Data Science: Machine Learning

Samantha Scott[a]

[a]*Stellenbosch University, Cape Town, South Africa*

Keywords: Machine Learning, Heart Disease Prediction, Random Forests

Email address: 20945043@sun.ac.za (Samantha Scott)

Table of Contents

1. Introduction

As Boehmke and Greenwell (2020) state, the machine learning process is very iterative and heuristic-based. It is difficult to establish which machine learning method would perform best as a predictive tool, with minimal knowledge of the problem or data at hand. This dilemma is known as the no free lunch theorem. The aim of this paper is to predict heart disease amongst patients, using the performing predictive tool.

The investigation consists of a comparison between two machine learning algorithms, namely Random Forests (RFs) and Support Vector Machines (SVMs), as prediction tools. In order to assess the predictive powers of the aforementioned methods, data on possible heart disease patients is used. The process of using machine learning algorithms to predict heart disease amongst patients has been broadly explored and this investigation serves to contribute to this body of literature. Ultimately, the results of this paper indicate that both RFs as well as SVMs showcase strong predictive powers, with a classification accuracy of above 70%. Further, the models do not gain predictive power once being fine tuned, indicating that the basic model is sufficient enough, given the data and problem.

The paper is structured as follows: section 2 provides an overview of the research question as well the data used to solve the problem at hand, section 3 provides information on the the methodology followed when generating the RF and SVM models, section 4 presents and discusses the results of the models, section 5 highlights the robustness checks implemented to verify the outcomes of the models, and lastly, section 6 contains the concluding remarks of the paper.

2. Research Question and Data

The main goal of the machine learning process is to find an algorithm that most accurately predicts future values based on a set of features. This paper aims to establish which machine learning algorithm performs best when predicting whether a patient has heart disease, or not. The problem at hand is expressed as a supervised binomial classification problem. The data used in this investigation is heart disease data from Kaggle and contains 180 observations. Kaggle is an online platform where data scientists and machine learning practitioners can access datasets as well as build portfolios. The benefit of obtaining data from this source is the usability score assigned to the dataset. This particular dataset is assigned a higher usability score.

The data used contains a number of health related indicators. For this investigation, the dependent variable is the presence of heart disease in a patient. This variable is a binomial, yes or no, variable. Using other indicators, such as age, whether a patient has a blood disorder called thalassemia as well as the type of chest pain a patient is experiencing, models are generated to predict if the patient has heart disease.

3. Methodology

For this investigation, code written in R is used. From this code, two machine learning algorithms are applied to the data. The first, a RF algorithm and the second, a SVM algorithm. To conduct these methods, the data is split into training and testing data. The ratio used is 70:30, respectively. This means that 70% of the data is used for training and the remaining 30% is for testing the model. The data is split using base R and the simple random sample method. This method is used as the responses do not vary much, with a ratio of 5.6 to 4.4. The training data is used to develop feature sets, train the selected algorithms, tune hyper parameters, compare models etc. The test data is used to estimate an unbiased assessment of the model's performance. For this paper, the classification accuracy between the models is compared. The classification accuracy measures the number of correct predictions made, divided by the total number of predictions made. This evaluation method is most popular for classification problems. The classification accuracy is presented in the form of a confusion matrix. A confusion matrix is a visual depiction of the correctly predicted responses, as well as the incorrectly predicted responses.

3.1. Random Forests

The RF algorithm is a modification of the bagged decision trees, which express better predictive performance. Using the randomForest package, a RF model (RFM 1) is generated using the training data. Using the testing data, a confusion matrix is presented. Although a RF model performs well, there are hyper parameters that may be implemented when tuning the model. To do this, the number of trees are adjusted, and the best m_{try} value is applied to the model. The number of trees needs to be large to stabilise the error rate. According to Boehmke and Greenwell (2020), it is suggested that the model starts with 10 times the number of features and in this case, there are 13 features. Starting with 10 times the number of features typically ensures the error estimate converges. However, once other hyper parameters are adjusted, more or less trees may be required. For the second RF model (RFM 2), the number of trees is set to the default, and performs slightly better. The next hyper parameter considered is m_{try}. Once the best m_{try} is computed, the value is inserted into the model. By doing this, the classification accuracy of the model is slightly improved, which is presented by the confusion matrix. Alternative methods of evaluating the models performance to the classification accuracy is also consulted, namely ROC curves as well as variable's importance. As a robustness check, the RMSE (Root Mean Square Error) scores are calculated for the RF models, and are compared to the classification accuracy.

