Initial medical data preparation

Tejasvini Mavuleti

2022-08-03

## Prepare for initial analysis

set.seed(123)  
library(lubridate)

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.1

library(randomForest)

## Warning: package 'randomForest' was built under R version 4.2.1

## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':  
##   
## margin

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:randomForest':  
##   
## combine

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(rpart)  
library(caret)

## Warning: package 'caret' was built under R version 4.2.1

## Loading required package: lattice

## Warning: package 'lattice' was built under R version 4.2.1

library(e1071)

## Warning: package 'e1071' was built under R version 4.2.1

library(corrplot)

## corrplot 0.92 loaded

## Functions to clean datasets

Read data sets from csv file

clean\_dataset <- function() {  
 datasetloc = "C:/Users/mavul/OneDrive/Desktop/Health care data.csv"  
 if (file.exists(datasetloc)) {  
 data <- read.csv(file=datasetloc, header = T)  
 }  
 return(data)  
}

Partition #The data were partitioned into a test and training set using a 70/30 split

set.seed(100)  
train <- sample(nrow(clean\_dataset()), 0.7\*nrow(clean\_dataset()), replace = FALSE)  
TrainSet <- clean\_dataset()[train,]  
TestSet <- clean\_dataset()[-train,]  
summary(TrainSet)

## id gender dob zipcode   
## Length:1400 Length:1400 Length:1400 Min. :10001   
## Class :character Class :character Class :character 1st Qu.:43221   
## Mode :character Mode :character Mode :character Median :60612   
## Mean :62877   
## 3rd Qu.:90008   
## Max. :94110   
## employment\_status education marital\_status children   
## Length:1400 Length:1400 Length:1400 Min. :0.000   
## Class :character Class :character Class :character 1st Qu.:1.000   
## Mode :character Mode :character Mode :character Median :2.000   
## Mean :2.227   
## 3rd Qu.:3.000   
## Max. :7.000   
## ancestry avg\_commute daily\_internet\_use available\_vehicles  
## Length:1400 Min. :-2.47 Min. :1.010 Min. :0.000   
## Class :character 1st Qu.:23.61 1st Qu.:4.070 1st Qu.:1.000   
## Mode :character Median :30.39 Median :5.020 Median :2.000   
## Mean :30.43 Mean :5.009 Mean :1.746   
## 3rd Qu.:37.18 3rd Qu.:5.945 3rd Qu.:3.000   
## Max. :63.73 Max. :8.640 Max. :4.000   
## military\_service disease   
## Length:1400 Length:1400   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##

summary(TestSet)

## id gender dob zipcode   
## Length:600 Length:600 Length:600 Min. :10001   
## Class :character Class :character Class :character 1st Qu.:43221   
## Mode :character Mode :character Mode :character Median :60612   
## Mean :64579   
## 3rd Qu.:90008   
## Max. :94110   
## employment\_status education marital\_status children   
## Length:600 Length:600 Length:600 Min. :0.000   
## Class :character Class :character Class :character 1st Qu.:1.000   
## Mode :character Mode :character Mode :character Median :2.000   
## Mean :2.358   
## 3rd Qu.:3.000   
## Max. :7.000   
## ancestry avg\_commute daily\_internet\_use available\_vehicles  
## Length:600 Min. : 4.63 Min. :1.250 Min. :0.000   
## Class :character 1st Qu.:23.30 1st Qu.:3.938 1st Qu.:1.000   
## Mode :character Median :29.91 Median :4.930 Median :2.000   
## Mean :30.26 Mean :4.958 Mean :1.747   
## 3rd Qu.:37.09 3rd Qu.:5.990 3rd Qu.:3.000   
## Max. :61.66 Max. :8.820 Max. :4.000   
## military\_service disease   
## Length:600 Length:600   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##

## Analysing the disease

The data set will predict the marital status with selected attributes that contributes to the analysis

disease\_TrainSet <- select(TrainSet, gender, employment\_status, education, marital\_status, ancestry, disease)  
disease\_TestSet <- select(TestSet, gender, employment\_status, education, marital\_status, ancestry, disease)  
disease\_TrainSet$disease <- as.factor(disease\_TrainSet$disease)

## Logistic Regression Model

The model was fit using a binomial logistic regression with the glm function in R, with family = binomial on the training data.

fit <- glm(disease~.,data=disease\_TrainSet,family=binomial())  
summary(fit)

##   
## Call:  
## glm(formula = disease ~ ., family = binomial(), data = disease\_TrainSet)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.3393 0.4403 0.5440 0.6441 1.0067   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.19000 0.37667 5.814 6.1e-09 \*\*\*  
## gendermale 0.03196 0.14765 0.216 0.828649   
## employment\_statusretired -0.59919 0.16538 -3.623 0.000291 \*\*\*  
## employment\_statusstudent -0.40097 1.13363 -0.354 0.723564   
## employment\_statusunemployed -0.17617 0.28038 -0.628 0.529790   
## educationhighschool -0.31052 0.20498 -1.515 0.129800   
## educationhighscool 13.38707 723.39477 0.019 0.985235   
## educationmasters -0.12839 0.21414 -0.600 0.548810   
## educationphd/md 0.08960 0.26839 0.334 0.738493   
## educationphD/MD 14.00046 481.18170 0.029 0.976788   
## marital\_statussingle 0.31020 0.18310 1.694 0.090238 .   
## ancestryBelgium -0.21598 0.46794 -0.462 0.644394   
## ancestryCzech Republic 0.16894 0.53153 0.318 0.750605   
## ancestryDenmark -0.83234 0.43726 -1.904 0.056972 .   
## ancestryEngland 0.32945 0.50949 0.647 0.517882   
## ancestryFinland -0.40610 0.47259 -0.859 0.390165   
## ancestryFrance -0.33601 0.47969 -0.700 0.483638   
## ancestryGermany -0.21203 0.46892 -0.452 0.651142   
## ancestryHungary -0.40526 0.48217 -0.840 0.400631   
## ancestryIreland -0.05155 0.46598 -0.111 0.911906   
## ancestryItaly -0.35404 0.47117 -0.751 0.452413   
## ancestryNetherlands 0.07750 0.48447 0.160 0.872910   
## ancestryPoland -0.35193 0.45597 -0.772 0.440208   
## ancestryPortugal 0.06311 0.48426 0.130 0.896308   
## ancestryRussia -0.86443 0.44835 -1.928 0.053852 .   
## ancestryScotland -0.66546 0.46906 -1.419 0.155982   
## ancestrySpain 0.01601 0.51393 0.031 0.975150   
## ancestrySweden 0.16839 0.49568 0.340 0.734076   
## ancestrySwitzerland -0.52548 0.43980 -1.195 0.232159   
## ancestryUkraine -0.13572 0.51885 -0.262 0.793645   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1257.3 on 1399 degrees of freedom  
## Residual deviance: 1214.1 on 1370 degrees of freedom  
## AIC: 1274.1  
##   
## Number of Fisher Scoring iterations: 14

confint(fit)

## Waiting for profiling to be done...

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## 2.5 % 97.5 %  
## (Intercept) 1.49310795 2.981983016  
## gendermale -0.25786241 0.321451399  
## employment\_statusretired -0.92785940 -0.278718811  
## employment\_statusstudent -2.31644858 2.593249869  
## employment\_statusunemployed -0.71917959 0.382573039  
## educationhighschool -0.70805338 0.096693127  
## educationhighscool -69.45973894 NA  
## educationmasters -0.54008833 0.301313496  
## educationphd/md -0.41765298 0.639333730  
## educationphD/MD -85.57307046 NA  
## marital\_statussingle -0.04143041 0.677631890  
## ancestryBelgium -1.15214453 0.701580323  
## ancestryCzech Republic -0.86444750 1.250416462  
## ancestryDenmark -1.72198660 0.006793701  
## ancestryEngland -0.66860072 1.356370401  
## ancestryFinland -1.35123141 0.520177091  
## ancestryFrance -1.29176670 0.608895559  
## ancestryGermany -1.14995333 0.707492641  
## ancestryHungary -1.36593697 0.544167167  
## ancestryIreland -0.98383541 0.862453401  
## ancestryItaly -1.29647330 0.569533833  
## ancestryNetherlands -0.88319757 1.038468463  
## ancestryPoland -1.26989871 0.534913406  
## ancestryPortugal -0.89729025 1.023564349  
## ancestryRussia -1.77291681 -0.000330003  
## ancestryScotland -1.60667582 0.249496457  
## ancestrySpain -0.99118858 1.050582435  
## ancestrySweden -0.80919936 1.158627641  
## ancestrySwitzerland -1.41776928 0.322136248  
## ancestryUkraine -1.15273134 0.907692356

exp(coef(fit))

