initial\_result\_code

Abdelrehim Sabri (500998232)

2019-11-04

## Introduction

Every ethnic group has a different diet and lifestyle. When people moved to North America, they brought their eating habits, level of education, and life style with them. The large size and highly diverse population resulted in different types of diseases that effected different age and ethnic groups. “Diagnosis has important implications for patient care. When a diagnosis is accurate and made in a timely manner, a patient has the best opportunity for positive health outcome because clinical decision making well be tailored to a correct understanding of the patient’s health problem.” [1] The research question is to define what factors are relevant to predicting the disease the patient might have. Also, who might be susceptible to which type of disease. After exploring the data, there is a possibility that a patient might have multiple diseases. Therefore, each instance can be assigned with multiple categories, as such, this type of problem needs to use multi-label classification.

## Install R packages

#install.packages("rpart")  
#install.packages("caret")  
#install.packages("e1071")  
#install.packages("randomForest")  
#install.packages("corrplot")

## Prepare for analyses

set.seed(1234)  
library(lubridate)

## Warning: package 'lubridate' was built under R version 3.5.3

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.5.3

library(randomForest)

## Warning: package 'randomForest' was built under R version 3.5.3

## randomForest 4.6-14

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

##   
## Attaching package: 'randomForest'

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

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.5.3

##   
## Attaching package: 'dplyr'

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

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

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

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

library(rpart)

## Warning: package 'rpart' was built under R version 3.5.3

library(caret)

## Warning: package 'caret' was built under R version 3.5.3

## Loading required package: lattice

library(e1071)

## Warning: package 'e1071' was built under R version 3.5.3

library(corrplot)

## Warning: package 'corrplot' was built under R version 3.5.3

## corrplot 0.84 loaded

## Includes functions to clean datasets

Read datasets from csv file

build\_clean\_dataset <- function() {  
 datasetloc = "C:/Users/abdel/Desktop/Ryerson University/capstone/capstone/R/Health\_Care\_History.csv"  
 if (file.exists(datasetloc)) {  
 alldata <- read.csv(file=datasetloc, header = T)  
 }  
 return(alldata)  
}

Convert the date to age and group them into four groups (0-25, 26-40, 41-50, 50-65, 65+)

age <- function(dob, age.day = today(), units = "years", floor = TRUE) {  
 calc.age = interval(dob, age.day) / duration(num = 1, units = units)  
 if (floor) return(as.integer(floor(calc.age)))  
   
 return(calc.age)  
}  
  
get\_age\_group <- function(a) {  
 ifelse(a<25,25, ifelse(a<40, 40, ifelse(a<50,50,65)))  
}

Group the countries of the patients based on ethnic groups

east\_europe <- c('Ukraine','Russia','Poland','Czech Republic','Hungary')  
west\_europe <- c('Austria','Belgium','France','Germany','Italy','Netherlands','Portugal','Spain','Switzerland')  
north\_europe <- c('Sweden', 'Finland', 'Denmark')  
british <- c('England','Scotland','Ireland')  
  
get\_ethnic\_group <- function(country) {  
 ifelse((country %in% east\_europe), 'east\_europe',  
 ifelse((country %in% west\_europe) ,'west\_europe',  
 ifelse((country %in% north\_europe), 'north\_europe',  
 ifelse((country %in% british), 'british',  
 country))))  
}

Read the dataset and remove patient ids from the analysis

patients <- build\_clean\_dataset()  
  
#remove the patient ids from the dataset  
patients <- patients[,-1]  
str(patients)

## 'data.frame': 2000 obs. of 13 variables:  
## $ gender : Factor w/ 2 levels "female","male": 1 1 2 2 1 1 1 1 1 2 ...  
## $ dob : Factor w/ 1877 levels "1923-10-10","1924-03-28",..: 505 1502 1811 545 327 1120 628 1378 631 1176 ...  
## $ zipcode : int 89136 94105 89127 44101 89136 94105 60612 43221 89127 43210 ...  
## $ employment\_status : Factor w/ 4 levels "employed","retired",..: 2 1 1 2 2 4 2 1 2 1 ...  
## $ education : Factor w/ 6 levels "bachelors","highschool",..: 1 5 4 1 4 2 5 1 4 2 ...  
## $ marital\_status : Factor w/ 2 levels "married","single": 1 1 1 1 1 1 1 1 2 1 ...  
## $ children : int 1 4 2 2 3 2 0 2 2 7 ...  
## $ ancestry : Factor w/ 20 levels "Austria","Belgium",..: 14 18 8 4 1 1 9 10 1 20 ...  
## $ avg\_commute : num 13.4 15.2 23.6 19.6 36.5 ...  
## $ daily\_internet\_use: num 2.53 6.77 3.63 5 7.75 3.34 6.75 3.01 4.12 3.15 ...  
## $ available\_vehicles: int 2 2 1 3 1 0 2 3 1 1 ...  
## $ military\_service : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ disease : Factor w/ 13 levels "Alzheimer disease",..: 8 4 11 10 13 1 9 2 1 7 ...

Fix the education column values by fixing the misspelled words

patients$education <- ifelse(patients$education == 'highscool', as.character('highschool'), as.character(patients$education))  
patients$education <- ifelse(as.factor(patients$education) == 'phD/MD', as.character('phd/md'), as.character(patients$education))  
patients$education <- as.factor(patients$education)

Group the ancestry countries to ethnic groups

patients$ancestry <- as.factor(get\_ethnic\_group(patients$ancestry))

Convert the date of birth into age and group them into 25 40 50 65

patients$age <- age(patients$dob)  
#patients$age <- get\_age\_group(age(patients$dob))

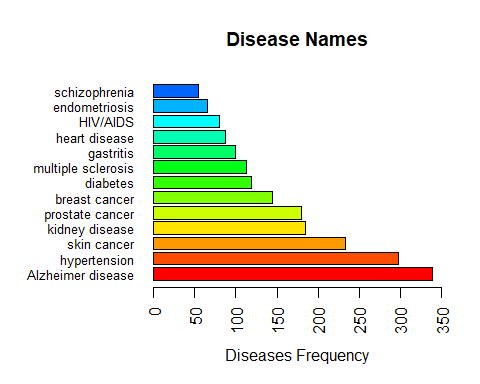
For the analysis purposes, move each disease to separate column with binary values, where 0: patient does not has the disease and 1: patient has the disease

get\_binary\_value <- function(value, compare\_to) {  
 ifelse(value==compare\_to,1,0)  
}  
patients$prostate\_cancer <- get\_binary\_value(patients$disease,'prostate cancer')  
patients$skin\_cancer <- get\_binary\_value(patients$disease,'skin cancer')  
patients$breast\_cancer <- get\_binary\_value(patients$disease,'breast cancer')  
patients$hiv\_aids <- get\_binary\_value(patients$disease,'HIV/AIDS')  
patients$diabetes <- get\_binary\_value(patients$disease,'diabetes')  
patients$heart\_disease <- get\_binary\_value(patients$disease,'heart disease')  
patients$hypertension <- get\_binary\_value(patients$disease,'hypertension')  
patients$endometriosis <- get\_binary\_value(patients$disease,'endometriosis')  
patients$multiple\_sclerosis <- get\_binary\_value(patients$disease,'multiple sclerosis')  
patients$schizophrenia <- get\_binary\_value(patients$disease,'schizophrenia')  
patients$kidney\_disease <- get\_binary\_value(patients$disease,'kidney disease')  
patients$gastritis <- get\_binary\_value(patients$disease,'gastritis')  
patients$alzheimer <- get\_binary\_value(patients$disease,'Alzheimer disease')  
str(patients)

