Machine Learning for Metabolomics

Abstract

The classification and regression using different algorithms were performed on different datasets to find the most optimised algorithm and better analytical technique from HPLC and Enose for the freshness of food. In this assignment the algorithms with best parameters were tested and the conclusion was written at the end.

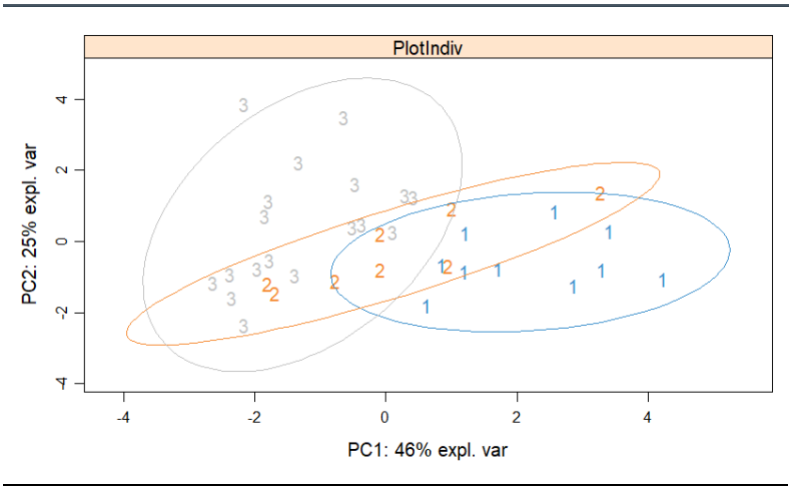
Objective 1

To use machine learning classification models on enose data (predictors) and sensory data (response) to find out the best algorithm for the prediction of sensory scores.

DATA PREPERATION

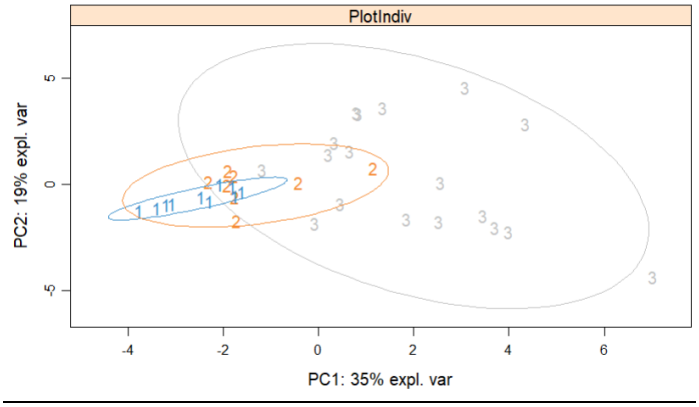
To do the data preparation and exploratory analysis first the required libraries were installed and loaded into the environment. Then the data (csv) was viewed after loading into the environment and it was found out that the HPLC and enose data had lesser data so with the help of merge function four data sets were created (two for HPLC and two for enose) and thus a smaller dataset was extracted.