3.2. Support Vector Machine

Using the Caret package, a SVM model (SVM 1) is generated. SVMs provide a direct approach to binary classification. A SVM is a supervised machine learning algorithm that is used to classify

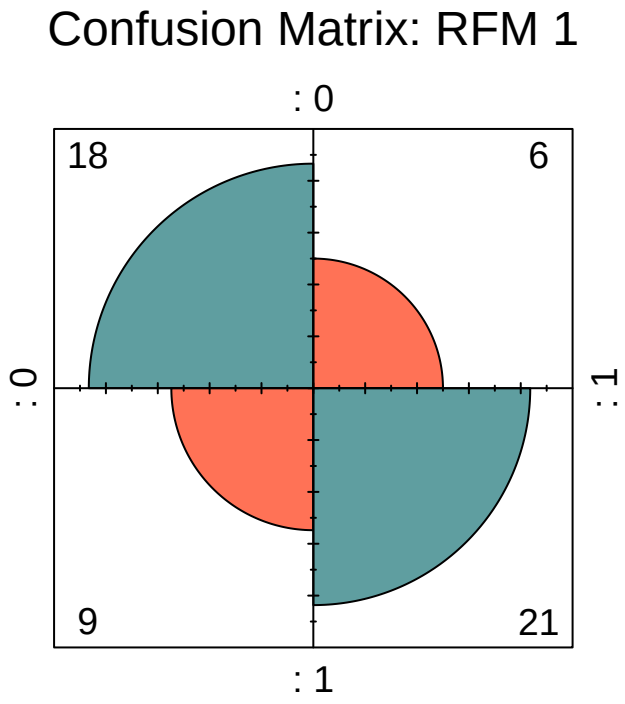
data into different classes. This approach makes use of a hyperplane - which is a decision boundary between the classes. Before training the data, the traincontrol() is implemented. This allows the train() function to be used under the caret package. The list returned by the traincontrol() method is placed into the train() method. A confusion matrix is presented, depicting the accuracy of this base model. Once this is done, the model is fine tuned by assigning values to the penalty parameter of the error term (C). This is the degree of correct classification that the algorithm has to meet. Once again, the results are presented in the form of a confusion matrix.

4. Results and Discussion

The next section of the paper presents the results of the data manipulation and methods used. The results are presented as plots and diagrams to better illustrate the outcomes. The classification accuracies as well as the RMSE scores are found in the Appendix.