## 'data.frame': 2000 obs. of 27 variables:  
## $ gender : Factor w/ 2 levels "female","male": 1 1 2 2 1 1 1 1 1 2 ...  
## $ dob : Factor w/ 1877 levels "1923-10-10","1924-03-28",..: 505 1502 1811 545 327 1120 628 1378 631 1176 ...  
## $ zipcode : int 89136 94105 89127 44101 89136 94105 60612 43221 89127 43210 ...  
## $ employment\_status : Factor w/ 4 levels "employed","retired",..: 2 1 1 2 2 4 2 1 2 1 ...  
## $ education : Factor w/ 4 levels "bachelors","highschool",..: 1 4 3 1 3 2 4 1 3 2 ...  
## $ marital\_status : Factor w/ 2 levels "married","single": 1 1 1 1 1 1 1 1 2 1 ...  
## $ children : int 1 4 2 2 3 2 0 2 2 7 ...  
## $ ancestry : Factor w/ 4 levels "british","east\_europe",..: 4 3 4 3 4 4 2 1 4 2 ...  
## $ avg\_commute : num 13.4 15.2 23.6 19.6 36.5 ...  
## $ daily\_internet\_use: num 2.53 6.77 3.63 5 7.75 3.34 6.75 3.01 4.12 3.15 ...  
## $ available\_vehicles: int 2 2 1 3 1 0 2 3 1 1 ...  
## $ military\_service : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ disease : Factor w/ 13 levels "Alzheimer disease",..: 8 4 11 10 13 1 9 2 1 7 ...  
## $ age : int 75 53 38 74 80 63 73 56 73 61 ...  
## $ prostate\_cancer : num 0 0 1 0 0 0 0 0 0 0 ...  
## $ skin\_cancer : num 0 0 0 0 1 0 0 0 0 0 ...  
## $ breast\_cancer : num 0 0 0 0 0 0 0 1 0 0 ...  
## $ hiv\_aids : num 0 0 0 0 0 0 0 0 0 1 ...  
## $ diabetes : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ heart\_disease : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ hypertension : num 1 0 0 0 0 0 0 0 0 0 ...  
## $ endometriosis : num 0 1 0 0 0 0 0 0 0 0 ...  
## $ multiple\_sclerosis: num 0 0 0 1 0 0 0 0 0 0 ...  
## $ schizophrenia : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ kidney\_disease : num 0 0 0 0 0 0 1 0 0 0 ...  
## $ gastritis : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ alzheimer : num 0 0 0 0 0 1 0 0 1 0 ...

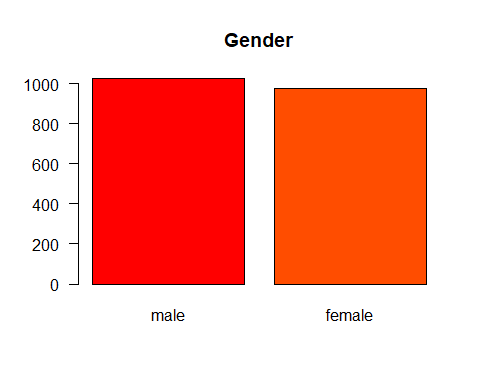
Draw a bar plot to count the total number of diseases in the dataset

par(las=2) # make label text perpendicular to axis  
par(mar=c(5,8,4,2)) # increase y-axis margin.  
  
disease\_counts <- table(patients$disease)  
barplot(sort(disease\_counts, decreasing = TRUE), main="Disease Names",   
 xlab="Diseases Frequency",   
 col=rainbow(20),  
 horiz=TRUE,  
 cex.names=0.8,  
 xlim = c(0, 350))



Observation : Male are more sick than Female

gender\_counts <- table(patients$gender)  
barplot(sort(gender\_counts, decreasing = TRUE), main="Gender",   
 col=rainbow(20), las=1)

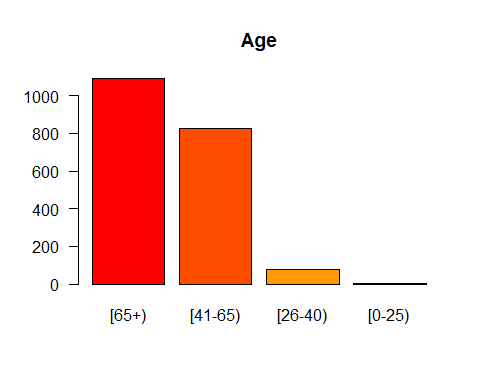


Observation : age group that are more sick

age\_breaks <- c(0,25,40,65,100)  
tags <- c("[0-25)","[26-40)", "[41-65)", "[65+)")  
age\_group\_tags <- cut(patients$age,   
 breaks=age\_breaks,   
 include.lowest=TRUE,   
 right=FALSE,   
 labels=tags)  
summary(age\_group\_tags)

