Performance of Parametric and Non-parametric Models on Heart Disease Prediction

Rachel Flodin and Drew Wise

June 10, 2022

Contents

Abstract

Predicting whether a patient has heart disease is a vital task with important implications for the medical community and public health. We compare the performance of several classification models, namely support vector machine, random forests, bagging, boosting, clustering with logistic regression, and logistic regression, on a heart disease data set from the UC Irvine Machine Learning Repository. Most of the models have prediction accuracy around 80%, with the cluster analysis performing the best of all. We finish with a discussion of variables that have a significant impact on the likelihood of heart disease. In particular, each model finds chest pain (`cp`) and "ST depression induced by exercise relative to rest" (`oldpeak`) highly important in predicting heart disease.

1 Introduction

Heart disease is the leading cause of death for men and women in the U.S., and the second leading cause of death in Washington state.[2,16] There are an estimated 659,000 deaths occurring each year in the U.S., and an estimated cost of $363 billion per year from 2016-2017, which includes costs related to health care (including medications) and lost productivity.[2] Heart disease includes many different heart conditions, however we will use the criteria which classifies heart disease as whether there is a narrowing of more than 50% in any major vessel in the heart.

There are many risk factors for heart disease, including age, sex, cholesterol level, blood pressure, and diabetes.[4] In 2016, an estimated 15% of adults in Washington state were 65 or older, and it is expected that this proportion will grow to 22% by 2030.[16] Since heart disease death rates are highest in this age group, being able to accurately predict whether someone may have heart disease from known factors if they have not yet received that diagnosis will help provide better estimates of the population with heart disease, which will aid in planning for future health care needs and costs.[16]

In addition, being able to provide a quicker diagnosis, and thus quicker access to treatment, could lead to a better quality of life for the patient. A heart attack occurs when blood flow to the heart is blocked; the longer the blockage occurs, the greater the damage that can occur to the heart.[3] A major symptom of heart attack is chest pain, however many women do not experience the same symptoms of heart disease as men.[3] Being able to discover what other symptoms/clinical findings, which may not be major symptoms/clinical findings, may help with diagnosis when a patient's heart vessel blockage is showing atypical presentation.

In order to find the best model for predicting heart disease from various risk factors and clinical presentation/findings, we are comparing the prediction accuracy of six different models using the Heart Disease data set from the UC Irvine Machine Learning Repository.[6] Four models are built using the following non-

parametric methods : support vector machine (SVM), random forests, bagging, and boosting. One model combines the non-parametric method of cluster analysis to pre-process the data, before then fitting each cluster with separate (parametric) logistic regression models. Our final model, uses the parametric method of logistic regression for comparison, and to see if any of the non-parametric methods outperform this parametric method.

2 Methods

2.1 Data

2.1.1 Data Background

The Heart Disease data sets were provided to the UCI Machine Learning Repository by Andras Janosi, M.D. at the Hungarian Institute of Cardiology in Budapest, William Steinbrunn, M.D. at the University Hospital in Zurich, Switzerland, Matthias Pfisterer, M.D. at the University Hospital in Basel, Switzerland, and Robert Detrano, M.D., Ph.D. at the V.A. Medical Center in Long Beach and the Cleveland Clinic Foundation.[6] The files are made available for download at https://archive.ics.uci.edu/ml/datasets/heart+disease.

The Cleveland data file contains 303 rows, the Hungarian data file contains 294 rows, the Switzerland data file contains 123 rows, and the Long Beach VA data file contains 200 rows.[6] Each file has also been modified by UCI Machine Learning Repository to only contain 14 variables, out of the 76 original variables provided.[6] We combined all four files into a single file with 920 rows.

The variable we are attempting to predict is the `num` variable, and it contains integers from 0 (indicating no presence of heart disease) to 4. Values 1, 2, 3, and 4 all indicate varying presence of heart disease, however the exact details on these four levels were not given. Because of this we re-coded `num` to values 0 indicates no presence of heart disease (defined as $< 50\%$ diameter narrowing of any major vessel in the heart), and 1 indicates presence of heart disease (defined as $> 50\%$ diameter narrowing of any major vessel in the heart).

