STAT 440 Case Study

Guangji Zou (Ray)

Student ID: 301216239

Group: Ray Zou, Helen Fu and Terry Liu

Email address: guangjiz@sfu.ca

Executive Summary

In this report, we focus on Inflammatory bowel disease (IBD), which is comprised of the two disease entities of Crohn’s disease (CD) and ulcerative colitis (UC) in the world.

In this analysis, we want to find out the relationship between genes and Inflammatory Bowel Disease (IBD) patients. The best model is created by LASSO, and the result is that we could use the data features to predict the disease state of individuals.

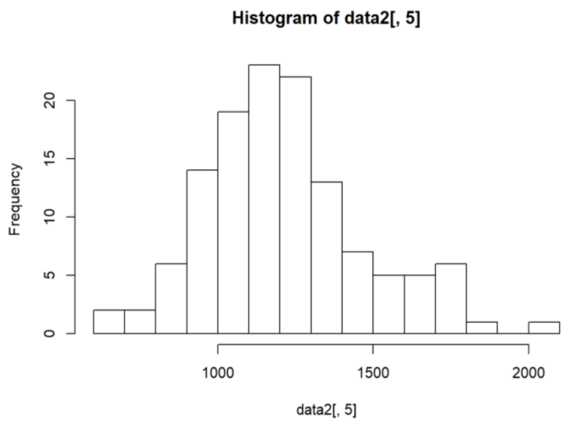
Introduction:

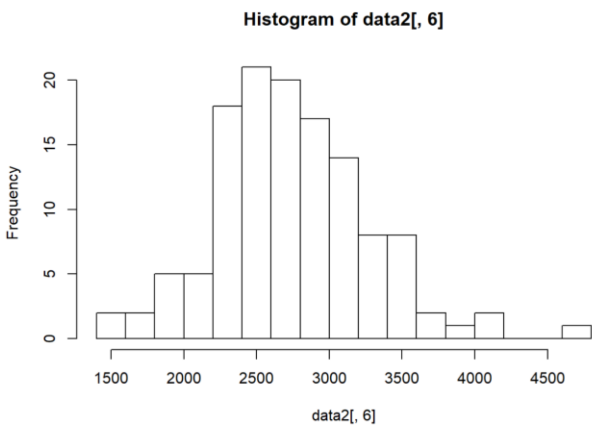
In this research, we will introduce the disease: Inflammatory bowel disease (IBD). The Inflammatory bowel disease (IBD) is a group of inflammatory conditions of the colon and small intestine. Crohn's disease (CD) and ulcerative colitis are the principal types of inflammatory bowel disease, it can affect the mouth, esophagus, stomach and the anus whereas ulcerative colitis primarily affects the colon and the rectum. Inflammatory bowel diseases fall into the class of autoimmune diseases, in which the body's own immune system attacks elements of the digestive system.

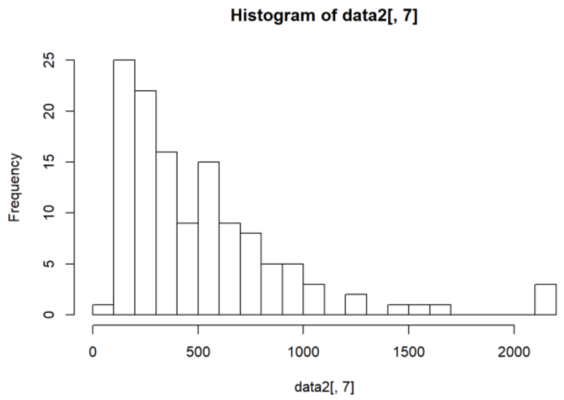
In this method section, we used the three different methods to describe the dataset of Inflammatory bowel disease, the three methods are following: generalized linear model with lasso, random forest and regression tree. In addition, we decided to show the plot of three methods, the graphs can showed the prediction of the methods, and finally used the result of accuracy lasso, accuracy random forest and accuracy regression tree to summarize what we did and describe the each meaning of the lasso, random forest and regression tree.