## (Intercept) gendermale   
## 8.935227e+00 1.032472e+00   
## employment\_statusretired employment\_statusstudent   
## 5.492537e-01 6.696732e-01   
## employment\_statusunemployed educationhighschool   
## 8.384750e-01 7.330646e-01   
## educationhighscool educationmasters   
## 6.515256e+05 8.795114e-01   
## educationphd/md educationphD/MD   
## 1.093741e+00 1.203159e+06   
## marital\_statussingle ancestryBelgium   
## 1.363698e+00 8.057475e-01   
## ancestryCzech Republic ancestryDenmark   
## 1.184052e+00 4.350307e-01   
## ancestryEngland ancestryFinland   
## 1.390197e+00 6.662430e-01   
## ancestryFrance ancestryGermany   
## 7.146188e-01 8.089375e-01   
## ancestryHungary ancestryIreland   
## 6.668019e-01 9.497524e-01   
## ancestryItaly ancestryNetherlands   
## 7.018501e-01 1.080580e+00   
## ancestryPoland ancestryPortugal   
## 7.033265e-01 1.065146e+00   
## ancestryRussia ancestryScotland   
## 4.212932e-01 5.140393e-01   
## ancestrySpain ancestrySweden   
## 1.016138e+00 1.183392e+00   
## ancestrySwitzerland ancestryUkraine   
## 5.912699e-01 8.730862e-01

exp(confint(fit))

## Waiting for profiling to be done...

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## 2.5 % 97.5 %  
## (Intercept) 4.450907e+00 19.7268966  
## gendermale 7.727015e-01 1.3791280  
## employment\_statusretired 3.953992e-01 0.7567527  
## employment\_statusstudent 9.862322e-02 13.3731621  
## employment\_statusunemployed 4.871518e-01 1.4660520  
## educationhighschool 4.926022e-01 1.1015223  
## educationhighscool 6.823680e-31 NA  
## educationmasters 5.826968e-01 1.3516330  
## educationphd/md 6.585907e-01 1.8952177  
## educationphD/MD 6.856267e-38 NA  
## marital\_statussingle 9.594161e-01 1.9692089  
## ancestryBelgium 3.159585e-01 2.0169376  
## ancestryCzech Republic 4.212842e-01 3.4917969  
## ancestryDenmark 1.787108e-01 1.0068168  
## ancestryEngland 5.124251e-01 3.8820773  
## ancestryFinland 2.589212e-01 1.6823255  
## ancestryFrance 2.747849e-01 1.8383999  
## ancestryGermany 3.166515e-01 2.0288977  
## ancestryHungary 2.551415e-01 1.7231727  
## ancestryIreland 3.738744e-01 2.3689656  
## ancestryItaly 2.734946e-01 1.7674429  
## ancestryNetherlands 4.134587e-01 2.8248873  
## ancestryPoland 2.808601e-01 1.7073004  
## ancestryPortugal 4.076729e-01 2.7830970  
## ancestryRussia 1.698369e-01 0.9996701  
## ancestryScotland 2.005532e-01 1.2833790  
## ancestrySpain 3.711353e-01 2.8593160  
## ancestrySweden 4.452144e-01 3.1855585  
## ancestrySwitzerland 2.422538e-01 1.3800728  
## ancestryUkraine 3.157731e-01 2.4785962

#predict(fit, type="response")   
#residuals(fit, type="deviance")

## Performance

Probabilities for the response variable based on the test data were assigned using the predict function.

#probs <- predict(fit, test, type = "response")  
#pred <- predict(fit, newdata = TestSet)  
#pred

## Confusion Matrix

#confusionMatrix(pred, TestSet$disease)

## Random forest model

Apply randomforest model

# Fine tuning parameters of Random Forest model  
model2 <- randomForest(disease ~ ., data = disease\_TrainSet, importance = TRUE)  
model2

##   
## Call:  
## randomForest(formula = disease ~ ., data = disease\_TrainSet, importance = TRUE)   
## Type of random forest: classification  
## Number of trees: 500  
## No. of variables tried at each split: 2  
##   
## OOB estimate of error rate: 83.29%  
## Confusion matrix:  
## Alzheimer's disease breast cancer diabetes endometriosis  
## Alzheimer's disease 85 42 2 1  
## breast cancer 31 30 0 1  
## diabetes 33 7 1 1  
## endometriosis 10 15 2 0  
## gastritis 21 13 0 0  
## heart disease 17 8 1 0  
## HIV/AIDS 9 8 0 0  
## hypertension 52 34 0 1  
## kidney disease 34 20 0 1  
## multiple sclerosis 29 14 0 0  
## prostate cancer 60 0 0 0  
## schizophrenia 17 3 0 0  
## skin cancer 60 23 1 0  
## gastritis heart disease HIV/AIDS hypertension  
## Alzheimer's disease 1 2 1 31  
## breast cancer 0 1 0 24  
## diabetes 0 1 1 10  
## endometriosis 0 0 1 6  
## gastritis 0 0 1 12  
## heart disease 0 0 0 8  
## HIV/AIDS 1 0 0 6  
## hypertension 1 1 1 29  
## kidney disease 0 1 1 13  
## multiple sclerosis 1 1 0 6  
## prostate cancer 0 0 1 7  
## schizophrenia 2 0 0 7  
## skin cancer 0 0 1 17  
## kidney disease multiple sclerosis prostate cancer  
## Alzheimer's disease 5 2 37  
## breast cancer 6 1 0  
## diabetes 6 0 18  
## endometriosis 6 0 0  
## gastritis 2 0 14  
## heart disease 4 0 22  
## HIV/AIDS 4 0 21  
## hypertension 12 1 43  
## kidney disease 17 1 39  
## multiple sclerosis 4 0 14  
## prostate cancer 4 0 55  
## schizophrenia 0 1 8  
## skin cancer 7 0 41  
## schizophrenia skin cancer class.error  
## Alzheimer's disease 0 23 0.6336207  
## breast cancer 0 11 0.7142857  
## diabetes 0 5 0.9879518  
## endometriosis 0 5 1.0000000  
## gastritis 2 5 1.0000000  
## heart disease 0 4 1.0000000  
## HIV/AIDS 1 2 1.0000000  
## hypertension 1 21 0.8527919  
## kidney disease 0 10 0.8759124  
## multiple sclerosis 1 7 1.0000000  
## prostate cancer 0 4 0.5801527  
## schizophrenia 0 2 1.0000000  
## skin cancer 0 17 0.8982036

# Predicting on train set  
predTrain <- predict(model2, disease\_TrainSet, type = "class")  
  
# Checking classification accuracy  
table(predTrain, disease\_TrainSet$disease)

##   
## predTrain Alzheimer's disease breast cancer diabetes endometriosis  
## Alzheimer's disease 147 13 31 12  
## breast cancer 23 65 7 13  
## diabetes 1 0 9 0  
## endometriosis 0 0 0 2  
## gastritis 0 0 0 0  
## heart disease 2 0 0 0  
## HIV/AIDS 0 0 1 1  
## hypertension 17 13 9 7  
## kidney disease 4 3 5 5  
## multiple sclerosis 0 0 0 0  
## prostate cancer 25 0 17 0  
## schizophrenia 0 0 0 0  
## skin cancer 13 11 4 5  
##   
## predTrain gastritis heart disease HIV/AIDS hypertension  
## Alzheimer's disease 21 13 6 47  
## breast cancer 12 9 9 20  
## diabetes 0 0 0 0  
## endometriosis 0 0 0 0  
## gastritis 5 0 0 0  
## heart disease 0 11 0 1  
## HIV/AIDS 1 1 13 1  
## hypertension 13 6 4 86  
## kidney disease 1 2 1 3  
## multiple sclerosis 0 0 0 0  
## prostate cancer 12 18 16 28  
## schizophrenia 1 0 0 0  
## skin cancer 4 4 3 11  
##   
## predTrain kidney disease multiple sclerosis prostate cancer  
## Alzheimer's disease 31 28 44  
## breast cancer 18 15 0  
## diabetes 0 0 0  
## endometriosis 0 0 0  
## gastritis 0 0 0  
## heart disease 0 1 0  
## HIV/AIDS 1 0 0  
## hypertension 9 5 3  
## kidney disease 35 2 0  
## multiple sclerosis 0 6 0  
## prostate cancer 34 14 83  
## schizophrenia 0 0 0  
## skin cancer 9 6 1  
##   
## predTrain schizophrenia skin cancer  
## Alzheimer's disease 17 43  
## breast cancer 3 19  
## diabetes 0 0  
## endometriosis 0 0  
## gastritis 0 0  
## heart disease 0 0  
## HIV/AIDS 0 0  
## hypertension 7 16  
## kidney disease 0 6  
## multiple sclerosis 1 0  
## prostate cancer 7 36  
## schizophrenia 3 0  
## skin cancer 2 47

model2 <- na.omit(model2)   
  