EXPLORATORY DATA ANALYSIS



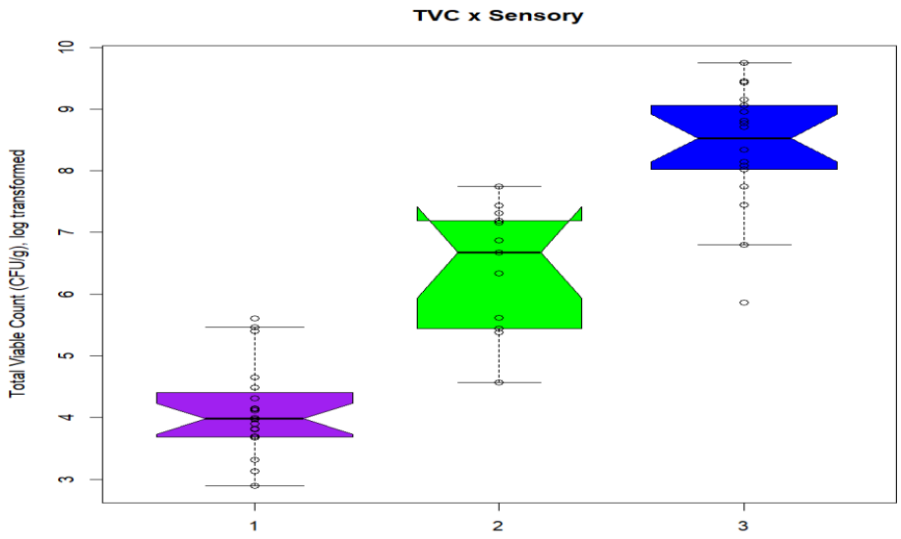
(enose)

The pca was done on the merged data of enose and merged data of HPLC, and 2d with 3d graph was plotted. From the 3d plot we can see more dimensions and see clear segregation according to different components and from the 2d plot we can see the overlapping of different sensory classes especially for “2” because it is a grey area with semi fresh properties. However the highest frequency was for the “3” class along with an outlier which we cannot remove as it is a sample in a comparatively smaller dataset (outliers are removed with humongous datasets). We also get the highest variance as 46 % and second highest as 25 %.



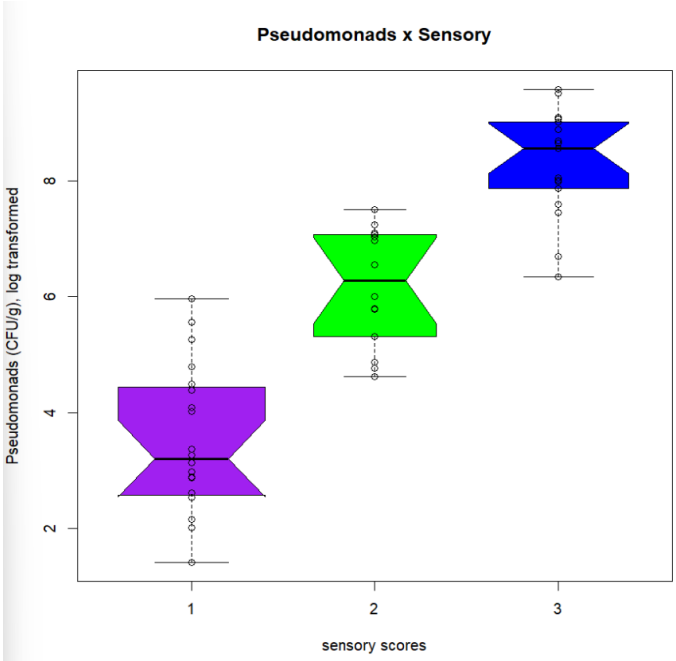
(HPLC)

From the HPLC data we again have the highest frequency for the “3” class and there is strong overlapping between classes which indicates they are not well separated in the original space and some reasons for it could be, variables are not good predictors, the sampling data is less or the characteristics are quite similar (eg, semi fresh samples have strong overlaps) . The highest variance recorded was 35 % with second highest being 19 %. Here also we have an outlier with very high variance (very stale) which we cant remove (part of small sample) Overall, enose has slightly better results dues to better separation of classes.



(TVC boxplot)

The boxplot with notch was plotted for TVC and pseudomonads against the sensory score, and for TVC the notch for second class was the least indicating a higher degree of confidence for the median class and from the width of boxplots. The TVC count is highest in class 3 along with an outlier. The sample points has almost an equal distribution around the median for all the classes.