Method:

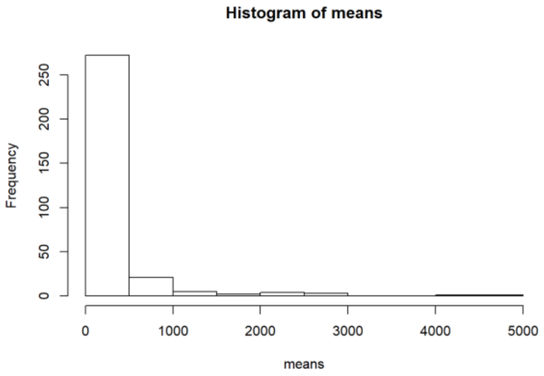
Our dataset is based on Statistical Society of Canada, 2017 Case Study competition, from Global gene expression data: Burczynski et al. (2006) generated genome-wide gene expression profiles for 41 healthy individuals (note that the processed data includes only 41 individuals although the original study included 42 individuals), 59 CD patients, and 26 UC patients. And two data files will be used for this case study, the two data include IBD Gene Expression (Sheet one) and IBD Matched Genes (Sheet two). In the sheet one, the data of IBD Gene Expression is represented 126 individuals, the probeset names, the group of disease, age, Ethnicity, sex and the first row contains the patient IDs. The IBD Matched Gene in sheet two is contained the names of the 309 probesets, gene symbols for 185 unique genes, some genes include two or more probesets. We loaded the data file name and used the methods of lasso, random forest and regression tree to manage the dataset and used the plot to show the distribution to get the result in the R studio.

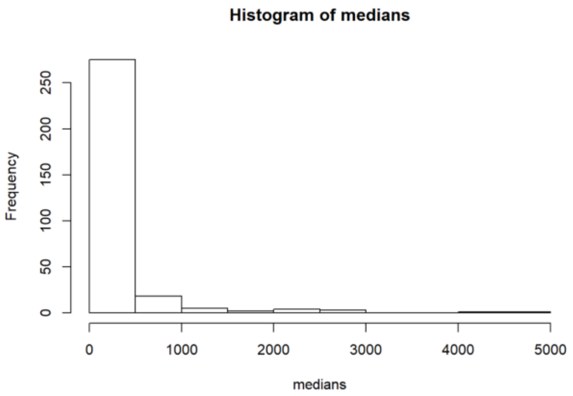


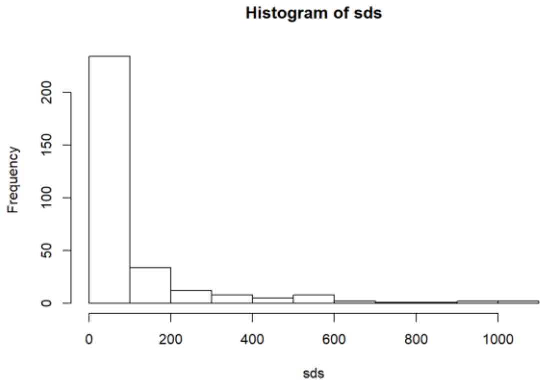


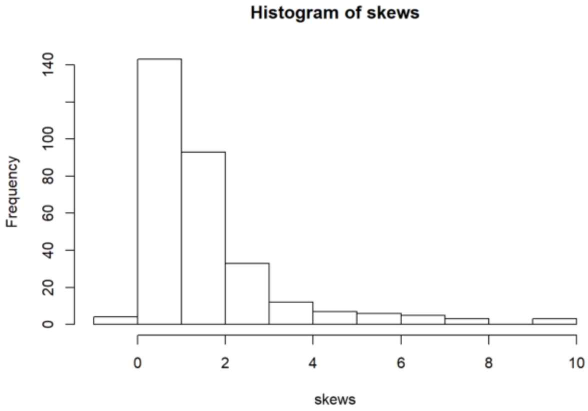


Secondly, the histograms (below the plots) for a distribution with extreme value with means, medians ,standard deviation and skew distribution, meaning is has more extreme value on one side.