# Predicting on Validation set  
predValid <- predict(model2, disease\_TestSet, type = "class")  
  
# Checking classification accuracy  
mean(predValid == disease\_TestSet$disease)

## [1] 0.1666667

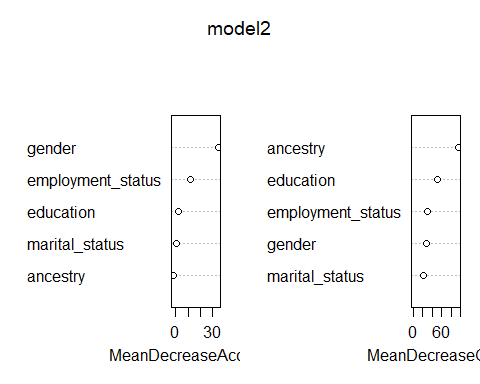
table(predValid,disease\_TestSet$disease)

##   
## predValid Alzheimer's disease breast cancer diabetes endometriosis  
## Alzheimer's disease 43 14 15 7  
## breast cancer 14 8 5 8  
## diabetes 1 0 0 0  
## endometriosis 0 0 0 0  
## gastritis 0 0 0 0  
## heart disease 0 0 0 0  
## HIV/AIDS 1 1 0 1  
## hypertension 8 7 10 5  
## kidney disease 7 4 1 0  
## multiple sclerosis 0 0 0 0  
## prostate cancer 26 0 4 0  
## schizophrenia 0 0 0 0  
## skin cancer 7 6 1 0  
##   
## predValid gastritis heart disease HIV/AIDS hypertension  
## Alzheimer's disease 10 10 2 31  
## breast cancer 3 0 3 17  
## diabetes 0 1 0 0  
## endometriosis 0 0 0 0  
## gastritis 0 0 0 0  
## heart disease 0 1 0 2  
## HIV/AIDS 0 0 0 2  
## hypertension 3 1 5 11  
## kidney disease 3 2 1 6  
## multiple sclerosis 0 0 0 1  
## prostate cancer 10 4 11 18  
## schizophrenia 0 0 0 0  
## skin cancer 1 4 6 13  
##   
## predValid kidney disease multiple sclerosis prostate cancer  
## Alzheimer's disease 19 14 15  
## breast cancer 5 3 0  
## diabetes 0 0 0  
## endometriosis 0 0 0  
## gastritis 0 0 0  
## heart disease 1 0 1  
## HIV/AIDS 0 0 0  
## hypertension 8 4 3  
## kidney disease 4 1 1  
## multiple sclerosis 0 0 0  
## prostate cancer 10 13 29  
## schizophrenia 0 0 0  
## skin cancer 1 1 0  
##   
## predValid schizophrenia skin cancer  
## Alzheimer's disease 6 29  
## breast cancer 2 11  
## diabetes 0 0  
## endometriosis 0 1  
## gastritis 0 0  
## heart disease 0 1  
## HIV/AIDS 1 0  
## hypertension 1 3  
## kidney disease 1 4  
## multiple sclerosis 0 0  
## prostate cancer 4 13  
## schizophrenia 0 0  
## skin cancer 0 4

# To check important variables  
importance(model2)

## Alzheimer's disease breast cancer diabetes endometriosis  
## gender -0.3069681 28.0718616 -1.380107 7.86012944  
## employment\_status 8.8371133 -0.2679983 1.728054 -3.34987951  
## education -8.1680115 1.3234619 2.799283 -1.92656883  
## marital\_status -0.8301061 -7.1404512 5.529093 -0.04084138  
## ancestry -4.4430299 -5.1554326 4.351547 -4.63476985  
## gastritis heart disease HIV/AIDS hypertension  
## gender -3.092683 4.3303144 2.8995935 1.5726714  
## employment\_status -2.737173 -0.4860694 5.9254210 4.7360929  
## education -7.936120 2.2773281 4.0985963 0.4401083  
## marital\_status 0.690441 4.2159322 -0.3571497 1.8884930  
## ancestry -2.603470 -0.4569619 5.6758207 1.9357525  
## kidney disease multiple sclerosis prostate cancer  
## gender 10.428706 1.6532290 37.528819  
## employment\_status 9.456225 -5.7025661 1.454414  
## education 8.865393 -1.5262814 2.838719  
## marital\_status 9.664055 -0.5272205 -6.919649  
## ancestry 7.473705 -0.2640322 -7.662150  
## schizophrenia skin cancer MeanDecreaseAccuracy  
## gender -7.7424957 5.196389 35.3158589  
## employment\_status 1.8606180 5.647893 12.0488859  
## education 2.9206449 4.813521 1.8113220  
## marital\_status -4.6329632 2.994119 0.9381829  
## ancestry -0.9734494 3.714359 -2.2621348  
## MeanDecreaseGini  
## gender 28.16827  
## employment\_status 30.39493  
## education 50.73786  
## marital\_status 21.08354  
## ancestry 96.29308

varImpPlot(model2)



## Naive Bayes Model

NBclassfier = naiveBayes(disease~., data=disease\_TrainSet)  
print(NBclassfier)