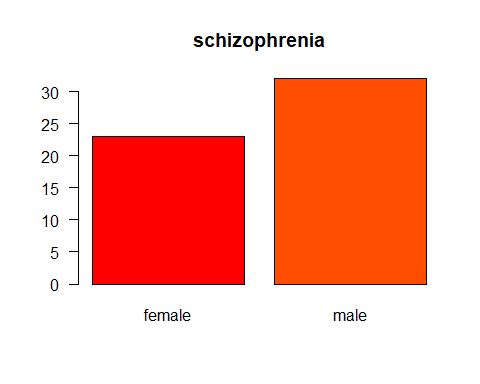
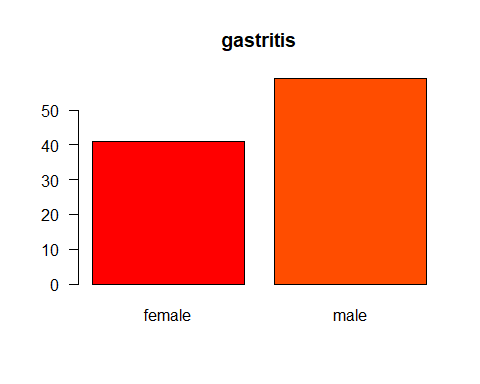
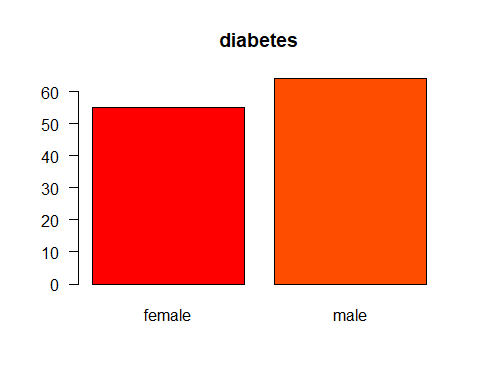
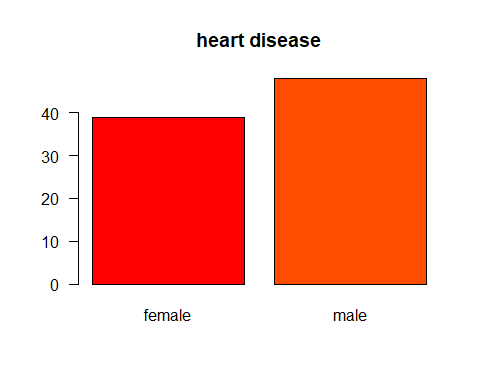
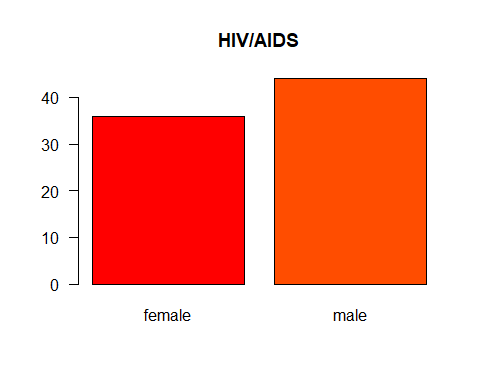
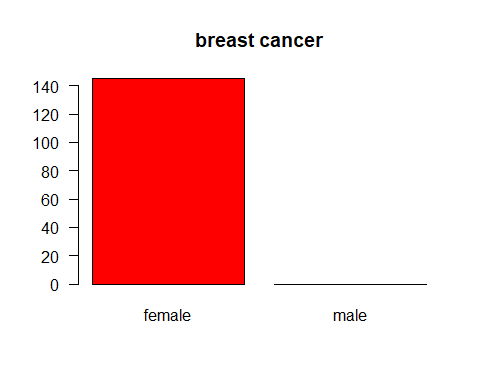
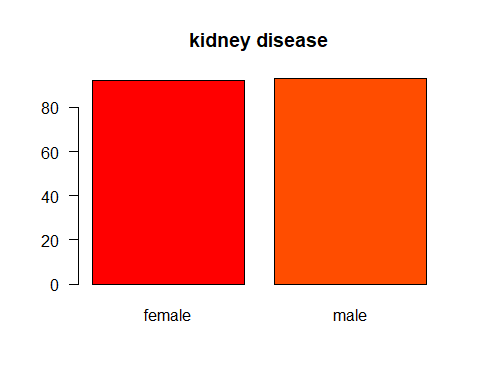
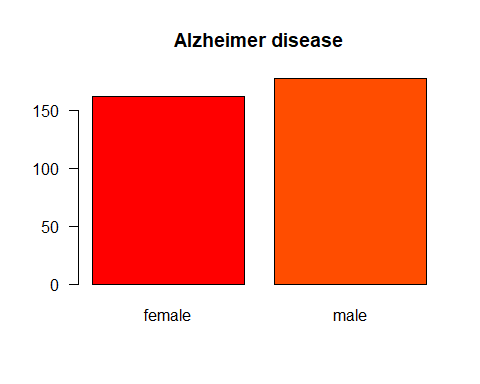
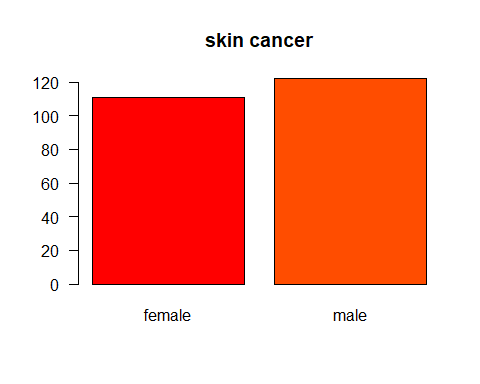
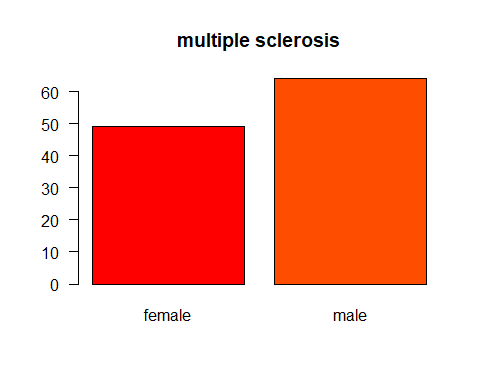
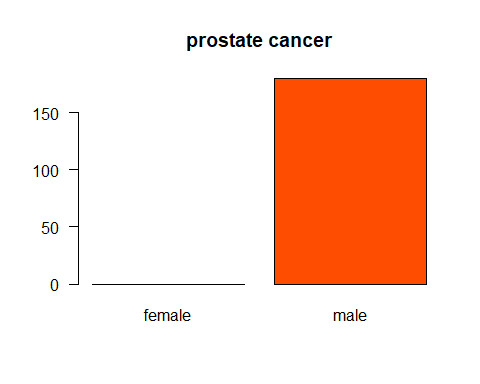
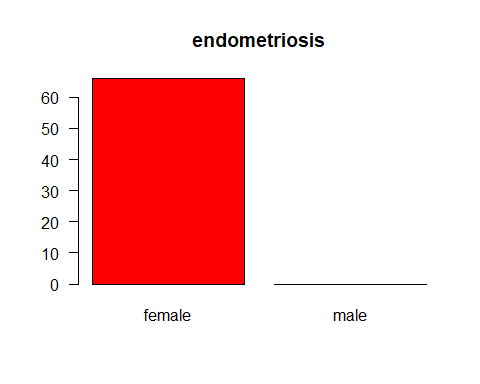
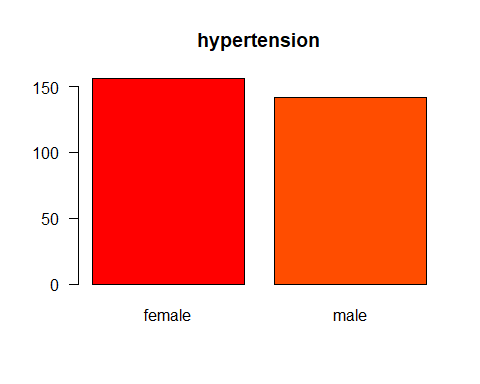
## [0-25) [26-40) [41-65) [65+)   
## 4 78 827 1091

#age\_counts <- table(patients$age)  
age\_counts <- table(age\_group\_tags)  
barplot(sort(age\_counts, decreasing = TRUE), main="Age",  
 col=rainbow(20), las=1)