(pseudomonas boxplot)

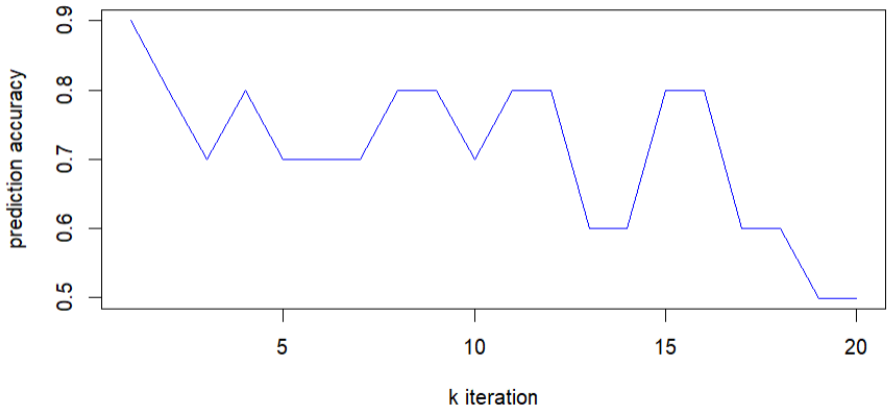
The maximum value in fresh class is nearly equal to the median of the semi-fresh class, and maximum concentrations in class"1" are above the median values. For class "2" and "3", the

concentrations are equally distributed about the median value with no outliers as such. The notch for “2” class is again the least which mean higher confidence values for median class.

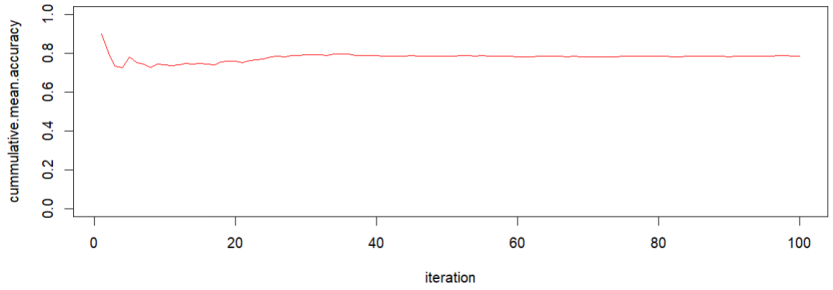
CLASSIFICATION

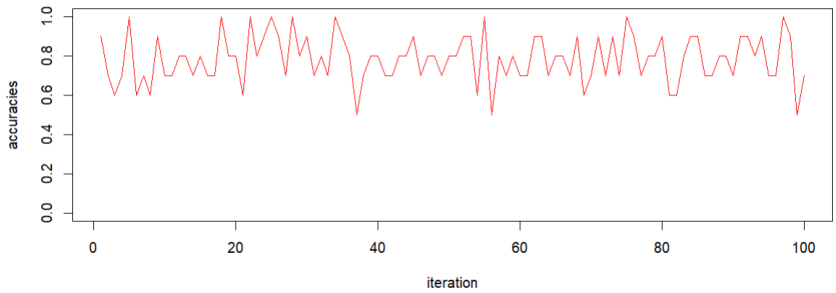
1) KNN classification

The merged data of enose was taken and then then classes were made as factors to perform the task followed by seeing the proportion of each (class 3 being max with 50 %). Then the data separation was done with 70% train and 30% test set to avoid overfitting and to test the parameters of k. The model was run with k=3 and a confusion matrix for overall was made with an overall accuracy of 0.8. Then different values for k was tried for tuning purpose (1-20) and the accuracy was checked, the highest being 0.9 for k=1. A graph was plotted for the same, and accuracy was decreasing over the iterations of k, and the reason could be a comparatively smaller dataset.

