NAFLD Study

Nan Zhang

August 14, 2015

Here just show the part of analysis for NAFLD Study. Random forest and Factor analysis method are adopted in this analsis .

Introduction

The distribution of steatosis in Nonalcoholic Fatty Liver Disease (NAFLD) can be categorized at Panacinar, Azonal, Zone 1 or Zone 3. Focal Zone 1 pathology, although rare in adults with NAFLD (<1%), has been reported in children with NAFLD. We hypothesize that focal Zone 1 steatosis and focal Zone 3 steatosis are two distinct subphenotypes of pediatric NAFLD.

1.1 Research Question

Aim: In order to better understand the potential differences in NAFLD histologic phenotypes in children, we performed a multi-center cohort study with the aim to determine the association between the zonality of steatosis and demographic, clinical, and histologic features of children with NAFLD.

1.2 data

A total of 813 children less than 18 years of age with biopsy-proven NAFLD enrolled in the NASH Clinical Research Network.165 charactistics were measured by different aspects.Liver histology was reviewed by the Central Pathology Committee using the NASH CRN scoring system. For each biopsy, the predominant location of the fat droplets was recorded as Zone 1, Zone 3.

1.2.1 Data issue

Over 50 measurements contains missing value exceeding 60%.So those measurements are excluded from data. There are over 60 irrelevant variables with diesease NAFLD, so those measurements are removed from analysis.Finally, 42 variables were used in following factor analysis.

2 Methods

Descriptive statistics have been genrated for the rest of 42 characteristics. But those results didn't contain here.

2.1 Factor analysis to reduce the matrix dimension

Factor analysis is used to reduce the dimension of predictors matrix.

2.2 Random Forest to find the final variables

Random forest method is used to find the most imporant variables to build up the final model.

2.3 Decision tree to build the priliminary classier for NAFLD

Decision tree is used to build up the priliminary model for classify the zonality partition.

3 Results

3.1 Factor analysis to reduce predictors matrix dimension

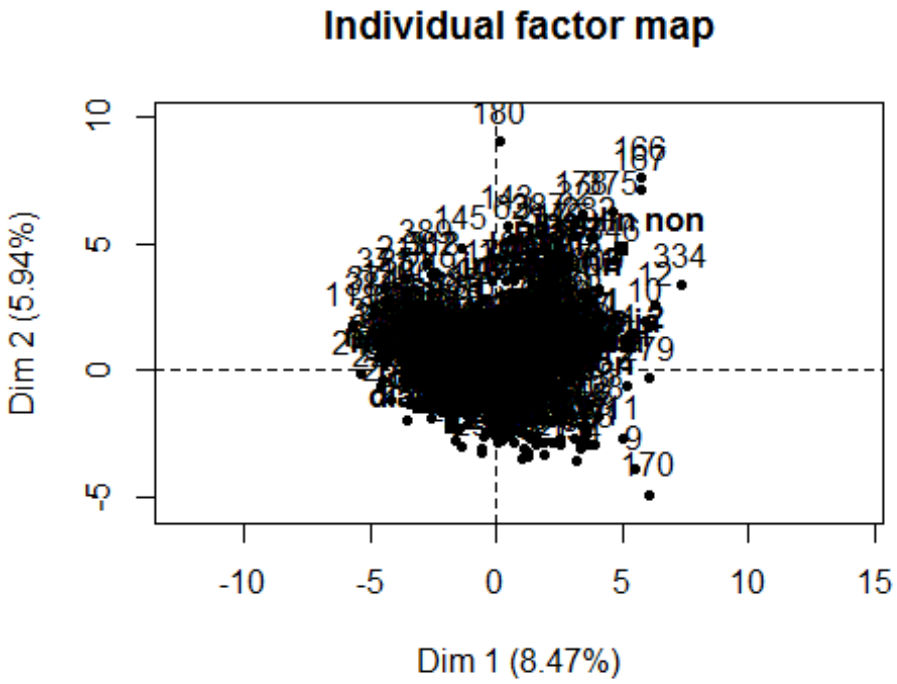
* The First two plots are individual factor map.