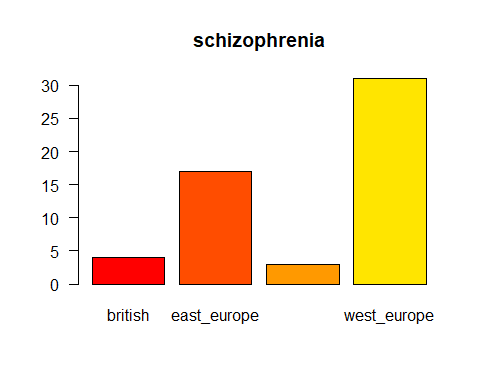
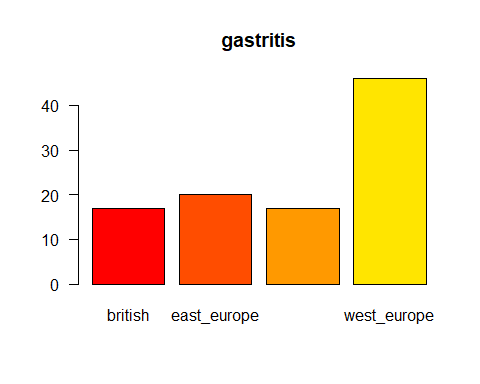
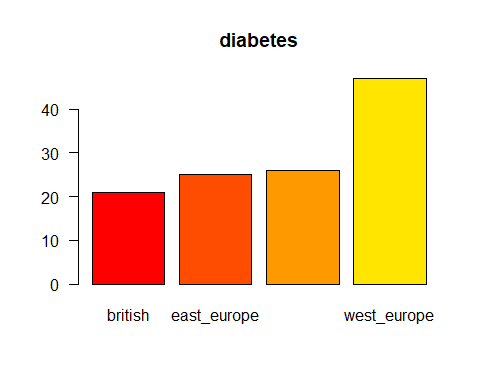
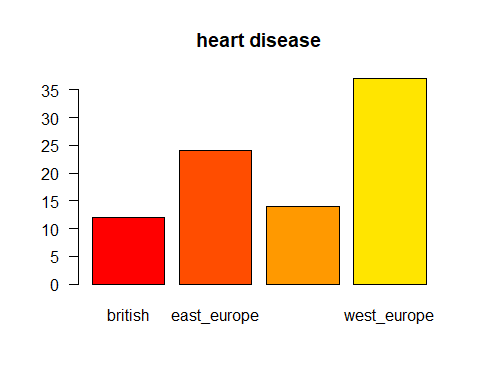
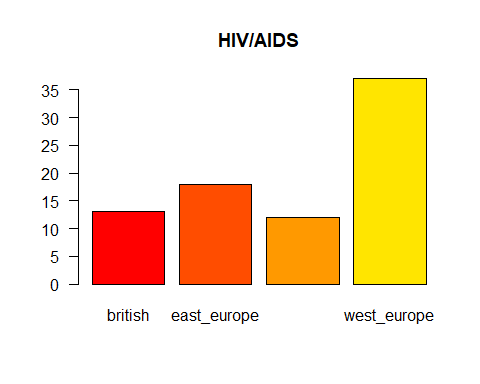
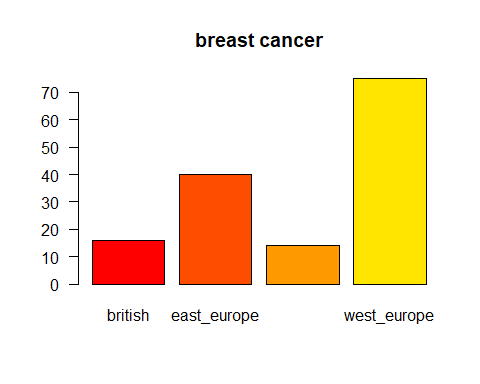
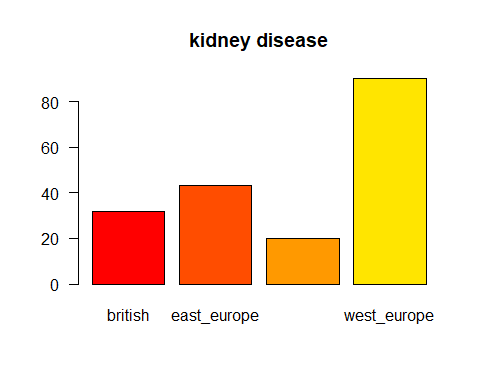
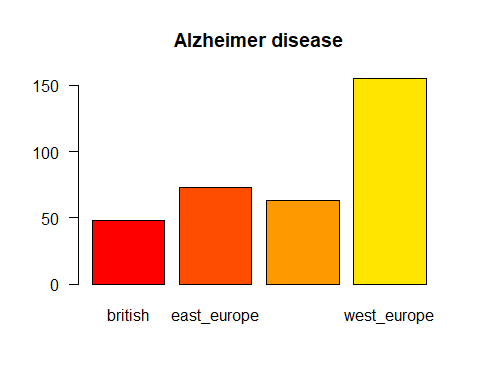
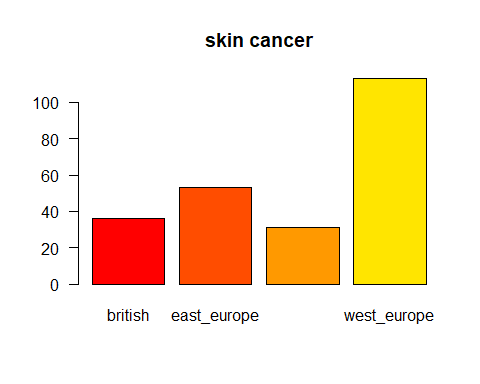
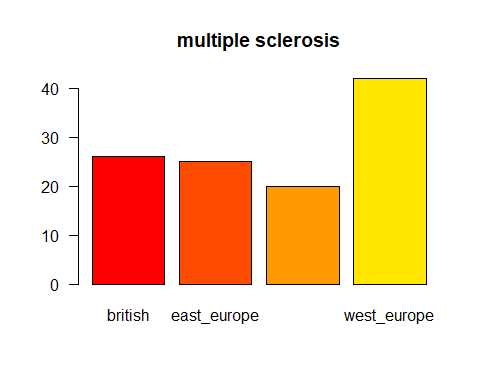
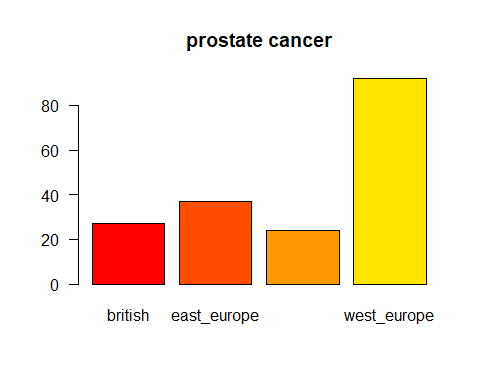
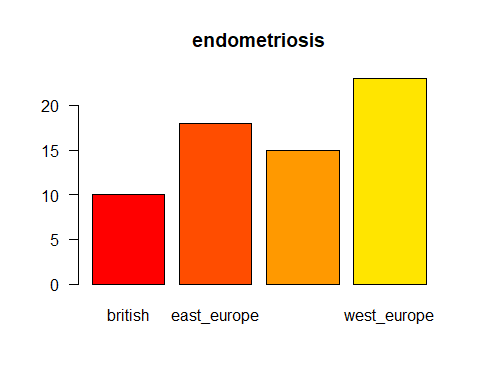
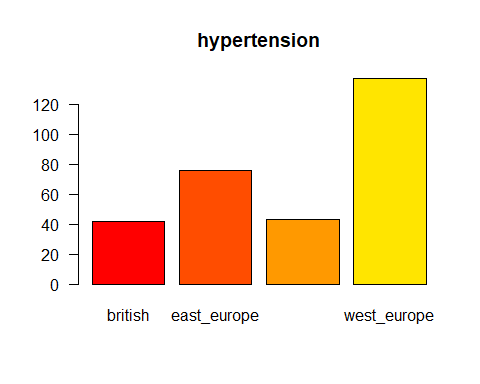
Observation : Disease and Gender distrubution

disease\_name = c(as.character(unique(patients$disease)))  
  
for (d in disease\_name) {  
 gender\_disease\_counts <- subset(patients, patients$disease == d)  
 gender\_disease\_counts <- table(gender\_disease\_counts$gender)  
 barplot(gender\_disease\_counts, main=d, col=rainbow(20), las=1)  
}



