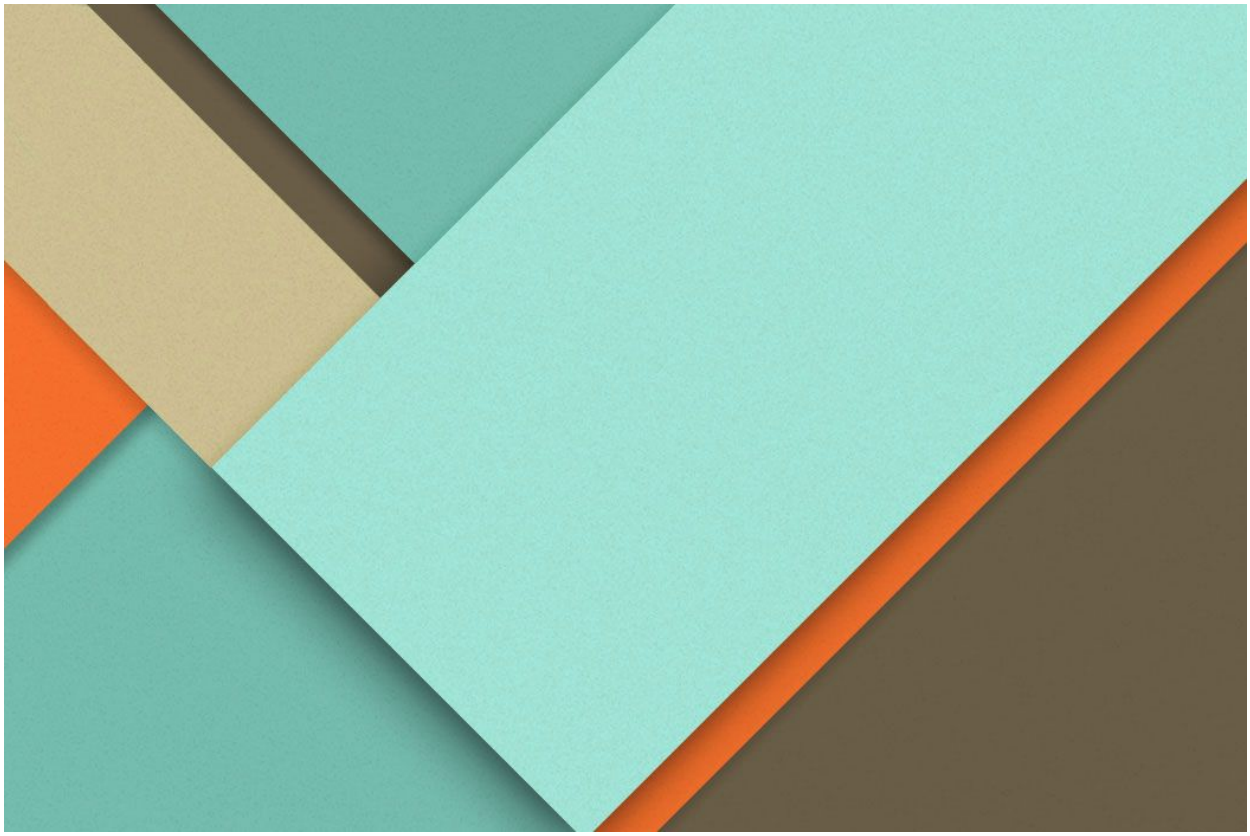


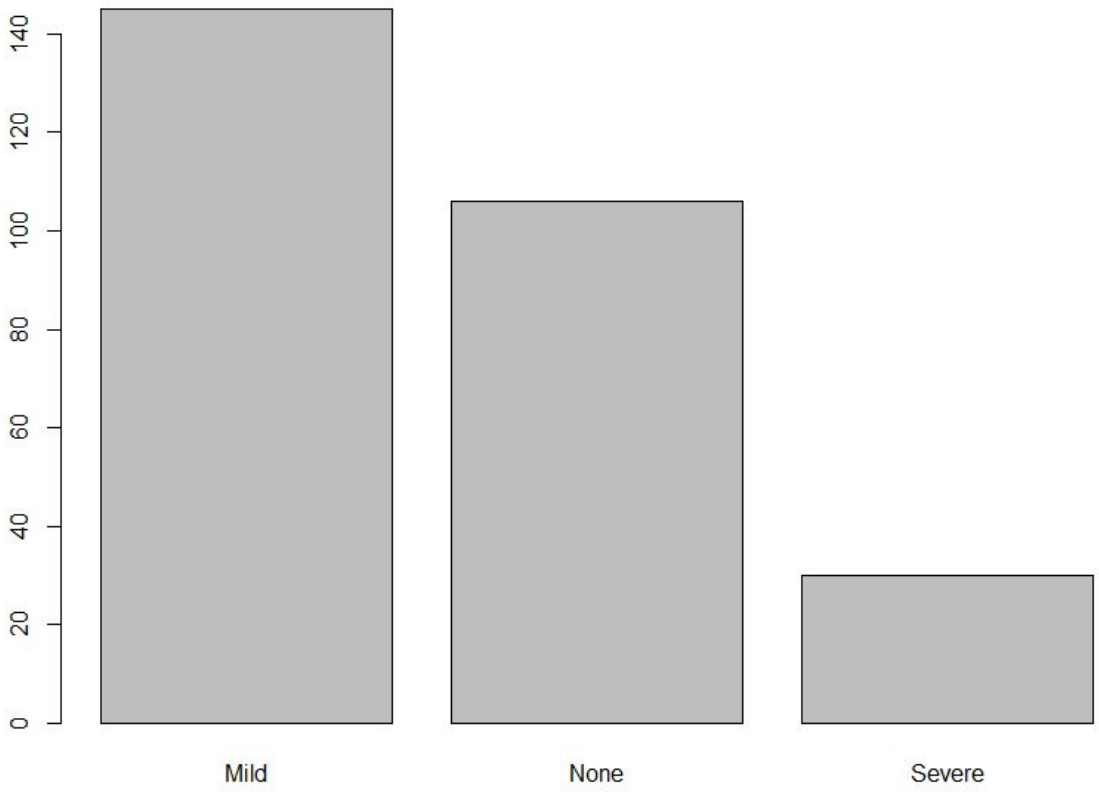
Homework 05

Supriya Bachal

Question:12.1

A.

The injury status shows very high imbalance in classes as seen in the barplot below.



Statistically the class wise count of each of the injuries is as below:

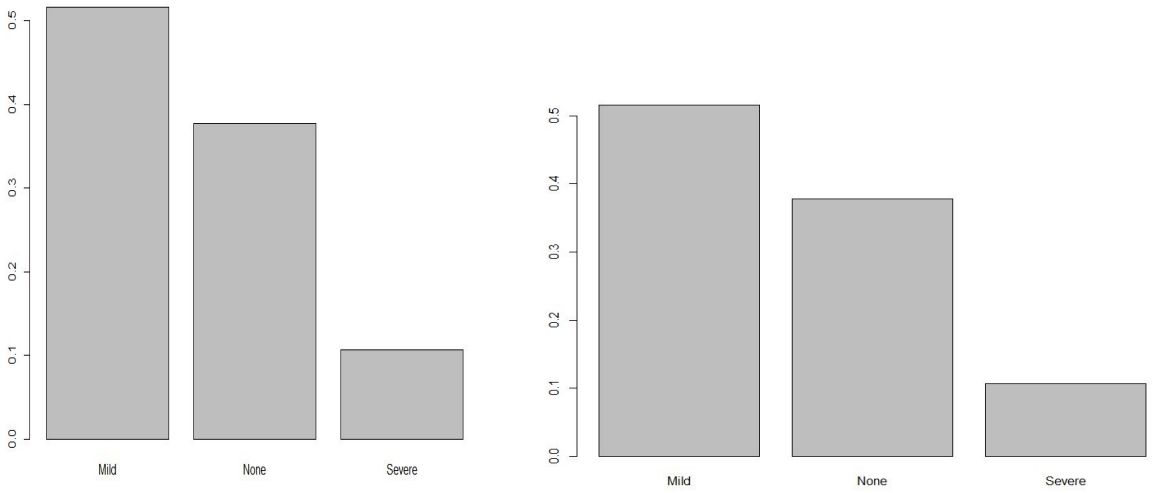
injury
Mild None Severe
145 106 30

Since the instances are not well distributed among the classes We use :

t = createDataPartition(injury, p=0.8)[[1]]

This creates one partition with 80 percent for training and 20 percent for testing data.

From this we conclude the class distribution: for the original and partitioned data,



From the graphs above we can conclude that the distribution for the classes after partitioning remains the same we repeat the procedure above with ten folds.

