Machine Learning Analysis of Heart Diseases

Miralireza Nabavi Bavil, Van McNulty, Amir Tavakkoli and Julie Larsen
School for Engineering of Matter, Transport & Energy
Arizona State University

1 Abstract

Heart disease is the leading cause of death globally, accounting for 32 percent of all deaths. There are many symptoms that can indicate the presence of heart disease. The aim of this project is to provide patients with a heart disease risk assessment based on a limited set of metrics regarding their health. In this project, various machine learning algorithms were utilized to predict heart diseases using clinically gathered data by the University of California Irvine Machine Learning repository. Aggregated from patients across several hospitals, the dataset includes patient features of: age, sex, chest pain, blood pressure, heart rate and more which can be used to diagnose heart diseases.

A ridge classifier model, K nearest neighbor (KNN) model, neural network (NN) model, support vector machine (SVM) model and decision tree model were all utilized in this project. These models all showed high accuracy, with many of these models provided 100% accuracy in determining the presence of heart disease. The NN model, decision tree model, SVM model, KNN model all yielded a 100% accuracy. The Ridge model provided a lower 83% accuracy.

This paper is organized as follows. The dataset and features are introduced in §2. Section §3 describes the exploratory analysis of the dataset. The results of the machine learning model development is covered in §4 followed by conclusions and suggested future works in §5.

2 Dataset Introduction

2.1 Features

The heart disease dataset selected for this study consists of 1026 observations. For each observation there are 13 features which describes the patients, their medical symptoms, and various medical test results. The features of the dataset are as follows:

1. **age**

   This numerical feature describes patient's age in years.

2. **sex**

   This categorical feature specifies the patient's sex, with 1 representing male and 0 representing female.

3. **cp**

   This categorical feature specifies the patient's chest pain type. Four types of chest pain are defined: typical angina, atypical angina, non-anginal pain, and asymptomatic. Typical angina pain is characterized by three factors: (1) a feeling of squeezing, pressure, or tightness in chest that is (2) provoked by some form of physical exertion and (3) can be relieved by rest or nitroglycerine. Atypical angina occurs when only two of the three factors that defines angina pain is present. Non-anginal pain can appear as any pain not involving the chest. Asymptomatic means the patient shows no symptoms of a cardiac event.

4. **trestbps**

   This numerical feature is the patient's resting blood pressure, measured in mm-Hg upon their admission to the hospital.

5. **chol**

   This feature is a numerical value of the serum cholesterol, measured in mg/dl. The serum cholesterol value represents the amount of total cholesterol in the patient's blood.

6. **fbs**

   This categorical feature specifies whether the patient's blood sugar is greater than or less than 120 mg/dl after an overnight fast, with 1 indicating levels greater than 120 mg/dl and 0 indicating less than 120 mg/dl. A fasting blood sugar level greater than 120 mg/dl classifies the patient as diabetic.

7. **restecg**

   This categorical feature describes the testing electrocardiograph (ECG) results. The categories of this feature are: normal (0), slight abnormality (1), or probable to definite signs of left ventricular hypertrophy, which is thickening of the hearts pumping chamber (2).

8. **thalach**

   This numerical feature is the patient's maximum heart rate achieved during testing.

9. **exang**

   This categorical feature specifies if the patient's angina pain (if present) was induced by exercise. A value of 1 represents 'yes', 0 represents 'no'.

10. **oldpeak**

    This categorical feature specifies if there is a ST depression induced by exercise relative to patient's resting condition. ST depression is a potential outcome on a patients ECG results that indicates a restriction in blood supply to the heart. A value of 1 indicates there is a ST depression.

11. **slope**

    The 'slope' numerical feature refers to the slope of the peak exercise ST segment on an ECG test. A normal result from this test slopes sharply upwards, while irregular results remain relatively flat.

12. **ca**

    This is a numerical feature that indicates the number of major vessels (0-3) colored by flourosopy. Fluoroscopy is a medical technique used to see the flow of blood through the coronary arteries. If a major blood vessel is not colored by flourosopy, there is an arterial blockage.

13. **thal**

    This categorical variable indicates whether the patient has the blood disorder, Thalassemia. The disorder causes the body to produce less hemoglobin than normal. Hemoglobin are what enables red blood cells to carry oxygen. Patients are classified as not having Thalassemia, having fixed case of Thalassemia, or having reversible case.

14. **target**

    The target feature is the presence of heart disease. A value of 1 indicating heart disease is present, and 0 meaning heart disease is not present.

2.2 Features Engineering

There are several categorical variables in the dataset. Variables 'sex', 'cp', 'fbs', 'displacement', 'exang', 'restecg' are categorical. These variables are converted to numerical values using one-hot encoding.

Thus, after conversion, there are 13 features available which determine the target feature, and price. The features are normalized using a standard scalar method which ensures the standard deviation of each variable is 1 and the mean is zero.

3 Exploratory Data Analysis

In this section, we will gain some general knowledge about the trends governing the dataset using visualization and exploratory data analysis.

One of the methods used in this project is called Decision Trees (DT). A Decision Tree is a type of supervised machine learning used to categorize or make predictions based on how a previous set of questions were answered. This method is further explained in part 4 where the models are developed. Using the Decision Tree Classifier, the importance of the features mentioned in the previous part was obtained.