##   
## Naive Bayes Classifier for Discrete Predictors  
##   
## Call:  
## naiveBayes.default(x = X, y = Y, laplace = laplace)  
##   
## A-priori probabilities:  
## Y  
## Alzheimer's disease breast cancer diabetes endometriosis   
## 0.16571429 0.07500000 0.05928571 0.03214286   
## gastritis heart disease HIV/AIDS hypertension   
## 0.05000000 0.04571429 0.03714286 0.14071429   
## kidney disease multiple sclerosis prostate cancer schizophrenia   
## 0.09785714 0.05500000 0.09357143 0.02857143   
## skin cancer   
## 0.11928571   
##   
## Conditional probabilities:  
## gender  
## Y female male  
## Alzheimer's disease 0.4870690 0.5129310  
## breast cancer 1.0000000 0.0000000  
## diabetes 0.4337349 0.5662651  
## endometriosis 1.0000000 0.0000000  
## gastritis 0.4714286 0.5285714  
## heart disease 0.3906250 0.6093750  
## HIV/AIDS 0.4423077 0.5576923  
## hypertension 0.5025381 0.4974619  
## kidney disease 0.5109489 0.4890511  
## multiple sclerosis 0.4545455 0.5454545  
## prostate cancer 0.0000000 1.0000000  
## schizophrenia 0.4250000 0.5750000  
## skin cancer 0.4670659 0.5329341  
##   
## employment\_status  
## Y employed retired student unemployed  
## Alzheimer's disease 0.288793103 0.586206897 0.004310345 0.120689655  
## breast cancer 0.390476190 0.485714286 0.009523810 0.114285714  
## diabetes 0.313253012 0.530120482 0.012048193 0.144578313  
## endometriosis 0.288888889 0.533333333 0.000000000 0.177777778  
## gastritis 0.428571429 0.371428571 0.042857143 0.157142857  
## heart disease 0.453125000 0.390625000 0.000000000 0.156250000  
## HIV/AIDS 0.673076923 0.115384615 0.019230769 0.192307692  
## hypertension 0.345177665 0.517766497 0.010152284 0.126903553  
## kidney disease 0.489051095 0.364963504 0.007299270 0.138686131  
## multiple sclerosis 0.363636364 0.493506494 0.025974026 0.116883117  
## prostate cancer 0.381679389 0.488549618 0.000000000 0.129770992  
## schizophrenia 0.350000000 0.500000000 0.075000000 0.075000000  
## skin cancer 0.407185629 0.526946108 0.000000000 0.065868263  
##   
## education  
## Y bachelors highschool highscool masters  
## Alzheimer's disease 0.508620690 0.245689655 0.000000000 0.159482759  
## breast cancer 0.580952381 0.190476190 0.000000000 0.152380952  
## diabetes 0.481927711 0.313253012 0.000000000 0.132530120  
## endometriosis 0.533333333 0.222222222 0.000000000 0.111111111  
## gastritis 0.500000000 0.242857143 0.000000000 0.157142857  
## heart disease 0.531250000 0.203125000 0.000000000 0.218750000  
## HIV/AIDS 0.346153846 0.307692308 0.038461538 0.192307692  
## hypertension 0.548223350 0.208121827 0.000000000 0.126903553  
## kidney disease 0.562043796 0.233576642 0.007299270 0.116788321  
## multiple sclerosis 0.649350649 0.155844156 0.012987013 0.103896104  
## prostate cancer 0.557251908 0.198473282 0.000000000 0.145038168  
## schizophrenia 0.475000000 0.225000000 0.000000000 0.100000000  
## skin cancer 0.556886228 0.185628743 0.000000000 0.137724551  
## education  
## Y phd/md phD/MD  
## Alzheimer's disease 0.086206897 0.000000000  
## breast cancer 0.076190476 0.000000000  
## diabetes 0.072289157 0.000000000  
## endometriosis 0.133333333 0.000000000  
## gastritis 0.071428571 0.028571429  
## heart disease 0.046875000 0.000000000  
## HIV/AIDS 0.096153846 0.019230769  
## hypertension 0.111675127 0.005076142  
## kidney disease 0.072992701 0.007299270  
## multiple sclerosis 0.064935065 0.012987013  
## prostate cancer 0.099236641 0.000000000  
## schizophrenia 0.125000000 0.075000000  
## skin cancer 0.119760479 0.000000000  
##   
## marital\_status  
## Y married single  
## Alzheimer's disease 0.7931034 0.2068966  
## breast cancer 0.8000000 0.2000000  
## diabetes 0.7590361 0.2409639  
## endometriosis 0.7777778 0.2222222  
## gastritis 0.7428571 0.2571429  
## heart disease 0.5937500 0.4062500  
## HIV/AIDS 0.6923077 0.3076923  
## hypertension 0.7664975 0.2335025  
## kidney disease 0.7153285 0.2846715  
## multiple sclerosis 0.7272727 0.2727273  
## prostate cancer 0.7175573 0.2824427  
## schizophrenia 0.7500000 0.2500000  
## skin cancer 0.7485030 0.2514970  
##   
## ancestry  
## Y Austria Belgium Czech Republic Denmark  
## Alzheimer's disease 0.04310345 0.05172414 0.03017241 0.08189655  
## breast cancer 0.06666667 0.04761905 0.05714286 0.03809524  
## diabetes 0.04819277 0.02409639 0.02409639 0.06024096  
## endometriosis 0.04444444 0.02222222 0.08888889 0.06666667  
## gastritis 0.10000000 0.02857143 0.01428571 0.05714286  
## heart disease 0.04687500 0.06250000 0.07812500 0.03125000  
## HIV/AIDS 0.03846154 0.05769231 0.03846154 0.00000000  
## hypertension 0.05076142 0.04568528 0.05076142 0.04568528  
## kidney disease 0.06569343 0.04379562 0.05839416 0.04379562  
## multiple sclerosis 0.02597403 0.05194805 0.02597403 0.07792208  
## prostate cancer 0.05343511 0.07633588 0.02290076 0.03816794  
## schizophrenia 0.05000000 0.05000000 0.05000000 0.02500000  
## skin cancer 0.04790419 0.08383234 0.05389222 0.04191617  
## ancestry  
## Y England Finland France Germany Hungary  
## Alzheimer's disease 0.03448276 0.05172414 0.04741379 0.05172414 0.04741379  
## breast cancer 0.04761905 0.03809524 0.04761905 0.02857143 0.04761905  
## diabetes 0.06024096 0.02409639 0.08433735 0.02409639 0.02409639  
## endometriosis 0.02222222 0.06666667 0.04444444 0.04444444 0.04444444  
## gastritis 0.05714286 0.01428571 0.04285714 0.05714286 0.01428571  
## heart disease 0.03125000 0.03125000 0.06250000 0.09375000 0.06250000  
## HIV/AIDS 0.03846154 0.07692308 0.00000000 0.07692308 0.07692308  
## hypertension 0.06598985 0.04060914 0.02538071 0.05076142 0.05076142  
## kidney disease 0.08029197 0.03649635 0.05109489 0.03649635 0.03649635  
## multiple sclerosis 0.07792208 0.05194805 0.07792208 0.02597403 0.07792208  
## prostate cancer 0.05343511 0.06870229 0.03053435 0.05343511 0.03816794  
## schizophrenia 0.05000000 0.02500000 0.00000000 0.10000000 0.02500000  
## skin cancer 0.05389222 0.04790419 0.05988024 0.05988024 0.04191617  
## ancestry  
## Y Ireland Italy Netherlands Poland Portugal  
## Alzheimer's disease 0.05172414 0.05172414 0.04310345 0.06034483 0.04310345  
## breast cancer 0.01904762 0.06666667 0.04761905 0.06666667 0.10476190  
## diabetes 0.10843373 0.08433735 0.07228916 0.07228916 0.02409639  
## endometriosis 0.06666667 0.02222222 0.02222222 0.04444444 0.00000000  
## gastritis 0.11428571 0.02857143 0.04285714 0.05714286 0.07142857  
## heart disease 0.07812500 0.09375000 0.06250000 0.04687500 0.01562500  
## HIV/AIDS 0.07692308 0.03846154 0.05769231 0.00000000 0.05769231  
## hypertension 0.05076142 0.02030457 0.05583756 0.08629442 0.06091371  
## kidney disease 0.04379562 0.05109489 0.05839416 0.05109489 0.04379562  
## multiple sclerosis 0.05194805 0.01298701 0.06493506 0.02597403 0.09090909  
## prostate cancer 0.04580153 0.09160305 0.05343511 0.03816794 0.07633588  
## schizophrenia 0.02500000 0.02500000 0.05000000 0.07500000 0.05000000  
## skin cancer 0.08383234 0.04790419 0.06586826 0.02994012 0.04790419  
## ancestry  
## Y Russia Scotland Spain Sweden Switzerland  
## Alzheimer's disease 0.07327586 0.05603448 0.03448276 0.03879310 0.07327586  
## breast cancer 0.06666667 0.01904762 0.02857143 0.04761905 0.06666667  
## diabetes 0.06024096 0.03614458 0.01204819 0.10843373 0.02409639  
## endometriosis 0.02222222 0.06666667 0.04444444 0.08888889 0.08888889  
## gastritis 0.01428571 0.02857143 0.01428571 0.11428571 0.07142857  
## heart disease 0.03125000 0.03125000 0.01562500 0.04687500 0.03125000  
## HIV/AIDS 0.01923077 0.07692308 0.05769231 0.05769231 0.07692308  
## hypertension 0.04060914 0.03045685 0.07106599 0.05583756 0.05583756  
## kidney disease 0.05109489 0.04379562 0.05839416 0.05109489 0.07299270  
## multiple sclerosis 0.03896104 0.07792208 0.03896104 0.02597403 0.05194805  
## prostate cancer 0.03816794 0.04580153 0.05343511 0.04580153 0.04580153  
## schizophrenia 0.10000000 0.00000000 0.15000000 0.02500000 0.10000000  
## skin cancer 0.02994012 0.02994012 0.04191617 0.05389222 0.05389222  
## ancestry  
## Y Ukraine  
## Alzheimer's disease 0.03448276  
## breast cancer 0.04761905  
## diabetes 0.02409639  
## endometriosis 0.08888889  
## gastritis 0.05714286  
## heart disease 0.04687500  
## HIV/AIDS 0.07692308  
## hypertension 0.04568528  
## kidney disease 0.02189781  
## multiple sclerosis 0.02597403  
## prostate cancer 0.03053435  
## schizophrenia 0.02500000  
## skin cancer 0.02395210

## Performance

Probabilities for the response variable based on the test data were assigned using the predict function.

probs <- predict(fit, type = "response")  
pred <- predict(fit, newdata = TestSet)  
pred

