Ex13

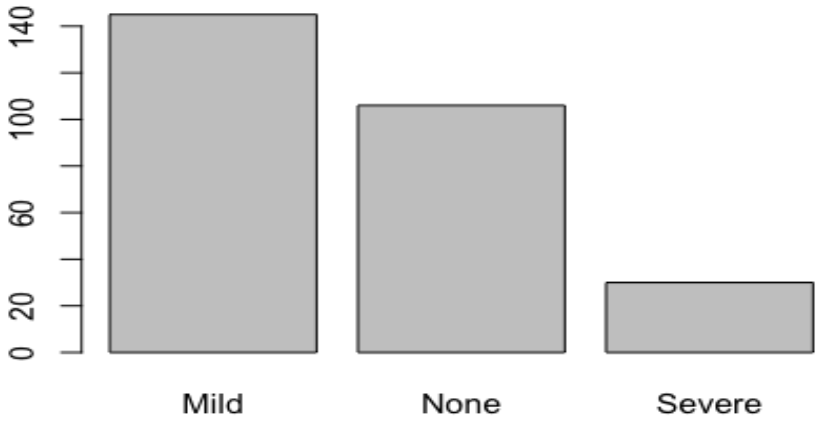
Chathrua Gunasekara

1.a Preprocessing steps done in chapter 12 are repeated in the same way in this exercise too.

Preprocessing done on both bio and chem and combined data sets.

i. Remove nearzero variance predictors
ii. Remove high correlated predictors
iii. Remove linear combination predictors
iv. Splitting data set using stratified sampling

Following diagram illustrates the class distribution in predictor variable.



1. Mixture Discriminant Analysis

```
225 samples
 96 predictor
  3 classes: 'Mild', 'None', 'Severe'

Pre-processing : Center and Scale
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 225, 225, 225, 225, 225, 225, ...
Resampling results across tuning parameters:

  subclasses  Accuracy   Kappa       Accuracy SD  Kappa SD
  1           0.4145599  0.05659400  0.05389300   0.06594883
  2           0.4239917  0.06274809  0.03231601   0.04932816
  3           0.4361365  0.07152077  0.03558633   0.05913592
  4           0.3832163  0.01625659  0.02123588   0.02585894
  5           0.4129537  0.05560859  0.06633504   0.08922207
```

```
Kappa was used to select the optimal model using  the largest value.
The final value used for the model was subclasses = 3.

Confusion Matrix and Statistics



          Reference
Prediction Mild None Severe
    Mild     19   12      3
    None      6    9      2
    Severe    4    0      1

Overall Statistics for Testing set

               Accuracy : 0.5179
                 95% CI : (0.3803, 0.6534)
    No Information Rate : 0.5179
    P-Value [Acc > NIR] : 0.5537

                  Kappa : 0.1424
 Mcnemar's Test P-Value : 0.2464

Statistics by Class:

                     Class: Mild Class: None Class: Severe
Sensitivity               0.6552      0.4286       0.16667
Specificity               0.4444      0.7714       0.92000
```