We observe that even with ten folds the class distribution is maintained. Hence we can say that we can use data partition as a valid method for creating training and testing sets.

To bring a balance to the dataset we can either oversample the data or undersample it. By oversampling the data we will have to add more instances of severe cases and by undersampling it we delete instances of mild and none cases. In this case we cannot undersample the data as that would shrink our data and we do not have enough data to afford to shrink it.

Oversampling of data maybe a viable option however we can compensate the imbalance of classes by using different techniques to measure our model appropriateness instead of using this approach.

b) It is evident that the class distribution for this dataset is highly imbalanced. If we use accuracy as to measure the appropriateness of the model we will not be able to get to a definitive conclusion, because even if all the instances of severe get misclassified the overall impact of this will be very meagre as the severe cases form about 1 perc of the data.

IWe can then use Kappa or cohens kappa to measure the performance of the model as this metric takes into account the distribution of the classes as well while determining the precision of the model.However we can also use accuracy of individual predictors to establish a conclusion. In this case it is more important to make sure that the cases that are mild and severe are classified correctly as compared to the cases that are none,since it is very important to determine a sick person as compare to misclassify a non-sick person.

ROC curves can also be used as they take into account specificity and sensitivity of the models.The ROC curve is only defined for two-class problems but has been extended to handle three or more classes. Hand and Till (2001), Lachiche and Flach (2003), and Li and Fine (2008) use different approaches extending the definition of the ROC curve with more than two classes.

c)Preprocessing the data:

First we drop the near zero variance and zero variance predictors from the bio data.Dropping 82 zero variance columns from 184 (fraction= 0.445652).

Then we drop near zero and zero variance predictors from the chem data.Dropping 58 zero variance columns from 192 (fraction= 0.302083).

The next step is to find any correlation between the data;

Then we divided the data into training and testing sets for bio and chem predictors using :

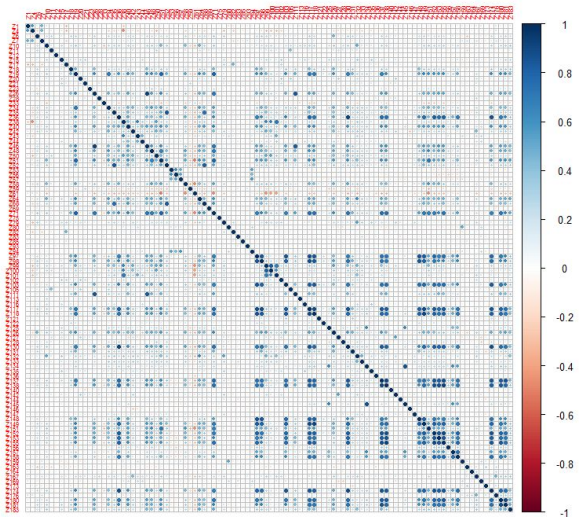
training_bio <- X[t,]

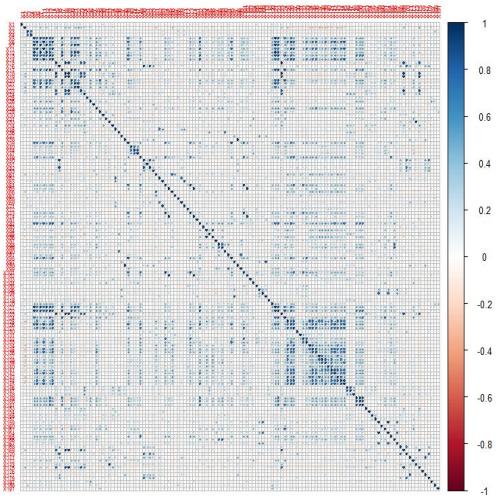
test_bio <- X[-t,]

training_chem <- Y[t,]

test_chem <- chem[-t,]

Further we found the correlation between the predictors to check for linearly dependent columns.The left figure is depicting the correlation in bio predictors and the right one in chem predictors.





We see that in both the datasets there are some very evident correlations. We have the cut off to 90 percent meaning that the correlations greater than 90 percent have been dropped.

This is the list of predictors for the bio predictors that has been dropped:

68 88 15 75 61 76

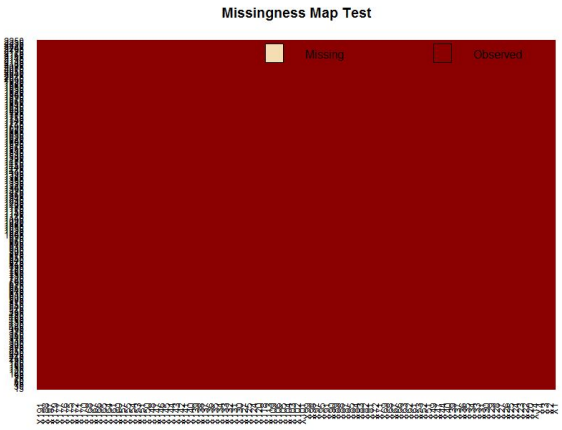
6 predictors have been dropped,96 are used to build the model.

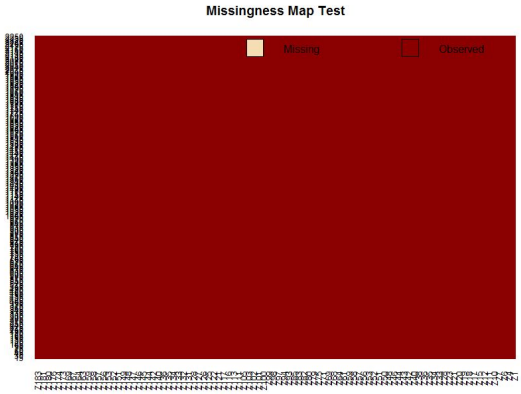
This is the list of predictors for chem which has been dropped

9 10 56 62 82 83 84 87 127 5 6 8 12 15 14 7 18 16 13 40 47 34 55 48 108 115 123

27 predictors have been dropped,107 predictors are used to build the model.

Moving on we try to find the characteristics of the data:





We observe no missing values in either of the datasets.

From the histograms we observe that the data is skewed. We carry,out certain transformation to curate the data.

We use the preprocess function to center and scale the data. We also use box cox transformation on the data.

Lambda estimates for Box-Cox transformation: is 0.1 for the chem data and Lambda estimates for Box-Cox transformation:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2	-2	-2	-2	-2	-2

For the bio data.

So now we have the curated data and we can model it in diffrent ways:

First We implemented the Linear Discriminant Analysis Model using the train function.Kappa metric was used and we used 10 fold cross validation considering Class Probabilities:

For Chemical dataset the results were as follows

Linear Discriminant Analysis

225 samples

73 predictor

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

Pre-processing: centered (73), scaled (73)

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 202, 203, 203, 203, 202, 202, ...

Resampling results:

logLoss AUC prAUC Accuracy Kappa Mean_F1 Mean_Sensitivity Mean_Specificity

2.317796 0.5433545 0.3486974 0.4573123 0.06407637 0.477451 0.3722222 0.6869756

Mean_Pos_Pred_Value Mean_Neg_Pred_Value Mean_Precision Mean_Recall Mean_Detection_Rate

0.3893543 0.6877642 0.3893543 0.3722222 0.1524374

Mean_Balanced_Accuracy

0.5295989

Linear Discriminant Analysis for BIO dataset

225 samples

81 predictor

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

Pre-processing: centered (81), scaled (81)

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 202, 203, 203, 203, 202, 202, ...

Resampling results:

logLoss AUC prAUC Accuracy Kappa Mean_F1 Mean_Sensitivity Mean_Specificity

2.26933 0.5635636 0.361539 0.4531621 0.06909249 0.3658304 0.3744108 0.6904067

Mean_Pos_Pred_Value Mean_Neg_Pred_Value Mean_Precision Mean_Recall Mean_Detection_Rate

0.3698715 0.6916326 0.3698715 0.3744108 0.151054

Mean_Balanced_Accuracy

0.5324088

The next model mentioned in the textbook is the Logistic regression model however the logistic regression model works only for 2 classes,since we have 3 classes present in this dataset.So we use the method as multinom instead of glm.

The results for the same are as follows:

Penalized Multinomial Regression

225 samples

73 predictor

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

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 202, 203, 202, 203, 203, 204, ...