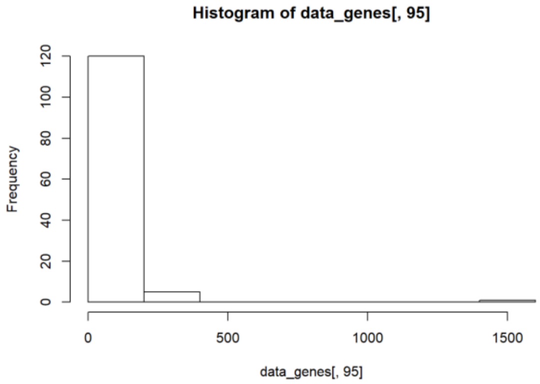


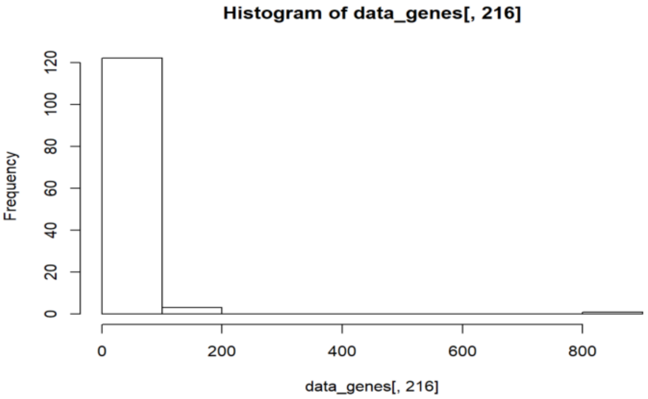


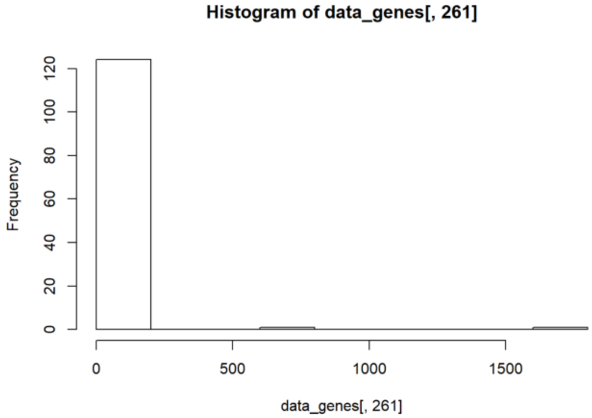


Next, in the data cleaning steps, we found that those which had skews greater than 8 had outliers so that we decided ignore them and clean the outliers and problems of the dataset. These steps as follow:

We changed the whole data format, create a new data variable name and used the tranpose and paste0 function clear to see character variables of the dataset, it means we chane the original data, data type and variable. Then we found that there is any outlier within the dataset by checking skewness of each variable. The value of skewness is greater than 8, and easier to show the variable “205479_s_at”, “211668_s_at”, “215101_s_at” have outliers. The outliers has a little influence for the whole data. If we set them as missing values, it might messed up the whole dataset.



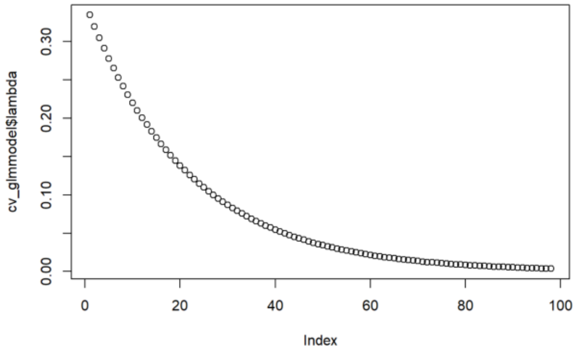


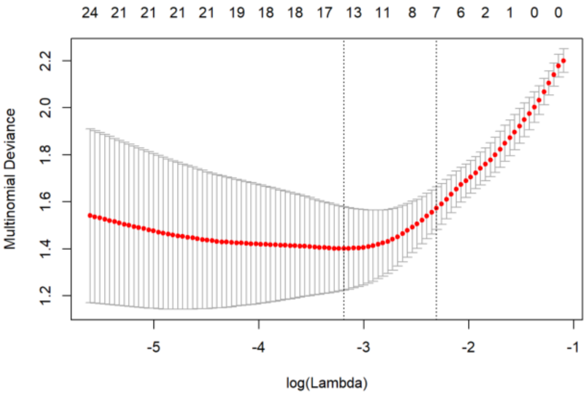


The least absolute shrinkage and selection operator (LASSO) method

We installed the package “glmnet” directly from CRAN, and fits a generalized linear model via penalized maximum likelihood. Sometimes LASSO is interpreted as linear regression with a budget of some fixed amount t that the L1 Norm of beta cannot exceed. As below the graph, the curve corresponds to a variable. It shows the path of its coefficient against the l-norm of the whole coefficient vector at as lambda varies. The axis indicates the number of nonzero coefficients at the current lambda, which is the effective degrees of freedom (df) for the LASSO.