2. Neural Network

```
225 samples
 96 predictor
  3 classes: 'Mild', 'None', 'Severe'

Pre-processing: spatial sign transformation, scaled, centered
Resampling: Bootstrapped (25 reps)

Summary of sample sizes: 225, 225, 225, 225, 225, 225, ...



Resampling results across tuning parameters:

  size  decay  Accuracy   Kappa          Accuracy SD  Kappa SD
   1    0.0    0.3912201  -0.0005849393  0.07975896   0.05673328
   1    0.1    0.4407795  -0.0262458622  0.04372911   0.07310105
   1    1.0    0.4600638  -0.0362919794  0.04409184   0.06066712
   1    2.0    0.4921698  -0.0068195868  0.07045113   0.01971180
   2    0.0    0.3856094  -0.0004402377  0.08293128   0.07808454
```

```
   5    0.0    0.4096316  -0.0014802390  0.05190740   0.07650497
   5    0.1    0.4270451   0.0019751218  0.04324519   0.06628027
   5    1.0    0.4633821  -0.0181128791  0.04928725   0.07184842
   5    2.0    0.4947387  -0.0061781221  0.06236068   0.01782982
   6    0.0    0.4095587   0.0010321192  0.04626995   0.05163436
   6    0.1    0.4321282   0.0098848259  0.04536148   0.07110344
   6    1.0    0.4618823  -0.0210948650  0.04883736   0.07177771
   6    2.0    0.4947387  -0.0061781221  0.06236068   0.01782982
   7    0.0    0.4185346   0.0122778107  0.05351421   0.08608156
   7    0.1    0.4320596   0.0091119355  0.04101778   0.06640802
   7    1.0    0.4628187  -0.0189470605  0.04884592   0.07115908
   7    2.0    0.4947387  -0.0061781221  0.06236068   0.01782982
   8    0.0    0.4288824   0.0154870583  0.04395671   0.06638126
   8    0.1    0.4280180   0.0038667622  0.04409914   0.06890253
   8    1.0    0.4623368  -0.0200885878  0.04870877   0.07126549
   8    2.0    0.4947387  -0.0061781221  0.06236068   0.01782982
   9    0.0    0.4200026  -0.0016920521  0.04874879   0.07801793
   9    0.1    0.4280546   0.0062862751  0.04730466   0.07316046
   9    1.0    0.4617518  -0.0207221916  0.04745530   0.07010182
   9    2.0    0.4947387  -0.0061781221  0.06236068   0.01782982

Kappa was used to select the optimal model using  the largest value.
The final values used for the model were size = 8 and decay = 0.

Confusion Matrix and Statistics Testing set

          Reference
Prediction Mild None Severe
    Mild     20   13      5
    None      7    7      1
    Severe    2    1      0

Overall Statistics

               Accuracy : 0.4821
                 95% CI : (0.3466, 0.6197)
    No Information Rate : 0.5179
    P-Value [Acc > NIR] : 0.7482

                  Kappa : 0.0453


Statistics by Class:

                     Class: Mild Class: None Class: Severe
Sensitivity               0.6897      0.3333       0.00000
Specificity               0.3333      0.7714       0.94000
```

3. Flexible Discriminant Analysis

```
  225 samples
   96 predictor
    3 classes: 'Mild', 'None', 'Severe'

  Pre-processing : Center and Scale
  Resampling: Bootstrapped (25 reps)

  Summary of sample sizes: 225, 225, 225, 225, 225, 225, ...

  Resampling results across tuning parameters:

    nprune  Accuracy   Kappa        Accuracy SD  Kappa SD
     2      0.4861072  0.006071507  0.05023715   0.05087418
    35      0.4402642  0.046216816  0.04940299   0.08463509
    69      0.4361744  0.048250463  0.05527767   0.08147733

  Tuning parameter 'degree' was held constant at a value of 1
  Kappa was used to select the optimal model using  the largest value.
  The final values used for the model were degree = 1 and nprune = 69.
```

Confusion Matrix and Statistics **Testing set**

Prediction	Reference Mild	None	Severe
Mild	25	16	5
None	1	4	1
Severe	3	1	0

Overall Statistics

Accuracy	0.5179
95% CI	(0.3803, 0.6534)
No Information Rate	0.5179
P-Value [Acc > NIR]	0.553730
Kappa	0.0847
Mcnemar's Test P-Value	0.003289

Statistics by Class:

	Class: Mild	Class: None	Class: Severe
Sensitivity	0.8621	0.19048	0.00000
Specificity	0.2222	0.94286	0.92000

4. Support Vector Machines with Radial Basis Function Kernel

```
225 samples
 96 predictor
  3 classes: 'Mild', 'None', 'Severe'

Pre-processing : Center and Scale
Resampling: Bootstrapped (25 reps)

Summary of sample sizes: 225, 225, 225, 225, 225, 225, ...

Resampling results across tuning parameters:

  C        Accuracy   Kappa         Accuracy SD  Kappa SD
   0.0625  0.5121283   0.000000000  0.04002546   0.000000000
   0.1250  0.5121283   0.000000000  0.04002546   0.000000000
   0.2500  0.5097409   0.000987425  0.03592039   0.008526046
   0.5000  0.4977915  -0.009749648  0.03509011   0.028841709
   1.0000  0.4932844   0.005611551  0.04567346   0.076477763
   2.0000  0.4863447   0.012146986  0.04611067   0.097487928
   4.0000  0.4837670   0.035061047  0.03573951   0.073209728
   8.0000  0.4860284   0.063532606  0.03820134   0.071332936
  16.0000  0.4700797   0.048854218  0.04206646   0.076239954

Tuning parameter 'sigma' was held constant at a value of 0.002492319
Kappa was used to select the optimal model using  the largest value.
The final values used for the model were sigma = 0.002492319 and C = 8.
```