Resampling results across tuning parameters:

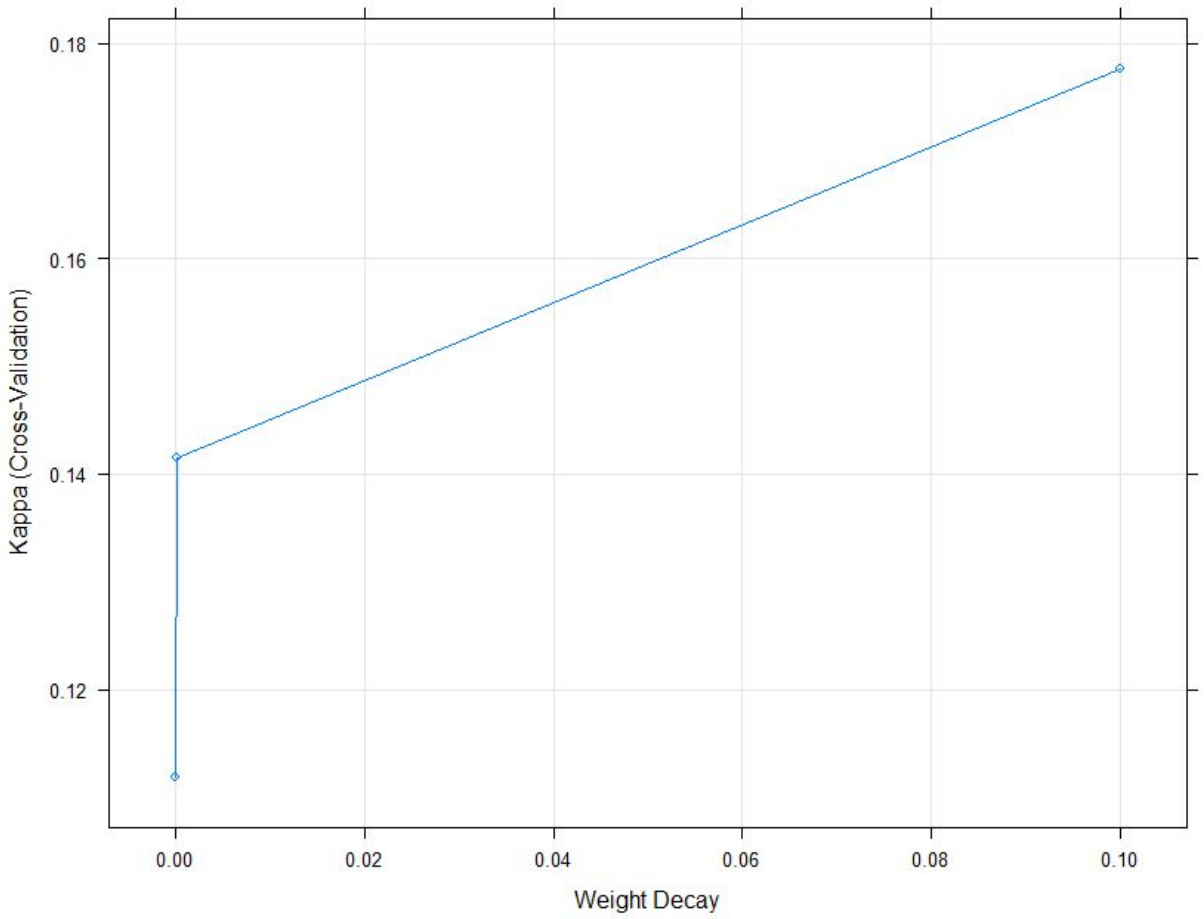
decay	logLoss	AUC	prAUC	Accuracy	Kappa	Mean_F1	Mean_Sensitivity
0e+00	14.632861	0.5901916	0.3090460	0.4581875	0.1118217	0.3748149	0.3812290
1e-04	10.350479	0.5925149	0.3714689	0.4813476	0.1414621	0.4210462	0.3981481
1e-01	2.306294	0.5918309	0.3833459	0.5216968	0.1776414	0.4956854	0.4140152

Mean_Specificity	Mean_Pos_Pred_Value	Mean_Neg_Pred_Value	Mean_Precision	Mean_Recall
0.7063251	0.4200482	0.7071258	0.4200482	0.3812290
0.7173407	0.4393976	0.7165716	0.4393976	0.3981481
0.7293010	0.4220483	0.7305849	0.4220483	0.4140152

Mean_Detection_Rate	Mean_Balanced_Accuracy
0.1527292	0.5437770
0.1604492	0.5577444
0.1738989	0.5716581

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

The final value used for the model was decay = 0.1.



Penalized Multinomial Regression

225 samples

81 predictor

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

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 203, 202, 203, 202, 202, 202, ...

Resampling results across tuning parameters:

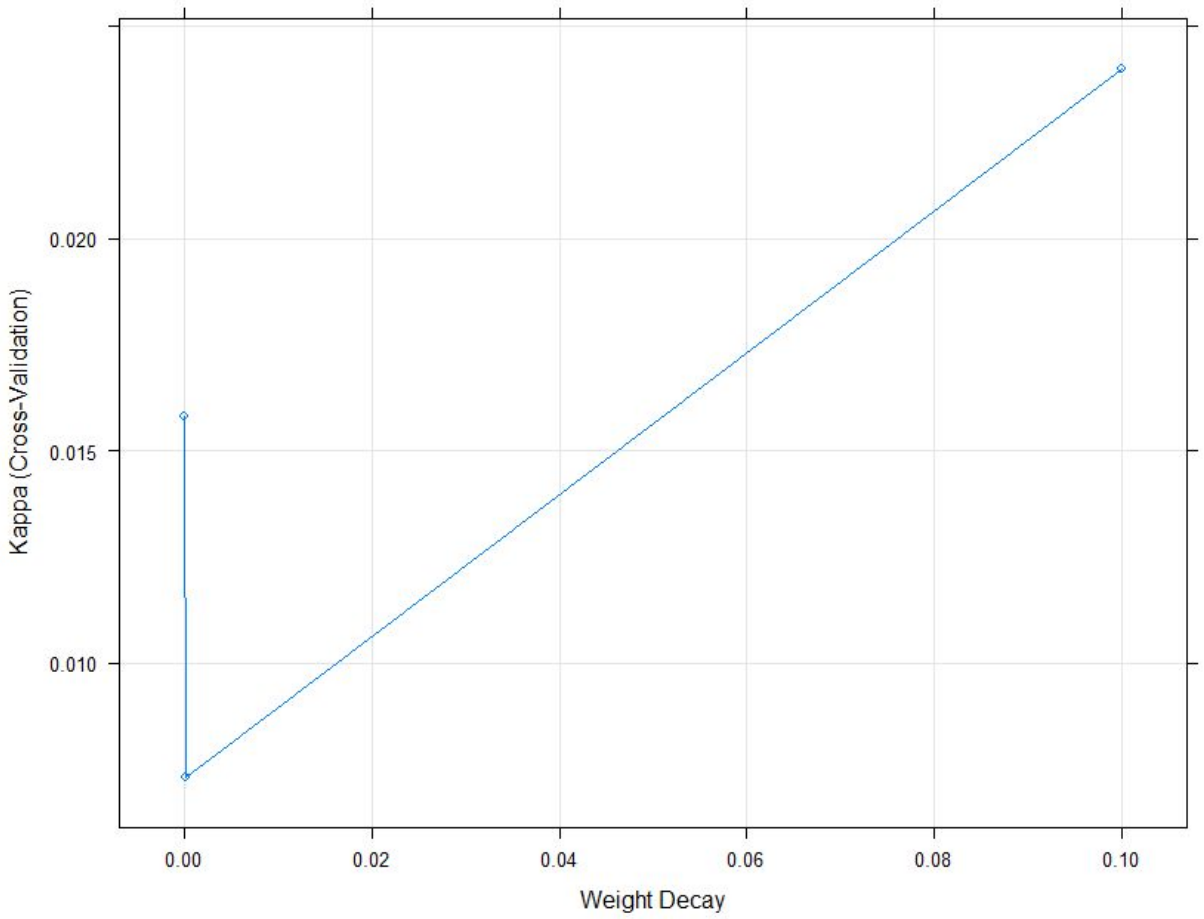
decay	logLoss	AUC	prAUC	Accuracy	Kappa	Mean_F1	Mean_Sensitivity
0e+00	13.404804	0.4967315	0.3168765	0.3778656	0.015818318	0.3327130	0.3164562
1e-04	10.442906	0.5051844	0.3334929	0.3822134	0.007295325	0.3837283	0.3055976
1e-01	1.745513	0.5649840	0.3527305	0.4444664	0.023997201	0.5134556	0.3371212

Mean_Specificity	Mean_Pos_Pred_Value	Mean_Neg_Pred_Value	Mean_Precision	Mean_Recall
0.6772703	0.3344635	0.6733671	0.3344635	0.3164562
0.6751679	0.3316656	0.6691918	0.3316656	0.3055976
0.6758563	0.3542643	0.6814519	0.3542643	0.3371212

Mean_Detection_Rate	Mean_Balanced_Accuracy
0.1259552	0.4968633
0.1274045	0.4903828
0.1481555	0.5064887

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

The final value used for the model was decay = 0.1


Kappa (Cross-Validation)
0.020
0.015
0.010
0.00
0.02
0.04
0.06
0.08
0.10
Weight Decay

PLSDA model:

Chemical dataset

Partial Least Squares

225 samples

73 predictor

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

Pre-processing: centered (73), scaled (73)

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 202, 203, 203, 203, 202, 202, ...