We used the following independent variables to predict `num`.

- `age` (numeric): patient's age in years.
- `sex` (binary): patient's sex, coded as 1 for male and 0 for female.
- `cp` (categorical): presence or absence of chest pain in patient, coded as 1 for "typical angina," 2 for "atypical angina," 3 for "non-anginal pain," and 4 for "asymptomatic."
- `trestbps` (numeric): patient's resting blood pressure, measured in mm Hg (millimeters of mercury) on admission to the hospital.

- `chol` (numeric): patient's serum cholesterol, measured in mg/dl.
- `fbs` (binary): patient's fasting blood pressure, coded as 1 for greater than 120 mg/dl, and 0 for less than 120 mg/dl.
- `restecg` (categorical): patient's resting electrocardiographic result, coded as 0 for "normal," 1 for "ST-T wave abnormality," and 2 for "showing probable or definite left ventricular hypertrophy by Estes' criteria."
- `thalach` (numeric): patient's maximum heart rate achieved during exercise test.
- `exang` (binary): exercise induced angina, coded as 1 for present and 0 for absent.
- `oldpeak` (numeric): measure of patient's "ST depression induced by exercise relative to rest."

2.1.2 Data Cleaning

The data set contained missing values encoded as "?", which we re-coded to NA. We found that the variables `slope`, `ca`, and `thal` had over 30% missing values (33.6%, 66.4%, and 52.8%, respectively), which was much higher than any of the other variables. We decided to remove these three variables from the analysis because of this. All other variables had less than 10% missing (see Table 1), with no missing variables in `age`, `sex`, `cp`, and `num`.

	Count	Percent
trestbps	59	6.4
chol	30	3.3
fbs	90	9.8
restecg	2	0.2
thalach	55	6.0
exang	55	6.0
oldpeak	62	6.7

Table 1: Missing values in predictors.

We also found that the `chol` and `trestbps` contained values of 0 (172 and 1 0 values, respectively), which is not a possible clinical value for cholesterol or blood pressure (unless the patient was no longer living). In fact, the `chol` column for the Switzerland data set only contained 0 values. We treated these 0 values as NAs in the data.

Because there are so many missing values in the data, even after removing the `slope`, `ca`, and `thal` columns, we decided to do a sensitivity analysis on the data by creating two data sets. For the low likelihood (LL) data set we imputed all missing values with the minimum value for each variable (i.e. the value that would be least likely to imply heart disease, see Table 2), and for the high likelihood (HL) data set we imputed all missing values with the maximum value for each variable (i.e. the value that would be most likely to imply heart disease, see Table 2). The two imputed data sets each contain 11 columns and 920 rows.

	Low Likelihood	High Likelihood
trestbps	80	200
chol	85	603
fbs	0	1
restecg	0	2
thalach	202	60
exang	0	1
oldpeak	-2.6	6.2

Table 2: Imputed values for sensitivity analysis.

We also created a third data set which contains only the complete cases. The complete case analysis data set contains 11 columns and 661 rows. (Note that because the `chol` column in the Switzerland data set contained all 0 values, which were treated as NAs, the complete case analysis no longer contains the Switzerland data.)

2.2 Models

2.2.1 Tuning Parameters

We tuned each of our non-parametric models using the complete case data set. Unless specified otherwise below, we used 10-fold cross validation with the metric of prediction accuracy to select model parameters. That is, given n test data points with response values $y_1, y_2, \ldots, y_n$ and predicted values $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$, we sought the parameters that maximize the sum

$$\frac{1}{n} \sum_{i=1}^{n} I\left(y_i = \hat{y}_i\right).$$

In the SVM model, we compared the performance of linear, polynomial, radial, and sigmoid kernels,

namely

$$K_{\text{linear}}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T\mathbf{v},$$

$$K_{\text{polynomial}}(\mathbf{u}, \mathbf{v}) = \left(\gamma \cdot \mathbf{u}^T\mathbf{v} + c_0\right)^d,$$

$$K_{\text{radial}}(\mathbf{u}, \mathbf{v}) = \exp\left(-\gamma \cdot \|\mathbf{u} - \mathbf{v}\|_2^2\right),$$