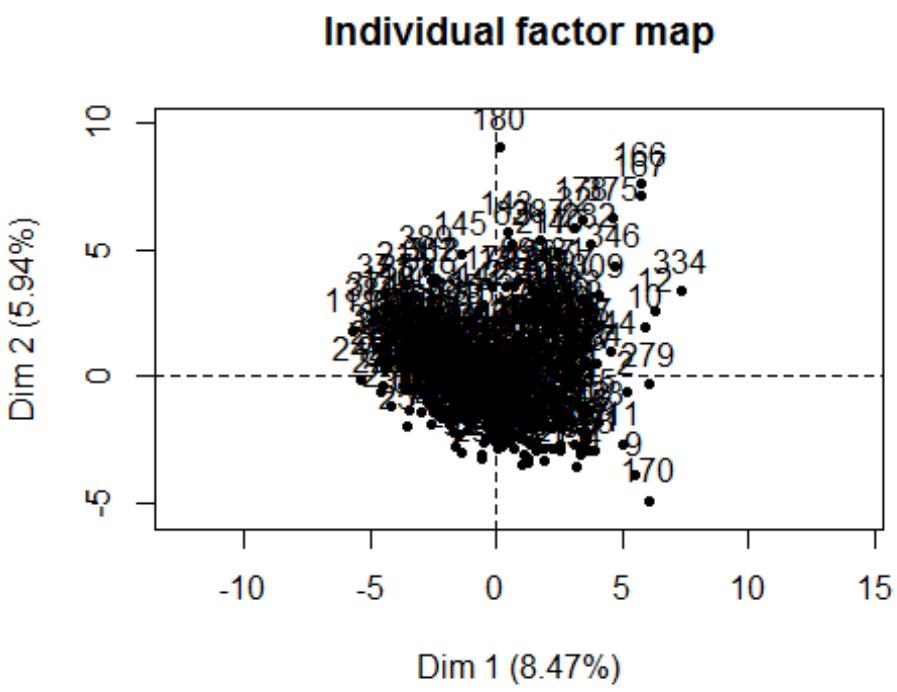
* The third plot is all variables projection in dim1 and dim2.

* The fourth plot is numeric variables in dim1 and dim2.

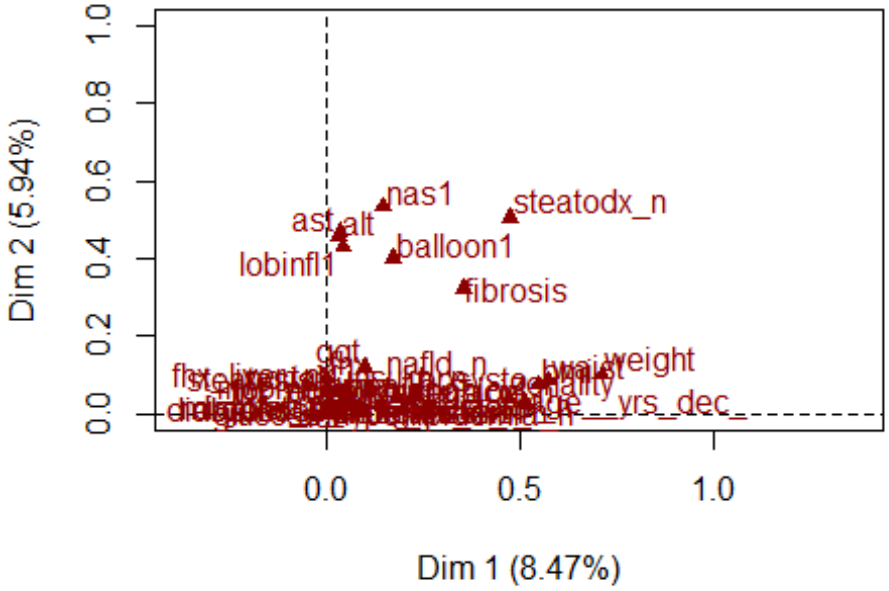
* The fifth plot is categorical variables in dim1 and dim2.

* The sixth plot is scree plot which shows ~92% of the informations (variances) contained in the data are retained by the first two principal components.