Resampling results across tuning parameters:

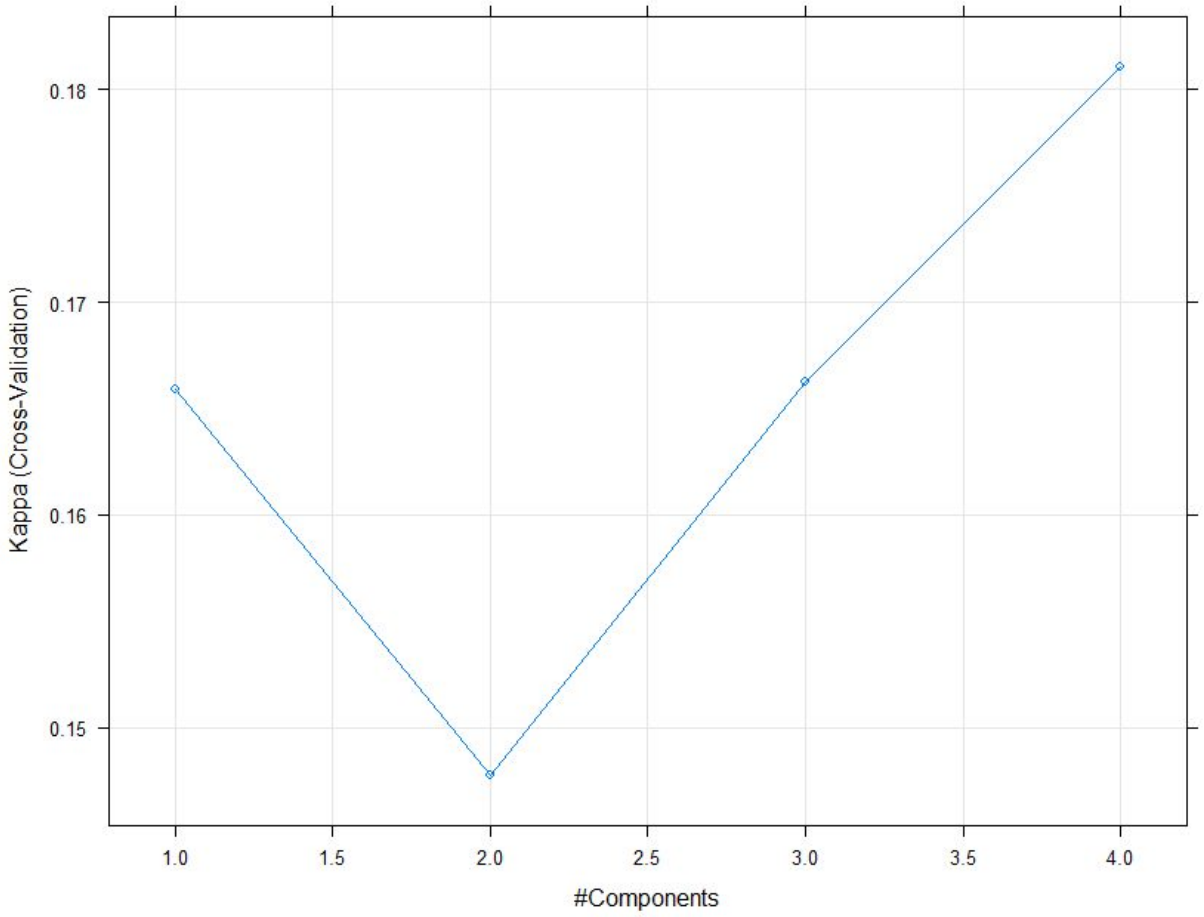
ncomp	logLoss	AUC	prAUC	Accuracy	Kappa	Mean_F1	Mean_Sensitivity
1	1.006914	0.6139002	0.4024027	0.5549407	0.1658657	NaN	0.4017256
2	1.007198	0.6031455	0.3888870	0.5422925	0.1477240	NaN	0.3969276
3	1.004185	0.6127274	0.3932583	0.5470356	0.1662514	0.4814815	0.4154461
4	1.002420	0.6273494	0.4147051	0.5553360	0.1810787	0.5570370	0.4214646

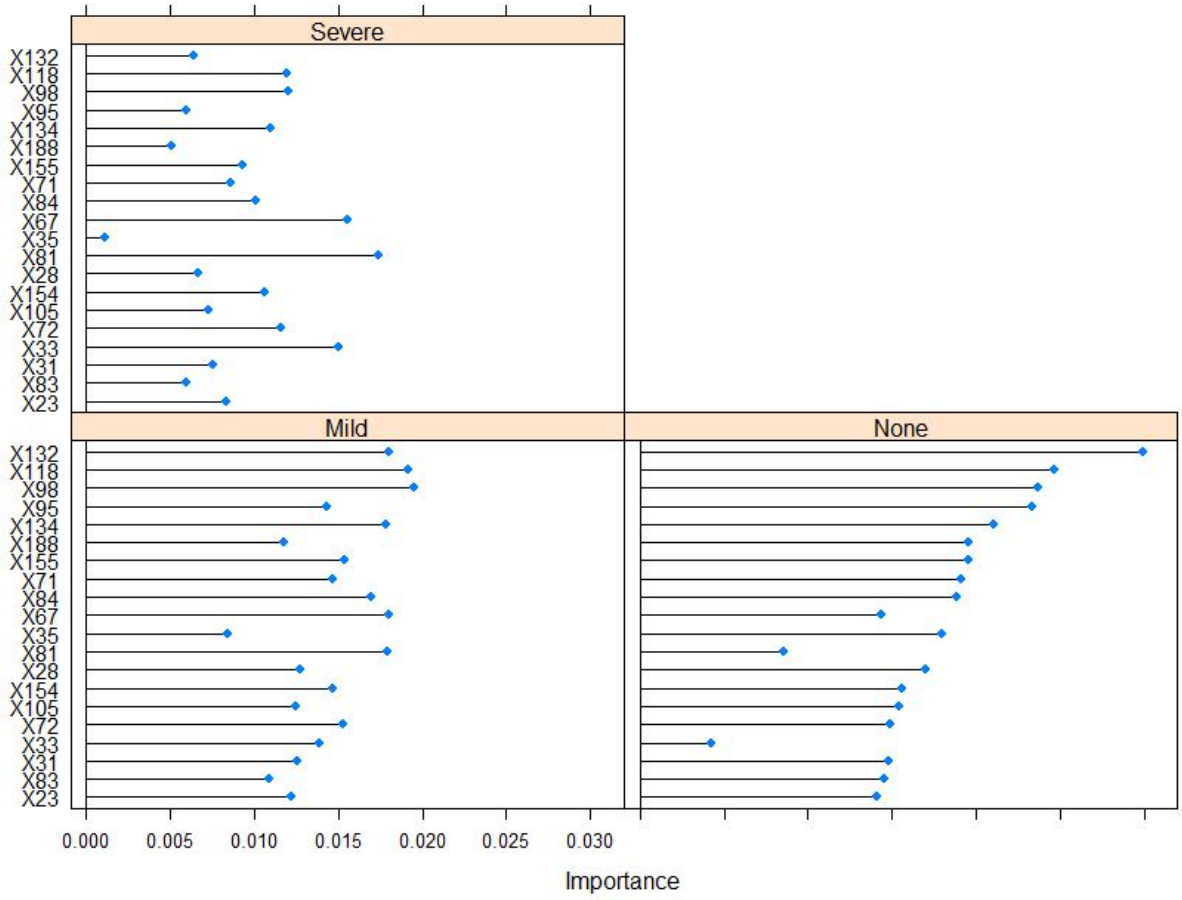
Mean_Specificity	Mean_Pos_Pred_Value	Mean_Neg_Pred_Value	Mean_Precision	Mean_Recall
0.7198374	NaN	0.7323086	NaN	0.4017256
0.7147824	NaN	0.7196978	NaN	0.3969276
0.7207337	0.4483165	0.7266959	0.4483165	0.4154461
0.7258191	0.4740913	0.7311344	0.4740913	0.4214646

Mean_Detection_Rate	Mean_Balanced_Accuracy
0.1849802	0.5607815
0.1807642	0.5558550
0.1823452	0.5680899
0.1851120	0.5736419

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

The final value used for the model was ncomp = 4.