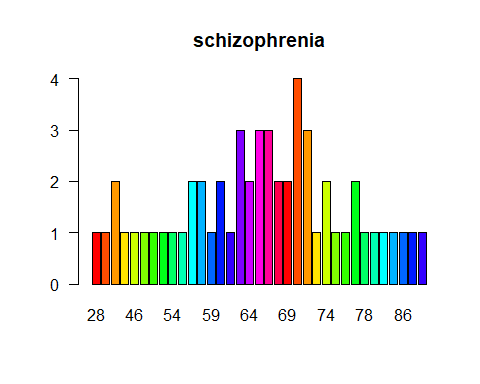
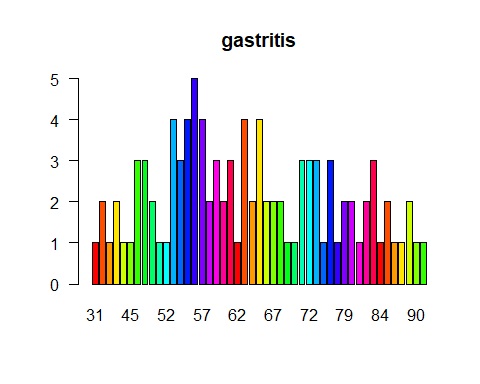
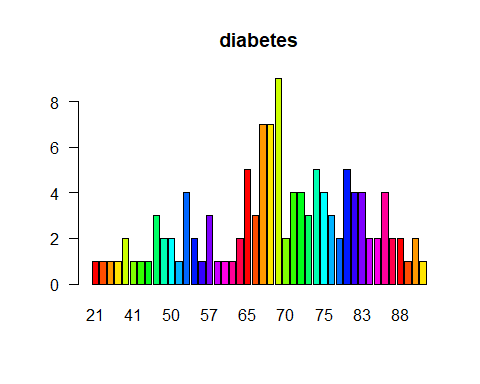
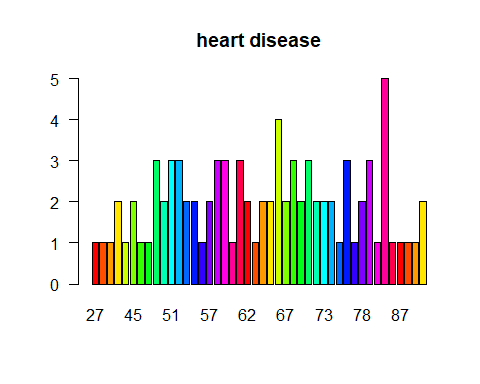
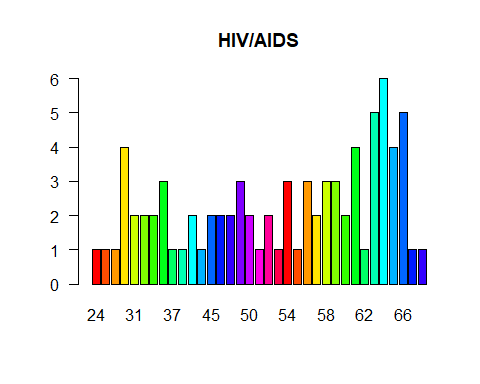
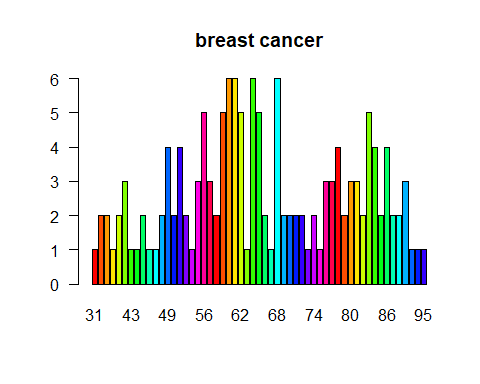
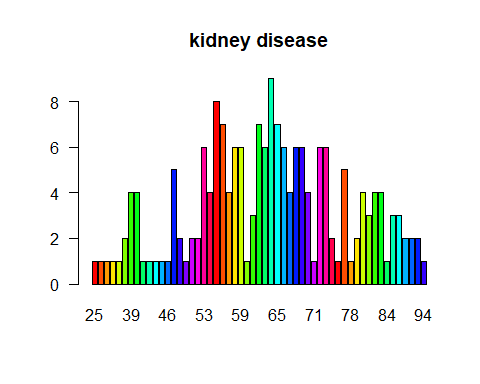
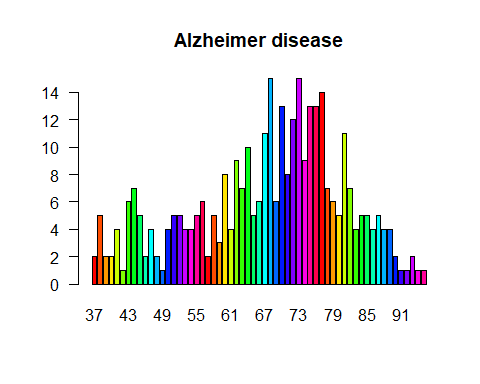
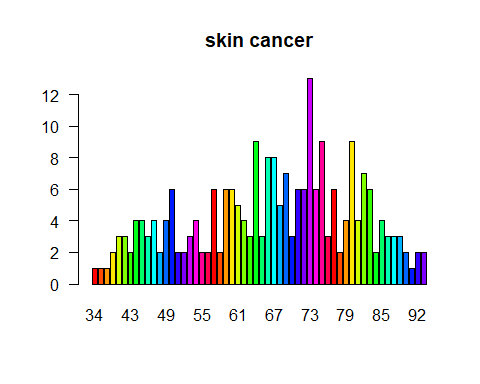
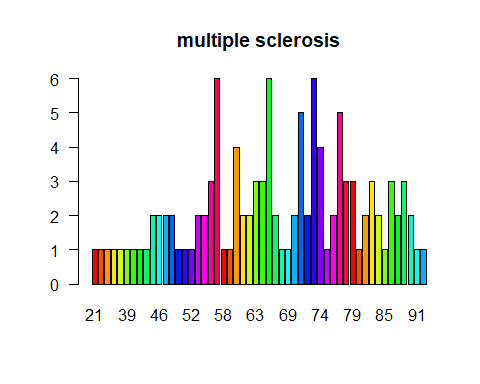
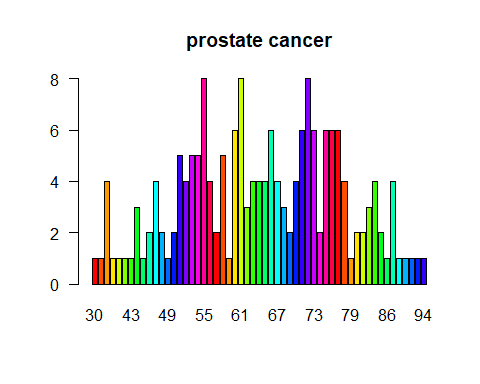
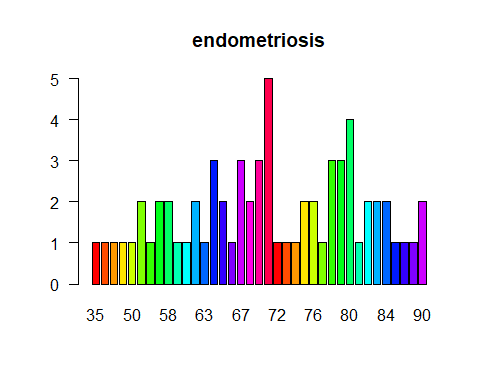
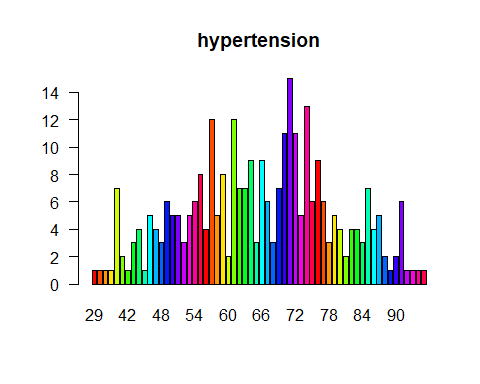
Observation : Disease and ancestry distrubution

for (d in disease\_name) {  
 ancestry\_disease\_counts <- subset(patients, patients$disease == d)  
 ancestry\_disease\_counts <- table(ancestry\_disease\_counts$ancestry)  
 barplot(ancestry\_disease\_counts, main=d, col=rainbow(20), las=1)  
}