The following table shows these features with their importance factors.

Feature	Importance Factor
age	0.102
sex	0.027
trestbps	0.068
chol	0.081
fbs	0.012
restecg	0.039
thalach	0.031
exang	0.043
oldpeak	0.149
ca	0.063
cp_0	0.269
cp_1	0.010
cp_2	0.005
cp_3	0.010
$thal_0$	0.000
$thal_1$	0.003
$thal_2$	0.027
$thal_3$	0.006
$slope_0$	0.000
$slope_1$	0.046
$slope_2$	0.001

Table 1: Importance factor of each feature.

It was found that the most influential features are oldpeak, age, cp_0, and chol. However, still the other features are considered impactful on the prediction and their importance is unavoidable but the main focus can be on the most influential features.

Moreover, in this study there were two sex categories defined. The female was defined as the number 0 and the male was defined as the number 1. The number of counts for each one of these was obtained and a histogram was obtained as follows. In addition, another factor named as target was provided. This factor indicates whether someone has heart disease or not (1=yes, 0=no). Another histogram for the number of counts for each one of these categories was obtained as follows.

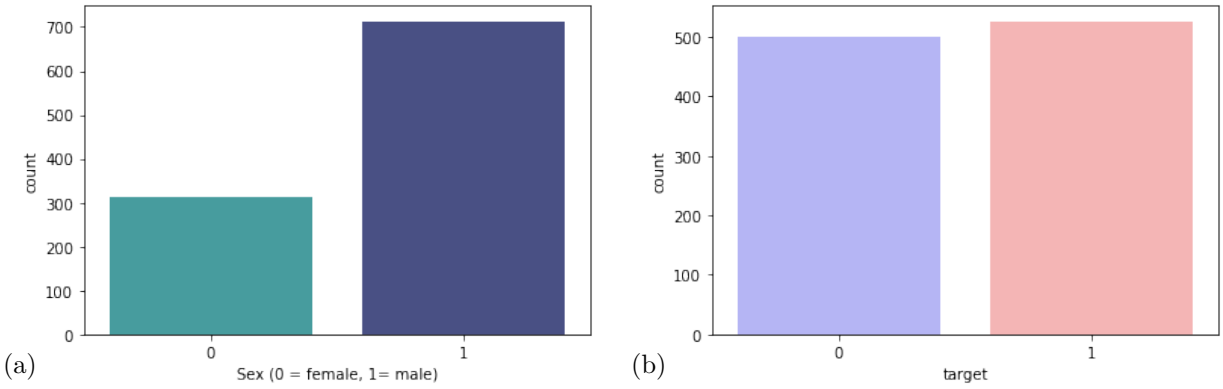


Figure 1: The histogram plots of counts versus (a) sex and (b) target

Figure 1 (a) shows the histogram plot of counts versus sex. We could infer from the histogram that the number of males in the data set provided is considerably higher compared to the number of females. However, in Figure 1 (b), we observe that the number of people who have heart disease is very close to the number of people who do not have heart disease where the number of those with heart disease is slightly higher.

As discussed before, the age feature was found as one of the most influential features in the data set. A histogram of heart disease versus age was obtained and it is shown as follows.

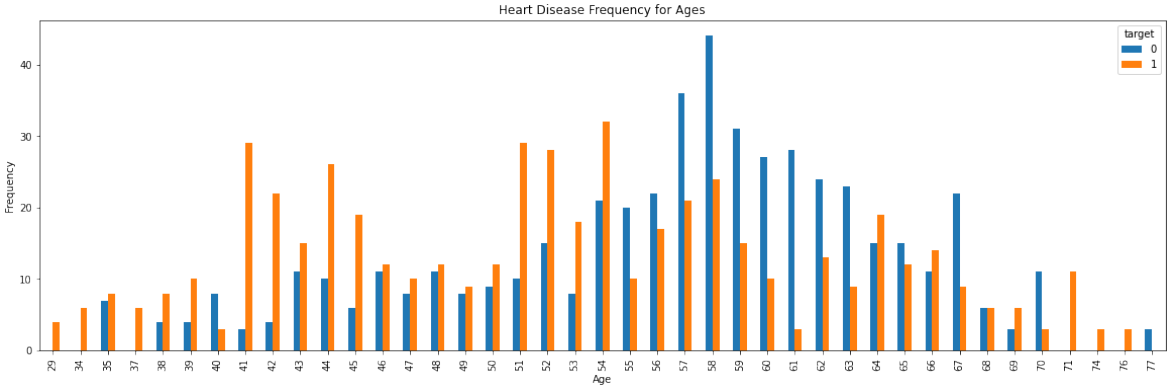
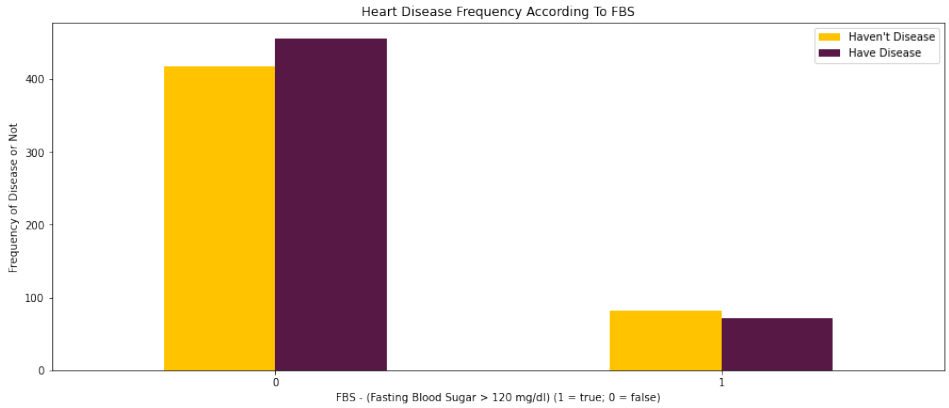