## 7 8 9 10 14 19 20   
## 1.2751480 2.1384476 1.7726179 1.7757150 2.2674994 1.4067777 1.9960744   
## 22 24 25 27 33 39 40   
## 2.3910580 1.7395332 2.2674994 2.4806032 1.0643005 0.6467856 0.7580147   
## 41 43 44 45 46 49 52   
## 1.7591920 1.3255753 1.6288563 1.7579081 2.1438841 1.6641972 0.8709709   
## 54 57 58 60 68 71 74   
## 1.8679219 1.6068155 1.6964746 1.6627590 2.0066189 1.3263748 1.6402029   
## 75 80 84 86 89 93 99   
## 1.7908262 2.0614740 1.1855444 1.7917053 1.7911478 1.9779680 1.9960744   
## 103 104 108 111 112 113 117   
## 2.2089253 1.0891115 1.8334317 2.3583868 1.6539188 2.0776216 1.9036506   
## 119 121 122 126 132 134 135   
## 1.1424398 2.2216358 0.8750230 1.2548008 0.9762352 0.7584680 1.1765202   
## 138 139 140 141 148 151 152   
## 2.3580652 1.3961573 1.9779680 1.9170154 2.2379661 1.5245460 0.9283514   
## 160 162 167 168 171 172 178   
## 1.4505382 1.3984728 1.1581396 1.4107290 1.4107290 2.0693918 1.5392528   
## 181 185 188 189 190 191 195   
## 1.5712086 1.8700234 1.3576628 2.5952692 1.0470097 2.2850694 1.9522077   
## 199 206 210 217 218 220 231   
## 1.3266971 1.2464332 1.1847059 1.0150539 1.5490726 1.7754781 1.3444043   
## 235 236 238 239 243 245 252   
## 1.2684055 2.2060103 2.3903426 1.8495793 1.2548008 0.7584680 2.0776216   
## 260 265 269 271 275 280 284   
## 1.2388728 1.2675068 1.9740167 1.8359661 2.2379661 1.3433974 1.9179300   
## 292 294 295 303 306 307 311   
## 0.7583363 1.2464332 2.2144080 1.7575627 1.9740167 0.9573070 1.7591920   
## 312 319 320 323 326 330 331   
## 2.3583868 1.8494526 1.9054600 1.4107290 2.2147296 2.3644804 1.9736951   
## 333 343 344 358 360 361 367   
## 1.2708286 1.9522077 2.7005424 2.3275696 1.5565018 1.6288563 1.3892971   
## 368 369 370 373 377 379 380   
## 1.4806264 1.5192806 2.2850694 2.0454652 1.8700234 1.6618970 1.8158565   
## 385 388 391 392 399 405 407   
## 1.8279262 1.1583679 1.5565018 1.1403384 1.6674466 1.1855444 1.7283748   
## 414 415 417 424 426 427 438   
## 1.6858746 1.2548008 1.2355510 2.0862364 1.7591920 1.7847392 1.8014759   
## 439 443 446 450 451 452 453   
## 1.2367713 1.6357751 0.8750230 1.4486706 1.3896186 1.0470097 1.3300031   
## 462 475 483 485 491 494 496   
## 1.6964746 2.0454652 1.7808074 2.2156598 1.7757150 2.2305556 1.5565018   
## 497 498 500 507 509 512 515   
## 0.9603072 2.1704034 1.2291423 1.4108641 2.4037685 2.2379661 2.3909001   
## 524 525 531 533 534 539 540   
## 1.2175002 1.3748219 1.8493530 1.2166617 1.5392528 1.6387713 0.9253512   
## 541 542 545 547 552 555 563   
## 1.8815351 2.2994552 1.5490726 1.5908068 1.5275462 15.6935962 1.4651273   
## 565 569 570 572 579 580 581   
## 1.3263748 0.6148298 1.3575311 2.1781217 2.1019051 2.4486474 2.2600069   
## 584 588 590 593 597 603 606   
## 1.9889338 1.6683046 2.5952692 0.6148298 1.7983775 1.7911478 1.8166950   
## 607 613 615 617 618 622 624   
## 1.2612299 2.0862364 1.7978462 1.6850217 2.3429546 2.2379661 1.8539956   
## 625 627 632 633 634 636 637   
## 1.7209288 1.8463299 1.6855530 2.2994552 1.4644254 1.5199825 1.9641186   
## 652 658 675 676 677 678 683   
## 1.7169775 1.9578477 1.5650006 2.0542806 1.3539974 1.6850217 2.3589443   
## 688 689 692 693 701 702 707   
## 2.0059725 0.9582057 2.7005424 1.6387713 1.5561802 2.0995273 16.2680823   
## 715 716 718 719 721 722 725   
## 1.3961573 1.1583679 1.1855444 1.6288563 2.0059725 1.5908068 1.5712086   
## 726 727 728 729 730 739 740   
## 1.9575254 0.4479467 2.1710664 1.4108641 1.6634953 2.1566807 1.7847392   
## 741 745 746 748 749 760 761   
## 1.9779680 0.7548026 1.7169775 1.2459804 2.0225659 1.2292741 0.8480716   
## 762 763 766 767 771 773 774   
## 1.5718716 0.8709709 0.7583363 2.4799461 1.6387713 1.8868449 1.8539956   
## 777 779 787 790 791 792 807   
## 1.8356445 2.0199807 1.1583679 1.3787732 2.0636202 1.6227626 0.5979918   
## 816 819 822 826 829 830 831   
## 1.9641186 1.5275462 1.6850217 1.6227626 2.2379661 1.7002604 1.5574859   
## 834 836 845 847 848 855 859   
## 1.4742178 1.4943738 1.5712086 0.6299476 1.5675887 1.6555120 0.9573070   
## 863 868 869 875 882 885 887   
## 1.0653240 0.8741845 2.5952692 1.5490726 2.2408811 1.2175002 2.3903426   
## 888 889 890 891 892 901 911   
## 1.6677309 2.2527920 2.2219574 1.8380676 1.6994024 1.5245460 1.3539974   
## 913 914 915 918 921 922 926   
## 1.3961573 1.3896186 1.6914298 1.9522077 1.2802854 1.8815351 1.2867566   
## 930 932 933 934 938 950 951   
## 1.2802854 1.7395332 2.3910580 2.2850694 1.4107290 1.0653240 0.7904238   
## 953 958 961 963 974 975 976   
## 2.2842165 1.9202519 2.1000637 1.0882730 2.1152446 1.4151788 1.3300031   
## 979 990 991 992 1002 1007 1010   
## 1.3745003 1.6227626 2.2600069 1.2471351 1.0972798 2.2619538 1.8337533   
## 1015 1016 1017 1020 1023 1024 1027   
## 1.7917053 1.5392528 2.2531136 1.2367713 1.4486706 1.8334317 1.4943738   
## 1029 1032 1035 1037 1042 1049 1054   
## 2.1883099 1.7983775 1.9960744 1.6393644 1.4108641 1.3753532 2.2994552   
## 1055 1058 1060 1064 1066 1069 1071   
## 1.5392528 1.6683046 1.6905311 2.0941005 1.6082471 1.4624181 1.7591920   
## 1075 1076 1077 1080 1082 1086 1087   
## 1.2708286 1.9415606 2.4805036 1.6683046 2.5162102 1.1847059 1.5392528   
## 1090 1091 1093 1094 1096 1097 1098   
## 1.7283748 0.9283514 1.1847059 1.4108641 1.9880249 2.3909001 1.2687271   
## 1101 1110 1113 1114 1129 1130 1134   
## 2.1900016 1.7591920 1.2287314 2.3749104 1.7847392 2.2138504 1.9720968   
## 1137 1140 1144 1150 1151 1158 1167   
## 1.6594740 1.0972798 1.9779680 1.8380676 1.2708286 2.1384476 0.4478149   
## 1183 1185 1189 1192 1196 1201 1203   
## 1.4873248 1.3266971 1.2548008 2.0099238 2.0910072 2.2850694 2.1384476   
## 1204 1205 1206 1211 1214 1220 1223   
## 1.3492742 0.9582057 0.8708391 1.8158565 1.8380676 1.7512742 2.2305556   
## 1224 1231 1233 1237 1239 1240 1243   
## 1.9880249 1.7664217 1.8356445 2.3589443 1.8815351 1.6563505 1.6883063   
## 1250 1251 1254 1258 1260 1261 1269   
## 2.4230138 2.0358067 0.9069788 2.1391106 2.0542806 1.0653240 1.6387713   
## 1273 1276 1280 1281 1284 1289 1291   
## 1.7352654 2.1818946 2.1597791 1.6313608 1.8380676 1.5754300 1.2743094   
## 1294 1303 1309 1318 1319 1330 1333   
## 1.9779680 2.1704034 0.4478149 0.9250296 1.7435223 1.5245460 2.3749104   
## 1341 1343 1352 1353 1355 1359 1360   
## 1.3787732 1.2166617 1.5925401 1.5712086 2.2531136 0.9573070 2.5194467   
## 1369 1371 1375 1376 1377 1386 1389   
## 2.0803787 1.1811707 1.5712086 2.1019051 1.6517556 1.5565018 1.2367713   
## 1390 1404 1405 1406 1409 1410 1412   
## 1.5810284 1.7597495 1.4949057 1.1847059 1.2131265 2.0776216 1.9042082   
## 1413 1418 1423 1425 1426 1429 1432   
## 1.9410030 1.4067777 1.7911478 1.7839007 1.7911478 1.2134481 1.5650006   
## 1435 1436 1442 1443 1445 1446 1455   
## 1.9187685 1.8166950 1.4870416 1.0891115 2.0803787 1.3859532 1.9736951   
## 1456 1457 1464 1466 1468 1469 1470   
## 2.0862364 2.3909001 1.0653240 1.6539188 2.0862364 1.6889730 1.0790972   
## 1476 1477 1480 1487 1492 1494 1496   
## 1.5277001 1.4492281 1.9740167 1.3284763 1.0653240 1.2544792 1.9187685   
## 1502 1507 1511 1512 1516 1521 1523   
## 2.0769431 1.6858746 0.6467856 2.1397899 1.9522077 2.1704034 1.7911478   
## 1526 1529 1535 1537 1538 1541 1542   
## 1.3896186 1.8158565 1.1765202 1.0653240 1.4870416 0.5979918 2.2624075   
## 1548 1551 1554 1557 1558 1559 1562   
## 1.8598820 1.5446893 2.2379661 1.2459804 1.6634953 2.5002014 1.4067777   
## 1564 1565 1568 1572 1574 1580 1581   
## 1.9036506 2.0910072 2.2060103 2.5162102 2.2060103 2.0862364 1.9880249   
## 1590 1592 1593 1594 1603 1604 1606   
## 1.8337533 1.3665170 1.6068155 1.6546041 1.6645188 1.5192806 1.9507243   
## 1610 1615 1616 1617 1622 1625 1626   
## 1.5574859 1.8294804 2.3746729 1.2388728 1.7075774 2.2850694 1.8722524   
## 1628 1630 1632 1635 1638 1640 1642   
## 0.4478149 0.7583363 2.0099238 2.3429546 1.5995445 1.5712086 0.7583363   
## 1643 1644 1646 1647 1650 1652 1654   
## 0.6148298 1.6645188 1.7033096 1.2175002 2.0910072 1.2367713 15.9639761   
## 1656 1658 1660 1664 1665 1667 1668   
## 2.0699493 0.7904238 2.4924160 2.0693918 1.3576628 1.2708286 1.5789269   
## 1671 1676 1683 1684 1688 1691 1693   
## 2.3589443 0.4799024 2.3240308 1.6068155 1.8166950 1.2867566 1.8815351   
## 1697 1699 1702 1703 1707 1710 1715   
## 1.0643005 1.1843843 0.4158591 1.5712086 1.4107290 1.7985481 1.5789269   
## 1716 1717 1719 1726 1736 1741 1747   
## 2.4799461 1.1855444 1.9779680 0.9603072 1.8158565 1.4550858 1.7597495   
## 1749 1750 1757 1759 1761 1766 1770   
## 1.4870416 1.8294804 1.5810284 1.4550858 1.3266971 2.1095774 0.9573070   
## 1774 1776 1777 1779 1788 1789 1796   
## 2.5194467 1.7033096 2.0100589 0.7584680 1.9735164 1.3896186 1.3255753   
## 1797 1800 1802 1807 1813 1814 1815   
## 2.1818946 1.0972798 1.5712086 2.7010999 1.8859514 1.8205824 2.0647106   
## 1818 1827 1829 1830 1831 1842 1848   
## 0.7548026 1.8166950 2.0776216 2.5002014 2.1461659 2.3583868 1.0653240   
## 1853 1856 1858 1860 1861 1865 1869   
## 1.2364497 2.0803787 2.5194467 2.5514025 1.0365803 2.3275696 0.8708391   
## 1870 1874 1876 1878 1881 1883 1894   
## 1.8775838 1.3255753 1.7002604 1.4949057 1.3539974 0.4478149 1.9740167   
## 1895 1897 1899 1900 1904 1905 1907   
## 2.0484229 1.9755549 1.8155349 0.9061403 1.1855444 1.2166617 1.6057089   
## 1908 1909 1911 1914 1922 1929 1930   
## 1.7597495 1.2687271 1.5192806 1.6674466 2.0099238 1.1490832 1.4074796   
## 1932 1935 1938 1940 1950 1951 1955   
## 1.4880267 2.0040526 1.3300031 2.0199807 0.9262499 1.6858746 1.5574005   
## 1958 1964 1973 1975 1977 1980 1982   
## 2.2408811 2.1802232 2.0358067 1.9779680 2.5194467 1.8238190 1.5712086   
## 1987 1988 1992 1996 1999   
## 1.8158565 1.6858746 2.3903426 0.9688911 1.2166617

