Homework 5

Calista Harris

Task 1: Conceptual Questions

Question 1

Cross-validation helps us estimate how well a random forest model is likely to perform on unseen data by repeatedly splitting the data into training and validation sets. This process reduces variability in the performance estimate that could arise from a single train/test split and allows for more reliable model assessment. Although random forests have built-in error estimation through out-of-bag observations, cross-validation can provide a complementary or alternative method, especially when comparing multiple models or tuning hyperparameters.

Question 2

The bagged tree algorithm, short for bootstrap aggregation, involves generating multiple bootstrap samples from the original dataset, fitting a decision tree to each sample, and aggregating their predictions. For regression tasks, the final prediction is the average of individual tree predictions, while for classification, it is typically determined by majority vote. This ensemble method helps reduce variance and improves predictive performance compared to a single tree model.

Question 3

A general linear model (GLM) is a statistical model where the expected value of the response variable is modeled as a linear combination of the explanatory variables. It has the form $E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$, and is used for tasks such as multiple linear regression and ANOVA. The model assumes the errors are normally distributed and have constant variance, making it a foundational approach for modeling relationships between variables.

Question 4

Adding an interaction term to a multiple linear regression model allows the effect of one explanatory variable on the response variable to depend on the level of another explanatory variable. This enables the model to capture relationships where the combined effect of two variables is not merely additive. In mathematical terms, it adds a term like $\beta_3 X_1 X_2$ to the model, allowing the slope of one variable to change based on the value of the other.

Question 5

Splitting the data into a training and test set allows us to evaluate how well a model generalizes to new, unseen data. The training set is used to fit the model, while the test set provides an unbiased estimate of the model's predictive performance. This helps guard against overfitting and ensures that the model is not simply memorizing the training data, but instead learning patterns that can apply more broadly.

Task 2: Data Prep

Packages and Data

```
#load require libraries
library(tidyverse)
library(tidymodels)
library(caret)
library(yardstick)

#read in the heart disease dataset as a tibble
heart <- read_csv("data/heart.csv") |>
  as_tibble()
```

Question 1

```
#summarize the data
summary(heart)
```

```
      Age            Sex            ChestPainType        RestingBP
 Min.   :28.00   Length:918         Length:918         Min.   :  0.0
 1st Qu.:47.00   Class :character   Class :character   1st Qu.:120.0
 Median :54.00   Mode  :character   Mode  :character   Median :130.0
 Mean   :53.51                                         Mean   :132.4
 3rd Qu.:60.00                                         3rd Qu.:140.0
 Max.   :77.00                                         Max.   :200.0
  Cholesterol      FastingBS       RestingECG            MaxHR
 Min.   :  0.0   Min.   :0.0000   Length:918         Min.   : 60.0
 1st Qu.:173.2   1st Qu.:0.0000   Class :character   1st Qu.:120.0
 Median :223.0   Median :0.0000   Mode  :character   Median :138.0
 Mean   :198.8   Mean   :0.2331                      Mean   :136.8
 3rd Qu.:267.0   3rd Qu.:0.0000                      3rd Qu.:156.0
 Max.   :603.0   Max.   :1.0000                      Max.   :202.0
 ExerciseAngina       Oldpeak          ST_Slope          HeartDisease
 Length:918         Min.   :-2.6000   Length:918         Min.   :0.0000
 Class :character   1st Qu.: 0.0000   Class :character   1st Qu.:0.0000
 Mode  :character   Median : 0.6000   Mode  :character   Median :1.0000
                    Mean   : 0.8874                      Mean   :0.5534
                    3rd Qu.: 1.5000                      3rd Qu.:1.0000
                    Max.   : 6.2000                      Max.   :1.0000
```

a. According to the summary() output, `HeartDisease` is currently treated as a numeric variable in R. This is evident from the statistical summaries displayed — Min, 1st Qu., Median, Mean, etc. — all indicators of a quantitative numeric type.

b. No, this does not make sense for modeling. The `HeartDisease` variable encodes binary outcomes — either 0 (no heart disease) or 1 (presence of heart disease). As described in Logistic Regression Models, binary outcomes should be treated as categorical when modeling classification problems. Using it as numeric may lead to inappropriate modeling choices, such as applying linear regression when logistic regression is more appropriate.