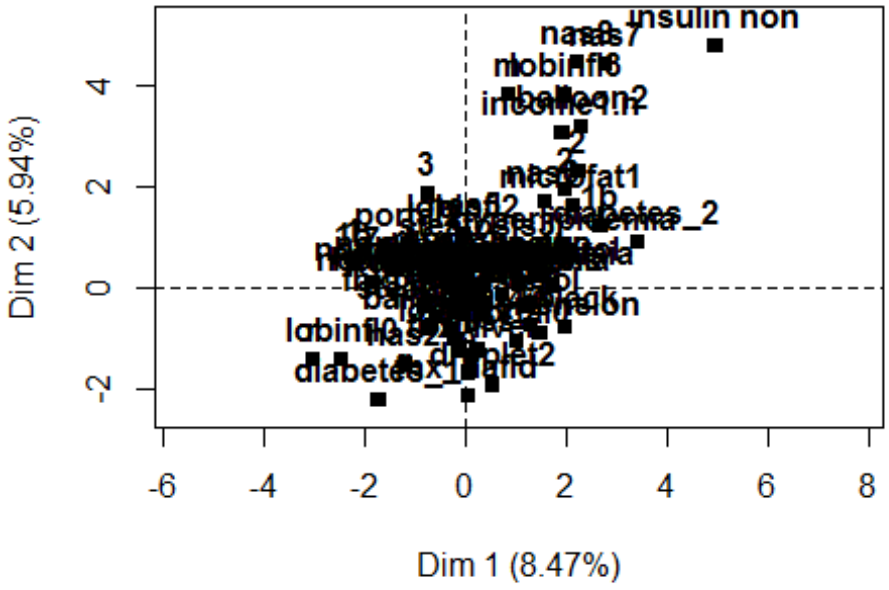

Individual factor map
Dim 2 (5.94%)
Dim 1 (8.47%)
180
166
334
170
non
10
5
0
-5
-10
15

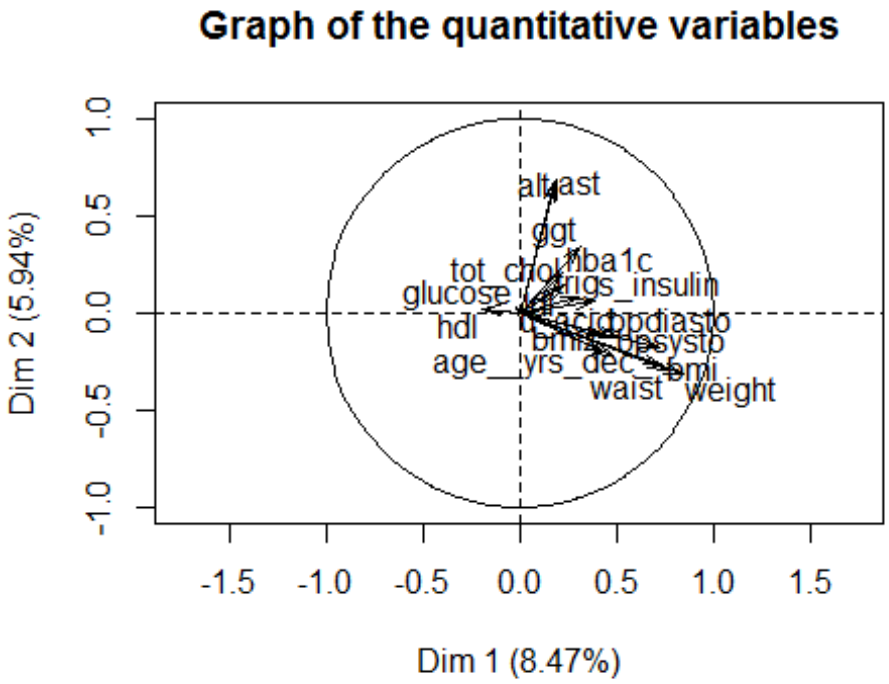

Individual factor map
Dim 2 (5.94%)
Dim 1 (8.47%)
180
166
334
279
170
346
10
5
0
-5
-10
15