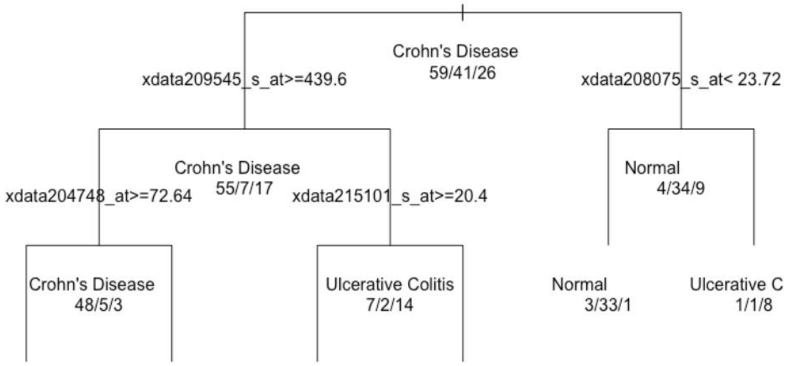
After that, except for viewing the selected values of lambda. It included the cross validation curve (red dotted line), and the data type of response variable was character, we decided to use multinomial within LASSO (error bars). As we know that the error is small, the model is better. Therefore, we selected -3.2 to -2.4.





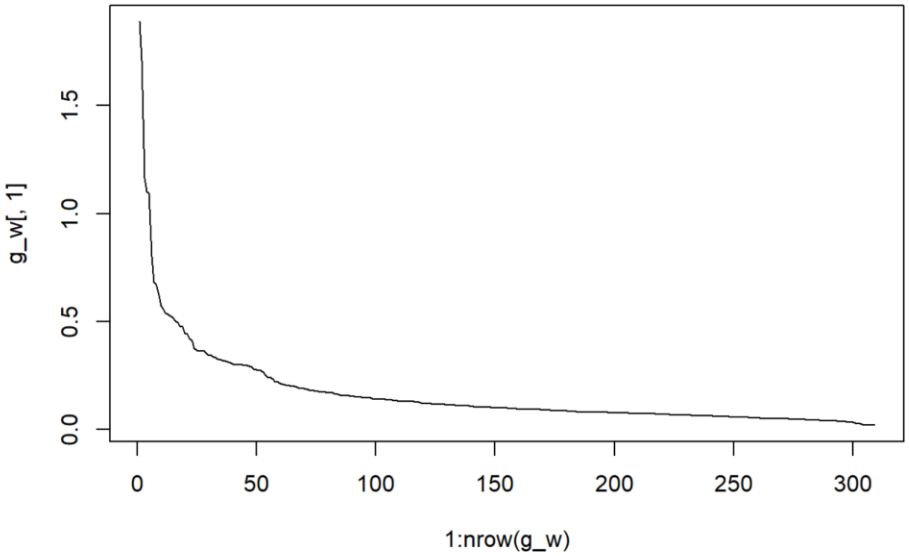
Regression Tree Model

With the idea of linear regression as a way of making quantitative predictions. In simple linear regression, a real-valued dependent variable Y is modeled as a linear function of a real-valued independent variable X. The predicted data of diseases is given beneath each leaf node.



Random Forest Model:

Random forest creates multiple different regression tree and average the result of all trees, we do not display this model. So we plot is the order of genes as below.