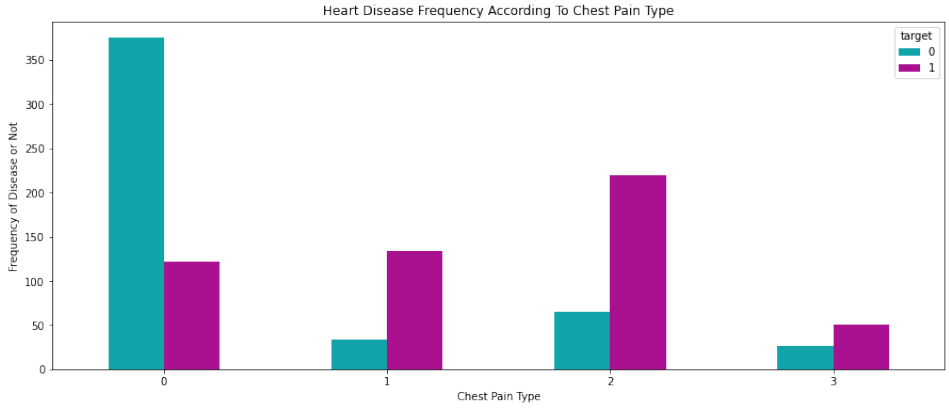
Figure 2: Heart disease versus age histogram

As it can be seen from the histogram provided for the frequency of heart disease for different ages, the highest number of people without heart disease is indicated for the age of 58 while the highest number of people diagnosed with heart disease is for the age categories of 54, 44, and 41.

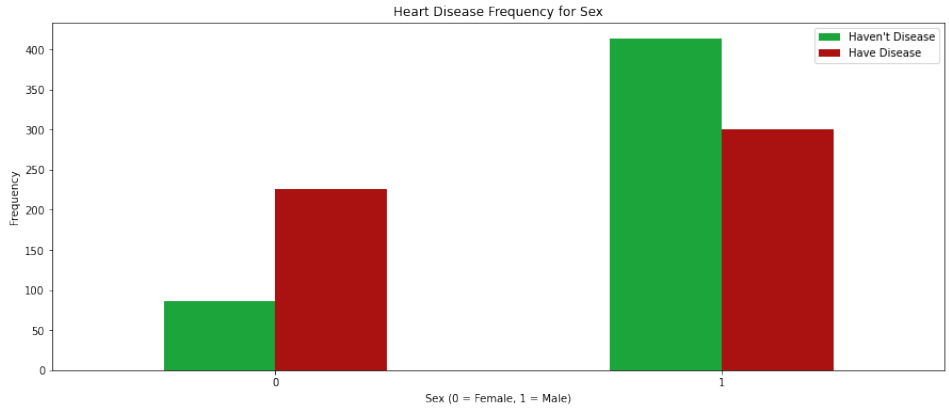
In addition, more plots were obtained to show the relation between the target feature and other features involved in this prediction. These plots are shown below.



(a)



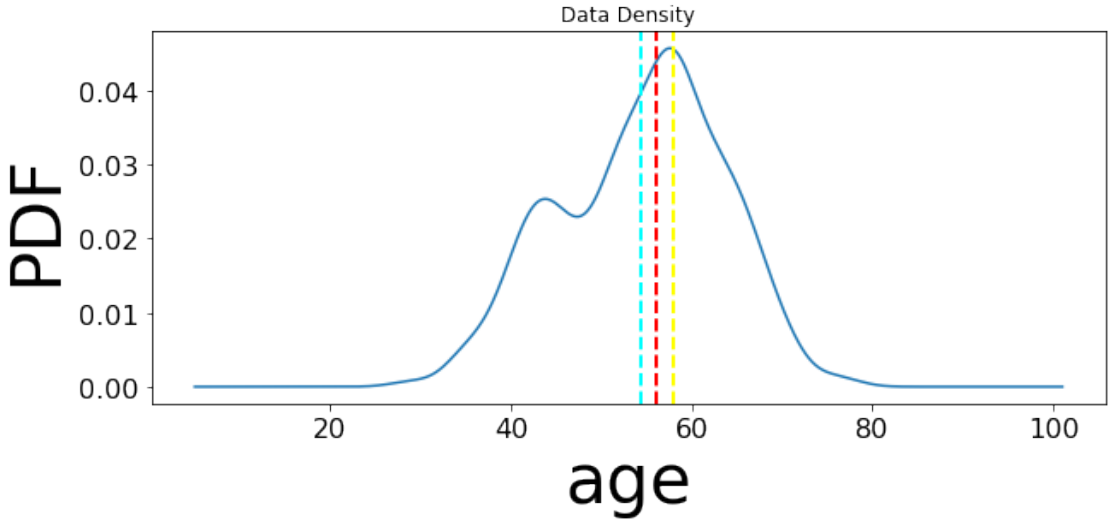
(b)