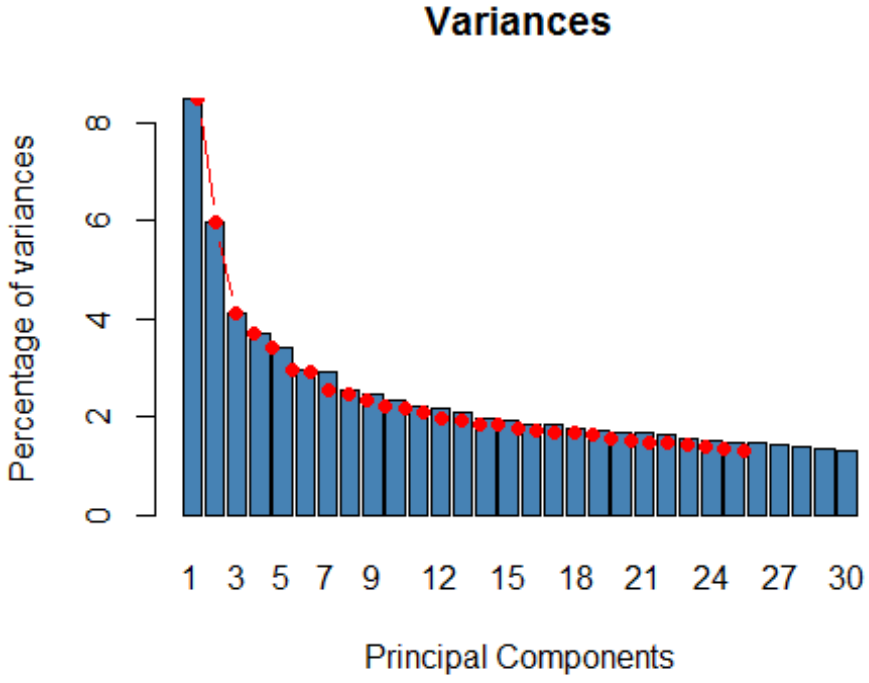
Graph of the variables



Individual factor map







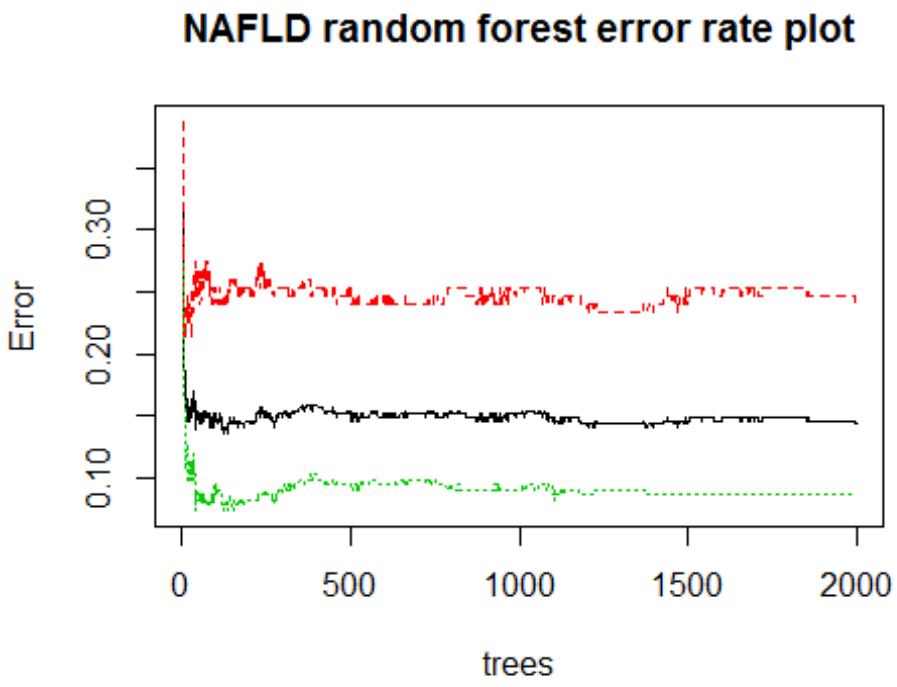
Below listing numeric variables and category variables are selected from the factor analysis. All those select variables p-values are less than 0.01.Due to necessity, there are 28 variables are kept in next analysis.Numeric variables (the 16 first selected):weight,

bpsysto, bpdiasto, uacid , hba1c, ast, alt, tot_chol, ldl, glucose, bmiz, age, ggt, waist, trig, hdl.Categories variables (the 14 first selected): gender,nas, race, hispanic, hba1c, steatosis, fibrosis, lobinfl, portal, balloon, sinsulin,nash.