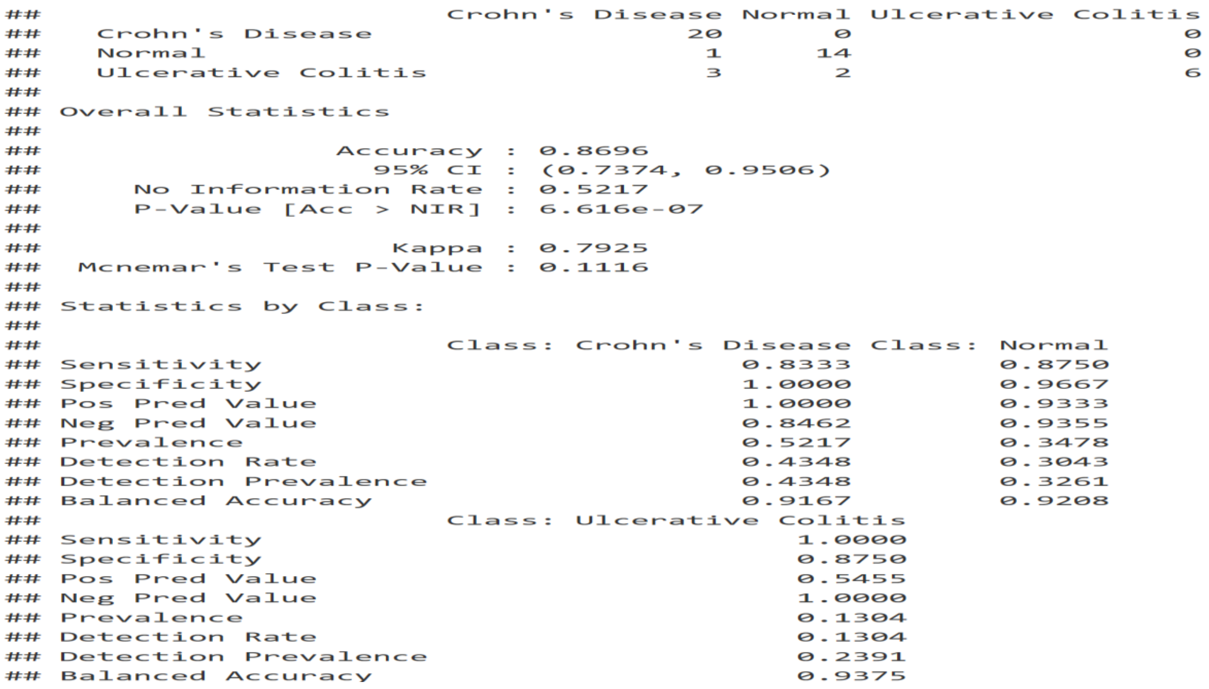


Result:

Three different accuracy of models used to predict the error rate in the result of confusion matrix