Question 2

```
#convert HeartDisease to a factor (categorical) variable and rename it
new_heart <- heart |>
  mutate(HeartDisease_status = factor(HeartDisease)) |>
  #drop the original numeric HeartDisease variable and the ST_Slope variable
  select(-ST_Slope, -HeartDisease)

#view the structure of the updated data set
glimpse(new_heart)
```

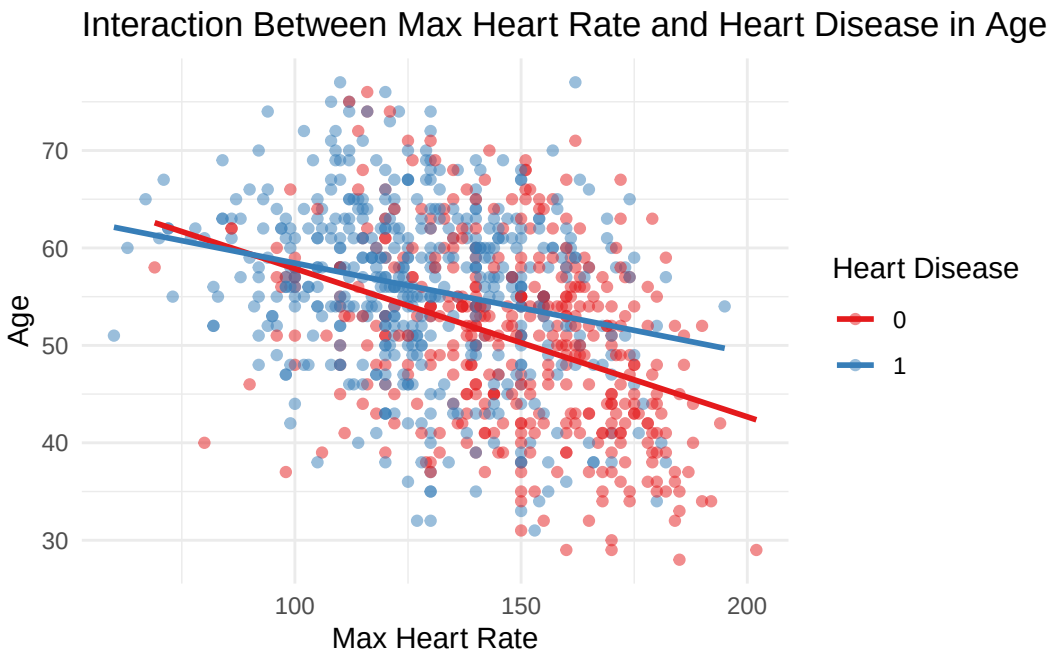
```
Rows: 918
Columns: 11
$ Age                <dbl> 40, 49, 37, 48, 54, 39, 45, 54, 37, 48, 37, 58, 39~
$ Sex                <chr> "M", "F", "M", "F", "M", "M", "F", "M", "M", "F", ~
$ ChestPainType      <chr> "ATA", "NAP", "ATA", "ASY", "NAP", "NAP", "ATA", "~
$ RestingBP          <dbl> 140, 160, 130, 138, 150, 120, 130, 110, 140, 120, ~
$ Cholesterol        <dbl> 289, 180, 283, 214, 195, 339, 237, 208, 207, 284, ~
$ FastingBS          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ RestingECG         <chr> "Normal", "Normal", "ST", "Normal", "Normal", "Nor~
$ MaxHR              <dbl> 172, 156, 98, 108, 122, 170, 170, 142, 130, 120, 1~
$ ExerciseAngina     <chr> "N", "N", "N", "Y", "N", "N", "N", "N", "Y", "N", ~
$ Oldpeak            <dbl> 0.0, 1.0, 0.0, 1.5, 0.0, 0.0, 0.0, 0.0, 1.5, 0.0, ~
$ HeartDisease_status <fct> 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1,~
```

Task 3: EDA

Question 1

```
#create the scatterplot with separate trend lines by heart disease
ggplot(new_heart, aes(x = MaxHR, y = Age, color = HeartDisease_status)) +
  geom_point(alpha = 0.5) +
  #remove the standard error bars
  geom_smooth(method = "lm", se = FALSE) +
  #Set1 is color-blind friendly
  scale_color_brewer(palette = "Set1", name = "Heart Disease") +
  labs(
    title = "Interaction Between Max Heart Rate and Heart Disease in Age",
    x = "Max Heart Rate",
    y = "Age"
  ) +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Question 2

The scatterplot shows non-parallel trend lines for people with and without heart disease, indicating that the relationship between Max Heart Rate and Age differs by Heart Disease. Specifically, the slope for individuals without heart disease is steeper than for those with heart disease. This suggests an interaction effect, where the impact of `MaxHR` on `Age` depends on `HeartDisease_status`. Recall Multiple Linear Regression, interaction terms allow the slope of one variable to change based on the level of another. Therefore, an interaction model is more appropriate than an additive model for this analysis.