Assignment 2.1

In this assignment, you will apply your knowledge and understanding of the topics presented in the module readings and materials.

Instructions:

1. Answer the five assignment Questions on the following pages. Read the questions carefully and respond to all prompts.
2. Create a single document that combines your answers to all five questions, using the Final Format for submission requirements below.

Final Format for Submission:

- Submit one assignment file that combines the answers to all questions:
- The file must be either a **Microsoft Word** or a **PDF** document file.
- **Do not** combine and submit files into a zipped compressed folder.
- Please use the following naming conventions for your **Word** or **PDF** assignment file:
  File Naming: **`LastName_FirstName_Assignment<Number>.pdf`**
  Example: **`Smith_James_Assignment2.1.pdf`**
- Answer all parts of a question in one place and answer questions in the order they appear in the assignment.
- For programming answers using **R**, it is recommended that the answers are written in R Markdown and 'knitted' to a Word/PDF file.
  - Do not print data frames in your submission, if you want to make a point about data you can use head(df) to print the first few rows.
  - Submit the code used to answer the questions in the assignment with your name on it, answers without code and appropriate results will not get full credit.
  - It is not a professional practice, but in case of difficulty, you can take screenshots of code and outputs and submit them in a Word/PDF file.
- Maximum number of pages should be 15. Any submissions that exceed 15 pages will **not be graded.**

Grading and Scoring:

- Use common sense to gauge the expectations of the answer to the number of points assigned to the question. See the Scoring Rubric in Blackboard for details.

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Assignment 2.1 - Questions

1. (20 points) The soybean data can also be found at the UC Irvine Machine Learning Repository. Data were collected to predict disease in 683 soybeans. The 35 predictors are mostly categorical and include information on the environmental conditions (e.g., temperature, precipitation) and plant conditions (e.g., left spots, mold growth). The outcome labels consist of 19 distinct classes. The data can be loaded via:
   library(mlbench)
   data(Soybean)
   # Use ?Soybean for details
   a. Investigate missing values for all the predictors, which predictors have the highest and the lowest number of missing values?Do the missing values depend on the outcome labels (summarize NAs by class)? Are any of the distributions degenerate in the ways discussed earlier in this chapter? (10 points)
   b. Compute what % of the predictor data is missing? Compute percent of missing data by outcome label and identify classes with highest percent missing values. An example of this calculation: if there are two predictors and 10 rows of data with 5 missing data points then % missing = 5 * 100 / (2*10) = 25% (10 points)

2. (10 points) The caret package contains a QSAR data set from Mente and Lombardo (2005). Here, the ability of a chemical to permeate the blood-brain barrier was experimentally determined for 208 compounds. 134 descriptors were measured for each compound.
   a. Start R and use these commands to load the data:
      library(caret)
      data(BloodBrain)
      # use ?BloodBrain to see more details
      The numeric outcome is contained in the vector logBBB while the predictors are in the data frame bbbDescr.
   b. Do any of the individual predictors have degenerate distributions? (5 points)
   c. Generally speaking, are there strong relationships between the predictor data? If so, how could correlations in the predictor set be reduced? Does this have a dramatic effect on the number of predictors available for modeling? (5 points)

3. (10 points) Consider the permeability data set described in Sect. 1.4. of the textbook. The objective for this data is to use the predictors to model compounds “permeability”.

Load the data as shown below:

```
library(AppliedPredictiveModeling)
data(permeability) # this creates two matrices fingerprints and permeability
permeabilitydf <- as.data.frame(fingerprints)
permeabilitydf$permeability <- permeability
```

a. What data splitting method(s) would you use for these data? Explain. (5 points)
b. Using tools described in this chapter, provide code for implementing your approach(es). (5 points)

4. (20 points) Partial least squares was used to model the yield of a chemical manufacturing process (Sect. 1.4). The data can be found in the AppliedPredictiveModeling package and can be loaded using

