MAT 303 Project Two Summary Report

Creston Getz
creston.getz@snhu.edu
Southern New Hampshire University

1. Introduction

The dataset we are exploring in this report contains various health indicators such as max heart rate, blood sugar, blood pressure, and cholesterol. The complete list of health indicators or variables will be discussed in the next section. This report will create two logistic regression models to predict if a person is at risk for heart disease. We will also create a classification random forest to predict the risk of heart disease and a regression random forest to predict maximum heart rate. The results of these models can be used to evaluate an individual's medical history and identify their potential risks. Models like these can be beneficial for doctors and other healthcare professionals when looking for subtle risk factors. The models also give doctors a quick overview of relevant aspects of a patient's medical history which can save time in necessary situations. It is important to remember that these models are just useful tools and not 100% accurate. These models should be used alongside human judgment, especially in healthcare contexts.

2. Data Preparation

As mentioned, this dataset which we will call heart data contains many variables or health indicators. The full list of variables in table form can be found in the notebook. The names of variables will be put in italics for this section. A variable is simply a way to store data, if we have a variable x it can hold any arbitrary value. This dataset uses variable names that are succinct and may not be easily interpreted. The values of the variables also contain various medical jargon which we will briefly explain in this section.

The first response variable we are trying to predict is *target* which is a yes or no variable telling us if the individual has heart disease or not. The second response is *thalach* which is a numerical value for the individual's maximum heart rate achieved. The other variables in the dataset we will be using as predictor variables or variables we can use to predict the response variable. The predictor variables include; *age* which is the person's age in years. *Sex* indicating if they are male or female. A categorical variable called *cp* stores the type of chest pain they experience if any. *Trestbps* represents the person's resting blood pressure. *Chol* measures their cholesterol in mg/dl. *Fbs* is a categorical variable indicating if their fasting blood sugar is higher than 120. *Restecg* is another categorical variable that indicates the type of resting electrocardiographic measurement (a test that measures the heart's electrical activity) (The Iowa Clinic, n.d.). *Thalach* or the max heart rate will also be used as a predictor in some models. *Exang* a yes or no variable reveals if the individual has exercise-induced angina (chest pain from exercise due to reduced blood flow) (Janosi et al., 1988). *Oldpeak* conveys ST depression induced by exercise compared to rest (the heart's electrical activity drops during exercise compared to resting, this indicates reduced blood flow) (Janosi et al., 1988). *Slope* signifies the slope of peak exercise ST segment (the shape of the heart's electrical signal after exercise) (Janosi et al., 1988). *Ca* represents the number of major vessels (0–3).

3. Model #1 - First Logistic Regression Model

In this section, we will create our first logistic regression model with target as the response variable. We will use age, resting blood pressure, exercise-induced angina, and max heart rate as our

predictor variables. The general form for this model is $E(y) = \pi(X) = \frac{e^{(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4)}}{1+e^{(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4)}}$. The prediction equation is $\hat{y} = \frac{e^{(\widehat{\beta_0}+\widehat{\beta_1}x_1+\widehat{\beta_2}x_2+\widehat{\beta_3}x_3+\widehat{\beta_4}x_4)}}{1+e^{(\widehat{\beta_0}+\widehat{\beta_1}x_1+\widehat{\beta_2}x_2+\widehat{\beta_3}x_3+\widehat{\beta_4}x_4)}}$. The prediction equation in terms of log odds is: $\ln(odds) = \widehat{\beta_0} + \widehat{\beta_1}x_1 + \widehat{\beta_2}x_2 + \widehat{\beta_3}x_3 + \widehat{\beta_4}x_4$. Where y is the predicted probability of target or heart disease. And x1 is age, x2 is blood pressure, x3 represents the binary variable exang, and x4 is max heart rate. $\widehat{\beta_i}$ where i is the number of betas are estimates of β_i.

The pi symbol represents the probability that the individual has heart disease given their health indicators: age, blood pressure, exang, and max heart rate. So, if pi is 0.6, then the model predicts a 60% chance of heart disease. $\frac{\pi}{1-\pi}$ is the odds of an individual having heart disease. It compares the probability of having heart disease to not having it.