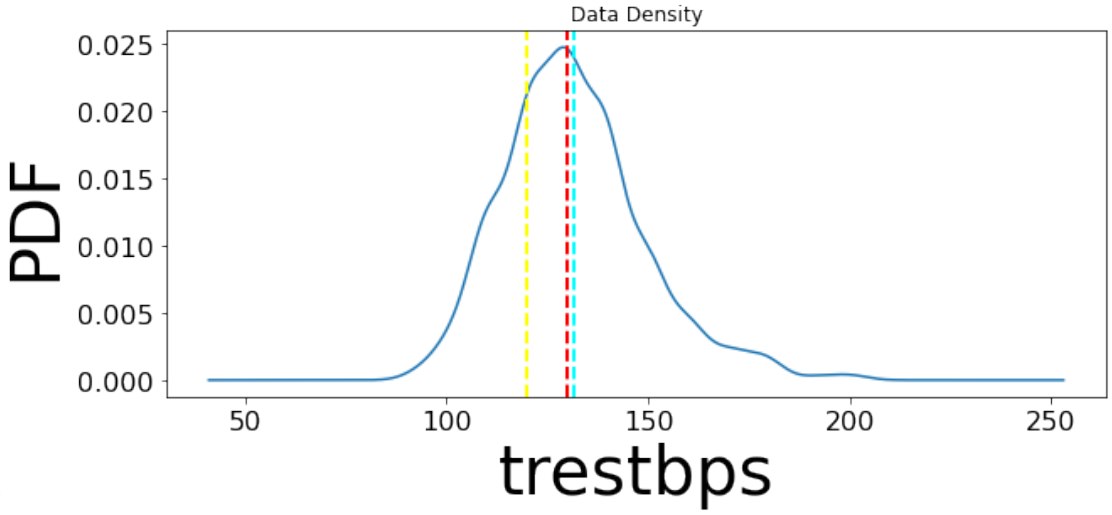
(c)

Figure 3: The histogram plots for frequency of heart disease versus (a) FBS, (b) Chest Pain Type, (c) Sex.

Finally, Figure 4 shows the probability density function (PDF) of the age and trestbps (which was explained in section 2). Mean, median and mode are shown with cyan, red and yellow dashed lines, respectively.



(a)



(b)

Figure 4: The probability density function (PDF) of (a) age and (b) trestbps. Mean, median and mode are shown with cyan, red and yellow dashed lines, respectively.

4 Machine Learning Prediction

In this section, several machine learning models will be developed and compared in terms of their ability to detect patients with heart disease. The metric used to evaluate the models is classification accuracy or simply the percentage of correctly identified cases. The dataset is divided into two sections: 1) the training set and 2) the testing set. 20 percent of the dataset is reserved for final evaluation, which means that the size of the training dataset is (820, 21), while the size of the test dataset is (205, 21). Note that the test dataset was not

used for model training. 10% of the training set was used for tuning the hyper-parameter of the models.

4.1 Ridge Classifier with Regularization

Linear regression is a linear model that assumes a linear relationship between the input variables (x) and the single output variable (y). This model is the simplest model that we could use to fit the data. For each feature, it provides an intercept and a slope. Furthermore, several regularization techniques are used to fit the Ridge classifier.

Regularization is an important concept that is used to avoid over-fitting the data, especially when the training and test data are much varying. Regularization is implemented by adding a "penalty" term to the best fit derived from the trained data, to achieve a lesser variance with the tested data and also restricts the influence of predictor variables over the output variable by compressing their coefficients [1].

In this study, a grid search study with 10 folds was performed to determine the best regularization constant. The number of constants used in cross-validation was 1000. Finally, the model was trained using the best regularization coefficient. The linear model was able to achieve an accuracy score of 82.94 %. This suggests that the price of the used cars is more or less linearly related to the features as the accuracy is quite high. Figure 5 shows the confusion matrix of the Ridge classifier.

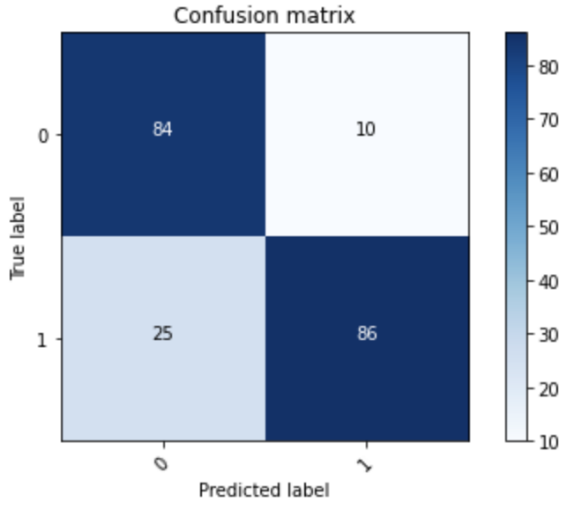


Figure 5: Confusion matrix for the Ridge classifier.

