initial Dataset

The dataset consists of four distinct data samples gathered from the Long Beach Veterans Administration Medical Center, the Cleveland Clinic, the Hungarian Institute of Cardiology, and Swiss university hospitals based in Zurich and Basel. All three study groups gathered their information between 1984 and 1987, with 303 patients from the Cleveland Clinic, 425 from the Hungarian Institute, 200 from Long Beach Medical Center, and 143 from the Swiss hospitals. The measurements taken at each facility were age, sex, chest pain characteristics, resting blood pressure, cholesterol, fasting blood sugar (fds), resting electrocardiogram results (restecg) , the maximum heart rate (thalach) , exercise-induced angina (chest pain induced during exercise related to heart health), old peak (ST depression difference between heartbeat intervals during rest and exercise), ST slope, number of blood vessels colored during fluoroscopy assessment, and exercise thallium defect results.

Based on all the information gathered and the patients' background, the target predictor in the dataset assigns a binary value to whether the subjects had heart disease or not, hence being the response variable used in the models. Patients from the Cleveland Clinic had no history or electrocardiographic evidence of prior myocardial infarction or known valvular or cardiomyopathy disease. It is worth noting that the Cleveland Clinic data serves as a reference group for comparisons with the other groups' clinical data measurements. The other three clinics serve as test groups to analyze the cardiology tests where the patients may or may not be prone to having heart disease. The dataset also has a relatively even split between patients with no heart disease and patients with heart disease. 499 were classified as having no heart disease while 526 were classified as having heart disease. Having a balanced data set is important because an unbalanced data set(a data set where the response variable has more observations in one specific class than another) can lead to an accuracy paradox where the model obtains high accuracy by assigning observations to a majority class rather than correctly classifying the observation.

### 3.2 Normalization

Before analyzing the dataset, we modified the predictors by making a dummy variable matrix with the factored predictors and then normalized the whole dataset to homogenize the values for each predictor variable. Without normalization, all the calculations done within the models would be disproportionate because an unequal weight would be assigned to sections of the data which in actuality should not have that much of an influence. Normalization ensures that the model performance, as well as accuracy, are the values they should be. In order to obtain more accurate results and models, we employed all our methods using this normalized dataset.

### 3.3 Splitting training and test sets

We decided to split the dataset into training and test sets using 70% for training and 30% for testing. In order to normalize the dataset distribution and provide the models with enough samples that indicate disease and those that do not, we randomly sampled subjects out of all testing and reference groups to make the training and testing data for model making.

## 4 Discussion

### 4.1 Logistic Regression

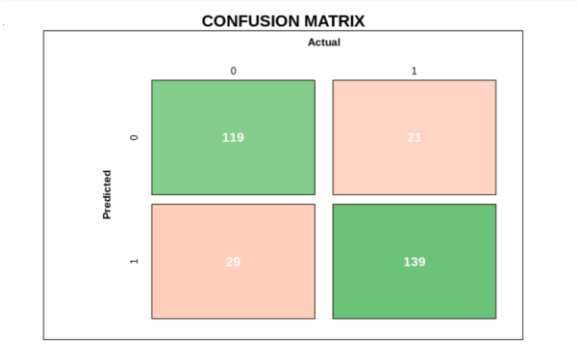


Figure 1: Logistic Regression Confusion Matrix 139 TP(True Positive) 119 TN(True Negative) 29 FP(False Positive) 21 FN(False Negative)

The logistic regression model learns and predicts the parameters in the data set using regression analysis. It requires the class variable to be binary and in our case, the target column has a binary value of 0 for a patient with no heart disease and 1 for a patient with heart disease. Logistic Regression then computes the ith patient's disease probability, Pi, using the formula: Pi = es/(1 + es), where fi is the linear combination of this patient's data, using the appropriate subset of coefficients. We fit a logistic regression model using the using target as a response variable and all predictors. Subsequently, we then performed backward stepwise selection on the resulting model to keep only the most significant variables in the model. After backward stepwise selection, we found the model with the minimum AIC was a model of 15 variables with an AIC of 445.43. The variables included in the model were sex0, cp1, cp2, cp3, trestbps , chol, thalach, exang1 , old peak, slope1 , ca1 , ca2 , ca3 , thal1, and thal2. The accuracy for the backward stepwise selection logistic regression model was 83.77% (Figure 1).