and

$$K_{\text{sigmoid}}(\mathbf{u}, \mathbf{v}) = \tanh\left(\gamma \cdot \mathbf{u}^T\mathbf{v} + c_0\right),$$

where $\mathbf{u}$ and $\mathbf{v}$ are vectors. We used the default value for polynomial degree, $d = 3$, and we further tuned the parameters γ and c_0 by checking the performance of each combination of $\gamma \in \{0.0001, 0.005, 10, 20, 50, 100\}$ and $c_0 \in \{0.1, 1, 50, 100, 150, 500\}$. We fitted our models using the function `svm` in the R package `e1071`[13].

We considered the random forest model and the bagging model together. We fitted our models using the function `randomForest` from the `randomForest`[11] R package. We tuned the following parameters: B, the number of trees used to fit the model; m, the number of predictors sampled for splitting at each iteration; and n, the minimum size of terminal nodes. We considered all combinations of $B \in \{50, 500, 1000, 2000, 5000, 10000\}$, $m \in \{2, 3, 5, 10\}$, and $n \in \{1, 5, 10, 25, 50, 75, 100\}$. Note that $m = 10$ corresponds to bagging. We used out-of-bag (OOB) error as our selection criterion in this case.

For boosting, we similarly tuned B and n, variables with the same interpretation as above. We also considered different values of λ, a shrinkage parameter that controls model complexity, and d, the maximum interaction depth of each tree. More specifically, we considered all combinations of $B \in \{50, 500, 1000, 2000, 5000, 10000\}, n \in \{1, 5, 10, 25, 50, 75, 100\}, \lambda \in \{0.001, 0.01\}$, and $d \in \{1, 2\}$. We used the Bernoulli loss function and fitted our models using the function `gbm` in the R package `gbm`[7].

Lastly, in the case of cluster analysis we tuned K, the number of clusters. We considered $K \in \{2, 3, \ldots, 110\}$. Instead of cross-validation, we fitted each model using 70% of our complete case data and tested its performance on the remaining 30%. We found issues fitting logistic regression models to small clusters, which motivated our decision to avoid 10-fold cross validation in this case. We clustered the data with the function `kmeans` from the R package `stats`[14].

2.2.2 Variable Importance

Although our primary objective is obtaining a model with high prediction accuracy, we are also interested in learning how the predictor variables affect the likelihood of heart disease. To this end, we turn our attention to how each model measures variable importance.

In the case of the SVM model, we used the function `Importance` in the R library `rminer`[5] with the

parameter `measure` = variance. This function uses a form of sensitivity analysis to measure how the output of the SVM changes as a given variable ranges through its values. High variance indicates that a variable has a strong impact on the model output, so large values correspond to high variable importance.

For the random forest and bagging models, we used the function `importance` from the R package `randomForest`[11]. This function calculates the mean decrease Gini for each variable, that is, the average amount that

$$\sum_{k=1}^{2} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

decreases by splits over the given variable, where $\hat{p}_{mk}$ is the proportion of data points in the mth region that are from the kth class. A large mean decrease Gini corresponds to high variable importance.

For boosting, we measured variable importance with the `relative.influence` function from the R package `gbm`[7]. The approximate relative influence of a variable x_j is given by

$$\hat{J}_j^2 = \sum_{\text{splits on } x_j} I_t^2,$$

with I_t^2 being the empirical improvement by splitting at x_j. Here again, large values correspond to high variable importance.

Lastly, for logistic regression and logistic regression with clustering, we measured variable importance by considering the p-value and coefficient size associated with any given variable.

3 Results

3.1 Tuned Parameters

Table 3 shows the tuned parameters for each model under consideration. When we saw very similar results between a simpler model and a more complex model within each method tested, we chose the more parsimonious model to avoid overfitting to the training data.

Since we found $m = 2$ is the ideal parameter for the random forest model, we will no longer consider bagging.

Model	Tuned Parameters
SVM	radial kernel, $\gamma = 0.005$
Random Forest/Bagging	$B = 500, m = 2, n = 25$
Boosting	$B = 5000, n = 25, d = 1, \lambda = 0.01$
Clustering	$K = 2$

Table 3: Tuned parameters for each model.