After creating the model in R and using our prediction equation, our model is: $\ln(odds) = -1.021 - 0.01755(age) - 0.01488(trestbps) - 1.62498(exang) + 0.031095(thalach)$. The coefficient for thalach or max heart rate is 0.031095. Our equation is in log odds form which allows us to linearize the relationship between max heart rate and heart disease. Holding all other variables constant for each 1 beat per min increase in max heart rate, the log odds of heart disease increases by about 0.0311. In odds form $e^{0.031095} \approx 1.0315$, the odds increase by about 3.15% for each bpm increase.

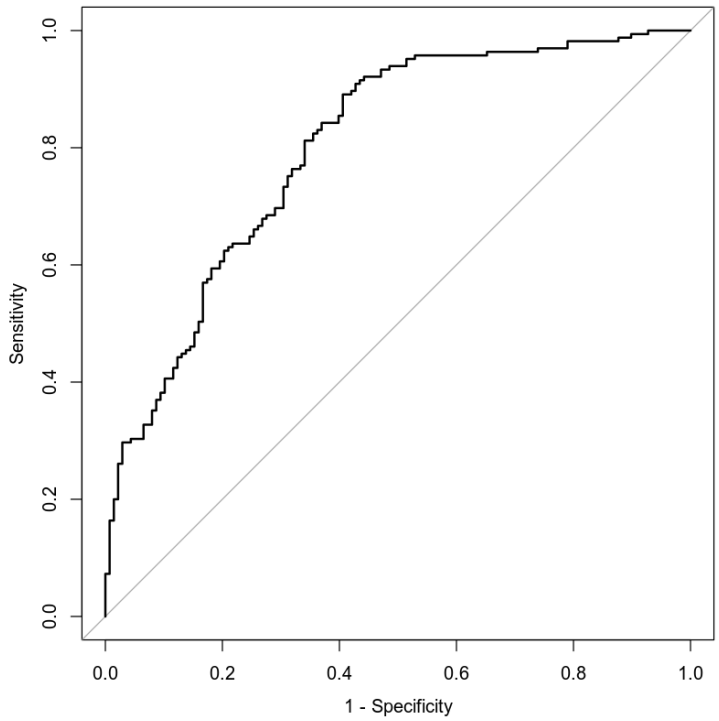
Evaluating Model Significance

It is important that we evaluate the significance of our model. We will first use a Hosmer-Lemeshow goodness of fit test (GOF test) to test our model. This test tells us if the model's predictions are close to the observed values of target or heart disease. The null hypothesis is that the model fits the data or mathematically H_0: the model fits the data. The alternative hypothesis is that the model does not fit the data or H_a: the model does not fit the data. We will be using a 5% level of significance for the test. The test statistic is 44.622 and our p-value is 0.612. Given our p-value is greater than our significance level, we fail to reject the null hypothesis. We do not have sufficient evidence to support the model is a poor fit for the data.

Next, we will use a Wald's test again with a 5% level of significance so we can determine which of the predictors are significant. For each test we will use the same null and alternative hypothesis. The null hypothesis is the term or parameter coefficient is 0, or $H_0: \beta_1 = 0$. While the alternative hypothesis is that the term or parameter coefficient is not 0 or mathematically $H_a: \beta_1 \neq 0$. Age has a p-value of 0.5671 which is higher than our significance level, we fail to reject the null hypothesis. Trestbps or resting blood pressure has a p-value of 0.0741 which is also higher than our significance level. So, we fail to reject the null hypothesis again. Neither of these terms have sufficient significance to reject the null hypothesis in favor of the alternative hypothesis. There is no statistically significant relationship between age and resting blood pressure with target. Exang1 or exercised induced angina's p-value is 1.07e-07 which is lower than our significance level. Therefore, we reject the null hypothesis in favor of the alternative hypothesis. Thalach or max heart rate's p-value is 1.92e-05 which again is lower than our significance level, so we reject the null hypothesis. Both exang and thalach have a statistically significant relationship with target.

Using a confusion matrix (see notebook), we will assess the performance of our model. Our matrix or model has 134 true positives, 89 true negatives, 49 false negatives, 31 false positives. Our model's accuracy is $\frac{134+89}{134+89+49+31} \approx 0.7557$ or 75.57%. The precision of the model is $\frac{134}{134+31} \approx 0.8121$ or about 81.21%. The recall is $\frac{134}{134+49} \approx 0.7332$ or 73.32%. We can measure the performance of a

classifier using an ROC curve. An ROC curve shows us the area under the curve or AUC which indicates how well the model distinguishes between yes or no in the response variable (zyBooks, n.d.). Given the ROC plot below the model performs moderately well distinguishing between if someone has heart disease or not. The ROC curve is well above the diagonal line and the model does a better job than random guessing. The AUC value is 0.8007 or 80.07% which is the probability the model will make the correct ranking.