4.2 K-Nearest Neighbors (KNN)

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. The KNN classifier could be trained with either uniform or distance-based weights. Furthermore, the number of neighbors is an influential parameter

in training the model. The grid search analysis was conducted using these two parameters and 10 cross-validation folds. The results show that with distance-based weights and 13 neighbors we have the best performance. The model was trained using these parameters and achieved an accuracy score of 100% on the test set. Figure 6 shows the confusion matrix for the KNN classifier.

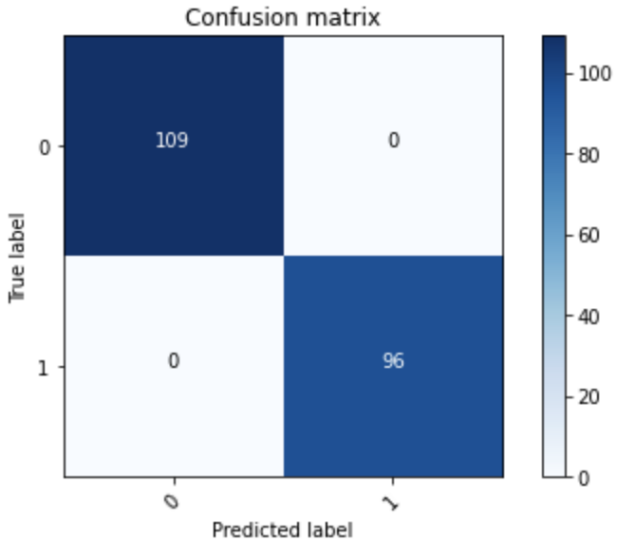


Figure 6: Confusion matrix for the KNN classifier.

4.3 Support Vector Machine

The support vector machine was utilized to develop a classification model for the heart disease dataset. For this model an appropriate kernel and regularization parameter (C) should be selected for ideal training. The options for kernel function are radial-based function (RBF), sigmoid, polynomial, and linear. The grid search analysis showed that the RBF function with $C = 40$ provided the best fit on the cross-validation test set. Figure 9 shows the accuracy on the cross-validation test set versus the regularization parameter, trained with RBF kernel function. The final model was trained with ideal parameters and achieved an accuracy score of 100% on the test set.

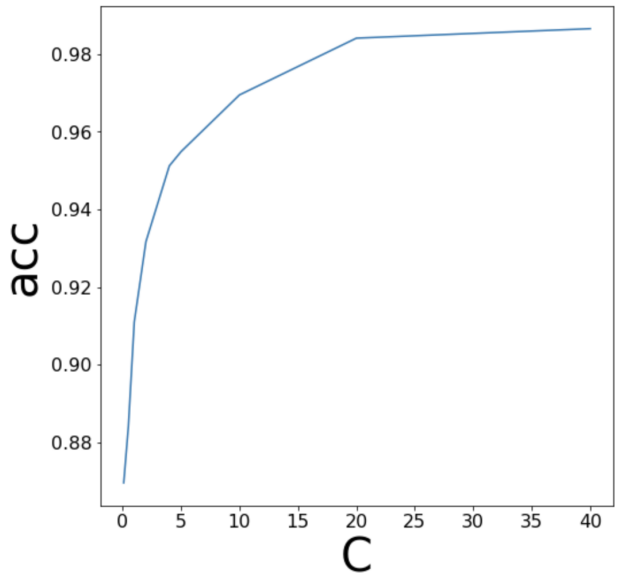


Figure 7: Accuracy of the support vector machine model on the cross-validation test set versus the regularization parameter, trained with RBF kernel function.

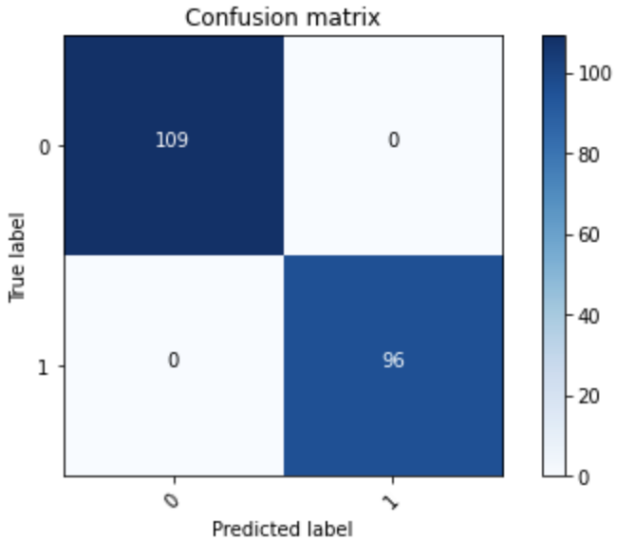


Figure 8: Confusion matrix for the SVM classifier.