### 4.2 Linear Discriminant Analysis & Naive Bayes

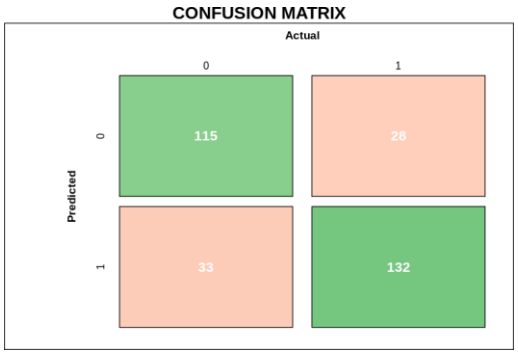


Figure 2: Naive Bayes Confusion Matrix 132 TP 115 TN 33 FP 28 FN

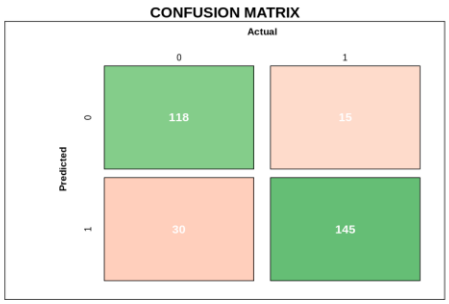


Figure 3: Linear Discriminant Analysis Confusion Matrix 145 TP 118 TN 30 FP 15 FN

The naive Bayes model classifies the data by assuming feature independence and then computing the probability of independent variables. It calculates the membership probability of each class and then assigns the class with the highest probability as the most likely class. LDA works similarly to naive Bayes and uses the Bayes theorem. A key difference is that LDA allows the features to be correlated. The naive Bayes model had an accuracy of 80.19% (Figure 2) while the linear discriminant analysis model had an accuracy of 85.39% (Figure 3).

## 4.3 Linear SVM

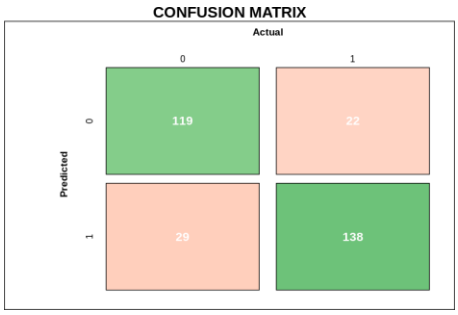


Figure 4: Linear SVM confusion matrix 138 TP 110 TN 29 FP 22 FN

SVM works by mapping data to a high-dimensional feature space. After the separator has been found, the data is transformed and can be drawn as a hyperplane which acts as a decision boundary. Anything that falls on one side of the boundary will be classified as one class and anything on the other side will be classified as the other class. The best hyperplane is one that maximizes the margins between the classes. The linear SVM model had an accuracy of 83.44%. (Figure 4) Using tuning we considered costs of 0.001, 0.01, 0.1,5,10,100 and found that a model with a cost of 0.01 had the highest accuracy. We decided to use these values of cost because they were exhaustive and covered the ranges where we were likely to find the optimal value that maximized our accuracy.

## 4.4 Radial SVM

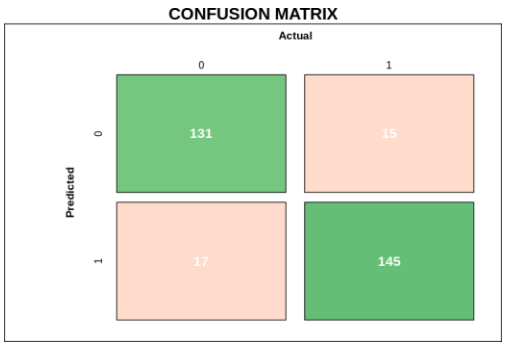


Figure 5: Radial SVM Confusion Matrix 145 TP 131 TN 17 FP 15 FN

Using tuning we fit different costs (0.1,1,10,100,1000) and gamma values(0.5,1,2,3,4) and were able to find the best radial model which had a cost of 10 and gamma of 0.5. As mentioned previously these values were exhaustive and covered the ranges where we were likely to find the optimal value that maximized our accuracy. The radial SVM model had an accuracy of 89.6% (Figure 5)which is a higher accuracy than the linear SVM model.