Making Predictions Using Model

Now that we have created and tested our model, we will use it to make some predictions. Our model predicts an individual who is 50 years old, has a resting blood pressure of 122, has exercise-induced angina, and has a max heart rate of 140 has a 0.2716 or 27.16% probability of having heart disease. The odds of this person having heart disease is $\frac{0.2716}{1-0.2716} \approx 0.37287$ to 1. For every 1 person that has heart disease there will be about 2 or 3 that do not. The chance that someone with similar health metrics has heart disease is unlikely to occur. An individual who is also 50 but has a blood pressure of 130, does not have exercise-induced angina, and has a max heart rate of 165 has a 0.7853 or 78.53% probability of having heart disease. The odds of this event occurring are $\frac{0.7853}{1-0.7853} \approx 3.6576$ to 1. For every one person predicted to not have heart disease there will be about 3.6 who do have heart disease. Both predictions seem reasonable, a combination of higher resting blood pressure, higher max heart rate, and the absence of exercise-induced angina led to a higher predicted probability of heart disease. Even being the same age, the second individual had a substantially higher risk. It is important to note that the dataset's relationships may differ from expected medical intuition. While the second individual had a higher risk of heart disease despite not having exercise-induced angina it does not mean that it is a direct relationship. It is important to validate results with medical expertise.

4. Model #2 - Second Logistic Regression Model

Reporting Results

In this section, we will create the second logistic regression model using target as the response variable and age, resting blood pressure, type of chest pain, and max heart rate as the predictors. We will also include the quadratic term for age and the interaction term between age and max heart rate. Cp, or the type of chest pain, is a categorical variable, so this model will require dummy variables. The general form for this model is: $E(y) = \frac{e^{(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x1^2+\beta_8x_1x_6)}}{1+e^{(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x1^2+\beta_8x_1x_6)}}$. And the prediction equation is $E(y) = \frac{e^{(\widehat{\beta_0}+\widehat{\beta_1}x_1+\widehat{\beta_2}x_2+\widehat{\beta_3}x_3+\widehat{\beta_4}x_4+\widehat{\beta_5}x_5+\widehat{\beta_6}x_6+\widehat{\beta_7}x1^2+\widehat{\beta_8}x_1x_6)}}{1+e^{(\widehat{\beta_0}+\widehat{\beta_1}x_1+\widehat{\beta_2}x_2+\widehat{\beta_3}x_3+\widehat{\beta_4}x_4+\widehat{\beta_5}x_5+\widehat{\beta_6}x_6+\widehat{\beta_7}x1^2+\widehat{\beta_8}x_1x_6)}}$. Where y is the probability of heart disease, x1 is age, x2 is resting blood pressure, x3-x5 are dummy variables for cp, and x6 is max heart rate. And $\widehat{\beta_i}$ where i in the number of betas are estimates of β_i. The prediction equation in terms of the natural log of odds is $\ln(odds) = \widehat{\beta_0} + \widehat{\beta_1}x_1 + \widehat{\beta_2}x_2 + \widehat{\beta_3}x_3 + \widehat{\beta_4}x_4 + \widehat{\beta_5}x_5 + \widehat{\beta_6}x_6 + \widehat{\beta_7}x1^2 + \widehat{\beta_8}x_1x_6$. The natural log odds form allows us to linearize the beta terms. After making the model in R and using our prediction equation, the model equation is:
$\ln(odds) = -15.56 + 0.1744(age) - 0.01958(trestbps) + 1.913(cp1) + 2.037(cp2) + 1.777(cp3) + 0.1362(thalach) + 0.0008424(age^2) - 0.01867(age \cdot thalach)$.