3.2 Sensitivity Analysis Results

After tuning the models on the complete case data set, we tested each model on the two sensitivity analysis data sets. We used 10-fold cross validation to find the prediction accuracy, except for the logistic regression with clustering model, where we again used 70% of the data for training and 30% for testing. Table 4 displays the prediction accuracy of each tuned model on the high likelihood, low likelihood , and complete case data sets. Most models have accuracy around 80%, with slightly higher accuracy on the high likelihood data set than the low likelihood data set. The model that mixes cluster analysis and logistic regression performs best of all.

	Prediction Accuracy		
Model	High Likelihood	Low Likelihood	Complete Case
SVM	0.7967	0.7891	0.8093
Random Forest	0.812	0.7978	0.8094
Boosting	0.8043	0.8076	0.7987
Clustering	0.844	0.8431	0.7920
Logistic Regression	0.8022	0.7902	0.80801

Table 4: Prediction accuracy on high likelihood, low likelihood, and complete case data sets.

3.3 Variable Importance

We now present the measures of variable importance for the tuned models, fit to each of the complete case, high likelihood, and low likelihood data sets. The measures are as described in section **2.2.2**.

Table 5 shows the results for the SVM. On this basis, we consider `age`, `sex`, `cp`, `chol`, `thalach`, and `oldpeak` significant predictors of heart disease.

	Importance		
Variable	High Likelihood	Low Likelihood	Complete Case
age	0	0.16	0.1439
sex	0	0.56	0.3022
cp	0.6364	0.28	0.1511
trestbps	0	0	0
chol	0	0	0.1439
fbs	0	0	0
restecg	0	0	0
thalach	0	0	0.1727
exang	0	0	0
oldpeak	0.3636	0	0

Table 5: Variable importance for tuned SVM models.

Table 6 gives the variable importance for the final random forest models. Here we select `cp`, `oldpeak`, `exang`, and `thalach` as the most important variables.

	Importance		
Variable	High Likelihood	Low Likelihood	Complete Case
age	21.2141	23.9069	15.1020
sex	15.1336	15.9496	10.7701
cp	64.6819	69.4140	44.5913
trestbps	11.0959	12.5375	9.1254
chol	25.2149	23.3848	9.1998
fbs	7.8867	2.7296	2.4468
restecg	4.3177	5.0476	3.8985
thalach	29.7034	26.7035	20.1556
exang	33.1729	32.0136	29.6817
oldpeak	32.0024	32.0300	34.2516

Table 6: Variable importance for tuned random forest models.

Table 7 shows the relative influence of the variables in the final boosting models. For boosting, we choose `cp`, `oldpeak`, `age`, `exang`, `thalach`, and `chol` as our most important predictors.

	Importance		
Variable	High Likelihood	Low Likelihood	Complete Case
`age`	486.5852	545.0636	457.8693
`sex`	297.1414	339.8249	238.9545
`cp`	1543.1952	1541.0238	987.2114
`trestbps`	357.5589	374.4725	344.3683
`chol`	711.8569	794.5817	586.7375
`fbs`	115.9943	72.0490	67.9122
`restecg`	85.2702	60.9432	80.6314
`thalach`	683.8838	584.7236	475.9462
`exang`	522.7746	516.2361	539.7517
`oldpeak`	609.8126	552.6678	708.2419

Table 7: Variable importance for tuned boosting models.

Table 8 lists the coefficient estimates under logistic regression models. The associated p-values are given in parentheses. Judging by these results, we would consider `cp`, `sex`, `oldpeak`, `exang`, and `fbs` the most important variables.