## Confusion Matrix

#confusionMatrix(pred, TestSet$disease)

## Random forest model

Apply random forest model

# Fine tuning parameters of Random Forest model  
model2 <- randomForest(disease ~ ., data = disease\_TrainSet, importance = TRUE)  
model2

##   
## Call:  
## randomForest(formula = disease ~ ., data = disease\_TrainSet, importance = TRUE)   
## Type of random forest: classification  
## Number of trees: 500  
## No. of variables tried at each split: 2  
##   
## OOB estimate of error rate: 83.57%  
## Confusion matrix:  
## Alzheimer's disease breast cancer diabetes endometriosis  
## Alzheimer's disease 83 39 1 2  
## breast cancer 30 37 0 0  
## diabetes 31 9 0 1  
## endometriosis 13 13 1 0  
## gastritis 21 14 0 0  
## heart disease 15 9 0 1  
## HIV/AIDS 7 8 0 1  
## hypertension 59 32 0 0  
## kidney disease 30 23 1 2  
## multiple sclerosis 28 16 0 0  
## prostate cancer 60 0 0 0  
## schizophrenia 12 6 0 0  
## skin cancer 63 22 1 1  
## gastritis heart disease HIV/AIDS hypertension  
## Alzheimer's disease 0 4 0 28  
## breast cancer 0 0 0 17  
## diabetes 0 0 1 8  
## endometriosis 0 0 1 6  
## gastritis 0 0 1 13  
## heart disease 0 0 0 9  
## HIV/AIDS 1 0 1 3  
## hypertension 1 1 1 24  
## kidney disease 0 1 2 11  
## multiple sclerosis 1 0 0 6  
## prostate cancer 0 1 2 7  
## schizophrenia 2 0 0 6  
## skin cancer 0 1 1 16  
## kidney disease multiple sclerosis prostate cancer  
## Alzheimer's disease 8 2 40  
## breast cancer 8 1 0  
## diabetes 5 0 20  
## endometriosis 4 0 0  
## gastritis 1 0 15  
## heart disease 4 0 22  
## HIV/AIDS 3 0 22  
## hypertension 11 0 46  
## kidney disease 14 3 39  
## multiple sclerosis 3 0 16  
## prostate cancer 2 1 55  
## schizophrenia 0 1 10  
## skin cancer 6 0 40  
## schizophrenia skin cancer class.error  
## Alzheimer's disease 0 25 0.6422414  
## breast cancer 0 12 0.6476190  
## diabetes 0 8 1.0000000  
## endometriosis 0 7 1.0000000  
## gastritis 2 3 1.0000000  
## heart disease 0 4 1.0000000  
## HIV/AIDS 1 5 0.9807692  
## hypertension 1 21 0.8781726  
## kidney disease 0 11 0.8978102  
## multiple sclerosis 1 6 1.0000000  
## prostate cancer 0 3 0.5801527  
## schizophrenia 0 3 1.0000000  
## skin cancer 0 16 0.9041916