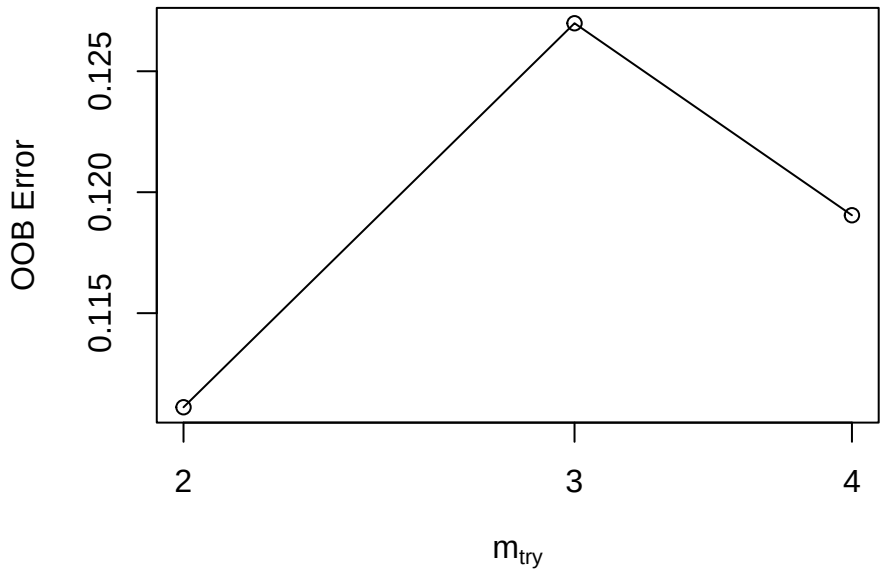
4.1. Random Forests

The figure below is a diagram of the confusion matrix for RFM 1. The graph depicts that a notable percentage of the "yes" and "no" (1 and 0, respectively) values are correctly predicted.

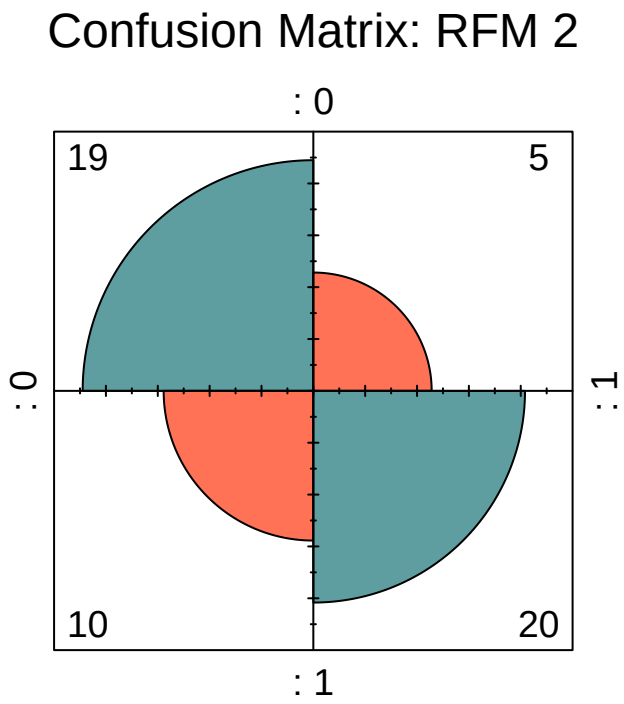


Below is a graphical representation of the optimal number assigned to m_{try}. As this is a machine learning algorithm, each time the code is run, a different optimal m_{try} may be revealed.

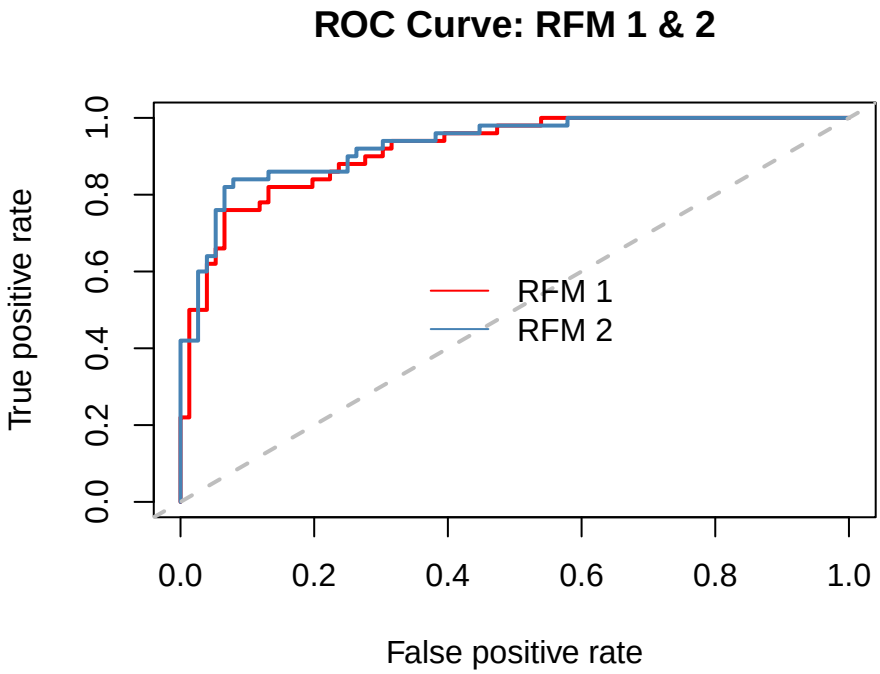
```
## mtry = 3  OOB error = 12.7%
## Searching left ...
## mtry = 2     OOB error = 11.11%
## 0.125 0.01
## Searching right ...
## mtry = 4     OOB error = 11.9%
## -0.07142857 0.01
```