	Importance		
Variable	High Likelihood	Low Likelihood	Complete Case
`age`	0.0378 (0.0006)	0.0479 (< 0.0001)	0.0230 (0.0893)
`sex`=1	1.2346 (< 0.0001)	1.3642 (< 0.0001)	1.4872 (< 0.0001)
`cp`=2	−1.1742 (0.0057)	−1.1626 (0.0047)	−0.2726 (0.5982)
`cp`=3	−0.4532 (0.2415)	−0.3330 (0.3713)	0.0731 (0.8801)
`cp`=4	1.1910 (0.0015)	1.1873 (0.0010)	1.5701 (0.0009)
`trestbps`	−0.006 (0.1860)	−0.0082 (0.0787)	0.0090 (0.1733)
`chol`	0.0034 (< 0.0001)	−0.0029 (0.0097)	0.0042 (0.0293)
`fbs`=1	0.6829 (0.0031)	0.5717 (0.0262)	0.5339 (0.0893)
`restecg`=1	−0.1196 (0.6377)	0.2874 (0.2456)	0.1456 (0.6819)
`restecg`=2	0.1827 (0.4341)	−0.0034 (0.9884)	0.2537 (0.3327)
`thalach`	−0.0010 (0.8004)	−0.0042 (0.2443)	−0.0103 (0.0396)
`exang`=1	1.2340 (< 0.0001)	1.2194 (< 0.0001)	0.9988 (0.0001)
`oldpeak`	0.1459 (0.0423)	0.2552 (0.0013)	0.7233 (< 0.0001)

Table 8: Estimated coefficients (with associated p-values in parentheses) for logistic regression models. We are omitting the estimate for the intercept.

Lastly, table 9 lists the coefficient estimates under the clustering with logistic regression models. The results for the first cluster are listed directly above the results for the second cluster. The associated p-values are given in parentheses. Judging by these results, we would consider `cp`, `sex`, `exang`, and `oldpeak` the most important variables.

	Importance		
Variable	High Likelihood	Low Likelihood	Complete Case
age	0.0466 (0.0002)	0.0136 (0.5199)	0.0351 (0.0561)
	0.0108 (0.7046)	0.0605 (< 0.0001)	0.0017 (0.9398)
sex=1	1.3562 (< 0.0001)	0.6011 (0.3142)	1.6048 (< 0.0001)
	0.3958 (0.6108)	1.566 (< 0.0001)	1.3795 (0.0005)
cp=2	−0.6857 (0.1411)	−3.5127 (0.0021)	0.0149 (0.9841)
	−18.2188 (0.9853)	−0.6816 (0.1470)	−1.0090 (0.2055)
cp=3	−0.1477 (0.7328)	−1.8699 (0.0666)	0.8188 (0.2040)
	−16.2738 (0.9869)	−0.0462 (0.9159)	−1.040 (0.2260)
cp=4	1.4566 (0.0005)	−0.2408 (0.8073)	2.343 (0.0003)
	−14.6850 (0.9882)	1.4123 (0.0009)	0.2636 (0.7348)
trestbps	−0.0071 (0.2055)	−0.0187 (0.0451)	0.0107 (0.2288)
	−0.0022 (0.8355)	−0.0052 (0.3784)	0.0043 (0.6727)
chol	0.0048 (0.0193)	−0.0250 (< 0.0001)	0.0043 (0.3840)
	0.0095 (0.3418)	0.0047 (0.0163)	0.0021 (0.5460)
fbs=1	0.4248 (0.1216)	0.8135 (0.2760)	0.4658 (0.2600)
	1.4785 (0.0050)	0.6041 (0.0335)	0.4223 (0.4101)
restecg=1	−0.0272 (0.9282)	0.1958 (0.6699)	0.1648 (0.7189)
	−0.2752 (0.6028)	0.5396 (0.0954)	0.1187 (0.8351)
restecg=2	0.1670 (0.4960)	2.3356 (0.0201)	−0.0694 (0.8444)
	0.5875 (0.5628)	−0.2022 (0.4260)	0.7818 (0.0680)
thalach	0.0006 (0.8908)	−0.0155 (0.0362)	−0.0098 (0.1415)
	−0.0084 (0.3662)	0.0028 (0.5375)	−0.0114 (0.1571)
exang=1	1.1990 (< 0.0001)	1.6181 (0.0025)	1.0799 (0.0007)
	1.4067 (0.0142)	1.1698 (< 0.0001)	1.1040 (0.0093)
oldpeak	0.2265 (0.0065)	0.1895 (0.2059)	0.7614 (< 0.0001)
	−0.1449 (0.3725)	0.3755 (0.0004)	0.7362 (0.0006)

Table 9: Estimated coefficients (with associated p-values in parentheses) for clustering with logistic regression models. Results from the first cluster are listed above the results from the second cluster. We are omitting the estimate for the intercept.

