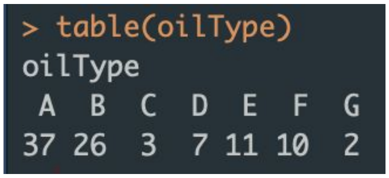
Assignment-3

Sai Hemanth Nirujogi
G01065588

Question 1

In Exercise 4.4, we described a data set which contained 96 oil samples each from one of seven types of oils (pumpkin, sunflower, peanut, olive, soybean, rapeseed, and corn). Gas chromatography was performed on each sample and the percentage of each type of 7 fatty acids was determined. We would like to use these data to build a model that predicts the type of oil based on a sample's fatty acid percentages.

a. *Like the hepatic injury data, these data suffer from extreme imbalance. Given this imbalance, should the data be split into training and test sets?*



```
> table(oilType)
oilType
 A  B  C  D  E  F  G 
37 26  3  7 11 10  2 
```

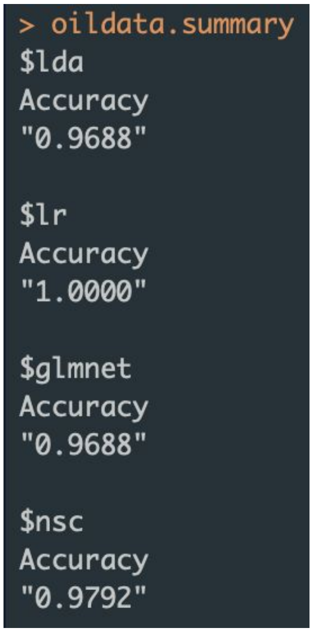
The data should not be split into training and test sets because the variables C, D, and G have less number of samples.

b. *Which classification statistic would you choose to optimize for this exercise and why?*

Accuracy would be the best classification statistic because it gives the percentage of the total items classified correctly.

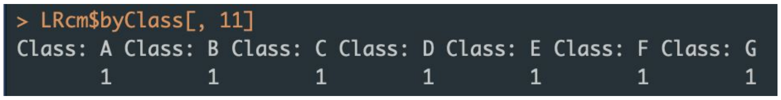
c. *Of the models presented in this chapter, which performs best on these data? Which oil type does the model most accurately predict? Least accurately predict?*

As we can see from the results on the right, Linear regression has the perfect performance. The Nearest Shrunken Centroids model also has the next better accuracy with some errors. The model that has the least accurate predictions is the LDA and GLMNET.



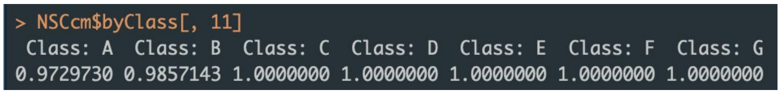
```
> oildata.summary
$lda
Accuracy
"0.9688"

$lr
Accuracy
"1.0000"

$glmnet
Accuracy
"0.9688"

$nsc
Accuracy
"0.9792"
```

1. The LR model has predicted all the oil types accurately. (LRcm$byClass[, 11])



```
> LRcm$byClass[, 11]
Class: A Class: B Class: C Class: D Class: E Class: F Class: G 
       1        1        1        1        1        1        1 
```

2. The NSC model has predicted the Type C, D, E, F, and G oil types more accurately. (NSCcm$byClass[, 11])



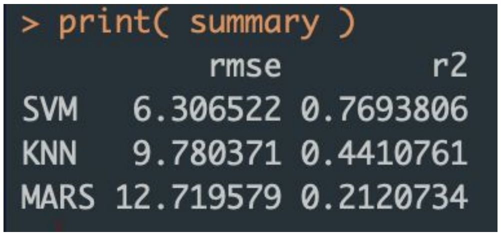
```
> NSCcm$byClass[, 11]
 Class: A  Class: B  Class: C  Class: D  Class: E  Class: F  Class: G 
0.9729730 0.9857143 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 
```

Question 2

Return to the permeability problem outlined in Exercise 6.2. Train several nonlinear regression models and evaluate the resampling and test set performance.

a. *Which nonlinear regression model gives the optimal resampling and test set performance?*

   RMSE was used to select the optimal model, which has the smallest value. From the summary results below, we can see that the SVM model performs the best compared to the KNN and MARS on the test set.

   ```
   > print( summary )
             rmse        r2
   SVM   6.306522 0.7693806
   KNN   9.780371 0.4410761
   MARS 12.719579 0.2120734
   ```

b. *Do any of the nonlinear models outperform the optimal linear model you previously developed in Exercise 6.2? If so, what might this tell you about the underlying relationship between the predictors and the response?*

   There is no linear relationship between the predictor and the response variable. So, that is the reason why the non-linear model i.e., SVM outperformed the optimal linear models.