Severe
Mild
None
X132
X118
X98
X95
X134
X188
X155
X71
X84
X67
X35
X81
X28
X154
X105
X72
X33
X31
X83
X23
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Importance

Bio dataset:

Partial Least Squares

225 samples

81 predictor

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

Pre-processing: centered (81), scaled (81)

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 202, 203, 202, 201, 203, 202, ...

Resampling results across tuning parameters:

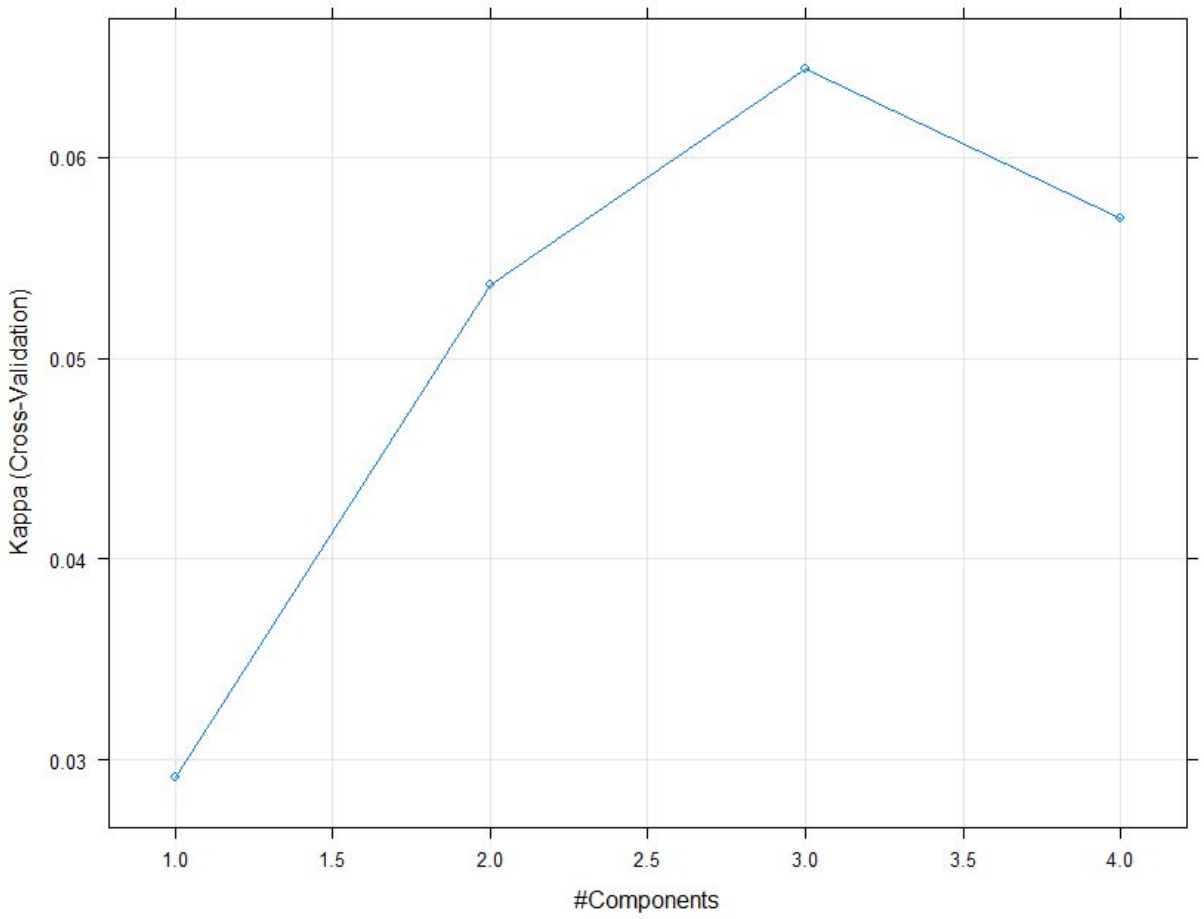
ncomp	logLoss	AUC	prAUC	Accuracy	Kappa	Mean_F1	Mean_Sensitivity
1	1.020856	0.5466447	0.3448259	0.5023457	0.02910376	NaN	0.3476852
2	1.027171	0.5506259	0.3486264	0.5076840	0.05368284	NaN	0.3564394
3	1.026380	0.5513661	0.3498378	0.5068370	0.06441814	NaN	0.3594276
4	1.027913	0.5453745	0.3412757	0.4975649	0.05695000	NaN	0.3545034

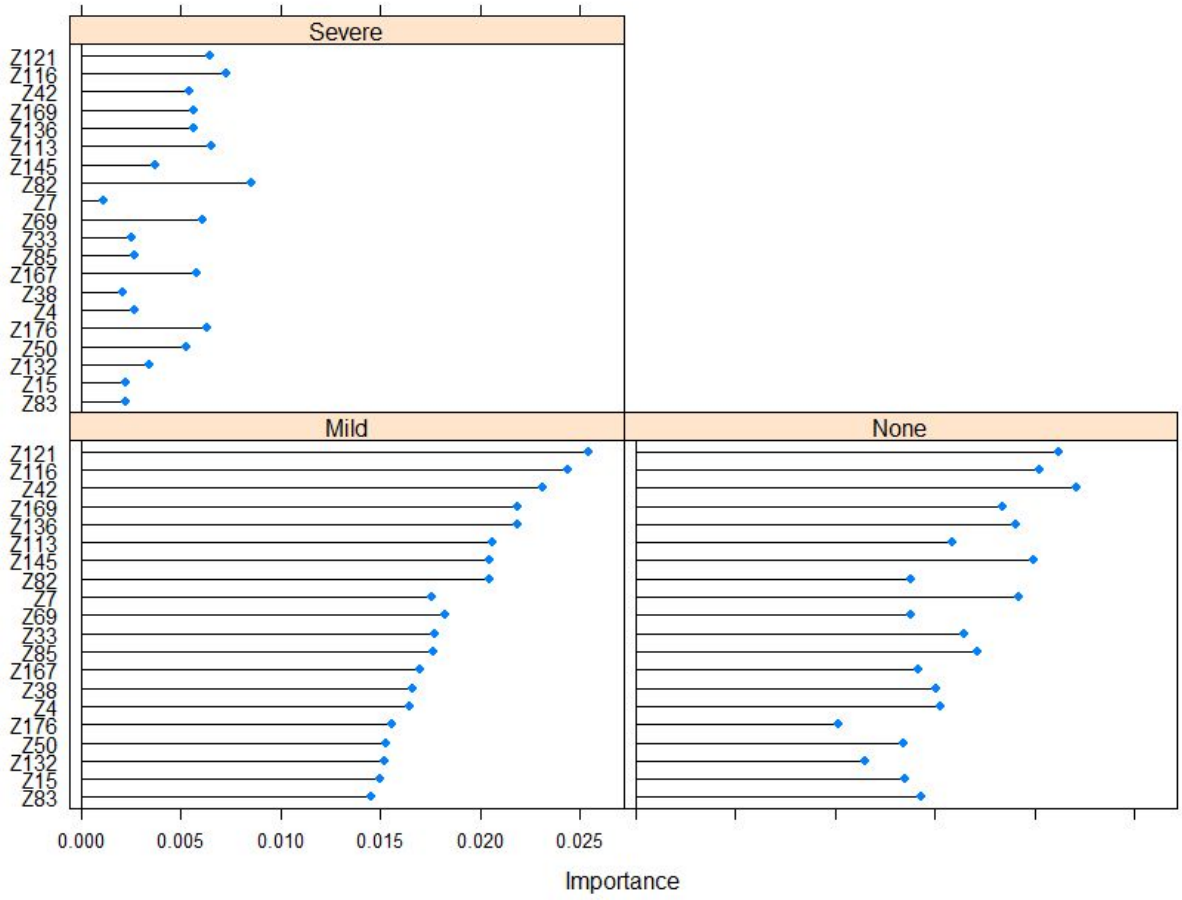
Mean_Specificity	Mean_Pos_Pred_Value	Mean_Neg_Pred_Value	Mean_Precision	Mean_Recall
0.6750128	NaN	0.6592795	NaN	0.3476852
0.6838800	NaN	0.6737007	NaN	0.3564394
0.6874820	NaN	0.6900418	NaN	0.3594276
0.6861333	0.3881944	0.6881043	0.3881944	0.3545034

Mean_Detection_Rate	Mean_Balanced_Accuracy
0.1674486	0.5113490
0.1692280	0.5201597
0.1689457	0.5234548
0.1658550	0.5203183

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

The final value used for the model was ncomp = 3.