# Predicting on train set  
predTrain <- predict(model2, disease\_TrainSet, type = "class")  
  
# Checking classification accuracy  
table(predTrain, disease\_TrainSet$disease)

##   
## predTrain Alzheimer's disease breast cancer diabetes endometriosis  
## Alzheimer's disease 137 14 32 11  
## breast cancer 29 67 8 14  
## diabetes 1 0 9 0  
## endometriosis 1 0 1 3  
## gastritis 0 0 0 0  
## heart disease 1 0 0 0  
## HIV/AIDS 0 0 1 1  
## hypertension 14 9 5 5  
## kidney disease 6 4 4 4  
## multiple sclerosis 0 0 0 0  
## prostate cancer 31 0 19 0  
## schizophrenia 0 0 0 0  
## skin cancer 12 11 4 7  
##   
## predTrain gastritis heart disease HIV/AIDS hypertension  
## Alzheimer's disease 20 14 4 48  
## breast cancer 12 9 10 24  
## diabetes 0 0 0 0  
## endometriosis 0 0 1 0  
## gastritis 6 0 0 0  
## heart disease 0 8 0 0  
## HIV/AIDS 1 1 13 1  
## hypertension 11 5 2 73  
## kidney disease 2 2 1 4  
## multiple sclerosis 0 0 0 0  
## prostate cancer 14 22 17 32  
## schizophrenia 0 0 0 0  
## skin cancer 4 3 4 15  
##   
## predTrain kidney disease multiple sclerosis prostate cancer  
## Alzheimer's disease 32 28 43  
## breast cancer 19 14 0  
## diabetes 0 0 0  
## endometriosis 0 0 0  
## gastritis 0 0 0  
## heart disease 0 0 0  
## HIV/AIDS 1 0 0  
## hypertension 9 4 2  
## kidney disease 34 3 0  
## multiple sclerosis 0 6 0  
## prostate cancer 35 16 85  
## schizophrenia 0 0 0  
## skin cancer 7 6 1  
##   
## predTrain schizophrenia skin cancer  
## Alzheimer's disease 13 43  
## breast cancer 6 19  
## diabetes 0 0  
## endometriosis 0 1  
## gastritis 1 0  
## heart disease 0 0  
## HIV/AIDS 0 0  
## hypertension 6 14  
## kidney disease 0 6  
## multiple sclerosis 1 0  
## prostate cancer 9 35  
## schizophrenia 2 0  
## skin cancer 2 49

model2 <- na.omit(model2)  
  
# Predicting on Validation set  
predValid <- predict(model2, disease\_TestSet, type = "class")  
  
# Checking classification accuracy  
mean(predValid == disease\_TestSet$disease)

## [1] 0.165

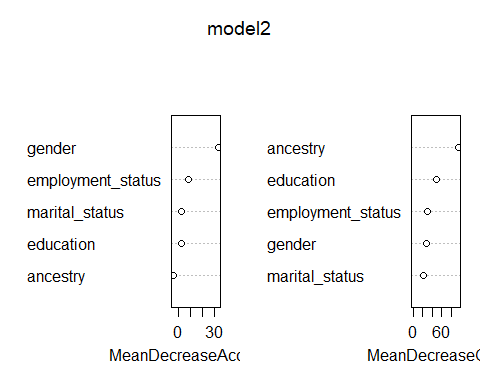
table(predValid,disease\_TestSet$disease)

##   
## predValid Alzheimer's disease breast cancer diabetes endometriosis  
## Alzheimer's disease 39 13 17 9  
## breast cancer 18 11 4 8  
## diabetes 1 0 0 0  
## endometriosis 0 0 0 0  
## gastritis 0 0 0 0  
## heart disease 0 0 0 0  
## HIV/AIDS 1 1 0 1  
## hypertension 6 4 9 2  
## kidney disease 7 5 1 0  
## multiple sclerosis 0 0 0 0  
## prostate cancer 28 0 4 0  
## schizophrenia 0 0 0 0  
## skin cancer 7 6 1 1  
##   
## predValid gastritis heart disease HIV/AIDS hypertension  
## Alzheimer's disease 10 9 2 27  
## breast cancer 3 1 3 21  
## diabetes 0 1 0 0  
## endometriosis 0 0 0 0  
## gastritis 1 0 0 0  
## heart disease 0 1 0 1  
## HIV/AIDS 0 0 0 2  
## hypertension 3 1 2 10  
## kidney disease 3 2 1 6  
## multiple sclerosis 0 0 0 1  
## prostate cancer 9 4 12 20  
## schizophrenia 0 0 0 0  
## skin cancer 1 4 8 13  
##   
## predValid kidney disease multiple sclerosis prostate cancer  
## Alzheimer's disease 17 12 15  
## breast cancer 8 5 0  
## diabetes 0 0 0  
## endometriosis 0 0 0  
## gastritis 0 0 0  
## heart disease 1 0 1  
## HIV/AIDS 0 0 0  
## hypertension 7 4 3  
## kidney disease 4 2 1  
## multiple sclerosis 0 0 0  
## prostate cancer 10 12 29  
## schizophrenia 0 0 0  
## skin cancer 1 1 0  
##   
## predValid schizophrenia skin cancer  
## Alzheimer's disease 6 27  
## breast cancer 3 13  
## diabetes 0 0  
## endometriosis 0 1  
## gastritis 0 1  
## heart disease 0 1  
## HIV/AIDS 1 1  
## hypertension 0 3  
## kidney disease 0 3  
## multiple sclerosis 0 0  
## prostate cancer 4 12  
## schizophrenia 0 0  
## skin cancer 1 4

# To check important variables  
importance(model2)

## Alzheimer's disease breast cancer diabetes endometriosis  
## gender -1.572030 29.337689 -2.372525 7.313963  
## employment\_status 8.261509 -3.560134 2.802048 -3.250488  
## education -7.859652 1.624276 3.229921 0.246096  
## marital\_status -1.170440 -4.579872 6.081446 3.695145  
## ancestry -4.582599 -7.807351 5.144771 -2.217989  
## gastritis heart disease HIV/AIDS hypertension  
## gender -0.7361897 1.920755 3.5379679 1.4733691  
## employment\_status -1.4860458 -2.759647 4.9418385 2.9941151  
## education -6.0412280 2.067650 3.3755598 1.2388433  
## marital\_status 0.8802722 5.436578 -0.8679253 1.6938141  
## ancestry -1.4805477 -2.740245 2.7803375 0.2844226  
## kidney disease multiple sclerosis prostate cancer  
## gender 9.248339 1.999961 38.3640487  
## employment\_status 10.049442 -9.329930 0.1440121  
## education 6.218238 -2.672903 4.7016289  
## marital\_status 9.817900 -2.472035 -5.8784037  
## ancestry 3.755459 -1.120006 -5.9362590  
## schizophrenia skin cancer MeanDecreaseAccuracy  
## gender -7.290833 5.416934 33.929733  
## employment\_status 2.122147 4.818350 8.330237  
## education 2.701801 4.801611 2.465961  
## marital\_status -5.176843 7.090811 2.618061  
## ancestry -1.959080 3.297698 -4.563110  
## MeanDecreaseGini  
## gender 27.81891  
## employment\_status 29.50241  
## education 48.83914  
## marital\_status 21.16534  
## ancestry 95.14737

varImpPlot(model2)



## Naive Bayes Model

NBclassfier = naiveBayes(disease~., data=disease\_TrainSet)  
print(NBclassfier)