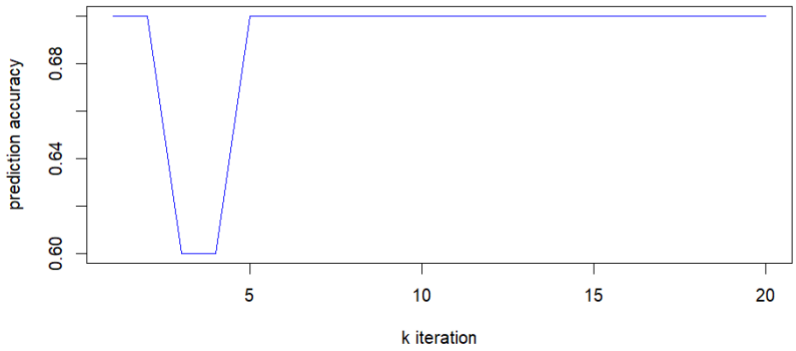


The model was iterated 100 times with the hyperparameter, k=1 and a graph was plotted with accuracies against number of iterations. A confusion matrix was made to analyse the prediction value against the observed value, and the overall accuracy was found to be approx. 0.7 with 95 % CI (0.34,0.93). The accuracies was stored in a vector and was printed to see the accuracies over 100 iterations. A mean of accuracy was also calculated which came to about 0.784. Then the mean cumulative accuracies was calculated and was printed with a graph being plotted against the iterations, which was almost stable to ~0.7 after 20 iterations.



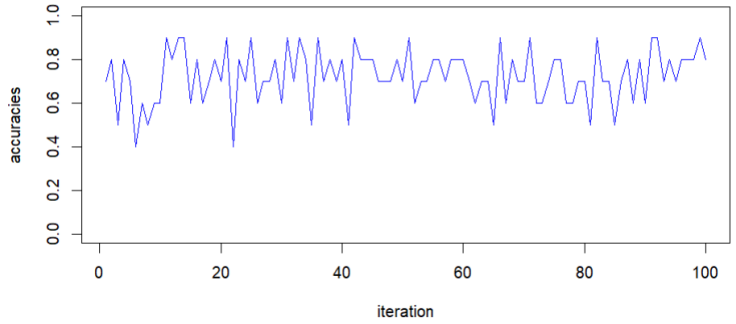


Also the varImp function is not applicable to knn modelling as knn is a non-parametric method and does not have inherent notion of variable importance, as they do not have any underlying assumptions about the relationships between the predictors and the response. Knn model was also run for the HPLC data and the initial steps for merging datasets, making classes as factors, proportions (of classes), data partition (70-30), train set, test set, train class, test class is same as the knn modelling for enose data. So again, the model was run for k=3 and an accuracy of 0.6 was obtained, along with an overall confusion matrix. The values of k was tested from 1-20 and max accuracy was found as 0.7 for many k values, but k=1 was selected, and the accuracies were printed.

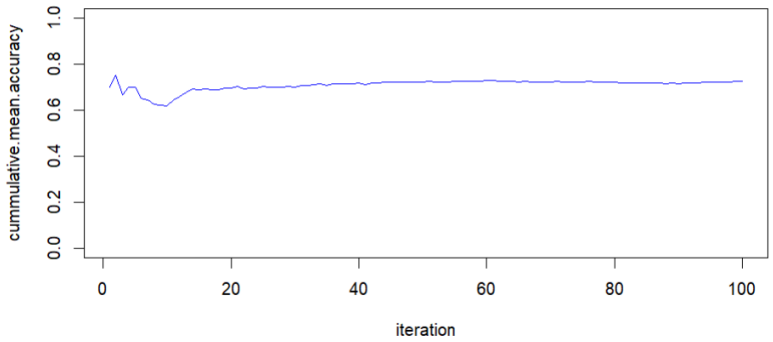


The plot for accuracies and k iterations till 20 was made, increasing the iterations would have no effect on the accuracy as knn model tends to find new data points similar to training

points, and for smaller k (this case) the model can be very sensitive to the specific training points and its possible to not being adjusted to new data. Then the data partition was done and model was iterated for 100 times along with the plotting of accuracies against its iterations.