Kappa (Cross-Validation)
0.03
0.04
0.05
0.06
1.0
1.5
2.0
2.5
3.0
3.5
4.0
#Components



For bio dataset :

Penalized Multinomial Regression

225 samples

81 predictor

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

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 202, 204, 203, 202, 202, 203, ...

Resampling results across tuning parameters:

decay	logLoss	AUC	prAUC	Accuracy	Kappa	Mean_F1	Mean_Sensitivity
0e+00	12.438252	0.5046334	0.3111728	0.3503294	-0.03114084	0.3763087	0.3096801
1e-04	9.461867	0.5147406	0.3269495	0.3594203	-0.03586444	0.3707294	0.3099747
1e-01	1.722344	0.5719163	0.3444219	0.4161020	-0.01114216	0.4566745	0.3312710

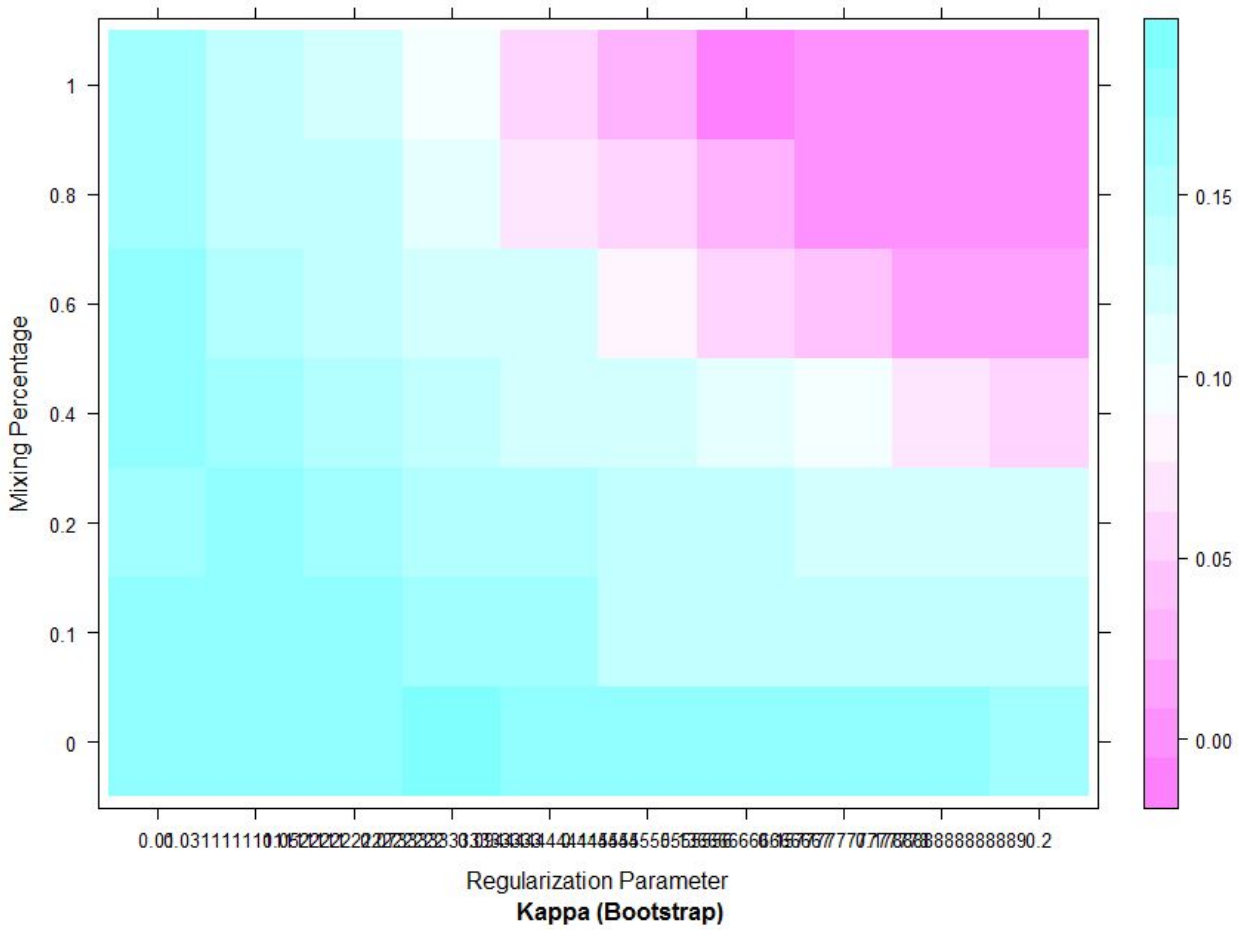
Mean_Specificity	Mean_Pos_Pred_Value	Mean_Neg_Pred_Value	Mean_Precision	Mean_Recall
0.6575998	0.3186340	0.6555460	0.3186340	0.3096801
0.6550396	0.3143040	0.6522409	0.3143040	0.3099747
0.6619773	0.3115981	0.6617119	0.3115981	0.3312710

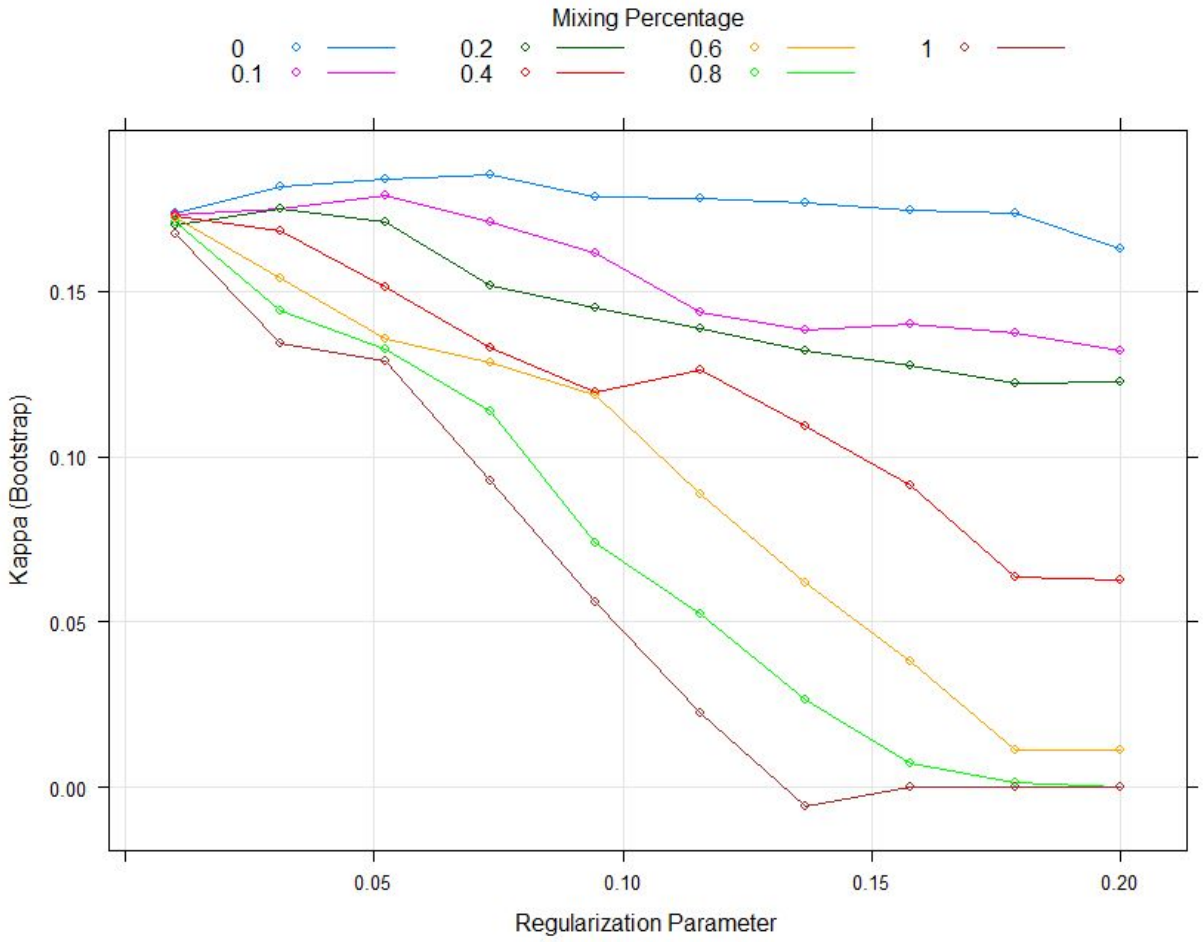
Mean_Detection_Rate	Mean_Balanced_Accuracy
0.1167765	0.4836400
0.1198068	0.4825072
0.1387007	0.4966242

