HW2

1. Due at midnight, 11/15/2022.
2. Organize your codes, output, and answer for each question in an R markdown file.
3. Submit the R markdown file and corresponding output PDF or HTML file to the Oaks.

Consider the dataset BostonHousing.csv. Use CAT.MEDV as the response variable. CAT.MEDV is a binary variable derived from MEDV (median value of owner-occupied homes in $1000s) so that CAT.MEDV = 1 if MEDV > 30 and CAT.MEDV = 0 otherwise. Only keep CAT.MEDV in your data and remove MEDV. The description of the rest of the variables are as below:

CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots of 25,000 sq ft
INDUS	proportion of nonretail business acres per town
CHAS	charles river dummy variable (= 1 if tract bounds river; = 0 otherwise)
NOX	nitric oxide concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centers
RAD	index of accessibility to radical highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil/teacher ratio by town
LSTAT	percentage lower status of the population

A. Split the data into training (80%) and test (20%) datasets.

1. Build a k-nearest neighbor model. Clearly show the model building steps for full credit. What is the optimal number of neighbors?
2. Evaluate the predictive performance of this model on the test data.

B. Use the BostonHousing data to accomplish the following modeling tasks.

1. Fit a classification tree using cost complexity pruning. Clearly show the model building steps for full credit.

2. Write down the results in terms of rules. (*i.e.* for each terminal node, write down the decision and list the conditions that must be followed to reach that decision)
3. Evaluate the predictive performance of this model on the test data.

C. Compare the performances of the two models that you built in this assignment.