##   
## Naive Bayes Classifier for Discrete Predictors  
##   
## Call:  
## naiveBayes.default(x = X, y = Y, laplace = laplace)  
##   
## A-priori probabilities:  
## Y  
## Alzheimer's disease breast cancer diabetes endometriosis   
## 0.16571429 0.07500000 0.05928571 0.03214286   
## gastritis heart disease HIV/AIDS hypertension   
## 0.05000000 0.04571429 0.03714286 0.14071429   
## kidney disease multiple sclerosis prostate cancer schizophrenia   
## 0.09785714 0.05500000 0.09357143 0.02857143   
## skin cancer   
## 0.11928571   
##   
## Conditional probabilities:  
## gender  
## Y female male  
## Alzheimer's disease 0.4870690 0.5129310  
## breast cancer 1.0000000 0.0000000  
## diabetes 0.4337349 0.5662651  
## endometriosis 1.0000000 0.0000000  
## gastritis 0.4714286 0.5285714  
## heart disease 0.3906250 0.6093750  
## HIV/AIDS 0.4423077 0.5576923  
## hypertension 0.5025381 0.4974619  
## kidney disease 0.5109489 0.4890511  
## multiple sclerosis 0.4545455 0.5454545  
## prostate cancer 0.0000000 1.0000000  
## schizophrenia 0.4250000 0.5750000  
## skin cancer 0.4670659 0.5329341  
##   
## employment\_status  
## Y employed retired student unemployed  
## Alzheimer's disease 0.288793103 0.586206897 0.004310345 0.120689655  
## breast cancer 0.390476190 0.485714286 0.009523810 0.114285714  
## diabetes 0.313253012 0.530120482 0.012048193 0.144578313  
## endometriosis 0.288888889 0.533333333 0.000000000 0.177777778  
## gastritis 0.428571429 0.371428571 0.042857143 0.157142857  
## heart disease 0.453125000 0.390625000 0.000000000 0.156250000  
## HIV/AIDS 0.673076923 0.115384615 0.019230769 0.192307692  
## hypertension 0.345177665 0.517766497 0.010152284 0.126903553  
## kidney disease 0.489051095 0.364963504 0.007299270 0.138686131  
## multiple sclerosis 0.363636364 0.493506494 0.025974026 0.116883117  
## prostate cancer 0.381679389 0.488549618 0.000000000 0.129770992  
## schizophrenia 0.350000000 0.500000000 0.075000000 0.075000000  
## skin cancer 0.407185629 0.526946108 0.000000000 0.065868263  
##   
## education  
## Y bachelors highschool highscool masters  
## Alzheimer's disease 0.508620690 0.245689655 0.000000000 0.159482759  
## breast cancer 0.580952381 0.190476190 0.000000000 0.152380952  
## diabetes 0.481927711 0.313253012 0.000000000 0.132530120  
## endometriosis 0.533333333 0.222222222 0.000000000 0.111111111  
## gastritis 0.500000000 0.242857143 0.000000000 0.157142857  
## heart disease 0.531250000 0.203125000 0.000000000 0.218750000  
## HIV/AIDS 0.346153846 0.307692308 0.038461538 0.192307692  
## hypertension 0.548223350 0.208121827 0.000000000 0.126903553  
## kidney disease 0.562043796 0.233576642 0.007299270 0.116788321  
## multiple sclerosis 0.649350649 0.155844156 0.012987013 0.103896104  
## prostate cancer 0.557251908 0.198473282 0.000000000 0.145038168  
## schizophrenia 0.475000000 0.225000000 0.000000000 0.100000000  
## skin cancer 0.556886228 0.185628743 0.000000000 0.137724551  
## education  
## Y phd/md phD/MD  
## Alzheimer's disease 0.086206897 0.000000000  
## breast cancer 0.076190476 0.000000000  
## diabetes 0.072289157 0.000000000  
## endometriosis 0.133333333 0.000000000  
## gastritis 0.071428571 0.028571429  
## heart disease 0.046875000 0.000000000  
## HIV/AIDS 0.096153846 0.019230769  
## hypertension 0.111675127 0.005076142  
## kidney disease 0.072992701 0.007299270  
## multiple sclerosis 0.064935065 0.012987013  
## prostate cancer 0.099236641 0.000000000  
## schizophrenia 0.125000000 0.075000000  
## skin cancer 0.119760479 0.000000000  
##   
## marital\_status  
## Y married single  
## Alzheimer's disease 0.7931034 0.2068966  
## breast cancer 0.8000000 0.2000000  
## diabetes 0.7590361 0.2409639  
## endometriosis 0.7777778 0.2222222  
## gastritis 0.7428571 0.2571429  
## heart disease 0.5937500 0.4062500  
## HIV/AIDS 0.6923077 0.3076923  
## hypertension 0.7664975 0.2335025  
## kidney disease 0.7153285 0.2846715  
## multiple sclerosis 0.7272727 0.2727273  
## prostate cancer 0.7175573 0.2824427  
## schizophrenia 0.7500000 0.2500000  
## skin cancer 0.7485030 0.2514970  
##   
## ancestry  
## Y Austria Belgium Czech Republic Denmark  
## Alzheimer's disease 0.04310345 0.05172414 0.03017241 0.08189655  
## breast cancer 0.06666667 0.04761905 0.05714286 0.03809524  
## diabetes 0.04819277 0.02409639 0.02409639 0.06024096  
## endometriosis 0.04444444 0.02222222 0.08888889 0.06666667  
## gastritis 0.10000000 0.02857143 0.01428571 0.05714286  
## heart disease 0.04687500 0.06250000 0.07812500 0.03125000  
## HIV/AIDS 0.03846154 0.05769231 0.03846154 0.00000000  
## hypertension 0.05076142 0.04568528 0.05076142 0.04568528  
## kidney disease 0.06569343 0.04379562 0.05839416 0.04379562  
## multiple sclerosis 0.02597403 0.05194805 0.02597403 0.07792208  
## prostate cancer 0.05343511 0.07633588 0.02290076 0.03816794  
## schizophrenia 0.05000000 0.05000000 0.05000000 0.02500000  
## skin cancer 0.04790419 0.08383234 0.05389222 0.04191617  
## ancestry  
## Y England Finland France Germany Hungary  
## Alzheimer's disease 0.03448276 0.05172414 0.04741379 0.05172414 0.04741379  
## breast cancer 0.04761905 0.03809524 0.04761905 0.02857143 0.04761905  
## diabetes 0.06024096 0.02409639 0.08433735 0.02409639 0.02409639  
## endometriosis 0.02222222 0.06666667 0.04444444 0.04444444 0.04444444  
## gastritis 0.05714286 0.01428571 0.04285714 0.05714286 0.01428571  
## heart disease 0.03125000 0.03125000 0.06250000 0.09375000 0.06250000  
## HIV/AIDS 0.03846154 0.07692308 0.00000000 0.07692308 0.07692308  
## hypertension 0.06598985 0.04060914 0.02538071 0.05076142 0.05076142  
## kidney disease 0.08029197 0.03649635 0.05109489 0.03649635 0.03649635  
## multiple sclerosis 0.07792208 0.05194805 0.07792208 0.02597403 0.07792208  
## prostate cancer 0.05343511 0.06870229 0.03053435 0.05343511 0.03816794  
## schizophrenia 0.05000000 0.02500000 0.00000000 0.10000000 0.02500000  
## skin cancer 0.05389222 0.04790419 0.05988024 0.05988024 0.04191617  
## ancestry  
## Y Ireland Italy Netherlands Poland Portugal  
## Alzheimer's disease 0.05172414 0.05172414 0.04310345 0.06034483 0.04310345  
## breast cancer 0.01904762 0.06666667 0.04761905 0.06666667 0.10476190  
## diabetes 0.10843373 0.08433735 0.07228916 0.07228916 0.02409639  
## endometriosis 0.06666667 0.02222222 0.02222222 0.04444444 0.00000000  
## gastritis 0.11428571 0.02857143 0.04285714 0.05714286 0.07142857  
## heart disease 0.07812500 0.09375000 0.06250000 0.04687500 0.01562500  
## HIV/AIDS 0.07692308 0.03846154 0.05769231 0.00000000 0.05769231  
## hypertension 0.05076142 0.02030457 0.05583756 0.08629442 0.06091371  
## kidney disease 0.04379562 0.05109489 0.05839416 0.05109489 0.04379562  
## multiple sclerosis 0.05194805 0.01298701 0.06493506 0.02597403 0.09090909  
## prostate cancer 0.04580153 0.09160305 0.05343511 0.03816794 0.07633588  
## schizophrenia 0.02500000 0.02500000 0.05000000 0.07500000 0.05000000  
## skin cancer 0.08383234 0.04790419 0.06586826 0.02994012 0.04790419  
## ancestry  
## Y Russia Scotland Spain Sweden Switzerland  
## Alzheimer's disease 0.07327586 0.05603448 0.03448276 0.03879310 0.07327586  
## breast cancer 0.06666667 0.01904762 0.02857143 0.04761905 0.06666667  
## diabetes 0.06024096 0.03614458 0.01204819 0.10843373 0.02409639  
## endometriosis 0.02222222 0.06666667 0.04444444 0.08888889 0.08888889  
## gastritis 0.01428571 0.02857143 0.01428571 0.11428571 0.07142857  
## heart disease 0.03125000 0.03125000 0.01562500 0.04687500 0.03125000  
## HIV/AIDS 0.01923077 0.07692308 0.05769231 0.05769231 0.07692308  
## hypertension 0.04060914 0.03045685 0.07106599 0.05583756 0.05583756  
## kidney disease 0.05109489 0.04379562 0.05839416 0.05109489 0.07299270  
## multiple sclerosis 0.03896104 0.07792208 0.03896104 0.02597403 0.05194805  
## prostate cancer 0.03816794 0.04580153 0.05343511 0.04580153 0.04580153  
## schizophrenia 0.10000000 0.00000000 0.15000000 0.02500000 0.10000000  
## skin cancer 0.02994012 0.02994012 0.04191617 0.05389222 0.05389222  
## ancestry  
## Y Ukraine  
## Alzheimer's disease 0.03448276  
## breast cancer 0.04761905  
## diabetes 0.02409639  
## endometriosis 0.08888889  
## gastritis 0.05714286  
## heart disease 0.04687500  
## HIV/AIDS 0.07692308  
## hypertension 0.04568528  
## kidney disease 0.02189781  
## multiple sclerosis 0.02597403  
## prostate cancer 0.03053435  
## schizophrenia 0.02500000  
## skin cancer 0.02395210