# **MAT 303 Project Two Summary Report**

[Full Name]

[SNHU Email]

Southern New Hampshire University

Note: Replace the bracketed text on page one (the cover page) with your personal information.

## **1. Introduction**

*Discuss the statement of the problem in terms of the statistical analyses that are being performed. Be sure to address the following.*

* *What is the data set that you are exploring?*
* *How will your results be used?*
* *What type of analyses will you be running in this project?*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

## **2. Data Preparation**

*There are some important variables that are used in this project. Identify and explain these variables.*

* *What are the important variables in this data set?*
* *How many rows and columns are present in this data set?*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

## **3. Model #1 - First Logistic Regression Model**

### **Reporting Results**

*Report the results of the regression model. Address the following questions in your analysis. Round all numbers to four decimal places*.

* *Write the general form of the multiple regression model for heart disease (target) using variables age (age), resting blood pressure (trestbps), and maximum heart rate achieved (thalach). Note that this general form should be written in terms of E(y), exponents and* beta sub i *(where i* equals *1, 2, ... ). Note: Use the equation editor to write the regression equation.*
* *Now write this model in terms of the natural log of odds to express the beta terms in linear form.   
  Note: Use the equation editor to write the regression equation.*
* *What do the following terms, from the general form of the model above, mean in terms of an individual defaulting on their credit?*

Top image: pi
Bottom image: pi over (1 minus pi)

* *Create this logistic regression model and write its equation in terms of E(y) and in terms of the natural log of odds.*
* *Interpret the estimated coefficient of the maximum heart rate achieved variable.*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

### **Evaluating Model Significance**

*Evaluate the significance of the regression model. Address the following questions in your analysis. Round all numbers to four decimal places.*

* *Perform the Hosmer-Lemeshow goodness of fit test to assess whether the model is appropriate for the data set. Identify the null and alternative hypotheses, the test statistic, and the P-value. Use a 5% level of significance.*
* *Which terms are significant in the model based on Wald’s test? Use a 5% level of significance.*
* *Obtain the confusion matrix and report the counts for true positives, true negatives, false positives, and false negatives.*
* *Report the following:*
  + *Accuracy*
  + *Precision*
  + *Recall*
* *Obtain the Receiver Operating Characteristic (ROC) curve. Interpret the graph and explain what it illustrates.*
* *What is the value of AUC? Interpret what this value represents.*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

### **Making Predictions Using Model**

*Make predictions using the regression model. Address the following questions in your analysis. Round all numbers to four decimal places***.**

* *What is the probability of an individual who is 50 years old, has a resting blood pressure of 122, and has maximum heart rate of 140 having heart disease? Find the odds of this event occurring.*
* *What is the probability of an individual who is 50 years old, has a resting blood pressure of 140, and has maximum heart rate of 170 having heart disease? Find the odds of this event occurring.*
* *Comment on the two predictions. What can be deduced based on the probabilities and the odds?*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

## **4. Model #2 - Second Logistic Regression Model**

### **Reporting Results**

*Report the results of the regression model. Address the following questions in your analysis. Round all numbers to four decimal places*.

* *Write the general form of the multiple regression model for heart disease (target) using variables maximum heart rate achieved (thalach), age of the individual (age), sex of the individual (sex), exercise-induced angina (exang), and type of chest pain (cp). Also include the quadratic term for age and the interaction term between age and maximum heart rate achieved. Note that this general form should be written in terms of E(y), exponents and* beta sub i *(where i* equals *1, 2, ... ). Note: Use the equation editor to write the regression equation.*
* *Now write this model in terms of the natural log of odds to express the beta terms in linear form.   
  Note: Use the equation editor to write the regression equation.*
* *Create this logistic regression model and write its equation in terms of E(y) and in terms of the natural log of odds.*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

### **Evaluating Model Significance**

*Evaluate the significance of the regression model. Address the following questions in your analysis. Round all numbers to four decimal places.*

* *Perform the Hosmer-Lemeshow goodness of fit test to assess whether the model is appropriate for the data set. Identify the null and alternative hypotheses, the test statistic, and the P-value. Use a 5% level of significance.*
* *Which terms are significant in the model based on Wald’s test? Use a 5% level of significance.*
* *Obtain the confusion matrix and report the counts for true positives, true negatives, false positives, and false negatives.*
* *Report the following:*
  + *Accuracy*
  + *Precision*
  + *Recall*
* *Obtain the Receiver Operating Characteristic (ROC) curve. Interpret the graph and explain what it illustrates.*
* *What is the value of AUC? Interpret what this value represents.*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

### **Making Predictions Using Model**

*Make predictions using the regression model. Address the following questions in your analysis. Round all numbers to four decimal places*.