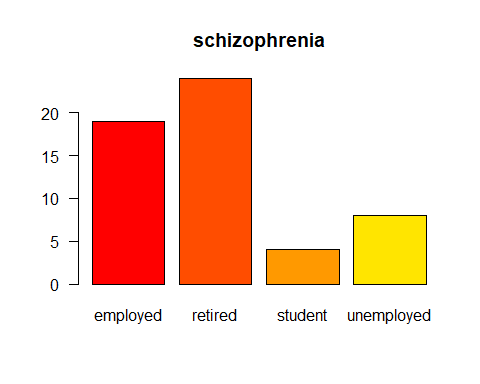
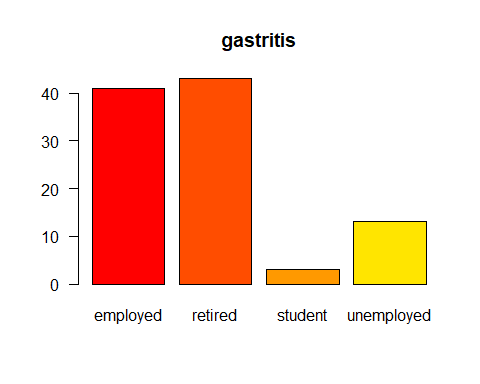
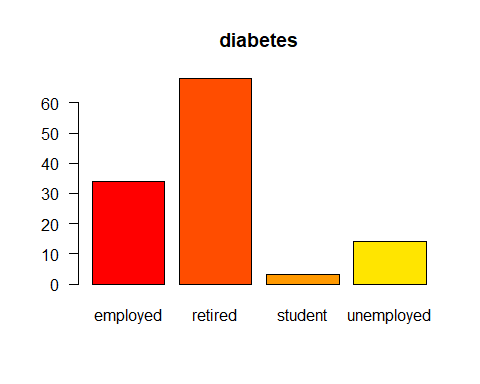
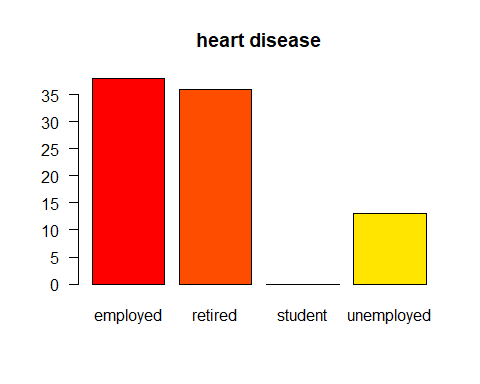
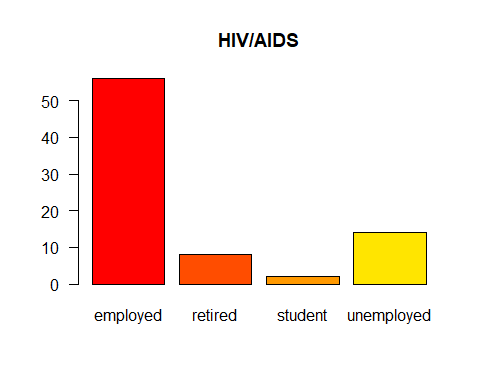
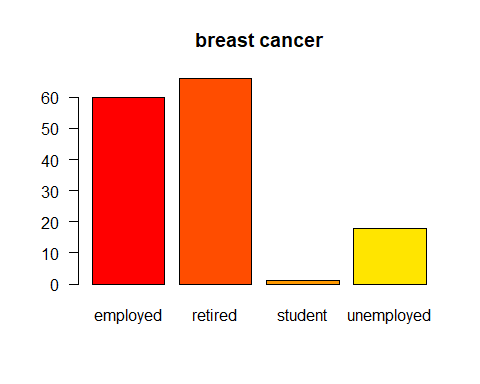
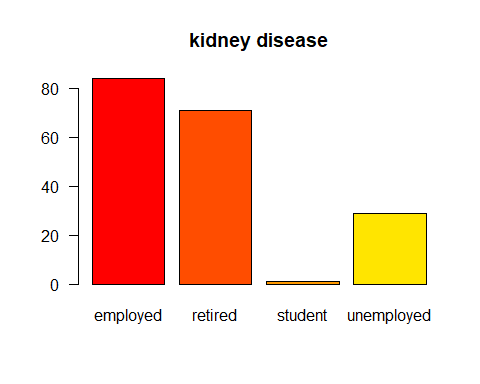
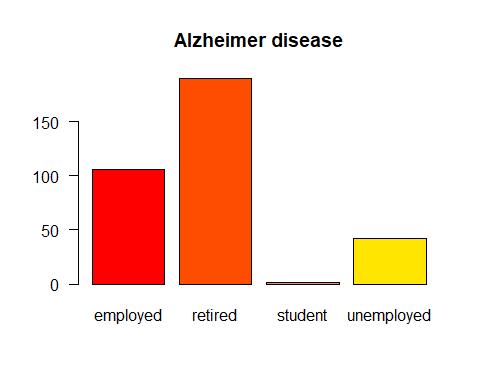
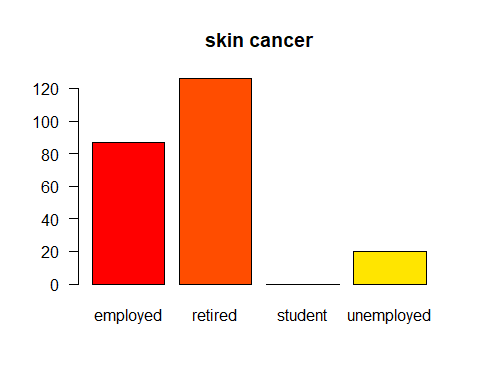
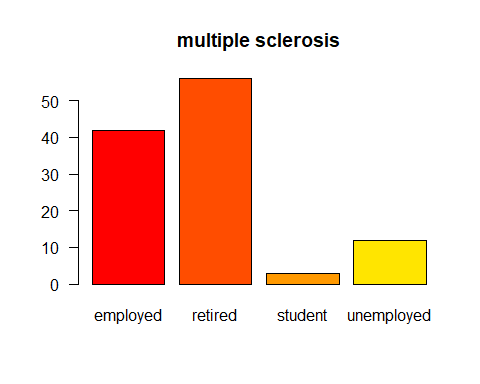
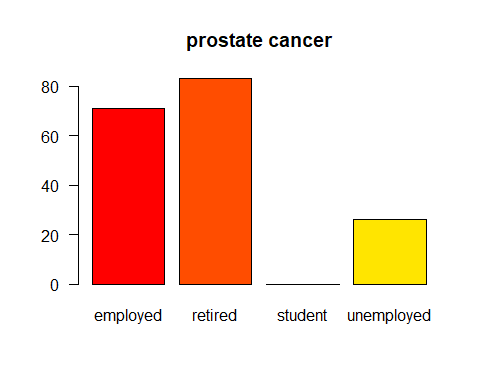
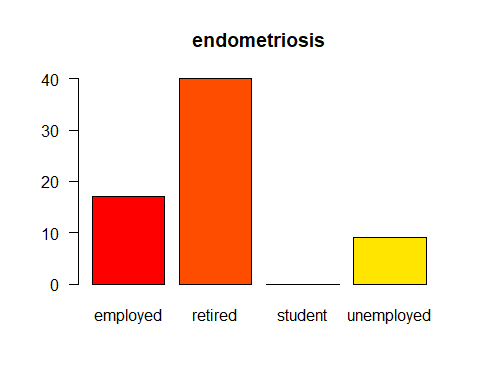
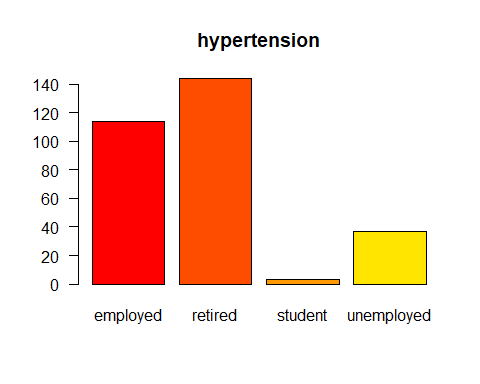
Observation : Disease and age distrubution

for (d in disease\_name) {  
 age\_disease\_counts <- subset(patients, patients$disease == d)  
 #age\_disease\_counts <- table(age\_disease\_counts$age\_group\_tags)  
 age\_disease\_counts <- table(age\_disease\_counts$age)  
 barplot(age\_disease\_counts, main=d, col=rainbow(20), las=1)  
}



Observation : Disease and employment status distrubution

for (d in disease\_name) {  
 emp\_disease\_counts <- subset(patients, patients$disease == d)  
 emp\_disease\_counts <- table(emp\_disease\_counts$employment\_status)  
 barplot(emp\_disease\_counts, main=d, col=rainbow(20), las=1)  
}

 ## Results

The data from the clened dataset were explored using basic R functions to determine reasonable candidate features for the model. The cleaned dataset contains 13 potential classes. However, only 5 were used to reduce computation.