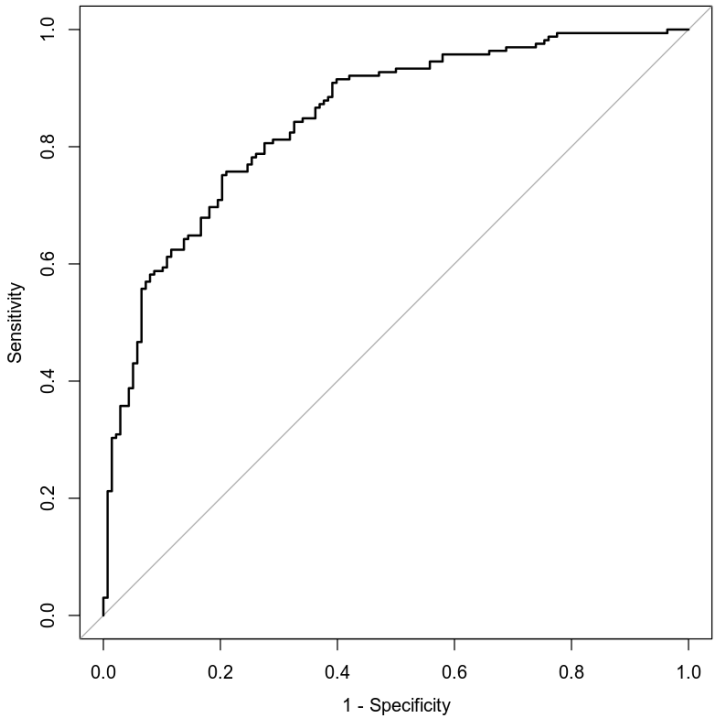
Evaluating Model Significance

Next, we will evaluate the significance of our second model. For every test in this section, we will be using a 5% level of significance. The Hosmer and Lemeshow GOF test has a p-value of 0.3209 and a test statistic of 52. The null hypothesis for the GOF test is that the model fits the data or H_0: the model fits the data. The alternative hypothesis is that the model does not fit the data, or H_a: the model does not fit the data. Our p-value of 0.3209 is higher than our significance level so we fail to reject the null hypothesis. We do not have enough evidence to conclude that the model is a poor fit for the data.

Using a Wald's test, we can determine which terms or predictors are significant to the model again using a 5% level of significance. Each test will use the same null and alternative hypothesis. The null hypothesis for each Wald's test is that the regression parameter is equal to zero or $H_0: \beta_1 = 0$. And the alternative hypothesis is that the regression parameter is not equal to zero or $H_a: \beta_1 \neq 0$. Our model has two terms that are not statistically significant; age and age squared. Age has a p-value of 0.51357 and age squared is 0.63025. We fail to reject the null hypothesis for both of these terms. The rest of the terms in our model are statistically significant to the model. Trestbps's p-value is 0.02916, cp1 has a p-value of 1.61e-05, cp2 has a p-value of 4.45e-09, cp3 has a p-value of 0.00117, thalach's p-value is 0.00775, and lastly the interaction term between thalach and age has a p-value of 0.03616. Every single one of these terms has a p-value lower than our significance level so we reject the null hypothesis in favor of the alternative hypothesis. These predictors have a statistically significant relationship with the probability of heart disease.

To analyze the performance of our model we may use a confusion matrix (see notebook). The matrix has 129 true positives, 102 true negatives, 36 false positives, and 36 false negatives. Therefore, the accuracy of our model is $\frac{129+102}{129+102+36+36} \approx 0.7624$ or 76.24%. The precision is $\frac{129}{129+36} \approx 0.7818$ which is

about 78.18%. Lastly the recall is $\frac{129}{129+36} \approx 0.7818$ or 78.18%. Now we will measure the performance of the classifier using an ROC graph (see below). Our ROC graph indicates our model performs well at distinguishing if someone has heart disease or not. The curve is well above the diagonal line and our area under the curve or AUC is 0.8478 which is higher than our first model. This model correctly ranks random positives over random negatives about 84.78% of the time. The AUC is the probability the model will make the correct guess or ranking.