## 4.5 KNN (K-nearest neighbors)

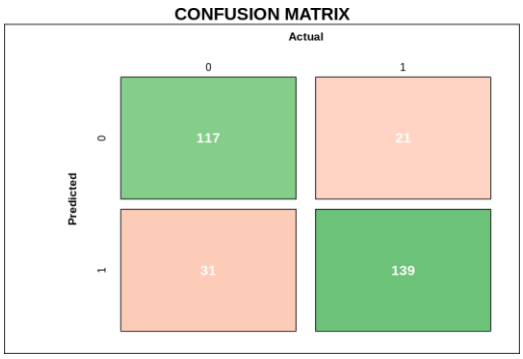


Figure 6: KNN Confusion Matrix 139 TP 117 TN 31 FP 21 FN

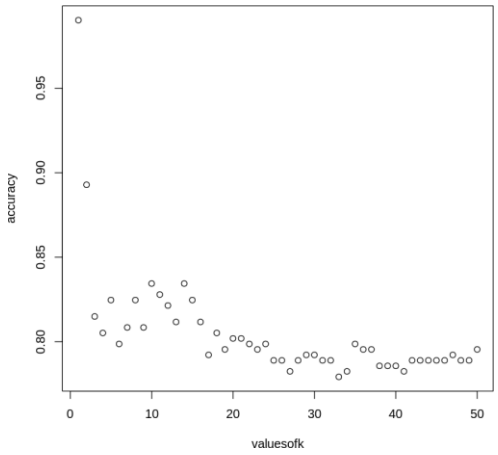


Figure 7: Values of K vs accuracy

KNN finds the distances between a query and all the observations in the data. The algorithm selects a K number of samples closest to the query, then classifies as the most frequent label in the case of classification or the average in the case of regression. In order to train our model properly without giving any bias, we removed the target predictor from the training data when building the model. We experimented using different k values ranging from 1-50. Figure 7 graphs the accuracy as a function of the values of k. The highest accuracy in this graph is a k value of 1 and a k value of 2 with accuracies of 99% and 89.28 percent respectively. There are many concerns with KNN when selecting a k value because a value of k that is too low may overfit the data and a value of k that is too high might result in underfitting. After k values of 1, and 2 there is a significant drop in accuracy. The highest accuracy after the decrease can be found at a k value of 14 which yielded an accuracy of 83.1% (Figure 6). After a k value of 14 the

accuracy only decreases. For these reasons we decided k=14 is a reasonable value.

## 4.6 Decision Tree

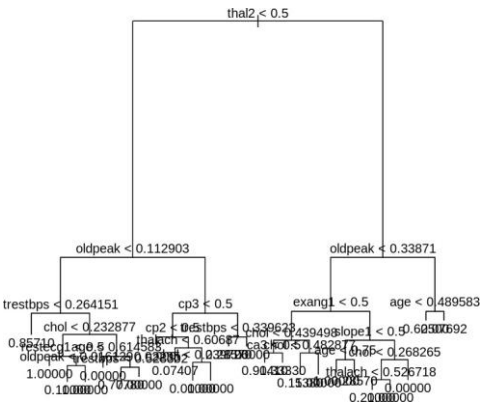


Figure 8: Tree Diagram

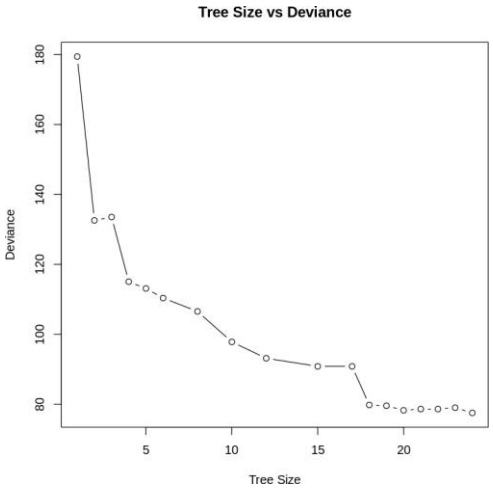


Figure 9: Tree size vs Deviance We can observe that the deviance continues to decrease until the size of the tree is equal to 24.