1. Accuracy LASSO (The result of confusion matrix for the LASSO model):

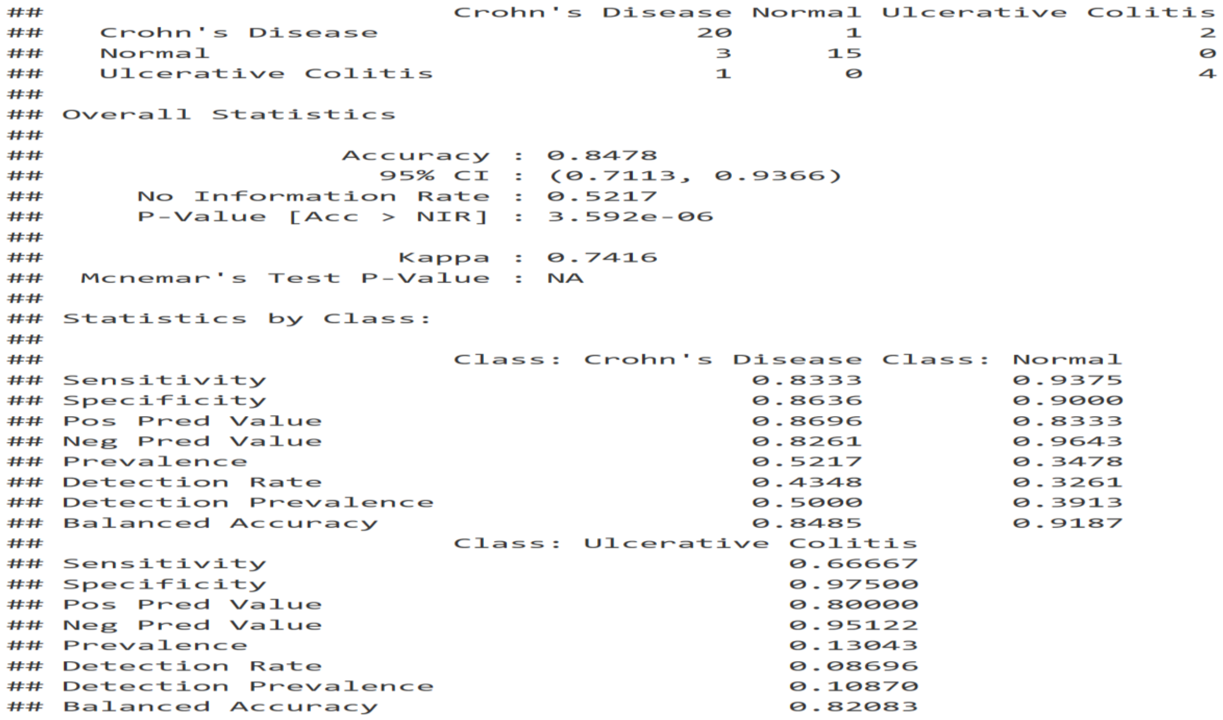
The LASSO is non-linear trend. The specificity and Pos Pred Value is good value data and higher accuracy of LASSO.



```
##                          Crohn's Disease Normal Ulcerative Colitis
##   Crohn's Disease                     20      0                  0
##   Normal                               1     14                  0
##   Ulcerative Colitis                   3      2                  6
##
## Overall Statistics
##
##                Accuracy : 0.8696
##                  95% CI : (0.7374, 0.9506)
##     No Information Rate : 0.5217
##     P-Value [Acc > NIR] : 6.616e-07
##
##                   Kappa : 0.7925
##  Mcnemar's Test P-Value : 0.1116
##
## Statistics by Class:
##
##                      Class: Crohn's Disease Class: Normal
## Sensitivity                          0.8333        0.8750
## Specificity                          1.0000        0.9667
## Pos Pred Value                       1.0000        0.9333
## Neg Pred Value                       0.8462        0.9355
## Prevalence                           0.5217        0.3478
## Detection Rate                       0.4348        0.3043
## Detection Prevalence                 0.4348        0.3261
## Balanced Accuracy                    0.9167        0.9208
##                      Class: Ulcerative Colitis
## Sensitivity                             1.0000
## Specificity                             0.8750
## Pos Pred Value                          0.5455
## Neg Pred Value                          1.0000
## Prevalence                              0.1304
## Detection Rate                          0.1304
## Detection Prevalence                    0.2391
## Balanced Accuracy                       0.9375
```

