The random forest method is used in the NAFLD study

Nan Zhang

August 14, 2015

1.Project Design

Backgroud: Zone 1 and Zone 3 Steatosis as distinct phenotypes of Nonalcoholic Fatty Liver Disease in Children and their Relationship with Overall Disease Severity.

Purpose: To determine the association between the zonality of steatosis and clinical, demographic, and histological features in children with NAFLD.

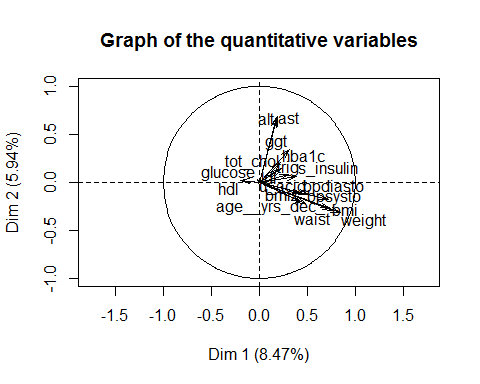
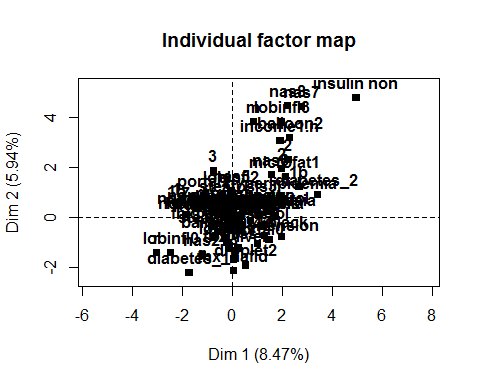
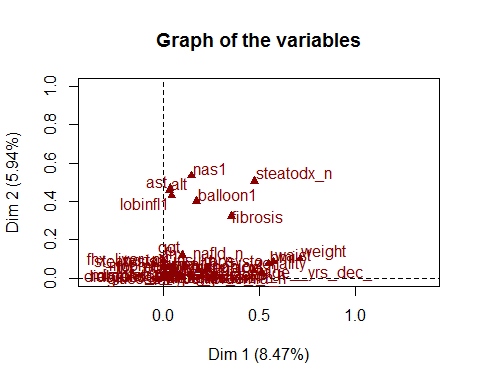
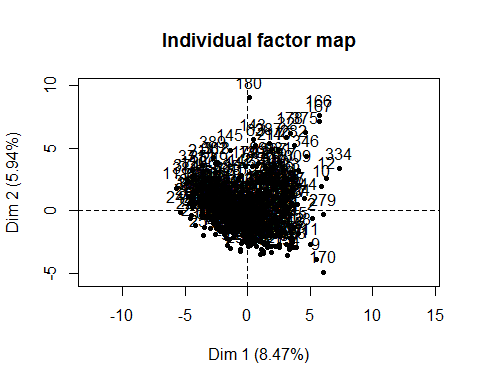
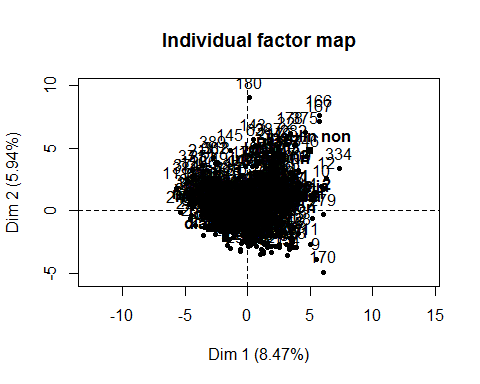
1. Import and maniplute data

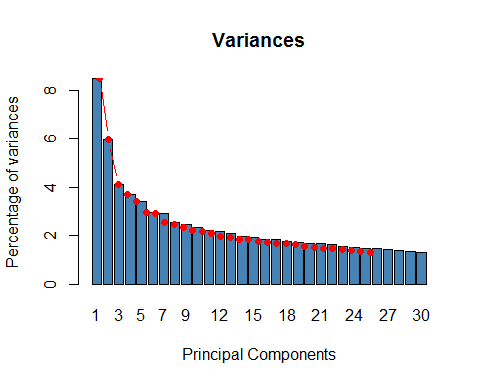
The original data contain 813 observations and 165 variables.Irrelevant variables for this study are removed , so 51 variables(16 numeric variables,35 category variables) are kept for checking the missing value or outlier. After data manipulation process 42 variables are kept into factor analysis.

## [1] "gender" "race" "hispanic"   
## [4] "bmi" "bpsysto" "bpdiasto"   
## [7] "u\_acid" "hba1c" "ast"   
## [10] "alt" "tot\_chol" "ldl"   
## [13] "glucose" "steatosis1" "fibrosis"   
## [16] "lobinfl1" "portal1" "balloon1"   
## [19] "bmiz" "age\_\_yrs\_dec\_" "zonality"   
## [22] "nas1" "ggt" "waist"   
## [25] "trig" "hdl" "s\_insulin"   
## [28] "income1" "diabetes\_1\_n" "diabetes\_2\_n"   
## [31] "hypertension\_n" "insulin\_n" "weight"   
## [34] "droplet\_n" "microfat\_n" "steatodx\_n"   
## [37] "hyperlipidemia\_n" "cholelithiasis\_n" "fhx\_liver\_n"   
## [40] "fhx\_nafld\_n" "fhx\_obesity\_n" "fhx\_cholesterol\_n"