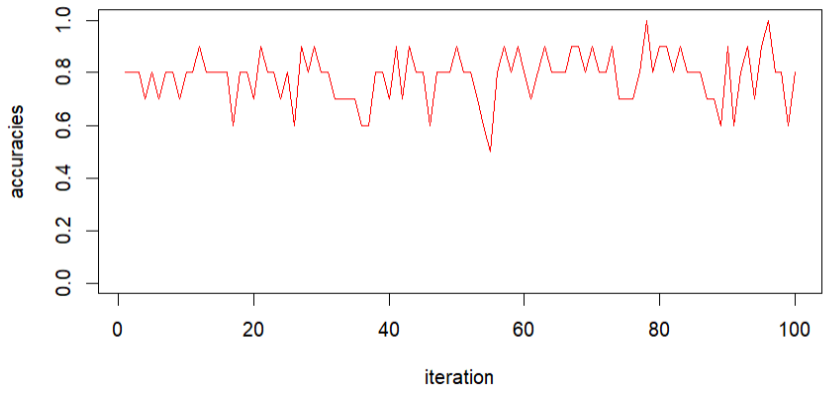


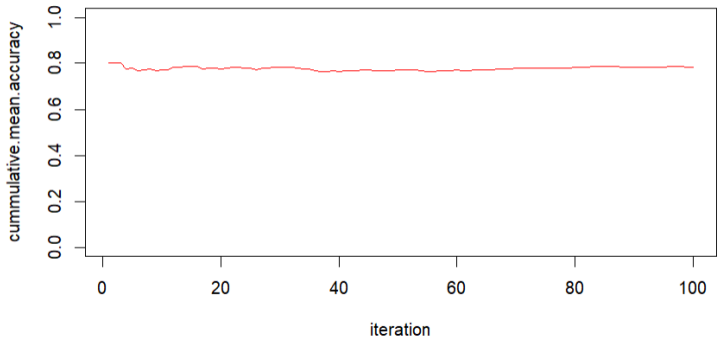
The mean accuracies was calculated and found out to be 0.725 and the graph for cumulative mean accuracies was plotted against iterations, which also got stable after ~20 iterations.