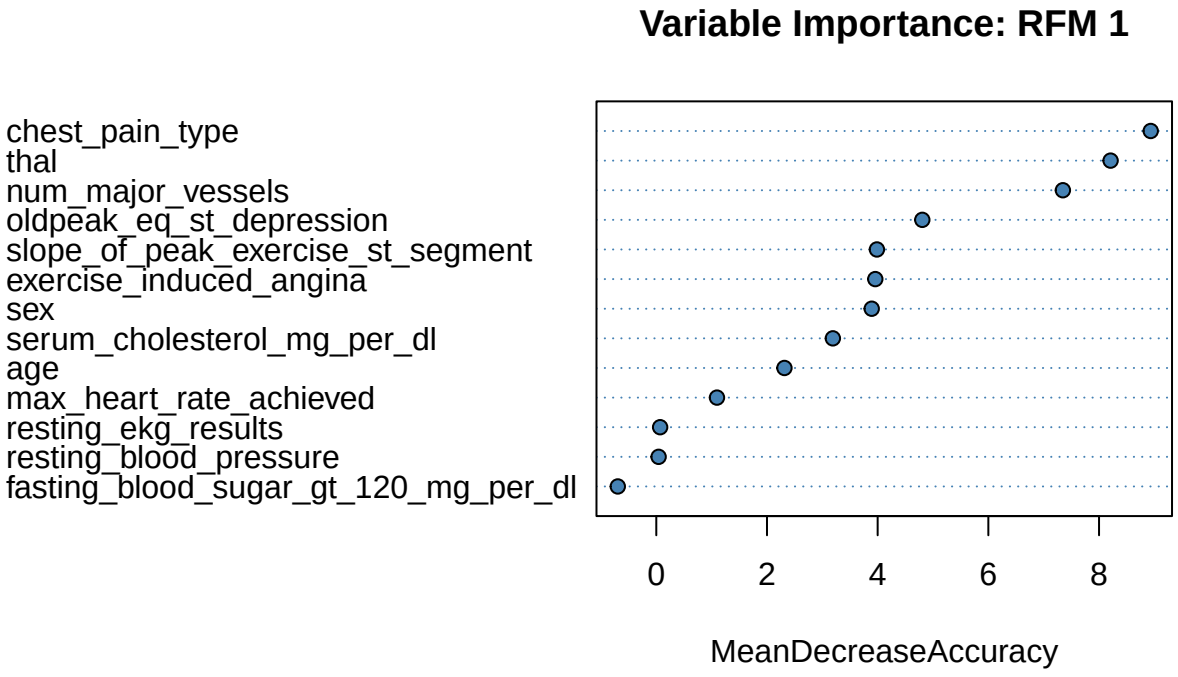
The confusion matrix of RFM 2 is depicted below. As shown in the diagram, RFM 2 performs equally as well as RFM 1. The hyper parameters did not impact the predictive power of the model.