4 Conclusions

The clustering model with logistic regression in the complete case analysis appeared to be very sensitive to the 70/30 split of the data, with different splits varying the prediction accuracy from around 78% to 84%. The complete case data set is approximately two-thirds the size of the high and low likelihood data sets. When the data is split into two clusters using the complete case data, one of the clusters only contains 250 observations, which leads to a test data set size of only 75. This may explain why we see the sensitivity to the split, and such a difference between the prediction accuracy in the complete case analysis compared to the high likelihood and low likelihood analysis.

Overall, the close prediction accuracy of the high likelihood, low likelihood, and complete case data sets was a good indication that even under the two different extreme assumptions about the bias due to the (mainly) missing cholesterol serum levels, the effect on the prediction accuracy was small.

Chest pain type (`cp`) is an important predictor variable for all of the models tested. In both the logistic regression model and the clustering with logistic regression model, we see that asymptomatic chest pain (`cp`= 4) has a very high association with heart disease. This is unexpected since typical angina chest pain is caused by a narrowing or blockage of a coronary artery.[8] However, around half of the observations had asymptomatic chest pain, with $75 - 80\%$ of the asymptomatic chest pain being diagnosed with heart disease for both the complete case and imputed data sets.

ST depression induced by exercise relative to rest (`oldpeak`) is another important predictor variable for all of the models tested. This is an expected result because ST depression is a known electrocardiographic (ECG) finding used to diagnose the narrowing of the arteries of the heart.[9]

Exercise induced angina (`exang`) was an important predictor variable in all of the models except SVM. This is also an expected finding. Since exercise increases demand for oxygen, if the patient has reduced blood flow to the heart caused by either narrow or blocked arteries in the heart, this can cause angina pain.[12]

Both sex and maximum heart rate at exercise (`thalach`) were important predictor variables in three models. Since the risk of heart disease changes for men and women depending on their age (possibly due to the amounts and effect differences of estrogen and testosterone) it makes sense that sex would be an important predictor of heart disease.[15] A lower maximum heart rate at exercise compared to the target rate of 220 minus age could be indicative of heart disease, however the association is unclear. It is possible that narrow/blocked arteries/vessels could cause a lower maximum heart rate at exercise.[10]

The serum cholesterol level was only an important variable in two of the models. This is surprising since high cholesterol (≥ 200) is a known risk factor for heart disease, and contributes to the narrowing of arteries/vessels. When cholesterol levels are high in the blood, it can build up in the walls of arteries, leading

to the narrowing of the arteries, and even to full blockages.[1] However, around 75% of the observations had readings above 200, which may be why we do not see this variable as an important predictor in more models.

It is also surprising that age was only an important predictor variable in two of the models, since age is a known risk factor for heart disease with risk increasing with age.[15] However, it is known that heart disease risk is much higher starting at age 65, and the complete case data and two imputed data sets only contain around $11 - 12\%$ of observations for people 65 and older, which may be contributing to this finding.

As mentioned above, the best performing model (assuming a larger imputed data set) was the combination of clustering with logistic regression. This model is a nice blend of both non-parametric and parametric methods. Being able to use the clustering to find similar observations, and then to build a logistic regression model on each set of similarly clustered observations worked very well, except when the cluster sizes dropped too low for logistic regression to be able to perform well. So, with large enough data, this method seems very promising for prediction.

It was somewhat surprising to see that logistic regression alone performed prediction as well or almost as well as all of the non-parametric methods. It does appear to be true that just because a method may have been around for a while, that does not mean that newer methods with cooler sounding names will outperform it.

4.1 Limitations

The fact that some of the more important predictor variables could have been important due to the proportion of values that also were associated with heart disease (i.e. that half of the observations had asymptomatic chest pain, but around three-quarters of those had heart disease) is a limitation of this analysis.

Since this data set only contained a binary variable for sex, it is also a limitation that we do not have any findings regarding non-binary or transgender individuals in order to build a good predictive model which works for all individuals.