Confusion Matrix and Statistics **Testing set**

```
          Reference
Prediction Mild None Severe
    Mild     22   13      5
    None      6    8      1
    Severe    1    0      0

Overall Statistics

               Accuracy : 0.5357
                 95% CI : (0.3974, 0.6701)
    No Information Rate : 0.5179
    P-Value [Acc > NIR] : 0.4475

                  Kappa : 0.1202
 Mcnemar's Test P-Value : 0.1003

Statistics by Class:

                     Class: Mild Class: None Class: Severe
Sensitivity               0.7586      0.3810       0.00000
Specificity               0.3333      0.8000       0.98000
```

5. k-Nearest Neighbors

```
225 samples
 96 predictor
  3 classes: 'Mild', 'None', 'Severe'
Pre-processing : Center and Scale
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 225, 225, 225, 225, 225, 225, ...
Resampling results across tuning parameters:
```

k	Accuracy	Kappa	Accuracy SD	Kappa SD
1	0.4798202	0.096591515	0.05429008	0.085075079
5	0.4644326	0.068348020	0.05141708	0.068566556
9	0.4749089	0.055596568	0.05450278	0.076484795
13	0.5054816	0.077005867	0.06561231	0.080430756
17	0.5148153	0.079358231	0.06213484	0.069301861
251	0.5256963	0.000000000	0.04019658	0.000000000
301	0.5256963	0.000000000	0.04019658	0.000000000
351	0.5256963	0.000000000	0.04019658	0.000000000
401	0.5256963	0.000000000	0.04019658	0.000000000
451	0.5256963	0.000000000	0.04019658	0.000000000

```
Kappa was used to select the optimal model using  the largest value.
The final value used for the model was k = 13.
```

Confusion Matrix and Statistics **Testing set**

Prediction	Reference Mild	None	Severe
Mild	27	16	5
None	2	5	0
Severe	0	0	1

```
Overall Statistics

               Accuracy : 0.5893
                 95% CI : (0.4498, 0.719)
    No Information Rate : 0.5179
    P-Value [Acc > NIR] : 0.1747
```

```
                  Kappa : 0.1904
   Mcnemar's Test P-Value : NA

  Statistics by Class:

                       Class: Mild Class: None Class: Severe
  Sensitivity               0.9310     0.23810       0.16667
  Specificity               0.2222     0.94286       1.00000
```

6. Naive Bayes

```
225 samples

 96 predictor

  3 classes: 'Mild', 'None', 'Severe'

Pre-processing : Center and Scale

Resampling: Bootstrapped (25 reps)

Summary of sample sizes: 225, 225, 225, 225, 225, 225, ...

Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa       Accuracy SD  Kappa SD

  FALSE            NaN         NaN         NA             NA

   TRUE      0.2618047  0.02044771  0.1022174    0.04616507


Tuning parameter 'fL' was held constant at a value of 0

Kappa was used to select the optimal model using  the largest value.

The final values used for the model were fL = 0 and usekernel = TRUE.


  Confusion Matrix and Statistics Testing set

            Reference
  Prediction Mild None Severe
      Mild      2    1      1
      None      3    4      0
      Severe   24   16      5

  Overall Statistics

                 Accuracy : 0.1964
                   95% CI : (0.1023, 0.3243)
      No Information Rate : 0.5179
      P-Value [Acc > NIR] : 1

                    Kappa : 0.0319
```

```
   Mcnemar's Test P-Value : 2.614e-08


Statistics by Class:

                     Class: Mild Class: None Class: Severe
  Sensitivity              0.06897     0.19048       0.83333
  Specificity              0.92593     0.91429       0.20000
```

From Ex 12 :

FOR Testing set:

LIEAR MODEL	Accuracy	Kappa
Logistic Reg (averaged)	0.5833	0.02
LDA	0.5179	0.102
PLSDA	0.5893	0.04
NSC	0.625	0.07

NON LIEAR MODEL	Accuracy	Kappa
MDA	***0.5179***	***0.1424***
NNet	0.4821	0.0453
FDA	0.5179	0.0847
SVM	0.5357	0.1202
KNN	0.5893	0.1904
Naïve Bayes	0.1964	0.0319

Best Models for Biological predictors is MDA model. Yes it does do a better job than all of the Linear models from chapter 12 for the biological data.