Making Predictions Using Model

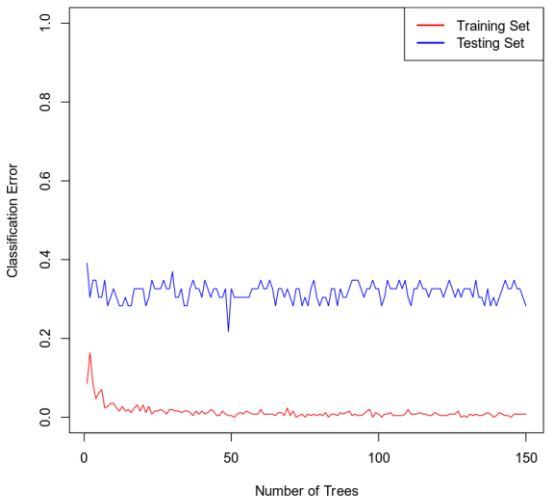
With our model created we may use it to create some predictions. An individual who is 50 years old, has a resting blood pressure of 115, no chest pain, and has a max heart rate of 133 has a 0.2188 or 21.88% probability of having heart disease. The odds of this event are $\frac{0.2188}{1-0.2188} \approx 0.28$ to 1. For every 1 person with heart disease, about 3.5 will not. This person is about 3.5 times more likely to not have heart disease. An individual who is also 50 but has a resting blood pressure of 125, experiences typical angina(cp=1), and has a max heart rate of 155 has a 0.8007 or 80.07% probability of having heart disease. The odds of this second event is $\frac{0.8007}{1-0.8007} \approx 4.02$ to 1. For every 4 people who do have heart disease there will be 1 that does not. This event is much more likely to occur than the first. This individual is 4 times more likely to have heart disease than to not.

These two predictions line up with the results we found in our first model. In this case, a combination of higher resting blood pressure, a higher maximum heart rate, and the presence of typical angina (cp=1) led to a much higher predicted probability of heart disease. In our second model we added chest pain as a predictor, which our Wald's test found to be highly significant. This helped increase the model's accuracy.

5. Random Forest Classification Model

Reporting Results

Now we will create a random forest classification model to predict heart disease. The original dataset has 303 rows, and we will use an 85% and 15% split for our training and testing sets. Our training set has 257 rows while our testing set is only 46. We will be using age, sex, chest pain, resting blood pressure, cholesterol, resting electrocardiographic measurement, exercise-induced angina, and the number of major vessels as predictors. The plot of classification error against the number of trees below indicates that around 20-60 trees the classification error is the lowest. That is a wide range to use so instead we can find the minimum number of trees by finding the minimum value in the list of classification errors we created. The number of trees with the lowest error is 51 so we will use that as the optimal number of trees.



Evaluating the Utility of the model

After we make the forest with the appropriate number of trees we need to evaluate its utility by finding the accuracy, precision and recall. The confusion matrix for the training set reports 112 true negatives, 128 true positives, 0 false negatives, and 2 false positives. The accuracy for the training model is $\frac{240}{242} \approx 0.9917$ or 99.17% accuracy, the precision is $\frac{64}{65} \approx 0.9846$ or 98.46%, and the recall is $\frac{128}{128} = 1$ or 100%. Now for the testing set our model had 19 true positives, 11 true negatives, 9 false negatives, and 7 false positives. The accuracy is $\frac{15+29}{15+19+9+9} \approx 0.7213$ or 72.13%, the precision is $\frac{29}{29+9} \approx 0.7631$ or 76.31%, and the recall is $\frac{29}{37} \approx 0.7838$ or 78.38%.

Our model is significantly more accurate with the training set than the testing set indicating it is likely overfitting the training data. This is common when trees are deep in the forest. To help prevent

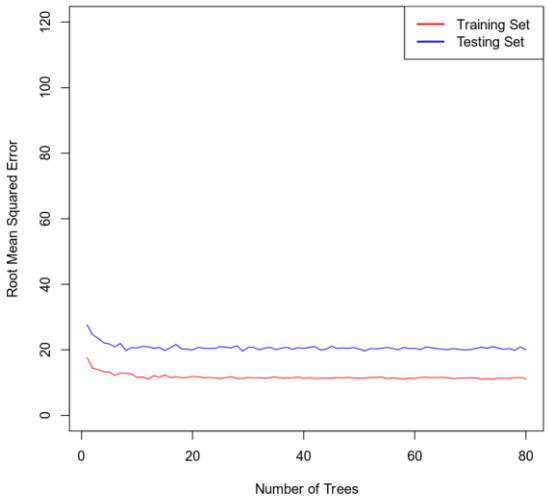
overfitting we could limit the depth of each tree so it would choose the best predictors. We may also remove weak variables or reduce the number of splits in the trees. If we limit the number of nodes the forest can have to 10, set the node size to 5 which can prevent the trees from going too deep, and limit the number of attributes the trees can use to 2 we do improve the accuracy of the testing set and reduce the accuracy in the training set. The forest with limits testing set accuracy is the same but the training sets accuracy is $\frac{215}{242} \approx 88.84\%$ compared to 99.17% with no limits. The smaller gap between the training and testing set accuracy indicates the forest is less overfit without hurting the testing set accuracy.

```
[1] "Matrix for training set 51 trees"

      Predicted
Actual   0   1
     0 112   2
     1   0 128

[1] "Matrix for testing set 51 trees"

      Predicted
Actual  0  1
     0 15  9
     1  8 29
```