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

The final value used for the model was decay = 0.1

Now we move to penalized models :


Mixing Percentage
1
0.8
0.6
0.4
0.2
0.1
0
0.15
0.10
0.05
0.00
Regularization Parameter
Kappa (Bootstrap)

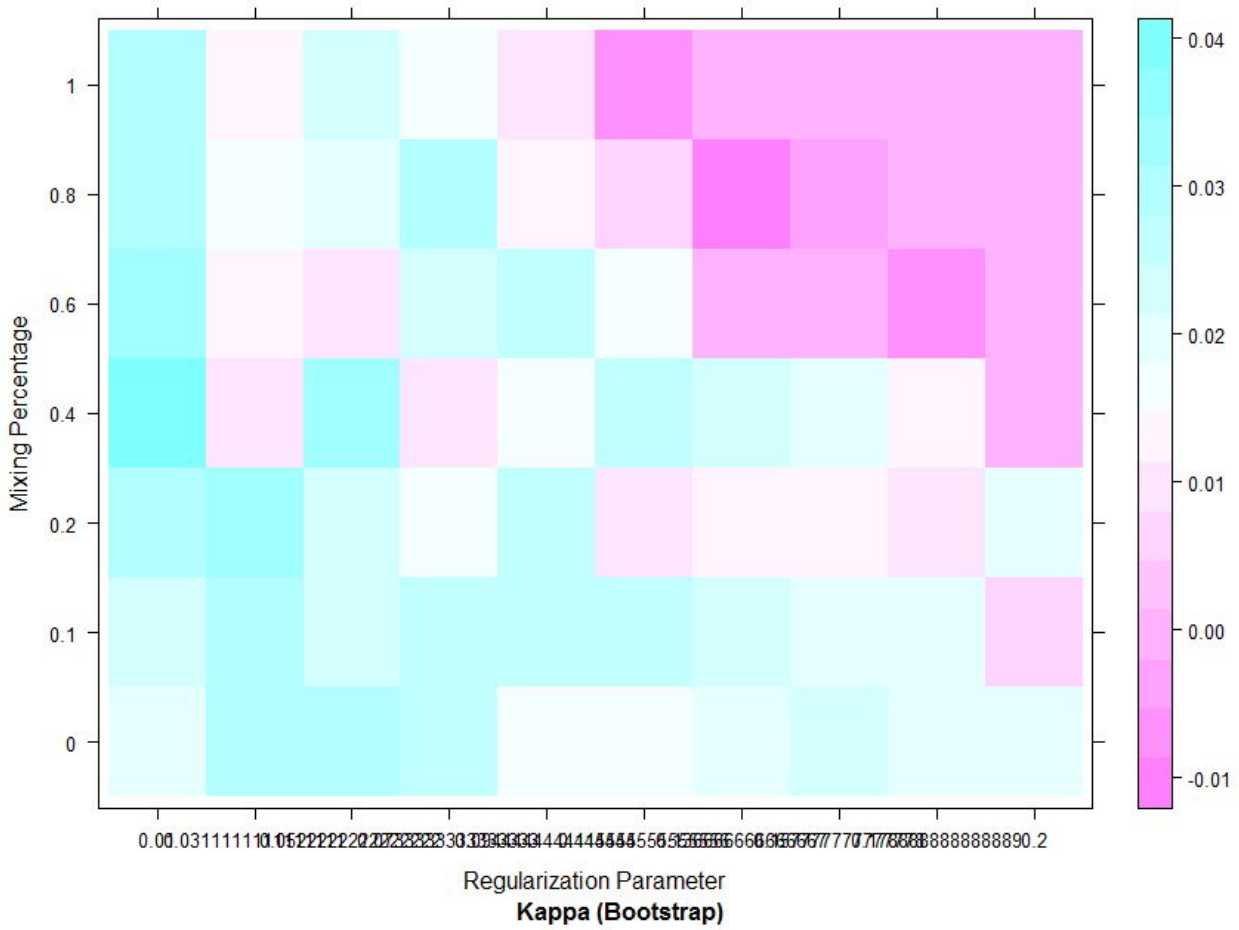


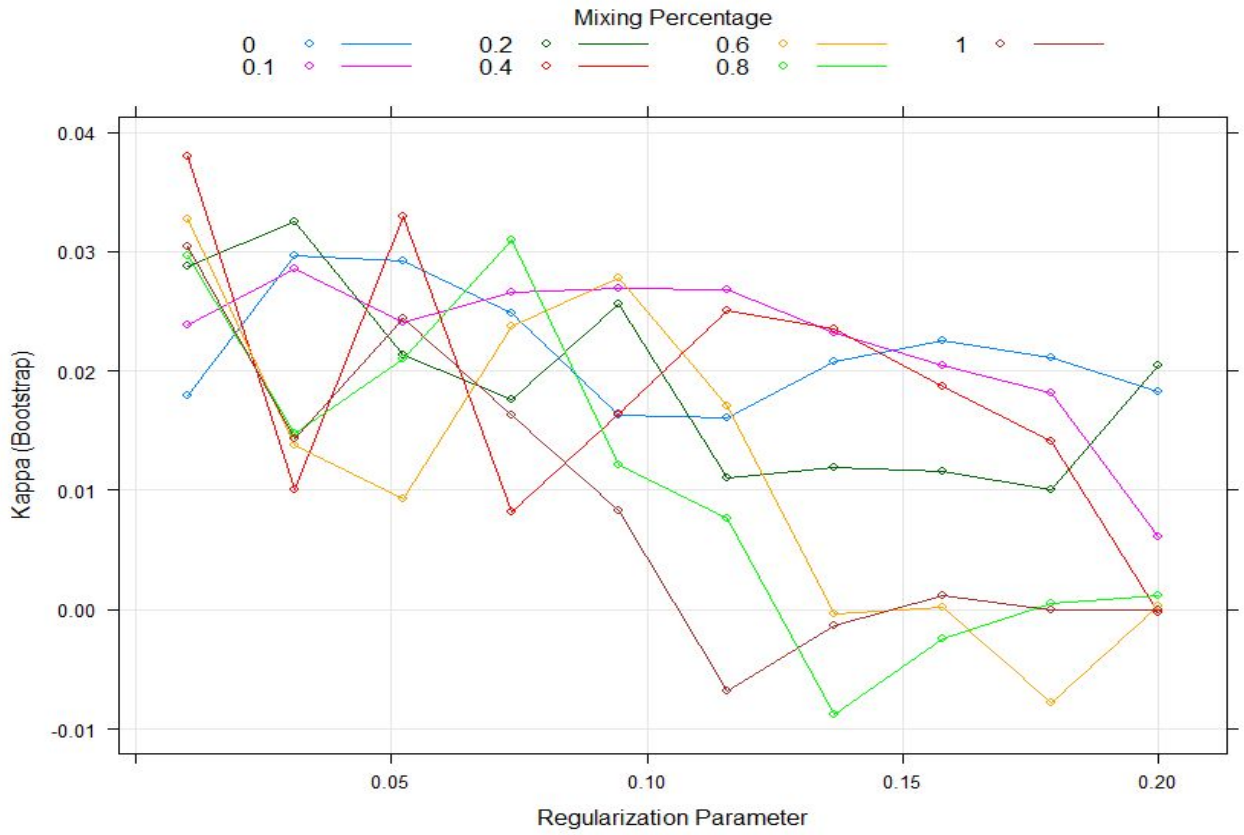
The above graph corresponds to the chemical dataset. Kappa was used to select the optimal model and a mutinomial classes were selected instead of binomial.Kappa was used to select the optimal model using the largest value.

The final values used for the model were alpha = 0 and lambda = 0.07333333.