* *What is the probability of a male individual having heart disease who is 30 years old; has a maximum heart rate of 145; experiences exercise-induced angina; and does not experience chest pain related to typical angina, atypical angina, or non-anginal pain? Find the odds of this event occurring.*
* *What is the probability of a male individual having heart disease who is 30 years old, has a maximum heart rate of 145, and does not experience exercise-induced angina but experiences typical angina? Find the odds of this event occurring.*
* *Comment on the two predictions. What can be deduced based on the probabilities and the odds?*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

## **5. Random Forest Classification Model**

### **Reporting Results**

*Report the results of the random forest classification model. Address the following questions in your analysis. Round all numbers to four decimal places.*

* *Split the heart disease data set into training and testing sets using 80% and 20% split, respectively. Use set.seed(511038). How many rows are in the original data set, and in the training and validation sets?*
* *Graph the training and testing error against the number of trees using a classification random forest model for the presence of heart disease (target) using variables age (age), sex (sex), chest pain type (cp), resting blood pressure (trestbps), cholesterol measurement (chol), resting electrocardiographic measurement (restecg), exercise-induced angina (exang), slope of peak exercise (slope), and number of major vessels (ca). Use a maximum of 200 trees. Use set.seed(511038).*
* *What is the optimal number of trees for the random forest model?*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

### **Evaluating the Utility of the model**

*Evaluate the utility of the random forest classification model. Address the following questions in your analysis. Round all numbers to four decimal places.*

* *Using the appropriate number of trees found, create a classification random forest model for the presence of heart disease (target) using variables age (age), sex (sex), chest pain type (cp), resting blood pressure (trestbps), cholesterol measurement (chol), resting electrocardiographic measurement (restecg), exercise-induced angina (exang), slope of peak exercise (slope), and number of major vessels (ca). Obtain the confusion matrix for the training set and report the accuracy, precision, and recall.*
* *Obtain the confusion matrix for the testing set and report the accuracy, precision, and recall.*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

## **6. Random Forest Regression Model**

### **Reporting Results**

*Report the results of the random forest regression model. Address the following questions in your analysis. Round all numbers to four decimal places.*

* *Split the heart disease data set into training and testing sets using 80% and 20% split, respectively. Use set.seed(511038). How many rows are in the original data set, and the training and validation sets?*
* *Graph the mean squared error against the number of trees for a random forest regression model for maximum heart rate achieved using age (age), sex (sex), chest pain type (cp), resting blood pressure (trestbps), cholesterol measurement (chol), resting electrocardiographic measurement (restecg), exercise-induced angina (exang), slope of peak exercise (slope), and number of major vessels (ca). Use a maximum of 80 trees. Use set.seed(511038).*
* *What is the optimal number of trees for the random forest model?*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

### **Evaluating the Utility of the Random Forest Regression Model**

*Evaluate the utility of the random forest regression model. Address the following questions in your analysis. Round all numbers to four decimal places.*

* *Using the appropriate number of trees found, create a random forest regression model for maximum heart rate achieved using age (age), sex (sex), chest pain type (cp), resting blood pressure (trestbps), cholesterol measurement (chol), resting electrocardiographic measurement (restecg), exercise-induced angina (exang), slope of peak exercise (slope), and number of major vessels (ca).*
* *What is the root mean squared error for the training set?*
* *What is the root mean squared error for the testing set?*

*Caution sign icon Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

## **7. Conclusion**

*Describe the results of the statistical analyses clearly, using proper descriptions of statistical terms and concepts. Fully describe what these results mean for your scenario.*

* *Which of the two logistic regression models would you choose to predict heart disease? Briefly summarize your findings in plain language.*
* *Would you recommend using the random forest classification model instead of the logistic regression model? Why or why not?*
* *What is the practical importance of the analyses that were performed?*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

## **8. Citations**

*You are* ***not*** *required to use external resources for this report. If you did not use any resources, you should remove this entire section. However, if you did use any resources to help you with your interpretation, you* ***must*** *cite them. Use proper APA format for citations.*

*Insert references here in the following format:*

Author's Last Name, First Initial. Middle Initial. (Year of Publication). Title of book: Subtitle of book, edition. Place of Publication: Publisher.