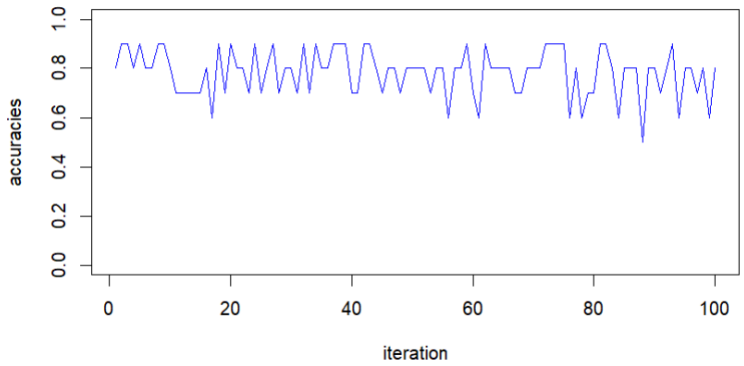
2) SVM CLASSIFICATION

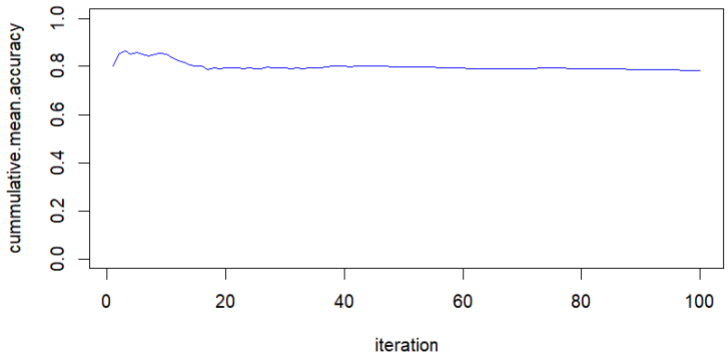
The merged data for enose with sensory was taken and the response variable (sensory) was converted to class. And then the training and test sets and classes were formed (70-30). Then the hyperparameter tuning was done with sampling as “cross” which would form many folds (3 in this case) in the dataset and repeat (10 times) to account for the variance in the dataset. However the hyperparameter tested was regularization hyperparameter (cost; C) from a list of values and c=1 was selected as given by summary(out_tune),in general a lesser c value corresponds to more robustness of noise. Then the model was run with c=1 and overall accuracy was found as 0.9. The model was iterated 100 times with c=1 and accuracies was stored with its mean being 0.784. The mean of cumulative accuracy was found to be 0.77 and was linear after ~15 iterations.





For HPLC the initial previous steps like merging, class as factor, proportions, train-test (70-30) were same . The values of hyperparameters were checked from a list of them (c) and the best value found was c=1 in this case too. And a cross table was constructed with a confusion matrix with an overall accuracy of 0.7 for the model. Then the model was iterated 100 times with a mean accuracy of 0.783.





Overall, Enose and HPLC had almost the same performance for this model due to theire mean accuracy but HPLC was very slightly (negligibly) better.

3) RANDOM FOREST CLASSIFICATION

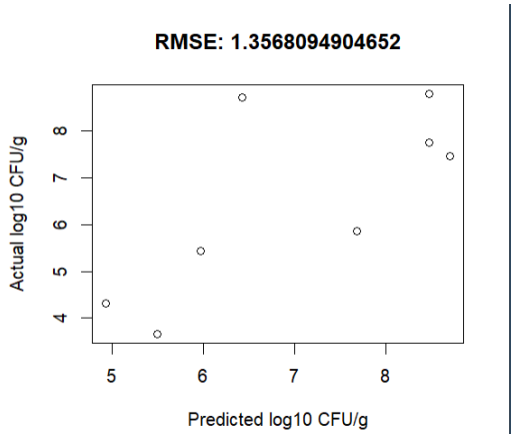
The library loading, data merging, factoring into class, and seeing proportion of each class was done. And then the model was constructed using the classification task as_task_classif function. A plot was made to see frequencies of classes, then the data partition was done (70-30) and the classification accuracy was found to be 0.7 and a confusion matrix was made. The model tuning was done using grid search and the best parameters found were ntree=200, mtry=2, nodesize=3 and maxnodes=16. Then the optimized model was run and an accuracy of 0.9 was found. The model was made and iterated 100 times (but wasn't being iterated). But the bar plot for prediction was plotted. For HPLC all the initial steps were carried out and 19 features were identified from the as_task_classif function, and a confusion matrix was made with 26.91 % error rate. And prediction plot was made with a classification accuracy of 0.8. Grid search was used, and the best parameters found were ntree=425, mtry=10, nodesize=4, maxnodes=11 and classification accuracy as 0.83 which were save in the model. And the optimized model generated accuracy of 1 (overfitting). The 100 iterations were tried, but it wasn't running. But the auto plot was made for prediction (model was overfitting).