1. Factor Analysis for Mix data

set.seed(3101318)  
t1<-FAMD(biopsy\_n, ncp =30)

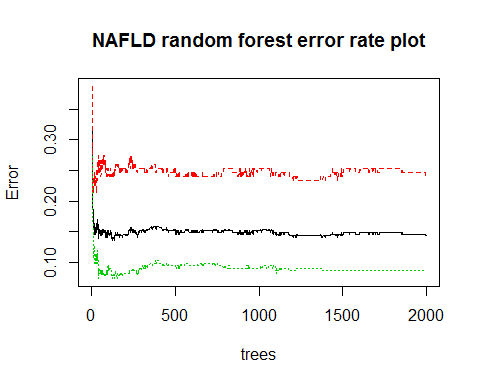


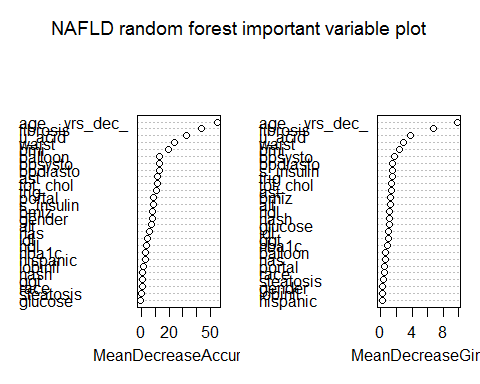


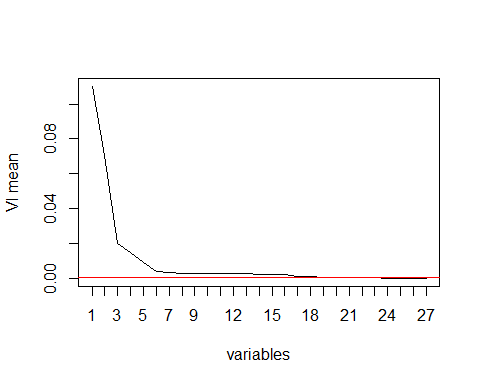
The third plot is all variables projection in dim1 and dim2. The fourth plot is numeric variables in dim1 and dim2. The fifth plot is categorical variables in dim1 and dim2. The sixth plot is scree plot which shows ~92% of the informations (variances) contained in the data are retained by the first two principal components.

## $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

From the factor analysis,continuous variables (the 16 first selected):Bmi, bpsysto, bpdiasto, uacid , hba1c, ast, alt, tot\_chol, ldl, glucose, bmiz, age, ggt, waist, trig, hdl.Categories variables (the 12 first selected):gender,nas, race, hispanic, hba1c, steatosis, fibrosis, lobinfl, portal, balloon, sinsulin,nash.

1. Random Forest method to use find out the most important predictors  From the plot above, we can get the average error rate is 0.15, the zone1 and zone3 error rate difference is 0.2.If there are more data to train this model, the difference of error rate will probably reduce.

 This plot shows the variables contribution for the MeanDecreaseAccurary and MeanDecreaseGini

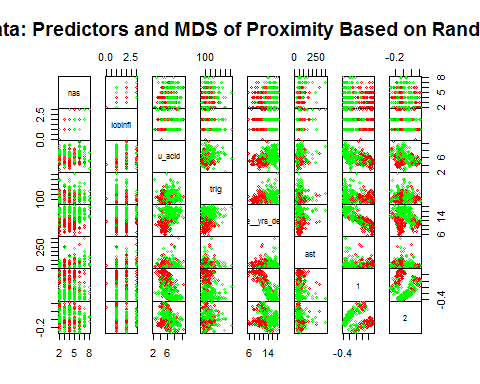


## [1] 16 27 4 19 1 15 6 2 8 20 3 22 13 14 7 17 24 5 9 21 18

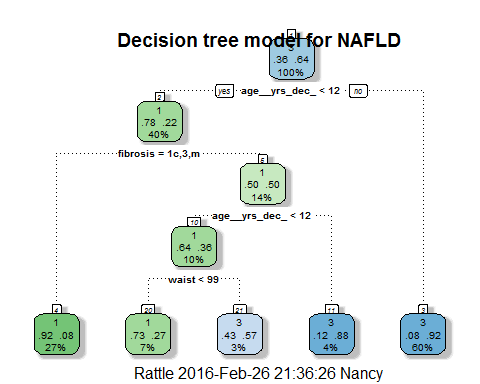
## [1] 16 27 4 19 1 15 6 2

## [1] 16 27 2

From analysis above, sixteen variables above threshold(the red line in the plot) are selected as important variables, and eight variables(age, fibrosis, waist, u\_acid, bmiz, ast,bmi,bpsysto) are picked as predictors for zonality partition. But bmi can be remove from here.



From the plot above, we can see how each important variable to distinguish the zonality 1 and zonality 3 partition.The plot can clearly shows that bmiz and ast can not distinguish these two zonalities. Four variables (age, fibrosis, u\_acid, waist) will be included in the decision tree.

5.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:

## n  
## 1 34

This model is built using the training dataset, and it was examined by the test dataset. 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) \*