2. Accuracy Random Forest (The result of confusion matrix for the R.F. model):

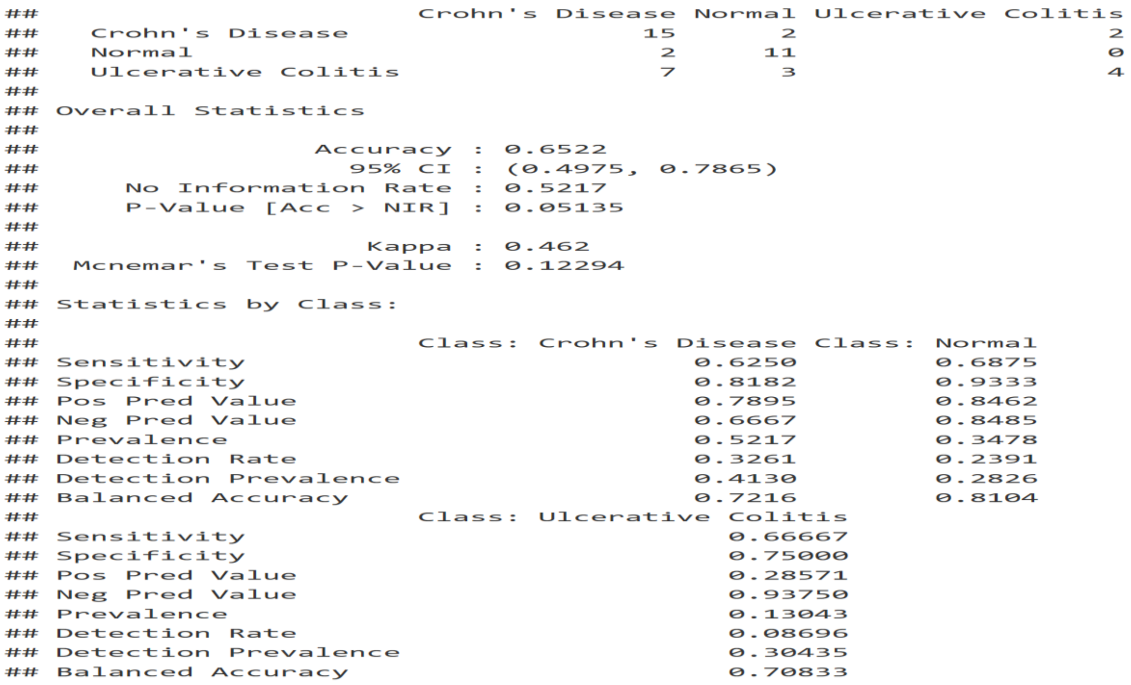
Unexcelled in accuracy among current algorithms, and the random forest method is based on regression tree.



```
##                          Crohn's Disease Normal Ulcerative Colitis
##   Crohn's Disease                     20      1                  2
##   Normal                               3     15                  0
##   Ulcerative Colitis                   1      0                  4
##
## Overall Statistics
##
##                Accuracy : 0.8478
##                  95% CI : (0.7113, 0.9366)
##     No Information Rate : 0.5217
##     P-Value [Acc > NIR] : 3.592e-06
##
##                   Kappa : 0.7416
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Crohn's Disease Class: Normal
## Sensitivity                          0.8333        0.9375
## Specificity                          0.8636        0.9000
## Pos Pred Value                       0.8696        0.8333
## Neg Pred Value                       0.8261        0.9643
## Prevalence                           0.5217        0.3478
## Detection Rate                       0.4348        0.3261
## Detection Prevalence                 0.5000        0.3913
## Balanced Accuracy                    0.8485        0.9187
##                      Class: Ulcerative Colitis
## Sensitivity                            0.66667
## Specificity                            0.97500
## Pos Pred Value                         0.80000
## Neg Pred Value                         0.95122
## Prevalence                             0.13043
## Detection Rate                         0.08696
## Detection Prevalence                   0.10870
## Balanced Accuracy                      0.82083
```

Accuracy Regression Tree (The result of confusion matrix for the R.T. model):

The Accuracy is lower and Neg Pred Value is pretty higher. And is a particular kind of nonlinear predictive model.



```
##                      Crohn's Disease Normal Ulcerative Colitis
##   Crohn's Disease                 15      2                  2
##   Normal                           2     11                  0
##   Ulcerative Colitis               7      3                  4
##
## Overall Statistics
##
##                Accuracy : 0.6522
##                  95% CI : (0.4975, 0.7865)
##     No Information Rate : 0.5217
##     P-Value [Acc > NIR] : 0.05135
##
##                   Kappa : 0.462
##  Mcnemar's Test P-Value : 0.12294
##
## Statistics by Class:
##
##                      Class: Crohn's Disease Class: Normal
## Sensitivity                          0.6250        0.6875
## Specificity                          0.8182        0.9333
## Pos Pred Value                       0.7895        0.8462
## Neg Pred Value                       0.6667        0.8485
## Prevalence                           0.5217        0.3478
## Detection Rate                       0.3261        0.2391
## Detection Prevalence                 0.4130        0.2826
## Balanced Accuracy                    0.7216        0.8104
##                      Class: Ulcerative Colitis
## Sensitivity                            0.66667
## Specificity                            0.75000
## Pos Pred Value                         0.28571
## Neg Pred Value                         0.93750
## Prevalence                             0.13043
## Detection Rate                         0.08696
## Detection Prevalence                   0.30435
## Balanced Accuracy                      0.70833
```

Conclusion

In our group, the best model is LASSO model, as we can show the error rate is the smallest and the P value is the lowest value (the P value is lower, the model is better). The accuracy of random forest model reaches 100% and accuracy regression tree not show the p value, that’s the reason why we didn’t choose this model. We used the above three different models can predict each person’s health and the information of gene.