4.2 Future Work and Extensions

In this analysis we only looked at the maximum heart rate achieved during exercise. An idea for further analysis would be to create a new variable which holds the value of the difference between the maximum heart rate during exercise and the individuals target heart rate, and to repeat the analysis using this new variable instead of the maximum heart rate variable. Additionally, since beta blockers can reduce an individual's maximum heart rate, having an indicator variable for beta blockers would also be beneficial to add to the new analysis.

It would also be interesting to reproduce this analysis using similar data which contained more of an equal spread of values amongst heart disease and no heart disease to see if/how the findings differed.

5 References

1. Beckerman, James. (2020, July 02). Heart Disease and Lowering Cholesterol. WebMD. https://www.webmd.com/heart-disease/guide/heart-disease-lower-cholesterol-risk.

2. Centers for Disease Control and Prevention. (2022, February 7). Heart Disease Facts. https://www.cdc.gov/heartdisease/facts.htm.

3. Centers for Disease Control and Prevention. (2021, January 11). Heart Attack Symptoms, Risk, and Recovery. https://www.cdc.gov/heartdisease/heart_attack.htm.

4. Centers for Disease Control and Prevention. (2019, December 19). Know Your Risk for Heart Disease. https://www.cdc.gov/heartdisease/risk_factors.htm.

5. Cortez, Paulo. (2020). rminer: Data Mining Classification and Regression Methods. R package version 1.4.6. https://CRAN.R-project.org/package=rminer

6. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

7. Greenwell, Boehmke, B., Cunningham, J. and GBM Developers (2020). gbm: Generalized Boosted Regression Models. R package version 2.1.8. https://CRAN.R-project.org/package=gbm

8. Hermann, Luke K., MD, Weingart, Scott D., MD, Yoon, Yong M., MD, Genes, Nicholas G., MD, PhD, Nelson, Bret P., MD, Shearer, Peter L., MD, Duvall, W. Lane, MD, and Henzlova, Milena J., MD, PhD. (2010). Comparison of Frequency of Inducible Myocardial Ischemia in Patients Presenting to Emergency Department With Typical Versus Atypical or Nonanginal Chest Pain. The American Journal of Cardiology, 105(11), 1561–1564. https://doi.org/10.1016/j.amjcard.2010.01.014.

9. Lanza, Mustilli, M., Sestito, A., Infusino, F., Sgueglia, G. A., and Crea, F. (2004). Diagnostic and prognostic value of ST segment depression limited to the recovery phase of exercise stress test. Heart (British Cardiac Society), 90(12), 1417–1421. https://doi.org/10.1136/hrt.2003.031260.

10. Lauer, Okin, P., Larson, M., Evans, J., and Levy, D. (1996). Impaired heart rate response to graded exercise - Prognostic implications of chronotropic incompetence in the Framingham Heart Study. Circulation (New York, N.Y.), 93(8), 1520–1526. https://doi.org/10.1161/01.CIR.93.8.1520.

11. Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. R News 2(3), 18–22.

12. Mayo Clinic. (2022, March 30). Angina. https://www.mayoclinic.org/diseases-conditions/angina/symptoms-causes/syc-20369373.

13. Meyer, Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2021). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-9. https://CRAN.R-project.org/package=e1071

14. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

15. Rodgers, Jones, J., Bolleddu, S. I., Vanthenapalli, S., Rodgers, L. E., Shah, K., Karia, K., and Panguluri, S. K. (2019). Cardiovascular Risks Associated with Gender and Aging. Journal of Cardiovascular Development and Disease, 6(2), 19. https://doi.org/10.3390/jcdd6020019.

16. Wiesman, J., and Lofy, K. (2018). Washington State Health Assessment. Washington State Department of Health. https://doh.wa.gov/sites/default/files/legacy/Documents/1000//2018SHA_FullReport.pdf.

6 Contributions from Each Group Member

Rachel pitched the idea for the project, including the models under consideration and the data set to be analyzed. Beyond that, we collaborated quite equitably: we wrote the code together, discussed ideas on what parameters to tune (and what values to test), and contributed equally to writing this final report.