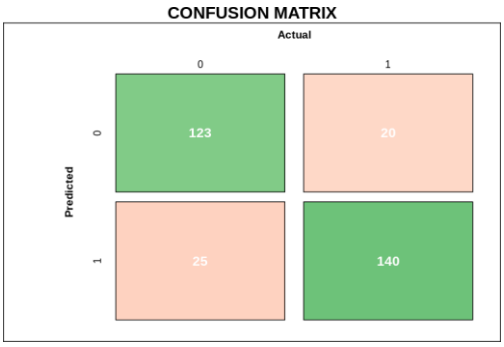


Figure 10: Tree confusion Matrix 140 TP 123 TN 25 FP 20 FN

The decision tree algorithm splits data features through a cost function. The decision tree is then grown and optimized through a process known as pruning, where branches containing insignificant features are removed. The result is the tree shown in Figure 8, which shows the most important indicator of heart disease to be thallium defect since it is the first branch that differentiates and splits the tree into two old peak variables based on whether thal2 was less than 0.5. We used 10-fold cross-validation to find the optimal tree size to obtain the lowest cross-validation error. After plotting the size of the tree to the number of cross-validation errors, we found that a tree of size 24(Figure 9) had the lowest amount of cross-validation errors. We then decided to prune the tree to obtain a higher test accuracy. After pruning the tree to 24 nodes we found that the accuracy was 85.3% (Figure 10). We also tried using a random forest with a split of 5 predictors and 10 trees as well as 100 trees. Random forest resulted in abnormally high accuracy. A random forest of 10 trees had a 95.45% accuracy and 98.05% accuracy with 100 trees. However, we chose not to include

random forests in our paper because the model is complex and becomes infeasible to interpret as the amount of trees increases.