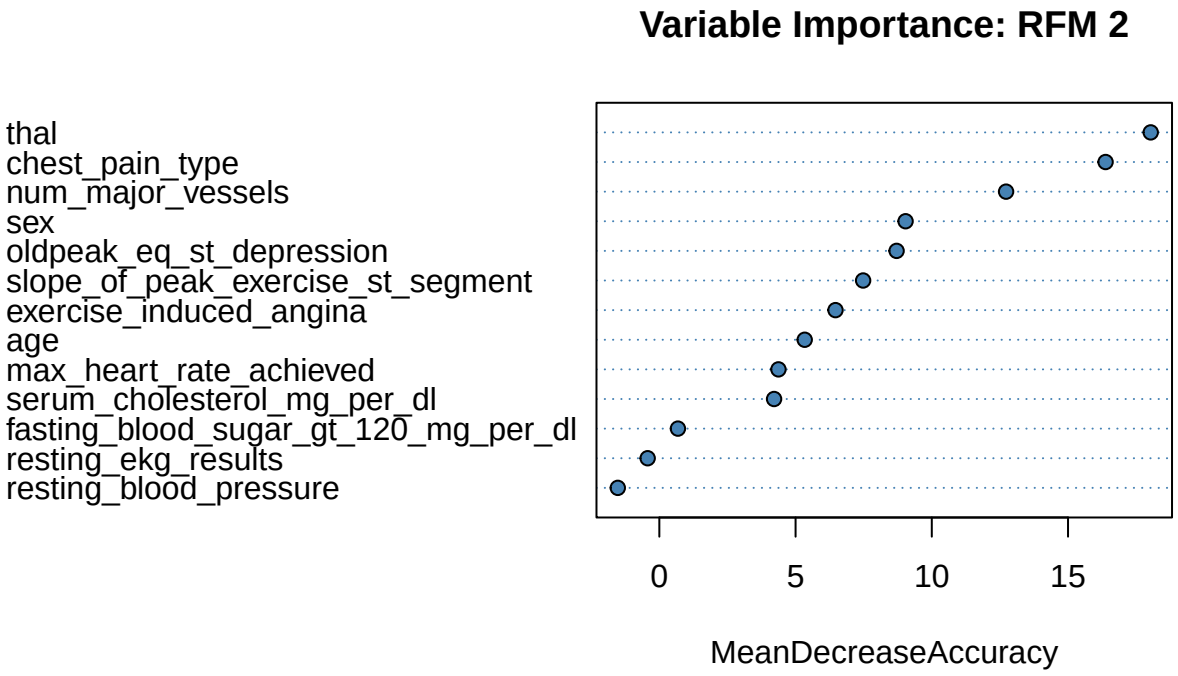


The ROC curves for the two RF models are presented below. Another way to assess the strength of the RF models is to compare the area under the ROC curves for each model. As the graphs depict, the area is similar, indicating similar predictive power between RFM 1 and RFM 2.

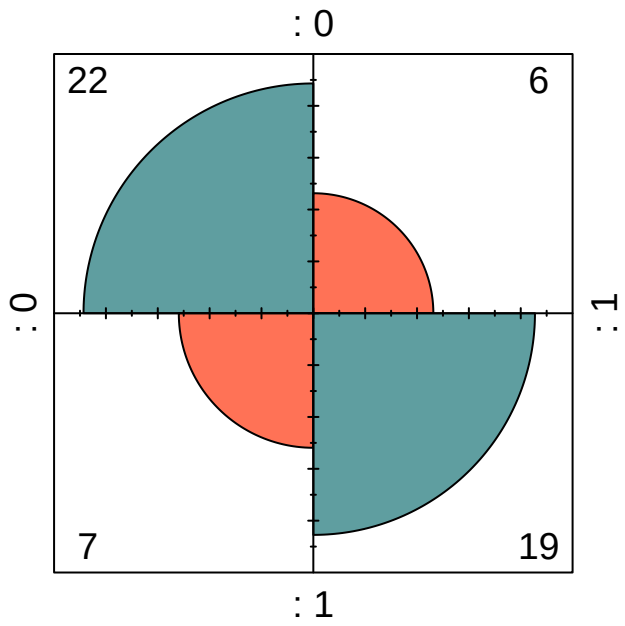


The variable importance of the two RF models are presented below. The mean decrease accuracy provides an estimate of the loss in prediction performance when that variable is omitted from the training dataset. In the variable importance graph for RFM 1, the variable that indicates if a patient has thalassemia is seen as the most important variable. For RFM 2, the variable indicating the chest pain type is the most important variable.