Overall, in this classification model

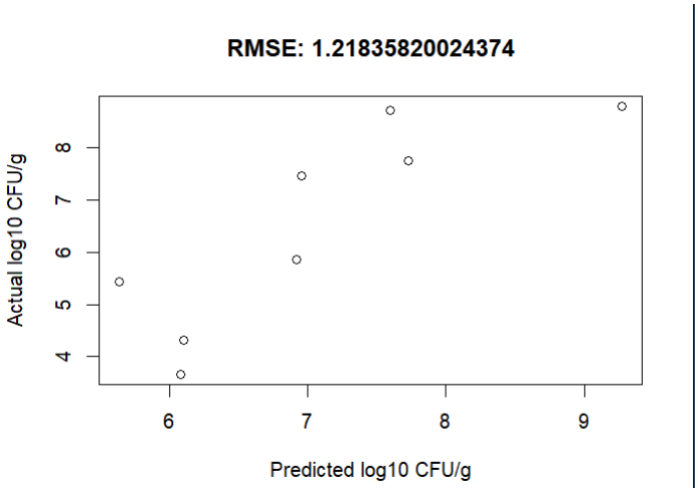
<u>REGRESSION</u>

1) KNN REGRESSION

The datasets were merged for enose and TVC and a plot was made for TVC count. Training and test sets (70-30) were created and the summary of model was found which gave best value of hyperparameter k=4. The graphs are all plotted after first iteration for all regression models as it will give the idea of "first hit". The data points from the graph is quite scattered from the x=y line (imaginary; not drawn). This concludes that the model was not able to predict the outcomes that accurately. As for accurate prediction the points should be along the linear line x=y (actual: predicted > actual/predicted=1 > ideal case). The model was iterated 100 times and the mean RMSE values (1.38), standard deviation (0.31), 95 % CI was calculated.



For dataset Pseudomonads merged with bacterial counts, the above initial process was done and the best value for hyperparameter found was k=7. And the graph (first iteration) for actual vs predicted score was plotted which was a little better than the previous one as the points were close to the x=y line (imaginary). And the mean RMSE value (1.47), standard deviation (0.31), and class interval (1.41 1.53) was calculated. A lower RMSE value (between 0-1) is considered a good fit and in this case it was not.

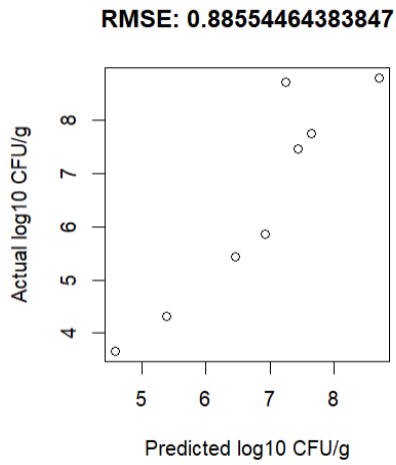


For TVC and HPLC, the hyperparameters were found to be k=5 and the predicted VS actual value was plotted and the model was iterated 100 times. From the graph the and rmse value we can understand that the points are not as near to the x=y line.(imaginary). The mean of rmse is 1.27, standard deviation=0.27, and CI= 1.22 1.33.

Then Pseudomonads and HPLC dataset was used and with same method, the hyper parameter found was k=5 and model was iterated 100 times with this value and the rmse value, standard deviation and CL was found. However the graph for this plot was very off grid as compared with x=y line.

2) PLS REGRESSION

The dataset for TVC was merged with Enose, and from the plot we can see that the RMSE value is very less and the points are in a very linear fashion (1st iteration) which means they are quite accurate as compared to the models of knn (until now) . ncomp=2 was selected (model.fit) as hyperparameter and model iterated 100 times generating RMSE value (1.21), standard deviation (0.25) and CI (1.15 1.26). The value for sd is also quite less which means the data is quite concentrated around the mean.



The dataset for pseudomonads was merged with HPLC and from plot we can see that the RMSE values are very high for this dataset (>2.0) and so the mean of RMSE calculated is 1.79 while standard deviation is 0.37 while CI being 1.71 1.86.