For the Bio dataset :

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


Mixing Percentage
1
0.8
0.6
0.4
0.2
0.1
0
0.04
0.03
0.02
0.01
0.00
-0.01
Regularization Parameter
Kappa (Bootstrap)



The final values used for the model were alpha = 0.4 and lambda=0.01

The nearest centroid shrinkage model:

For chem dataset:

Nearest Shrunken Centroids

225 samples

73 predictor

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

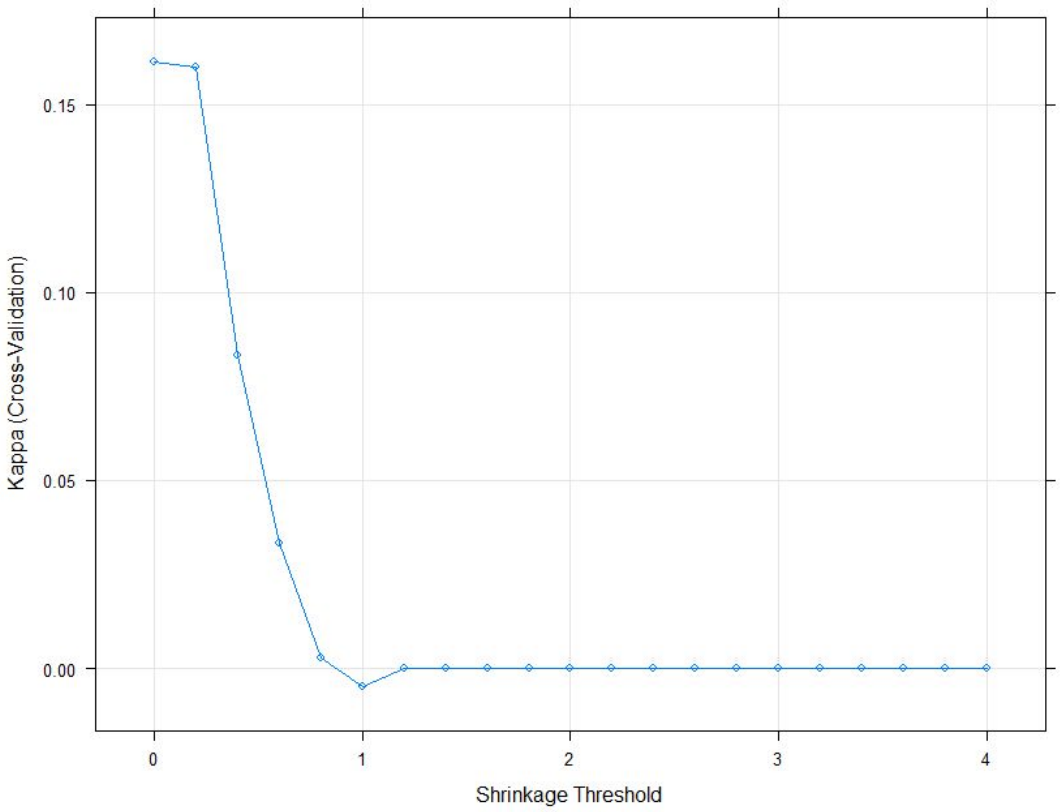
Pre-processing: centered (73), scaled (73)

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 204, 202, 202, 203, 203, 202, ...

Resampling results across tuning parameters:

threshold	logLoss	AUC	prAUC	Accuracy	Kappa	Mean_F1	Mean_Sensitivity
0.0	0.9676369	0.6490131	0.413404256	0.5595262	0.161505663	0.6221140	0.4160354



For bio:

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

The final value used for the model was threshold = 0.

225 samples

73 predictor

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

Pre-processing: centered (73), scaled (73)

Resampling: Cross-Validated (10 fold)

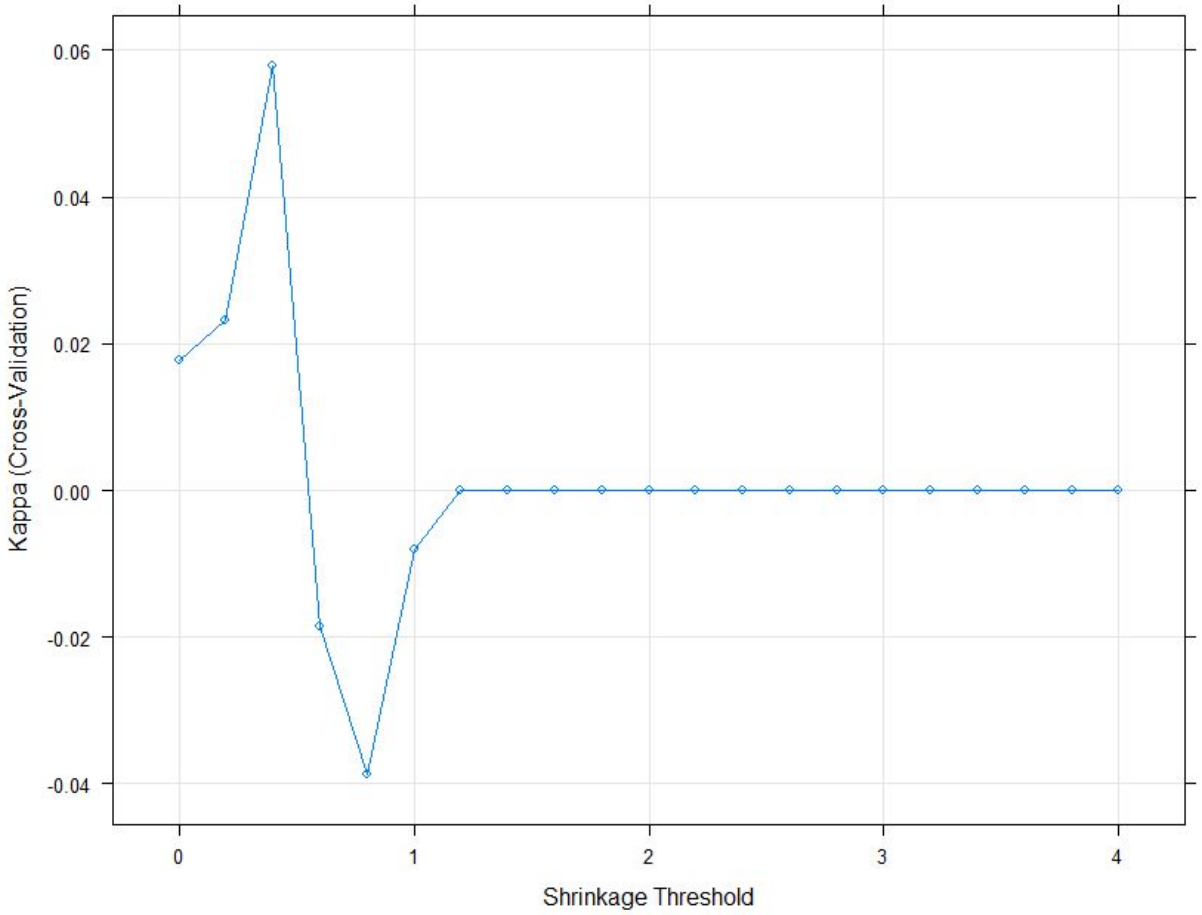
Summary of sample sizes: 204, 202, 202, 203, 203, 202, ...

Resampling results across tuning parameters:

threshold logLoss AUC prAUC Accuracy Kappa Mean_F1 Mean_Sensitivity

0.0 0.9676369 0.6490131 0.413404256 0.5595262 0.161505663 0.6221140 0.4160354

0.2 0.9418548 0.6533768 0.418261161 0.5636928 0.160060822 0.5973964 0.4153620



After building these Models let us evaluate them using the test data:

1.Logistic Regression: Chemical

Prediction Mild None Severe Kappa : 0.0218

Mild 15 12 4

None 8 8 1

Severe 6 1 1

For Bio Data:

Prediction Mild None Severe Kappa : -0.0858

Mild 18 15 3

None 10 5 3

Severe 1 1 0

2.LDA model:

Chemical :

Prediction Mild None Severe Kappa : 0.0148

Mild 18 14 4

None 7 6 1

Severe 4 1 1

Bio data:

Prediction Mild None Severe Kappa : -0.0877

Mild 14 14 3

None 11 6 2

Severe 4 1 1

3.PLSDA Model:

Bio Data

Prediction Mild None Severe Kappa : 0.1587

Mild 22 12 3

None 7 9 3

Severe 0 0 0

Chem model:

Prediction Mild None Severe Kappa : 0.017

Mild	18	14	5
None	9	7	0
Severe	2	0	1

4.GLM NET penalized models:

Chemical dataset

Prediction Mild None Severe Kappa : 0.0624

Mild	18	13	5
None	8	8	0
Severe	3	0	1

Bio dataset:

Prediction Mild None Severe Kappa : 0.0221

Mild	18	14	2
None	10	7	3
Severe	1	0	1

5.Near shrinkage centroid method:

Chemical dataset:

Prediction Mild None Severe Kappa : 0.014

Mild	21	15	5
None	7	6	1
Severe	1	0	0

Bio dataset:

Prediction	Reference: Mild	None	Severe
Mild	26	21	6
None	3	0	0
Severe	0	0	0

Kappa : -0.0937