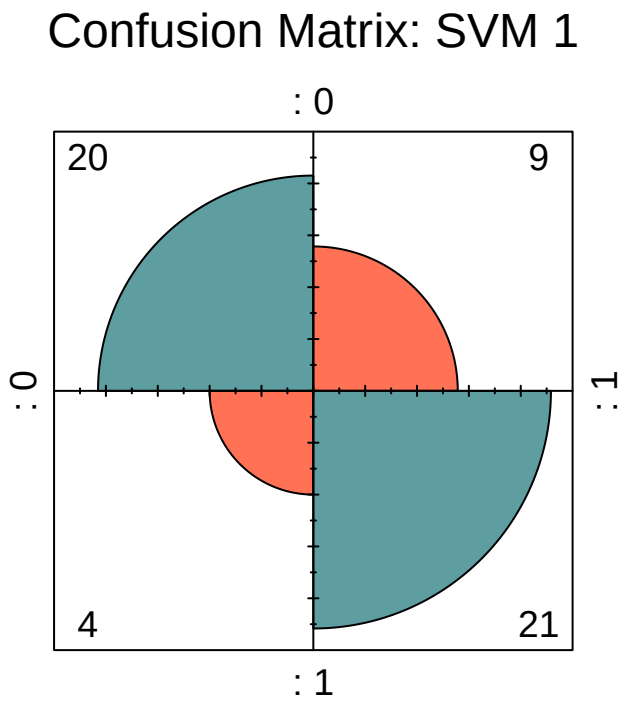


Below is the confusion matrix for a third RF model (RFM 3). In this model, the bottom three variables of the variable importance plots above are omitted. As shown in the confusion matrix, this does slightly impact the predictive power of the model. RFM 3 has the highest classification accuracy, however, the model also has the highest RMSE score. The RMSE score is below 0.5, showcasing that the model is still a somewhat accurate predictor.