4.4 Decision Tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning [2].

The hyper-parameters for this model are 1) 'criterion' which is the function to measure the quality of a split (available options: squared error, Friedman MSE, absolute error, and Poisson), the maximum leaf nodes of the model and maximum features which is the number of features to consider when looking for the best split (available options are auto, sqrt, and log2). The best hyper-parameters, according to the grid search analysis, were the Friedman MSE criterion, 200 maximum leaf nodes, and automatic maximum feature selection. The model was able to achieve an accuracy score of 100% on the test set.

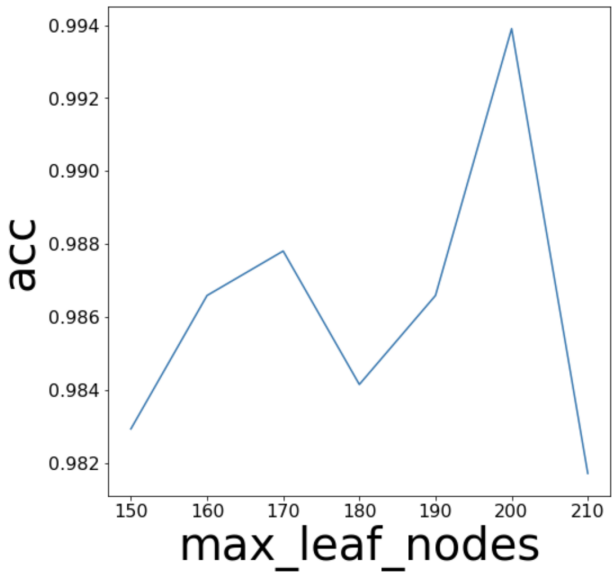


Figure 9: Accuracy of the decision tree model on the cross-validation test set versus the number of lead nodes, trained with RBF kernel function.

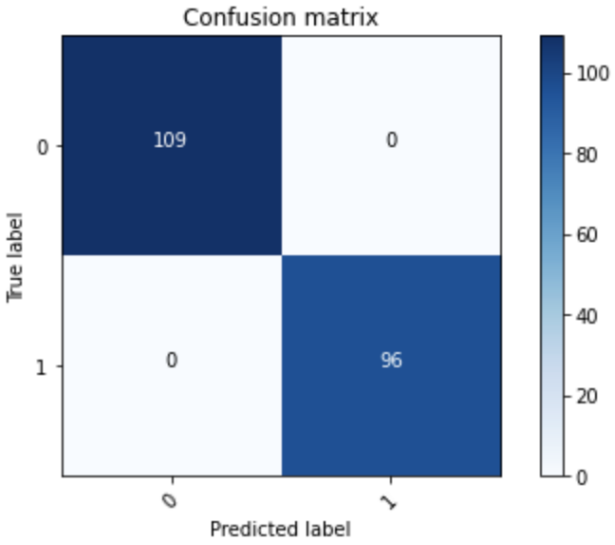


Figure 10: Confusion matrix for the decision tree classifier.

Using the decision tree model, the importance of each feature could be assessed. This provides priceless knowledge about the influential factors determining heart disease. Furthermore, it could be used to simplify the model by not considering insignificant features. The decision tree model provides the importance factor for each feature, where the sum of these factors is one over the entire feature set. The features with the highest feature importance factor are 'the slope of the peak exercise ST segment', 'number of major vessels, and 'age'. The feature importance factor was significantly smaller for other features, suggesting that training a model with only these three features would suffice.

4.5 Neural Network

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature [1].

Several neural networks with different architectures were trained to fit the data. Both ReLU and Tanh activation functions were tested to fit the database. The performance of the Tanh and ReLU activation functions was similar. The Neural Network was trained using Adam optimizer on 500 epochs and accuracy criteria as the metric. Figure 11 shows the architecture of the Neural Network used in this study. Furthermore, several dropout layers were added to the architecture. However, since the accuracy of the validation and training set is almost equal, dropout layers were unable to enhance the model's accuracy (there was no overfitting existent in the model). Also, training for longer epochs seems unnecessary as it didn't improve the accuracy as shown in Figure 12 (the loss function is almost zero on both the training and test set). Finally, the results were summarized in Table 2. This table only contains architecture with one and two hidden layers as increasing the number of hidden layers didn't improve the results. The best structure was able to achieve an accuracy score of 100% on the test set.

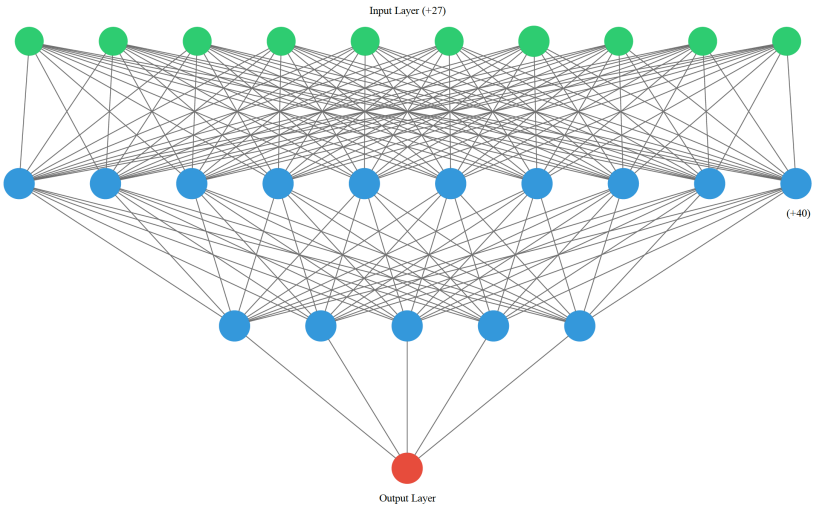


Figure 11: Architecture of the Neural Network applied in this study.