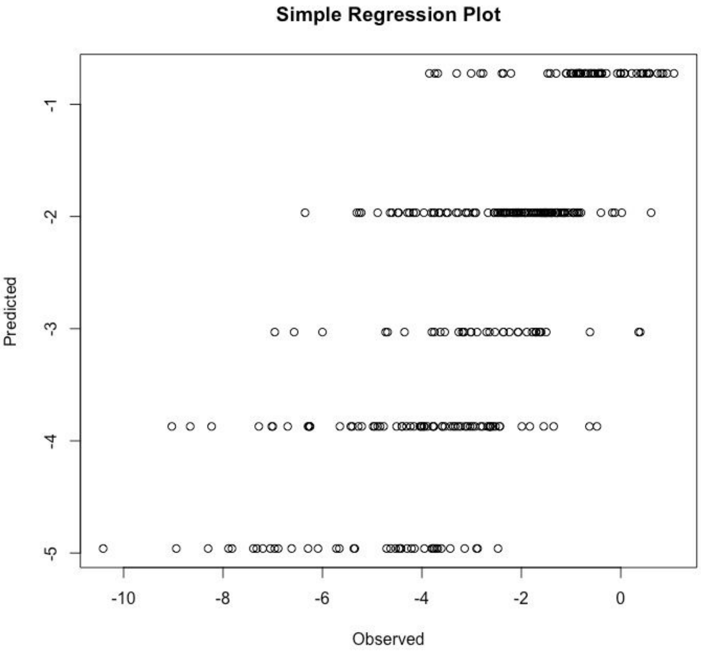
Question 3

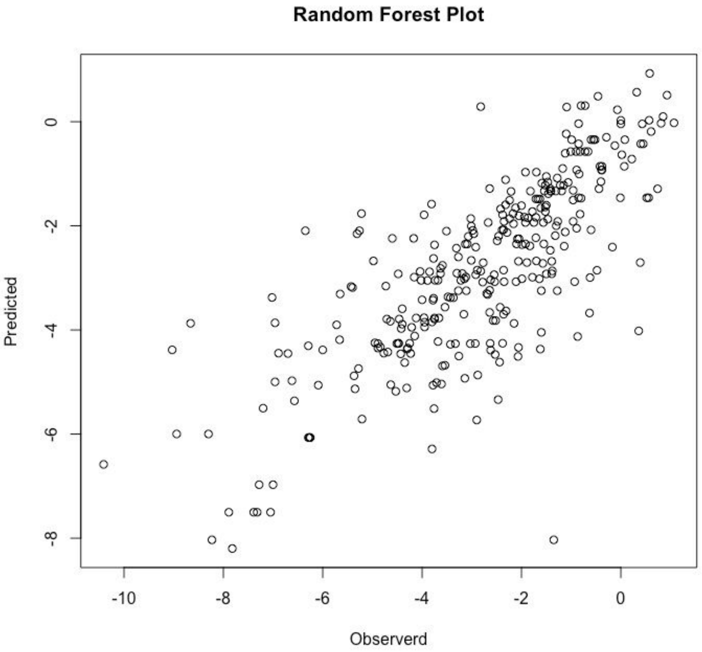
Use a single predictor in the solubility data, such as the molecular weight or the number of carbon atoms and fit several models:

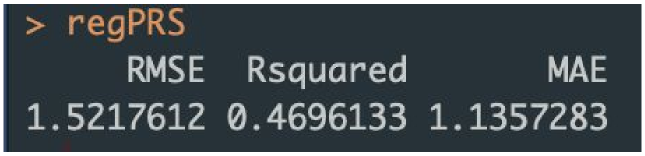
a. *A simple regression tree*
b. *A random forest model*

Plot the predictor data versus the solubility results for the test set. Overlay the model predictions for the test set. How does the model differ? Does changing the tuning parameter(s) significantly affect the model fit?

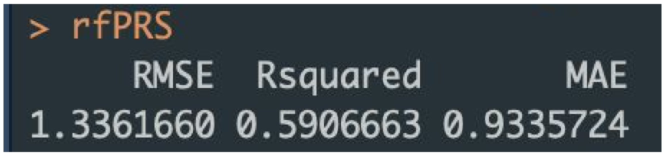
The Random Forest Model performs the best compared to the Regression Tree Model. By using the randomness, the Random Forest Model improves the predictive ability. No, changing the tuning parameter does not affect the fit of the model.







```
> regPRS
     RMSE  Rsquared       MAE
1.5217612 0.4696133 1.1357283
```



```
> rfPRS
     RMSE  Rsquared       MAE
1.3361660 0.5906663 0.9335724
```

Prediction results of the Simple Regression Model on the test set

Prediction results of the Random Forest Model on the test set