## Partitioning

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

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

## gender dob zipcode employment\_status  
## female:688 1946-02-22: 3 Min. :10001 employed :525   
## male :712 1960-08-01: 3 1st Qu.:43221 retired :678   
## 1934-01-28: 2 Median :60612 student : 15   
## 1934-06-19: 2 Mean :62719 unemployed:182   
## 1936-11-01: 2 3rd Qu.:90008   
## 1939-04-18: 2 Max. :94110   
## (Other) :1386   
## education marital\_status children ancestry   
## bachelors :762 married:1036 Min. :0.000 british :204   
## highschool:326 single : 364 1st Qu.:1.000 east\_europe :324   
## masters :185 Median :2.000 north\_europe:210   
## phd/md :127 Mean :2.294 west\_europe :662   
## 3rd Qu.:3.000   
## Max. :7.000   
##   
## avg\_commute daily\_internet\_use available\_vehicles military\_service  
## Min. :-2.47 Min. :1.010 Min. :0.000 no :1272   
## 1st Qu.:23.43 1st Qu.:4.048 1st Qu.:1.000 yes: 128   
## Median :30.32 Median :5.055 Median :2.000   
## Mean :30.37 Mean :5.039 Mean :1.738   
## 3rd Qu.:37.16 3rd Qu.:6.000 3rd Qu.:3.000   
## Max. :63.73 Max. :8.820 Max. :4.000   
##   
## disease age prostate\_cancer   
## Alzheimer disease:241 Min. :21.00 Min. :0.00000   
## hypertension :209 1st Qu.:56.00 1st Qu.:0.00000   
## skin cancer :150 Median :67.00 Median :0.00000   
## prostate cancer :127 Mean :65.76 Mean :0.09071   
## kidney disease :126 3rd Qu.:76.00 3rd Qu.:0.00000   
## breast cancer :106 Max. :96.00 Max. :1.00000   
## (Other) :441   
## skin\_cancer breast\_cancer hiv\_aids diabetes   
## Min. :0.0000 Min. :0.00000 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.00000   
## Median :0.0000 Median :0.00000 Median :0.00000 Median :0.00000   
## Mean :0.1071 Mean :0.07571 Mean :0.04357 Mean :0.06357   
## 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.00000   
## Max. :1.0000 Max. :1.00000 Max. :1.00000 Max. :1.00000   
##   
## heart\_disease hypertension endometriosis multiple\_sclerosis  
## Min. :0.00000 Min. :0.0000 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.00000   
## Median :0.00000 Median :0.0000 Median :0.00000 Median :0.00000   
## Mean :0.04214 Mean :0.1493 Mean :0.03071 Mean :0.05929   
## 3rd Qu.:0.00000 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.00000   
## Max. :1.00000 Max. :1.0000 Max. :1.00000 Max. :1.00000   
##   
## schizophrenia kidney\_disease gastritis alzheimer   
## Min. :0.00000 Min. :0.00 Min. :0.00 Min. :0.0000   
## 1st Qu.:0.00000 1st Qu.:0.00 1st Qu.:0.00 1st Qu.:0.0000   
## Median :0.00000 Median :0.00 Median :0.00 Median :0.0000   
## Mean :0.02571 Mean :0.09 Mean :0.05 Mean :0.1721   
## 3rd Qu.:0.00000 3rd Qu.:0.00 3rd Qu.:0.00 3rd Qu.:0.0000   
## Max. :1.00000 Max. :1.00 Max. :1.00 Max. :1.0000   
##

summary(ValidSet)

## gender dob zipcode employment\_status  
## female:287 1959-09-22: 3 Min. :10001 employed :244   
## male :313 1932-04-10: 2 1st Qu.:43221 retired :277   
## 1935-07-27: 2 Median :60612 student : 6   
## 1946-01-07: 2 Mean :64948 unemployed: 73   
## 1954-12-31: 2 3rd Qu.:90015   
## 1961-12-21: 2 Max. :94110   
## (Other) :587   
## education marital\_status children ancestry   
## bachelors :314 married:460 Min. :0.000 british :100   
## highschool:137 single :140 1st Qu.:1.000 east\_europe :145   
## masters : 95 Median :2.000 north\_europe: 92   
## phd/md : 54 Mean :2.203 west\_europe :263   
## 3rd Qu.:3.000   
## Max. :7.000   
##   
## avg\_commute daily\_internet\_use available\_vehicles military\_service  
## Min. :-0.74 Min. :1.400 Min. :0.000 no :545   
## 1st Qu.:23.59 1st Qu.:3.940 1st Qu.:1.000 yes: 55   
## Median :30.30 Median :4.905 Median :2.000   
## Mean :30.39 Mean :4.887 Mean :1.765   
## 3rd Qu.:36.98 3rd Qu.:5.832 3rd Qu.:3.000   
## Max. :56.27 Max. :8.340 Max. :4.000   
##   
## disease age prostate\_cancer   
## Alzheimer disease: 98 Min. :21.00 Min. :0.00000   
## hypertension : 89 1st Qu.:54.00 1st Qu.:0.00000   
## skin cancer : 83 Median :65.00 Median :0.00000   
## kidney disease : 59 Mean :64.19 Mean :0.08833   
## prostate cancer : 53 3rd Qu.:75.00 3rd Qu.:0.00000   
## breast cancer : 39 Max. :94.00 Max. :1.00000   
## (Other) :179   
## skin\_cancer breast\_cancer hiv\_aids diabetes   
## Min. :0.0000 Min. :0.000 Min. :0.00000 Min. :0.00   
## 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.00000 1st Qu.:0.00   
## Median :0.0000 Median :0.000 Median :0.00000 Median :0.00   
## Mean :0.1383 Mean :0.065 Mean :0.03167 Mean :0.05   
## 3rd Qu.:0.0000 3rd Qu.:0.000 3rd Qu.:0.00000 3rd Qu.:0.00   
## Max. :1.0000 Max. :1.000 Max. :1.00000 Max. :1.00   
##   
## heart\_disease hypertension endometriosis multiple\_sclerosis  
## Min. :0.00000 Min. :0.0000 Min. :0.00000 Min. :0.00   
## 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.00   
## Median :0.00000 Median :0.0000 Median :0.00000 Median :0.00   
## Mean :0.04667 Mean :0.1483 Mean :0.03833 Mean :0.05   
## 3rd Qu.:0.00000 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.00   
## Max. :1.00000 Max. :1.0000 Max. :1.00000 Max. :1.00   
##   
## schizophrenia kidney\_disease gastritis alzheimer   
## Min. :0.00000 Min. :0.00000 Min. :0.00 Min. :0.0000   
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.00 1st Qu.:0.0000   
## Median :0.00000 Median :0.00000 Median :0.00 Median :0.0000   
## Mean :0.03167 Mean :0.09833 Mean :0.05 Mean :0.1633   
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.00 3rd Qu.:0.0000   
## Max. :1.00000 Max. :1.00000 Max. :1.00 Max. :1.0000   
##