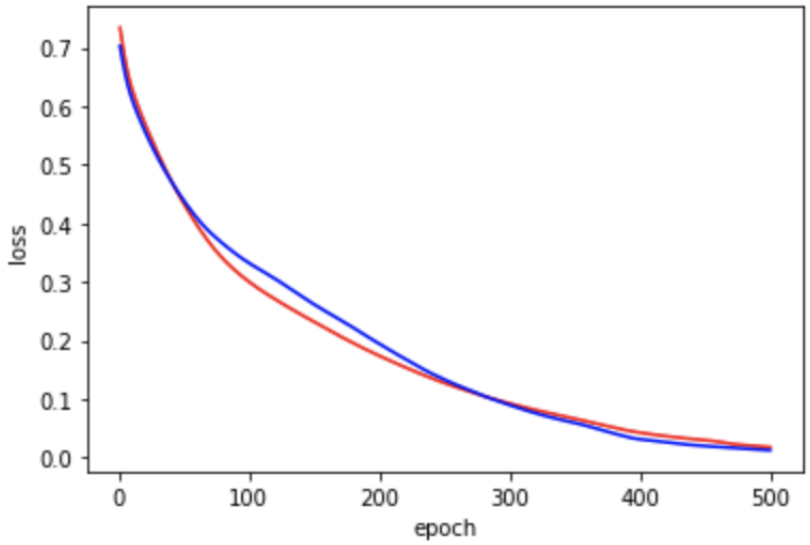


Figure 12: Binary cross-entropy loss of the Neural Network on both training (red) and validation sets (blue).

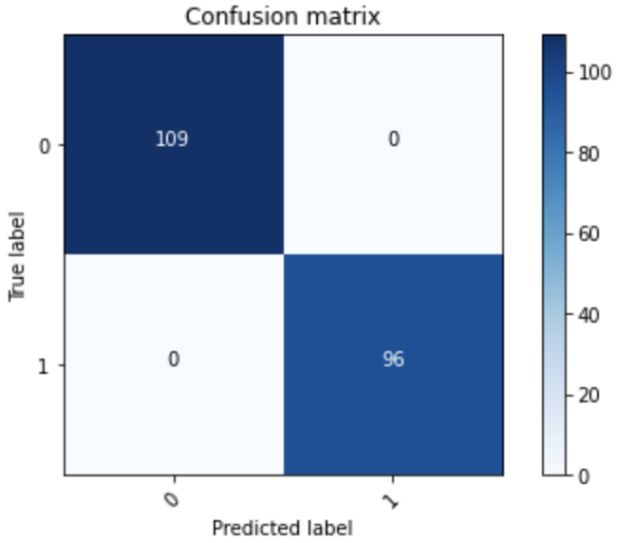


Figure 13: Confusion matrix for the NN classifier.

No	Number of layers	Activation	accuracy
1	$10 \times 10 \times 1$	Relu	88.78%
2	$20 \times 10 \times 1$	Relu	97.07%
3	$50 \times 10 \times 1$	Relu	100.00 %
4	$100 \times 10 \times 1$	Relu	100.00 %
5	$10 \times 10 \times 1$	Tanh	92.20%
6	$20 \times 10 \times 1$	Tanh	95.12%
7	$50 \times 10 \times 1$	Tanh	98.54%
8	$100 \times 10 \times 1$	Tanh	100.00 %

Table 2: Architecture of the Neural Networks fitted to the data.

4.6 Random Forest

The basis of a random forest model is the decision tree model. Random forests rely on a large grouping of small, varied decision trees to make a strong prediction. These weak decision trees are designed to have very little correlation with one another thus are individually worthless. When working with a large enough number a robust model is formed using overall trends. It is important to make sure that the trees are distinct from one another, without variance random incorrect trends will prevail. This is why the random forest name aptly identifies this model.

A random forest model is created by randomly sampling from the training data to create weak decision trees. These low level decision trees are formed using the same principles as described in section 4.4. This method of random sampling is commonly referred to as bagging. Since decision tree structures change greatly based on the features with the largest differences generating trees with different data samples ensures a low correlation between individual trees. The random forest model is often more efficient than pure decision trees for large data sets while generally no more complicated to implement when using libraries like Sci-Kitlearn. With little optimisation the random forest method was 96 percent accurate on the testing data set.