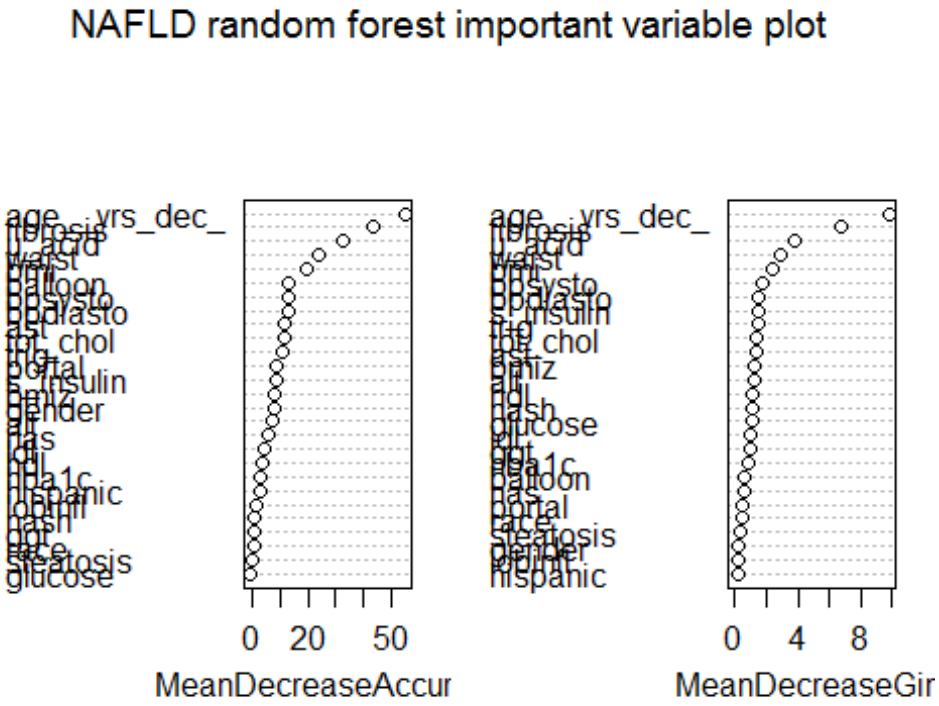
```
## $quanti
##                correlation       p.value
## weight           0.8442461 3.822462e-107
## waist            0.7583611  4.091437e-74
## bmi              0.7415995  2.917777e-69
## age__yrs_dec_    0.7169356  9.093304e-63
## u_acid           0.5130195  1.449202e-27
## bpsysto          0.4810663  5.553538e-24
## bpdiasto         0.4468289  1.543188e-20
## bmiz             0.4300402  5.499137e-19
## s_insulin        0.3872886  2.094763e-15
## ggt              0.3153869  1.868946e-10
## trig             0.3093852  4.262230e-10
## tot_chol         0.2343967  2.876385e-06
## hba1c            0.2268703  6.037083e-06
## ldl              0.2036934  5.072266e-05
## ast              0.1890050  1.737137e-04
## alt              0.1787241  3.899607e-04
## hdl             -0.1991580  7.489567e-05
##
## $quali
##                        R2      p.value
## zonality       0.47334031 5.562401e-56
## steatodx_n     0.47553270 8.740758e-54
## fibrosis       0.35464630 8.998332e-34
## balloon1       0.17448912 7.688870e-17
## hispanic       0.14437496 7.885964e-14
## nas1           0.14726840 2.415324e-11
## diabetes_2_n   0.09823338 2.452995e-10
## race           0.08639149 4.744673e-06
## portal1        0.05252270 2.924595e-05
## insulin_n      0.04186270 4.685685e-05
## hypertension_n 0.03584691 1.691719e-04
## lobinfl1       0.04566724 4.262764e-04
## steatosis1     0.02670442 5.313243e-03
## microfat_n     0.01771760 8.489750e-03
## income1        0.04342366 8.952898e-03
```