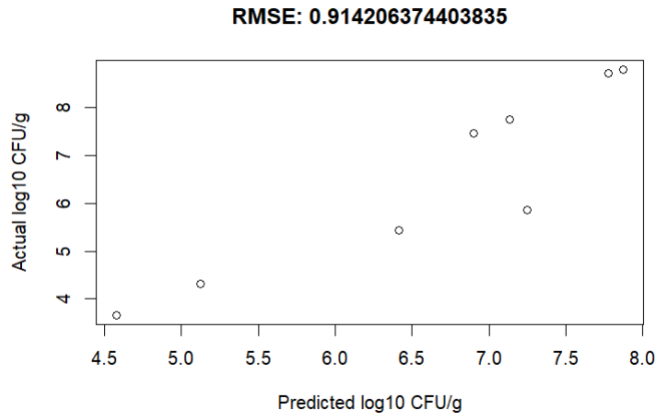
The dataset for TVC was merged with HPLC and the hyperparameter selected was ncomp=2 (smallest RMSE value) and graph was plotted with predicted vs actual and

RMSE was 1.26 while the mean RMSE was 1.3, standard deviation was 0.4 (quite less) and CI being 1.22 1.39.

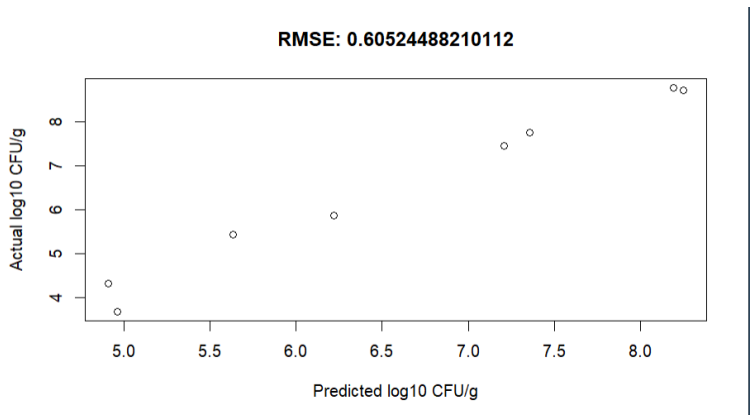
Then the Pseudomonads with HPLC dataset was merged and the best hyperparameter found was ncomp=2 (model.fit) while the graph were having RMSE values as not defined. Although the standard deviation was quite less (0.3)

3) RANDOM FOREST REGRESSION

The dataset of TVC was merged with Enose and the hyperparameter to tune was mtry which is the number of variables that randomly separate as candidates for splitting at decision node and the optimal mtry was mtry=2. Usually a higher mtry corresponds to a high predictors but in this case we only had 8 predictors. Then after the 1st iteration, the model was plotted for actual vs predicted. And RMSE value found was .914 (better). The points are almost linear.



For the data sets Pseudomonads and HPLC the hyperparameter selected was mtry=2 (best)



For the data sets of TVC and HPLC ,the best plot was obtained till now was this with rmse value 0.605 which almost forms a perfect straight line, the best hyperparameter was found to be mtry=2

bacterial type	model	rmse	stdv	95 % CL
	knn	1.387648	0.2934607	1.329419 1.445877
TVC	rf	1.244884	0.2287478	1.199495 1.290272
	plsr	1.122108	0.2901242	1.064541 1.179675
	knn	1.49168	0.2753724	1.43704 1.54632
Pseudomonads	rf	1.444295	0.2418822	1.396301 1.492290
	plsr	1.516802	0.2287219	1.471419 1.562185
	knn	1.247801	0.2782636	1.192587 1.303014
TVC	rf	1.007141	0.1732016	0.9727739 1.0415078
	plsr	1.313242	0.422967	1.229316 1.397168
	knn	1.30399	0.3014	1.244186 1.363794
Pseudomonads	rf	1.173471	0.2129427	1.131218 1.215723
	plsr	1.388749	0.3257893	1.324105 1.453393

(THE RMSE table was filled and the excel sheet is provided for the same in file "rmse_table")

CONCLUSION

From the classification models the best classification model was probably knn because the mean accuracy was much better than the other classification models and the analytical technique which was much better was Enose as the datasets run from this analytical technique gave better results for the algorithms. The best regression method was random forest as the RMSE value was the least for it as compared with other algorithms.