6. Random Forest Regression Model

Reporting Results

In this section we will create a random forest regression model to predict max heart rate. We will use age, sex, chest pain, blood pressure, cholesterol, resting electrocardiographic measurement, exercise-induced angina, and number of major vessels as predictor variables. For this forest our training data will be 80% of the total rows and the testing data will be 20%. So out of the 303 rows, the training set will have 242 rows and the testing set will have 61 rows. After graphing the mean squared error against the number of trees (see below) the mean squared error bottoms out around 15 trees. After 20 trees there is almost no difference in mean squared error so we will use 20 as the optimal number of trees.



Evaluating the Utility of the Random Forest Regression Model

Unlike our classification tree, the utility of a regression forest is typically measured using the Root Mean Squared Error (RMSE). For our model with 20 trees, the RMSE for the training set was 11.63, while the RMSE for the testing set was 21.11. This means that, on average, the model's predictions were about 21 beats per minute (bpm) off from the actual max heart rate in the testing set, compared to only about 11 bpm off in the training set. The max heart rate in this data set ranges from 71 to 202, the testing RMSE of 21.11 represents about a 14% average error.

This gap of 10 bpm could indicate that our model is overfitting the training set and the model may not perform well with new data. To help prevent overfitting we could limit the number of nodes, their size, and reduce the number of attributes the trees can use. Similar to what we did with the classification forest. For the regression forest we would want to make the trees slightly larger because the response variable is continuous.

7. Conclusion

In this report, we created two logistic regression models to predict heart disease. Of those two, I would recommend the second model. The first model still performs well; however, we added chest pain in the second model, which we found to be a significant predictor of heart disease. This increased the second model's accuracy slightly. The Hosmer-Lemeshow GOF test indicated that the model was a good fit, and we had several meaningful predictors from the Wald's test. The second model also has slightly better performance; not only was the accuracy higher, but so were the precision and recall compared to the first model. While the second model is a modest improvement, it would be more useful in practice than the first.

In the context of this dataset, I would recommend the logistic regression models over the classification random forest. Random forests tend to perform better on larger datasets, whereas the heart disease dataset contains only about 300 rows. Even after applying limitations to the forest, the

gap between training and testing accuracy was still larger than that of the logistic models. The logistic model also achieved higher testing accuracy to begin with. Additionally, logistic regression provides clear insight into how each variable affects the probability of heart disease, which is valuable in a medical context. This interpretability allows doctors and other healthcare professionals to understand why a model predicts a high risk. While a random forest may produce similar predictions, it does not offer the same level of interpretability.

The analyses performed in this project demonstrate how different models can be used to assess the risk an individual has for heart disease. We used logistic regression, which gave interpretable models we can use to determine the influence of each predictor. For example, in both sets of predictions we made, we found that a higher max heart rate and a higher resting blood pressure increased the risk of heart disease. As mentioned, this interpretability is key to medical contexts. Random forests may give similar results as a logistic model, but they are much less interpretable. They can help us identify non-linear relationships that would otherwise not be captured in logistic models. By evaluating the forest's utility, we gained insights into how the size of the dataset, variable importance, and node size can affect the accuracy of the forests with new data.

Any of these models can be used to support early screening and risk assessment in medical settings. They may also support various research settings, providing insights into risk factors. They can help identify high-risk patients who can benefit from more preventive measures, which may improve health outcomes. It can also help us optimize resource allocation to people who need it most.

8. Citations

Janosi, A., Steinbrunn, W., Pfisterer, M., Detrano, R. (1988). Heart disease data set [Data file and codebook]. Retrieved from https://archive.ics.uci.edu/ml/datasets/Heart+Disease

The Iowa Clinic. (n.d.). Resting EKG. https://www.iowaclinic.com/specialties/employer-health/executive-health/core-services/resting-ekg/#:~:text=Preparation:%20Before%20the%20EKG%2C%20you,chest%2C%20arms%2C%20and%20legs.

zyBooks. (n.d.). MAT 303: Applied Statistics II for Science. https://learn.zybooks.com/zybook/MAT-303-13299.202556-1