# STA130H1S – Fall 2022

Problem Set 9

() and STA130 Professors

## Instructions

Complete the exercises in this .Rmd file and submit your .Rmd and .pdf output through Quercus on Thursday, November 24th by 5:00 p.m. ET.

```
library(tidyverse)
library(rpart)
library(partykit)
library(knitr)
```

## Part 1: Binary Classification Decision Trees

### Question 1: Gallup World Poll

Using data from the Gallup World Poll (and the World Happiness Report), we are interested in predicting which factors influence life expectancy around the world. These data are in the file `happinessdata_2017.csv`.

```
happiness2017 <- read_csv("happiness2017.csv")
```

**(a) Begin by creating a new variable called `life_exp_category` which takes the value "Good" for countries with a life expectancy higher than 65 years, and "Poor" otherwise.**

```
# code you answer here
```

**(b) Divide the data into training (80%) and testing (20%) datasets. Build a classification tree using the training data to predict which countries have `Good` vs `Poor` life expectancy, using only the `social_support` variable as a predictor.**

```
set.seed(111)  # Use the last 3 digits of your student ID number for the random seed.
# code you answer here
```

**(c) Use the same training dataset created in (b) to build a second classification tree to predict which countries have good vs poor life expectancy, using `logGDP`, `social_support`, `freedom`, and `generosity` as potential predictors.**

```
# code you answer here
```

**(d) Use the testing dataset you created in (b) to calculate the confusion matrix for the trees you built in (b) and (c). Report the sensitivity (true positive rate), specificity (true negative rate) and accuracy for each of the trees. Here you will treat "Good" life expectancy as the positive response and prediction.**

```
# code you answer here for the tree created in part (b)
```

```
# code you answer here for the tree created in part (c)
```

**(e) Fill in the following table using the tree you constructed in part (c). Does the fact that some of the values are missing (NA) prevent you from making predictions for the life expectancy category for these observations?**

| | `logGDP` | `social_support` | `freedom` | `generosity` | Predicted life expectancy category |
|---|---|---|---|---|---|
| Obs 1 | 9.68 | 0.76 | NA | -0.35 | *REPLACE THIS TEXT WITH YOUR ANSWER* |
| Obs 2 | 9.36 | NA | 0.82 | -0.22 | *REPLACE THIS TEXT WITH YOUR ANSWER* |
| Obs 3 | 10.4 | 0.88 | 0.77 | 0.11 | *REPLACE THIS TEXT WITH YOUR ANSWER* |
| Obs 4 | 9.94 | 0.85 | 0.63 | 0.01 | *REPLACE THIS TEXT WITH YOUR ANSWER* |

**Hint: make a `tibble()` of this data and then use it with the `predict()` function.**

## Question 2: Confusion Matrices and Metrics (Accuracy, etc.)

Two classification trees were built to predict which individuals have a disease using different sets of potential predictors. We use each of these trees to predict disease status for 100 new individuals. Below are confusion matrices corresponding to these two classification trees.

**Tree A**

| | Disease | No disease |
|---|---|---|
| Predict disease | 36 | 22 |
| Predict no disease | 2 | 40 |

**Tree B**

| | Disease | No disease |
|---|---|---|
| Predict disease | 24 | 6 |
| Predict no disease | 14 | 56 |

**(a) Calculate the accuracy, false-positive rate, and false negative rate for each classification tree. Here, a "positive" result means we predict an individual has the disease and a "negative" result means we predict they do not.**

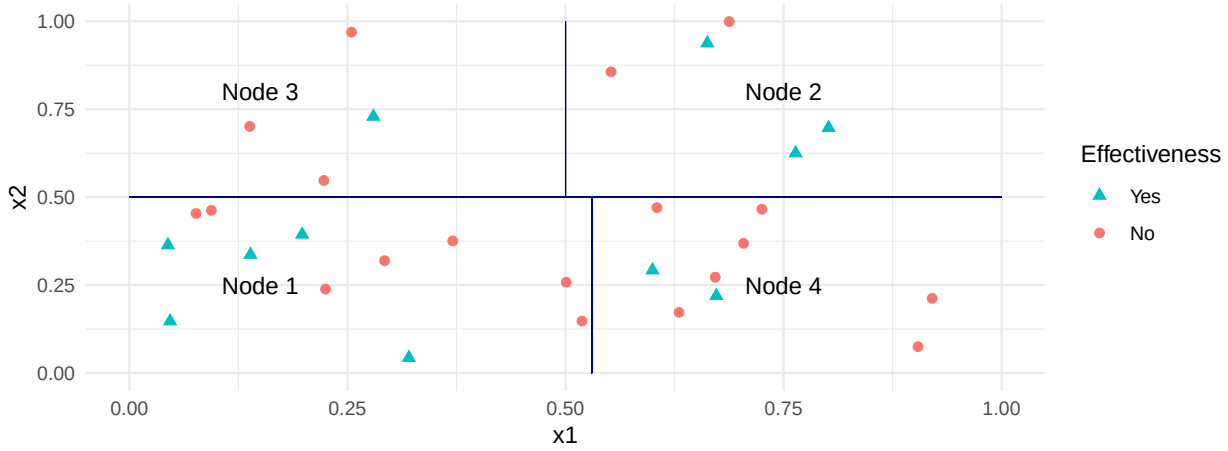
*REPLACE THIS TEXT WITH YOUR ANSWER*

**(b) Suppose the disease is very serious if untreated. Explain which classifier you would prefer to use.**

*REPLACE THIS TEXT WITH YOUR ANSWER*

## Question 3: Geometric Interpretation of Prediction

Data was collected on 30 cancer patients to investigate the effectiveness (Yes/No) of a treatment. Two quantitative variables, x1 and x2 (but taking values between 0 and 1), are thought to be important predictors of effectiveness. Suppose that the rectangles labeled as nodes in the scatter plot below represent nodes of a classification tree.




**(a) The diagram above is the geometric interpretation of a classification tree to predict drug effectiveness based on two predictors, x1 and x2. What is the predicted class of each node?**

| Node | Proportion of "Yes" values in each node | Prediction (assume we declare "effective" if more than 50% of the values are "Yes") |
|---|---|---|
| 1 | ????? | ????? |
| 2 | ????? | ????? |
| 3 | ????? | ????? |
| 4 | ????? | ????? |