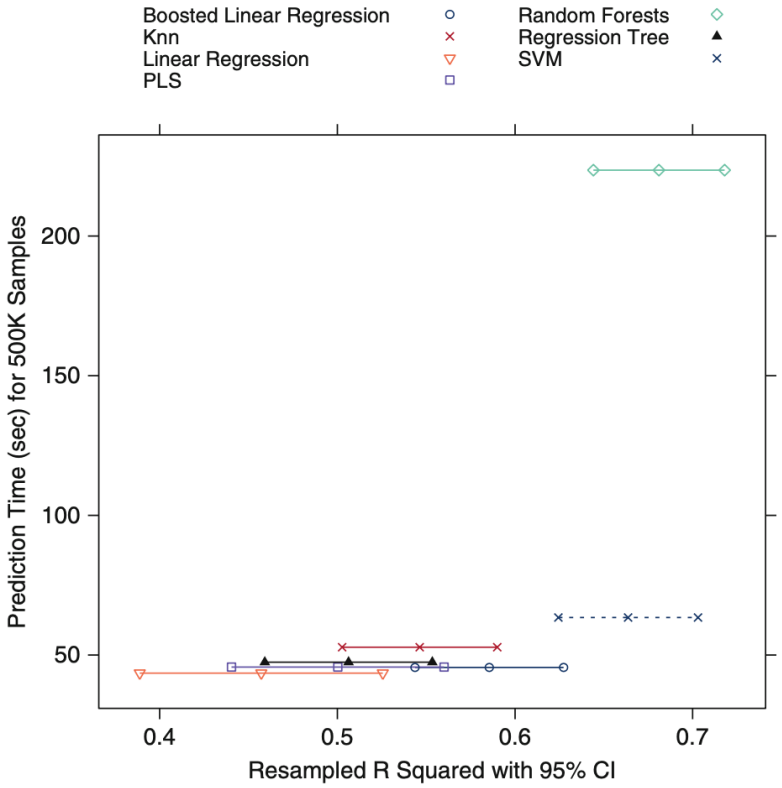
```
library(AppliedPredictiveModeling)
data(ChemicalManufacturingProcess)
```

The objective of this analysis is to find the number of PLS components that yields the optimal R^2 value. PLS models with 1 through 10 components were each evaluated using five repeats of 10-fold cross-validation and the results are presented in the following table:

	Resampled R^2	
Components	Mean	Std. Error
1	0.444	0.0272
2	0.500	0.0298
3	0.533	0.0302
4	0.545	0.0308
5	0.542	0.0322
6	0.537	0.0327
7	0.534	0.0333
8	0.534	0.0330
9	0.520	0.0326
10	0.507	0.0324

a. Using the “one-standard error” method, what number of PLS components provides the most parsimonious model? (7 points)
b. If a 10% loss in optimal R^2 is acceptable, then what is the optimal number of PLS components? (7 points)

c. Several other models with varying degrees of complexity were trained and tuned and the results are presented in Figure below. If the goal is to select the model that optimizes R^2, then which model(s) would you choose, and why? (3 points)



d. Prediction time, as well as model complexity are other factors to consider when selecting the optimal model(s). Given each model's prediction time, model complexity, and R2 estimates, which model(s) would you choose, and why? (3 points).

5. (30 points) Brodnjak-Vonina et al. (2005) develop a methodology for food laboratories to determine the type of oil from a sample. In their procedure, they used a gas chromatograph (an instrument that separates chemicals in a sample) to measure seven different fatty acids in an oil. These measurements would then be used to predict the type of oil in a food sample. To create their model, they used 96 samples of seven types of oils.

   These data can be found in the caret package using data(oil). The oil types are contained in a factor variable called oilType. The types are pumpkin (coded as A), sunflower (B), peanut (C), olive (D), soybean (E), rapeseed (F) and corn (G).

a. Use the sample function in base R to create a completely random sample of 60 oils. How closely do the frequencies of the random sample match the original samples? Repeat this procedure several times to understand the variation in the sampling process. (10 points)
b. Use the caret package function createDataPartition to create a stratified random sample. How does this compare to the completely random samples? (5 points)
c. With such a small sample size, what are the options for determining performance of the model? Should a test set be used? (5 points)
d. One method for understanding the uncertainty of a test set is to use a confidence interval. To obtain a confidence interval for the overall accuracy, the based R function binom.test can be used. It requires the user to input the number of samples and the number correctly classified to calculate the interval. For example, suppose a test set sample of 20 oil samples was set aside and 76 were used for model training. For this test set size and a model that is about 80 % accurate (16 out of 20 correct), the confidence interval would be computed using binom.test(16, 20)

```
        Exact binomial test
data: 16 and 20
number of successes = 16, number of trials = 20, p-value = 0.01182
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
0.563386 0.942666
sample estimates:
probability of success
          0.8
```

In this case, the width of the 95% confidence interval is 37.9%. Try different sample sizes and accuracy rates to understand the trade-off between the uncertainty in the results, the model performance, and the test set size. (10 points).