3.2 Random Forest method to use find out the most important predictors

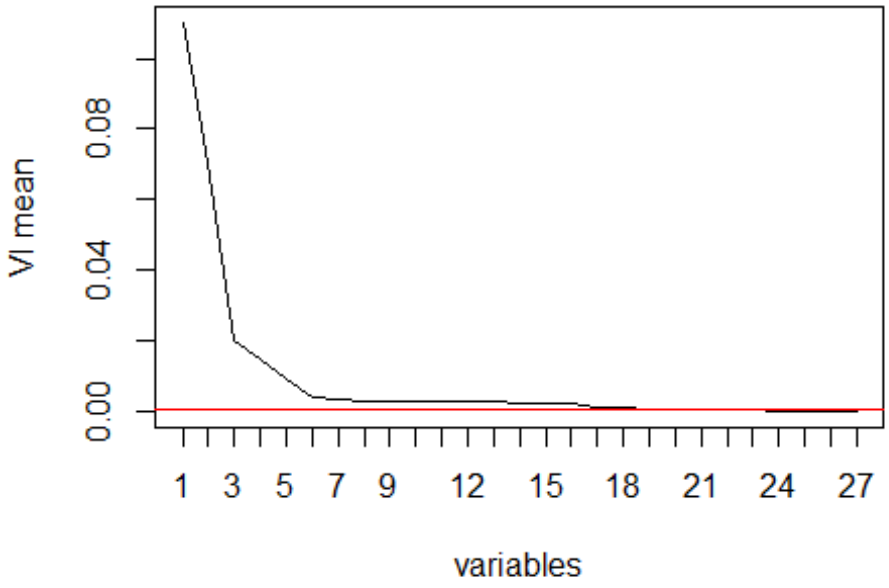
From the plot below, we can get the average error rate is 0.15, the difference of zone1 and zone3 error rate is 0.2. If there are more data to train this model, the difference of error rate will probably reduce.



This plot shows that the variables contribute tp the MeanDecreaseAccurary and MeanDecreaseGini. From the plot, when judging by eyes, we could see that the top four variables are significantly impact the accuracy and Gini score.

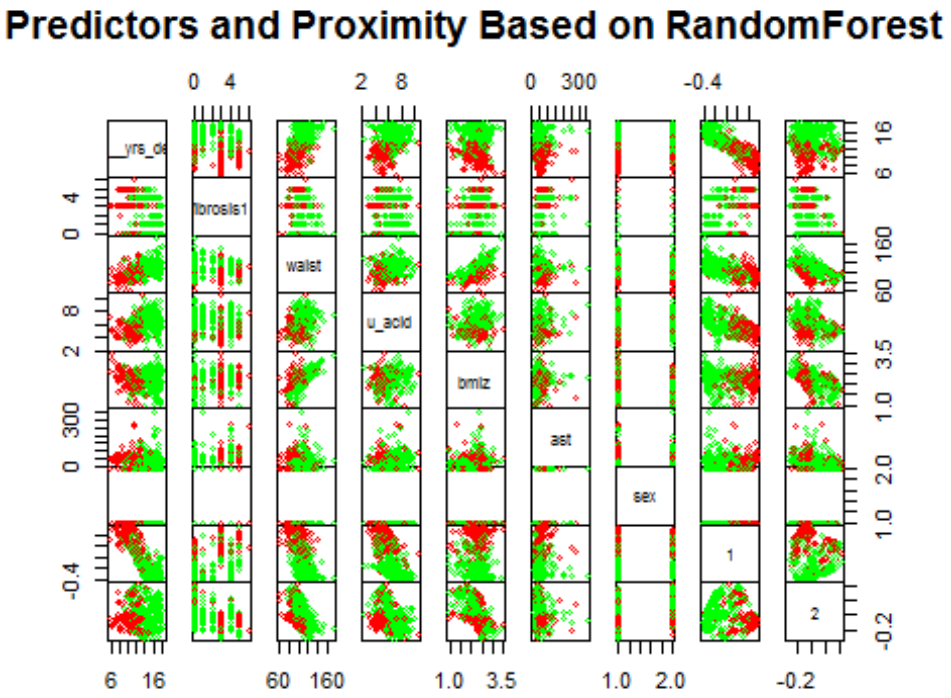


From analysis below, sixteen variables are automatically outputed and all those variables are above threshold(the red line in the plot) as important variables. And eight variables(age, fibrosis, waist, u_acid, bmiz, ast,bmi,bpsysto) are picked as predictors for zonality partition. Because bmiz is standard bmi, bmi can be remove from here. Random forest also output 3 variabls are the most important to predict the labels.




```
##  [1] 16 27  4 19  1 15  6  2  8 20  3 22 13 14  7 17 24  5  9 21 18
## [1] 16 27  4 19  1 15  6  2
## [1] 16 27  2
```

From the plot below, we could clearly to see correlation between each important variables, and how those 7 important variables to distinguish the zonality 1 and zonality 3 partition.The plot can clearly shows that bmiz, ast and u_acid can not distinguish these two zonalities. Four variables (age, fibrosis, sex, waist) will be included in the decision tree.