## Analysing the hypertension disease

The dataset will predict the hypertension disease with selected attributes that contributes to the analysis

hyper\_TrainSet <- select(TrainSet, gender, age, employment\_status, education, marital\_status, ancestry, hypertension)  
hyper\_ValidSet <- select(ValidSet, gender, age, employment\_status, education, marital\_status, ancestry, hypertension)  
hyper\_TrainSet$hypertension <- as.factor(hyper\_TrainSet$hypertension)

## 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(hypertension~.,data=hyper\_TrainSet,family=binomial())  
summary(fit) # display results

##   
## Call:  
## glm(formula = hypertension ~ ., family = binomial(), data = hyper\_TrainSet)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.7219 -0.5948 -0.5559 -0.5057 2.2905   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.598064 0.517515 -3.088 0.00202 \*\*  
## gendermale -0.161921 0.150647 -1.075 0.28245   
## age 0.003875 0.008270 0.469 0.63938   
## employment\_statusretired -0.256010 0.238934 -1.071 0.28396   
## employment\_statusstudent -0.746796 1.071625 -0.697 0.48588   
## employment\_statusunemployed 0.180992 0.266726 0.679 0.49741   
## educationhighschool -0.222720 0.206987 -1.076 0.28192   
## educationmasters -0.168767 0.238080 -0.709 0.47841   
## educationphd/md -0.245495 0.298308 -0.823 0.41053   
## marital\_statussingle -0.093976 0.182966 -0.514 0.60752   
## ancestryeast\_europe -0.064898 0.245987 -0.264 0.79191   
## ancestrynorth\_europe -0.235585 0.279753 -0.842 0.39972   
## ancestrywest\_europe -0.106049 0.220397 -0.481 0.63040   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1180.1 on 1399 degrees of freedom  
## Residual deviance: 1172.9 on 1387 degrees of freedom  
## AIC: 1198.9  
##   
## Number of Fisher Scoring iterations: 5

confint(fit) # 95% CI for the coefficients

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) -2.63062208 -0.60001322  
## gendermale -0.45806884 0.13311766  
## age -0.01213159 0.02031294  
## employment\_statusretired -0.72705408 0.21039324  
## employment\_statusstudent -3.67947133 0.96593490  
## employment\_statusunemployed -0.34968605 0.69837542  
## educationhighschool -0.63726228 0.17554844  
## educationmasters -0.65415958 0.28270854  
## educationphd/md -0.86581671 0.31121822  
## marital\_statussingle -0.46024152 0.25818123  
## ancestryeast\_europe -0.54320612 0.42389803  
## ancestrynorth\_europe -0.78905865 0.31176835  
## ancestrywest\_europe -0.52887951 0.33751264

exp(coef(fit)) # exponentiated coefficients

## (Intercept) gendermale   
## 0.2022878 0.8505084   
## age employment\_statusretired   
## 1.0038825 0.7741340   
## employment\_statusstudent employment\_statusunemployed   
## 0.4738823 1.1984052   
## educationhighschool educationmasters   
## 0.8003391 0.8447060   
## educationphd/md marital\_statussingle   
## 0.7823171 0.9103049   
## ancestryeast\_europe ancestrynorth\_europe   
## 0.9371634 0.7901085   
## ancestrywest\_europe   
## 0.8993810

exp(confint(fit)) # 95% CI for exponentiated coefficients

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 0.07203364 0.5488044  
## gendermale 0.63250393 1.1423844  
## age 0.98794171 1.0205207  
## employment\_statusretired 0.48333075 1.2341633  
## employment\_statusstudent 0.02523631 2.6272427  
## employment\_statusunemployed 0.70490936 2.0104839  
## educationhighschool 0.52873798 1.1918997  
## educationmasters 0.51987880 1.3267184  
## educationphd/md 0.42070782 1.3650871  
## marital\_statussingle 0.63113119 1.2945734  
## ancestryeast\_europe 0.58088289 1.5279058  
## ancestrynorth\_europe 0.45427222 1.3658383  
## ancestrywest\_europe 0.58926486 1.4014573

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

## Performace

Probabilites for the response varable based on the test data were assigned using the predict functin.

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

## Confusion Matrix

#confusionMatrix(pred, ValidSet$hypertension)

## Randomforest model

Apply randomforest model

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

##   
## Call:  
## randomForest(formula = hypertension ~ ., data = hyper\_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: 14.93%  
## Confusion matrix:  
## 0 1 class.error  
## 0 1190 1 0.0008396306  
## 1 208 1 0.9952153110

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

##   
## predTrain 0 1  
## 0 1191 193  
## 1 0 16

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

## [1] 0.8516667

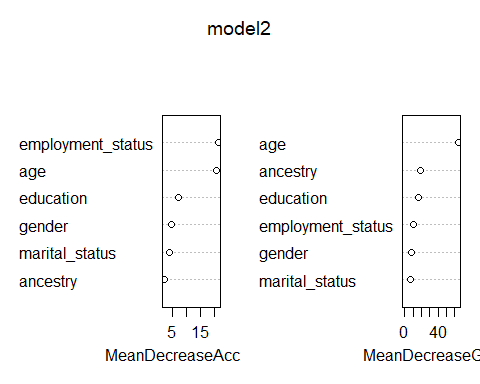
table(predValid,hyper\_ValidSet$hypertension)

##   
## predValid 0 1  
## 0 511 89  
## 1 0 0

# To check important variables  
importance(model2)

## 0 1 MeanDecreaseAccuracy  
## gender 2.993876 4.0978549 4.415259  
## age 21.552305 -14.6831814 20.817189  
## employment\_status 21.759440 -14.7451270 21.549575  
## education 7.903064 -1.4755122 6.995147  
## marital\_status 5.229751 -3.0194415 3.768667  
## ancestry 2.518510 -0.7734031 2.128883  
## MeanDecreaseGini  
## gender 7.570786  
## age 64.612706  
## employment\_status 10.175206  
## education 15.749273  
## marital\_status 7.063689  
## ancestry 18.210400

varImpPlot(model2)



## Naive Bayes Model

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

##   
## Naive Bayes Classifier for Discrete Predictors  
##   
## Call:  
## naiveBayes.default(x = X, y = Y, laplace = laplace)  
##   
## A-priori probabilities:  
## Y  
## 0 1   
## 0.8507143 0.1492857   
##   
## Conditional probabilities:  
## gender  
## Y female male  
## 0 0.4853065 0.5146935  
## 1 0.5263158 0.4736842  
##   
## age  
## Y [,1] [,2]  
## 0 65.82032 14.02212  
## 1 65.43541 13.12685  
##   
## employment\_status  
## Y employed retired student unemployed  
## 0 0.370277078 0.491183879 0.011754828 0.126784215  
## 1 0.401913876 0.444976077 0.004784689 0.148325359  
##   
## education  
## Y bachelors highschool masters phd/md  
## 0 0.53736356 0.23425693 0.13434089 0.09403862  
## 1 0.58373206 0.22488038 0.11961722 0.07177033  
##   
## marital\_status  
## Y married single  
## 0 0.7380353 0.2619647  
## 1 0.7511962 0.2488038  
##   
## ancestry  
## Y british east\_europe north\_europe west\_europe  
## 0 0.1435768 0.2300588 0.1528128 0.4735516  
## 1 0.1578947 0.2392344 0.1339713 0.4688995