## 4.7 Artificial Neural Networks(ANN)

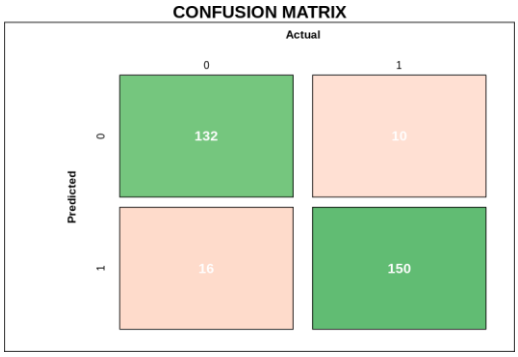


Figure 11: ANN Confusion Matrix 150 TP 132 TN 16 FP 10 FN

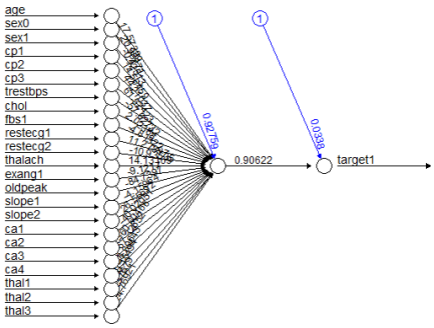


Figure 12: Neural Model 1 hidden layer

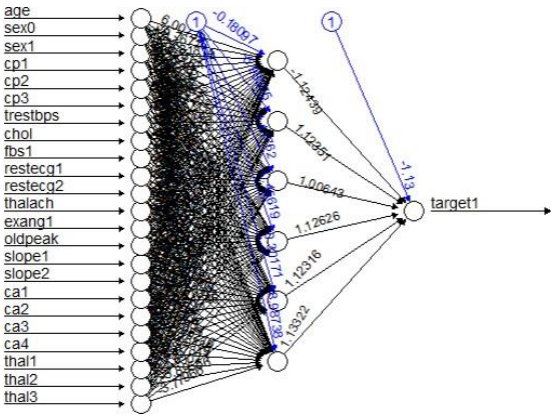


Figure 13: Neural model with 6 hidden layers

Neural nets are modeled on the neurons in a human brain consisting of many simple processing nodes that are interconnected. A simple neural network has three layers known as the input layer, the hidden layer, and the output layer. First is the input layer where the data enters the network to be processed and passed on to the next layer. Then there are hidden layers that take input from the input layer or other hidden layers, process the data further, and pass the data on to the next layer. Lastly, the output layer is where the final results of the data processing are achieved. In deep neural networks, there can be several hidden layers that take greater computational power the more layers there are. To each of the connections, a node will assign a weight and a different number when receiving data. The ANN then multiplies the number by the weight to obtain a single value. If the single value exceeds some threshold, the node will fire and pass on the data. The neural network trains by manipulating the weights and thresholds until the training data is similar to the outputs. The ANN shown in figure 12 was built using only one hidden layer of neurons, outputting 91.5% accuracy in predictions. When changing the number of hidden neurons, we see that increasing the number of neurons also increases the accuracy, getting 95% accuracy with 6 layers of neurons in the model. The more hidden layers the higher the accuracy. When comparing Figures 12(single layer) and 13(six layers) we can see a sizable increase in the complexity of the model which will make the six-layer model uninterpretable.

| Model | Accuracy |
|---|---|
| <chr> | <dbl> |
| Logistic Regression | 0.8376623 |
| LDA | 0.8538961 |
| Naive Bayes | 0.8019481 |
| Linear SVM | 0.8344156 |
| Radial SVM | 0.8961039 |
| KNN | 0.8311688 |
| Tree | 0.8538961 |
| ANN | 0.9155844 |

Figure 14: Table of model accuracies

## 5 Limitations of the model

Angiography is an examination by X-ray of blood or lymph vessels, carried out after the introduction of a radiopaque substance which is simply a substance that is opaque to X-rays and similar radiation. Since the samples were obtained from a population referred for angiography, the best performance of the models is expected from a population with the same characteristics. In the test groups, noninvasive test results were not withheld from treating physicians which would lead to referral bias because it would influence the decision to perform coronary angiography. In addition, the angiograms were read by clinical angiography through visual assessment. The readings were done without the knowledge of test results as intended. However, it is likely that the clinical angiographers were aware of the clinical symptoms of the patient as well as the age and sex of the patient.

## 6 Conclusions and Future Work

The fitting of models to the dataset in our project proved quite complicated upon finding that some of the more complex models were outputting accuracies of above 95% with some even having 100% accuracy. This led us to the decision to normalize the dataset to ensure correct weights were being placed for each factor and predictor. After validating the models with the testing set, the models that performed the best were the Artificial Neural Network and the radial kernel SVM. ANN had an accuracy rate of 91.55% and radial SVM had an accuracy of 89.61% as seen in figure 14. The predictions fit almost all of the testing data correctly while minimizing Type I and Type II errors. In machine learning-assisted medical diagnosis, it is crucial that misclassification is minimized since the implications of misdiagnosing someone who does have heart disease can make a difference in whether a patient's disease progresses undetected or suffers side effects from being given the wrong medications from misdiagnosis. This is why the significance of having accurate models assist in diagnosing patients correctly is so meaningful, not only does it make the costs of additional tests to get a full diagnosis decrease, but it also improves the efficiency with which hospitals and doctors could operate their clinics.

The models and analysis also showed that the most important predictors in evaluating heart disease were the thallium exercise test, old peak, and maximum heart rate achieved in testing. The data we used is quite old, so

revisiting the dataset with more powerful algorithms than the ones used in the reference research paper proved to make more accurate predictions as well. For future work, we would like to gather more data to ensure the accuracy of the models is correct and there is always room for improvement in the models where doing more cross-validation to make changes could provide better results and more insight into the importance of the predictors in heart disease diagnosis.

## 7 Contributions

We contributed equally to the project.

## References


[1] Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid JJ, Sandhu S, Guppy KH, Lee S, Froelicher V. International application of a new probability algorithm for the diagnosis of coronary artery disease. Am J Cardiol. 1989 Aug 1;64(5):304-10. doi: 10.1016/0002-9149(89)90524-9. PMID: 2756873.

[2] Nagavelli U, Samanta D, Chakraborty P. Machine Learning Technology-Based Heart Disease Detection Models. J Healthc Eng. 2022 Feb 27;2022:7351061. doi: 10.1155/2022/7351061. PMID: 35265303; PMCID: PMC8898839.

[3] Fahd Saleh Alotaibi, "Implementation of Machine Learning Model to Predict Heart Failure Disease" International Journal of Advanced Computer Science and Applications(IJACSA), 10(6), 2019.