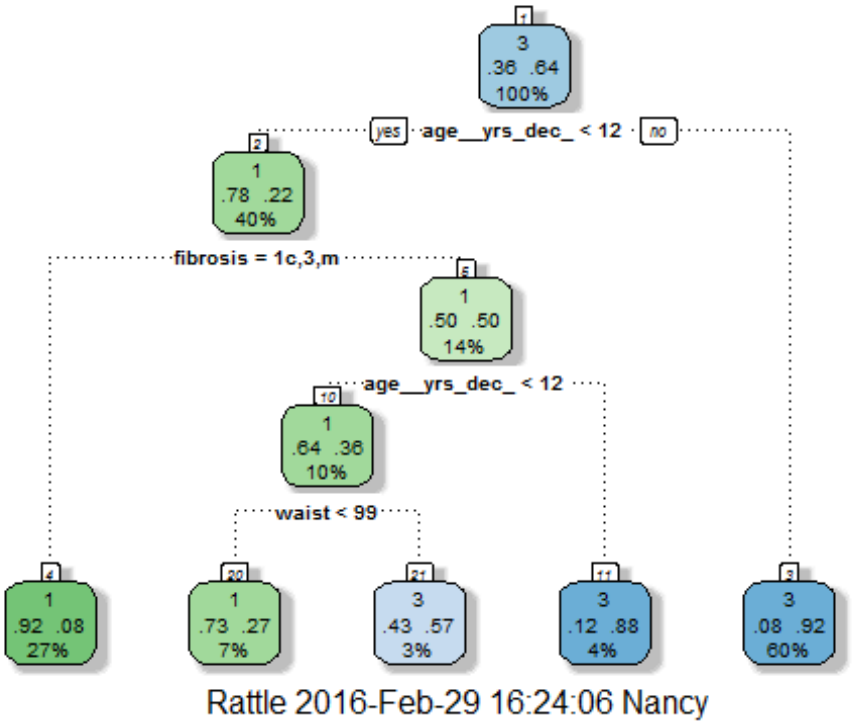
3.3.Build up a preliminary decision tree model

A decision tree model is build up. Since this is the preliminary model, it will be adjusted later.

- The formular of model is:

$$Zonality = Age + Fibrosis + Waist + error$$

The Decision tree plot below show that how the decision tree to split the data into zone1 and zone3.This model is built using the training dataset, and it was examined by the test dataset.




```
##    n
## 1 34
```

There are 34 observations between predicted zonality and zonality.The error rate for this model is 0.1588235. That means this model can reach almost 85% correction rate for predicted zonality comparing original zonality result in test dataset. The details for this model is below.

```
## n= 220
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 220 79 3 (0.35909091 0.64090909)
##    2) age__yrs_dec_< 12.44932 89 20 1 (0.77528090 0.22471910)
##      4) fibrosis=1c,3,m 59  5 1 (0.91525424 0.08474576) *
##      5) fibrosis=0,1a,1b,2 30 15 1 (0.50000000 0.50000000)
##       10) age__yrs_dec_< 11.75068 22  8 1 (0.63636364 0.36363636)
##         20) waist< 98.9935 15  4 1 (0.73333333 0.26666667) *
##         21) waist>=98.9935 7  3 3 (0.42857143 0.57142857) *
##       11) age__yrs_dec_>=11.75068 8  1 3 (0.12500000 0.87500000) *
##    3) age__yrs_dec_>=12.44932 131 10 3 (0.07633588 0.92366412) *
```

4 Discussion

The limitation includes that the sample size is not large enough, the bigger data the better for building up the classifier, and other point is lots of missing value in the dataset. Inorder to keep the accuracy, those variables are removed from the data.