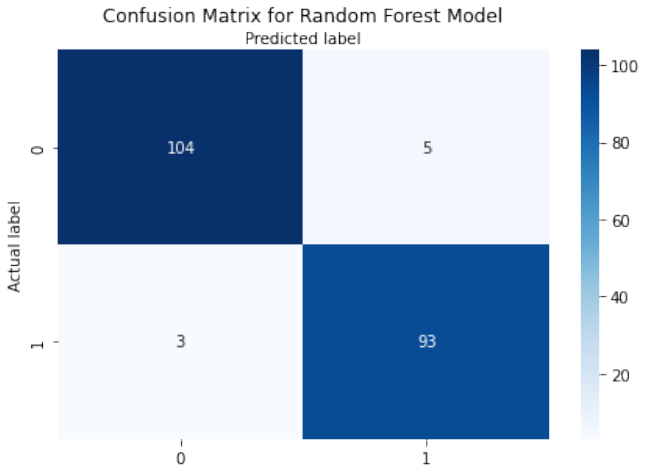


Figure 14: Confusion matrix for the RF classifier.

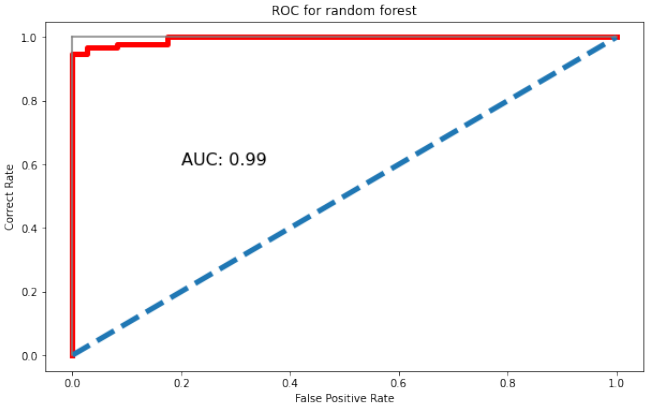


Figure 15: Receiver operating characteristic, ROC, for the RF classifier.

4.7 Gradient Boosting

Gradient boosting not to be confused with other methods like Ada boosting is a composite of many weak learning models. Gradient boosting learns by continuously minimizing the loss function every iteration by adding further small learners like decision trees to build on the weakest points of the best previous model using a modified data set. Each round the observations with the lowest scores have their weight increased for the next round. Converting weak learners to strong learners by adding trees. The gradient boosting method performed well, 99 percent accuracy on the testing dataset. Increasing the size of the weak learning trees and the number of iterations could produce an even more accurate model at the cost of speed.

Gradient boosting can be tuned using tree constraints to prevent the model from overgrowth. Setting the parameters like number of trees, tree depth, tree structure limits, and mini um improvement to loss can expedite gradient boosting. The weight updates are another area to customize the model's performance.

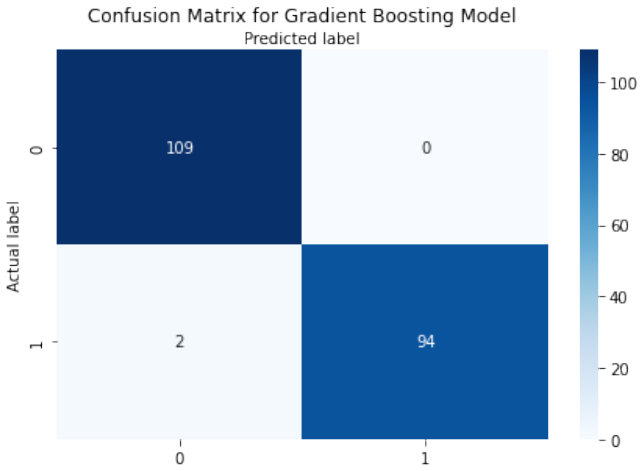


Figure 16: Confusion matrix for the RF classifier.

5 Conclusion

No	model	accuracy
1	Ridge	82.94%
2	Random Forest	96.10%
3	Gradient Boost	99.02%
4	KNN	100.00 %
5	SVM	100.00 %
6	Decision Tree	100.00 %
7	Neural Network	100.00 %

Table 3: Summary of the machine learning models and their accuracy.

The life saving value of simply applied machine learning is immense. Basic machine learning models using data already collected for many patients can reliably predict heart disease using only thirteen features. The application of the models created costs nearly nothing to run on a patient's existing data and can provide a recommendation for further tests or preventive treatment at a high confidence level.

The results of the seven models run in this study show the strong capabilities of current machine learning technology to form predictions based on datasets. Only the Ridge classification model under-performed. While many of the models relied on similar principals like decision trees their implementations and resources required vary greatly. A neural network model is generally far more complicated to implement than a random forest model with limited advantages on these simple applications. Methods like gradient boosting and random forest models utilize many small weak decision trees to build a strong model. The

largest flaw of this study is the ease at which these methods made models with near perfect accuracy. The extremely high results make it difficult to discuss each model's individual strengths based on the results alone because they hardly differ. When working with exponentially more complex data sets, more advanced models like support vector machines will outperform regular decision trees as available computing power becomes a limiting factor.

6 Future Work

There are several other models that could be used to fit the heart disease dataset, for instance, XGboost. Furthermore, there has been a lot of improvement in the structure of Neural Networks recently. The authors suggest the application of these novel architectures like Cascade feed-forward neural networks, Multi-layer perceptron neural networks, Radial basis neural networks, and Adaptive neuro-fuzzy inference systems.

Acknowledgement

The authors are grateful to Dr. Houlong Zhuang for fruitful discussions.

References

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[2] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.