Instructions:

- Due date: June 27, 2019.
- Total points = 60.
- Submit a typed report.
- It is OK to discuss the homework with others, but you must write your own code and answers.
- Do a good job.
- You must use the following template for your report:

  Homework #
  Name

Section 1. Answers to the specific questions asked. This section should not exceed 4 pages and only those output from R should be included that are actually referred to in the report.

Section 2: R code. Your code must be annotated. No points may be given if a brief look at the code does not indicate what it is doing.

1. Consider the Caravan dataset which is a part of the ISLR package. The dataset consists of Purchase, indicating whether or not a given individual purchases a caravan insurance policy, and 85 predictors that measure demographic characteristics of 5,822 individuals. Read the analysis reported on pages 165-167 of the ISL book.

   (a) Based on what you have read, explain whether or not standardizing the predictors before performing any analysis is a good idea. Regardless of your answer, standardize the predictors.

   (b) Create a test set consisting of the first 1,000 observations and training set consisting of the remaining observations.

   (c) Fit a logistic regression model. Rather than using 0.5 as the cut-off, predict that a person will make the purchase if the estimated probability of purchase is greater than 0.2. Compute the confusion matrix and misclassification rate based on the test data.

   (d) Fit a tree. Unless the number of terminal nodes is large, display the tree graphically and explicitly describe the regions corresponding to the terminal nodes that provide a partition of the predictor space (i.e., provide expressions for the regions $R_1, \ldots, R_J$). Report the confusion matrix and misclassification rate for the unpruned tree based on the test data.

   (e) Use cross-validation to determine whether pruning is helpful and determine the optimal size for the pruned tree. Compare the pruned and un-pruned trees. Report the confusion matrix and misclassification rate for the pruned tree based on the test data. Which predictors seem to be the most important?

   (f) Use a bagging approach to analyze the data with $B = 1000$. Compute the confusion matrix and misclassification rate based on the test data. Which predictors seem to be the most important?

   (g) Repeat (f) with a random forest approach with $B = 1000$ and $m \approx \sqrt{p}$.

   (h) Repeat (f) with KNN with $K$ chosen optimally from training data using 10-fold cross-validation.
(i) Fit a support vector classifier to the training data with cost parameter chosen optimally using 10-fold cross-validation. Summarize key features of the fit. Compute the confusion matrix and misclassification rate based on the test data.

(j) Repeat (i) using a support vector machine with a polynomial kernel of degree two.

(k) Repeat (i) using a support vector machine with a radial kernel with both $\gamma$ and cost parameter chosen optimally.

(l) Compare the results in (c)-(k). Which method would you recommend? Justify your answer.