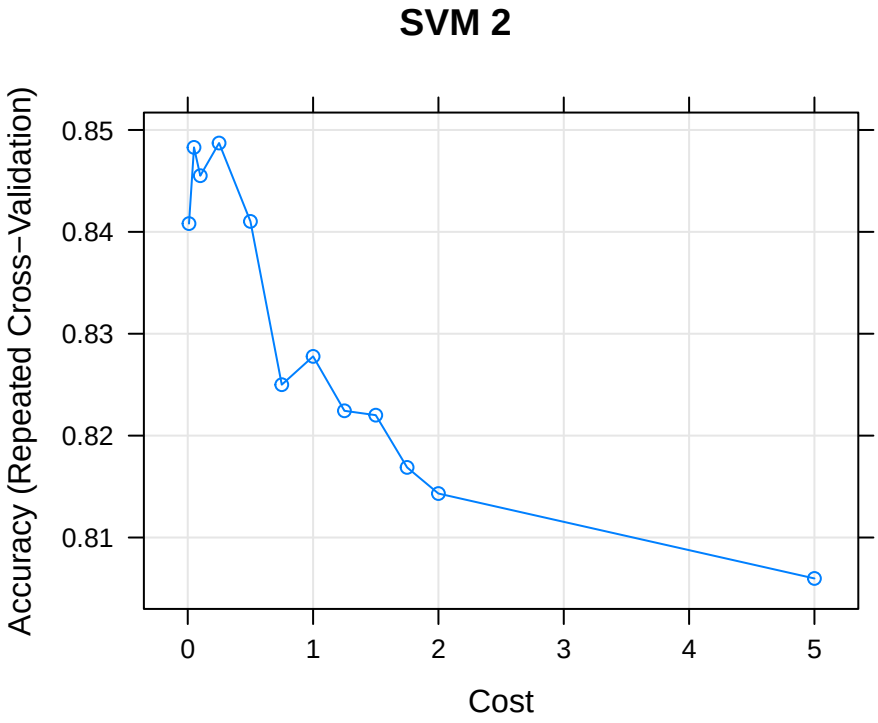


4.2. Support Vector Machine

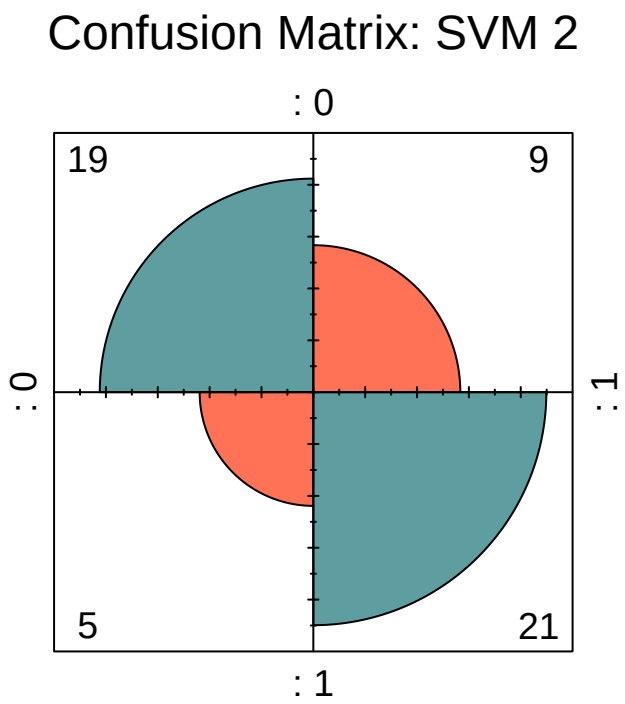
Below is the confusion matrix for the first SVM model, SVM 1. As the figure depicts, the majority of the responses are correctly predicted.



The graph below depicts the changes made to the C parameter of the SVM model. Once this is applied to the model, there is no significant change in the classification accuracy of the model.



The figure below is the confusion matrix for the second SVM model, SVM 2. As the graph depicts, the model has predicted a majority of the responses correctly.



As the results of the investigation establish, the two machine learning algorithms (RFs and SVMs) are both adequate predictive tools when attempting to depict heart disease in patients in their base form as well as after fine tuning. The classification accuracies of the algorithms remain similar to one another before and after the addition of hyper parameters.

5. Robustness Checks

Due to the similar results (namely the classification accuracy) presented by the models, an attempt to assess the validity of the models is implemented. To do this, a robustness check in which the dataset is altered, is conducted. The binomial dependent variable in the data is manually altered, whereby 56 values are changed. Once the changes are made, the code is run. The outcomes show a classification accuracy of around 50% for both the RF and SVM models. This indicates that even with a dataset with lower predictability, the models perform on a similar level. As another robustness check, an alternative package for RFs is used. In line with Boehmke & Greenwell (2020), the ranger package is applied. The results indicate that the RF models have RMSE scores of between 3 and 4, with the latter being the lowest.

6. Conclusion

In conclusion, this paper reveals that the SVM algorithm slightly outperforms the RF algorithm when predicting the presence of heart disease amongst patients. After the addition of hyper parameters for both algorithms, the classification accuracies are not significantly altered. As such, this indicates that the base models and the fine tuned models possess similar predictive powers. It is important to note that both algorithms, base as well as fine tuned of each, provide an acceptable classification accuracy of above 70%. After conducting robustness checks, it is established that the outcomes of the algorithms are reliable.

7. Reference List

Boehmke, B. & Greenwell, B. 2020. *Hands-On Machine Learning with R.* CRC Press

8. Appendix

8.1. Classification Accuracies

8.1.1. RFM 1:

```
## [1] 0.7222222
```

8.1.2. RFM 2:

```
## [1] 0.7222222
```

8.1.3. RFM 3

```
## [1] 0.7592593
```

8.1.4. SVM 1:

```
## [1] 0.7592593
```

8.1.5. SVM 2:

```
## [1] 0.7407407
```

8.2. RFM 1, 2 & 3: Ranger Results

```
## [1] 0.3563483
```

```
## [1] 0.3779645
```

```
## [1] 0.4454354
```