Predict the risk of breast cancer

Zheman Zhao

2023-11-27

```
rm(list=ls())
setwd("/Users/zhemanzhao/Downloads")
```

```
adult18=read.csv("adult18.csv")
adult17=read.csv("adult17.csv")
adult16=read.csv("adult16.csv")
adult15=read.csv("adult15.csv")
adult14=read.csv("adult14_new.csv")
adult13=read.csv("adult13.csv")
adult12=read.csv("adult12.csv")
adult11=read.csv("adult11.csv")
adult10=read.csv("adult10.csv")
adult09=read.csv("adult09.csv")
adult08=read.csv("adult08.csv")
adult07=read.csv("adult07.csv")
adult06=read.csv("adult06.csv")
adult05=read.csv("adult05.csv")
```

clean the dataset adult05

```
# 1 have breast cancer; 2 don't have breast cancer
adult05$breast_cancer <- adult05$CNKIND5
adult05$breast_cancer[is.na(adult05$breast_cancer)] <- 2  #people who didn't have cancer are coded into
```

```
# 1 have hypertension ; 2 don't have hypertension
adult05$hypertension <- adult05$HYPEV
```

```
# 01 White 02 Black/African American 03 AIAN 04 Asian 05 Unknow 06 other race
adult05$race <- adult05$RACERPI2
```

```
adult05$age <-adult05$AGE_P
```

```
# 1 Married 2 Widowed 3 Divorced 4 Separated 5 Never married 6 Living with partner 7 Unknow
adult05$marital_status <- adult05$R_MARITL
adult05$marital_status[adult05$marital_status %in% c(1,2,3)] <- 1
adult05$marital_status[adult05$marital_status == 4] <- 2
adult05$marital_status[adult05$marital_status == 5] <- 3
adult05$marital_status[adult05$marital_status == 6] <- 4
```

```
adult05$marital_status[adult05$marital_status == 7] <- 5
adult05$marital_status[adult05$marital_status == 8] <- 6
adult05$marital_status[adult05$marital_status == 9] <- 7
```

```
adult05$BMI <- adult05$BMI*0.01
```

```
adult05$sex <- adult05$SEX
```

```
# 1 Current smokers 2 former smokers 3 never smokers 4 don't know
adult05$smoking_status <- adult05$SMKSTAT2
adult05$smoking_status[adult05$smoking_status %in% c(1,2)] <- 1
adult05$smoking_status[adult05$smoking_status==3] <- 2
adult05$smoking_status[adult05$smoking_status==4] <- 3
adult05$smoking_status[adult05$smoking_status %in% c(5,9)] <- 4
```

```
adult2005 <- adult05[,c("breast_cancer","hypertension","race","age","marital_status","BMI","smoking_sta
```

clean the dataset adult06

```
# 1 have breast cancer; 2 don't have breast cancer
adult06$breast_cancer <- adult06$CNKIND5
adult06$breast_cancer[is.na(adult06$breast_cancer)] <- 2  #people who didn't have cancer are coded into
```

```
# 1 have hypertension ; 2 don't have hypertension
adult06$hypertension <- adult06$HYPEV
```

```
# 01 White 02 Black/African American 03 AIAN 04 Asian 05 Unknow 06 other race
adult06$race <- adult06$RACERPI2
```

```
adult06$age <-adult06$AGE_P
```

```
# 1 Married 2 Widowed 3 Divorced 4 Separated 5 Never married 6 Living with partner 7 Unknow
adult06$marital_status <- adult06$R_MARITL
adult06$marital_status[adult06$marital_status %in% c(1,2,3)] <- 1
adult06$marital_status[adult06$marital_status == 4] <- 2
adult06$marital_status[adult06$marital_status == 5] <- 3
adult06$marital_status[adult06$marital_status == 6] <- 4
adult06$marital_status[adult06$marital_status == 7] <- 5
adult06$marital_status[adult06$marital_status == 8] <- 6
adult06$marital_status[adult06$marital_status == 9] <- 7
```

```
adult06$BMI <- adult06$BMI
```

```
adult06$sex <- adult06$SEX
```

```
# 1 Current smokers 2 former smokers 3 never smokers 4 don't know
adult06$smoking_status <- adult06$SMKSTAT2
adult06$smoking_status[adult06$smoking_status %in% c(1,2)] <- 1
```

```r
adult06$smoking_status[adult06$smoking_status==3] <- 2
adult06$smoking_status[adult06$smoking_status==4] <- 3
adult06$smoking_status[adult06$smoking_status %in% c(5,9)] <- 4
```

```r
adult2006 <- adult06[,c("breast_cancer","hypertension","race","age","marital_status","BMI","smoking_sta
```

clean the dataset adult07

```r
# 1 have breast cancer; 2 don't have breast cancer
adult07$breast_cancer <- adult07$CNKIND5
adult07$breast_cancer[is.na(adult07$breast_cancer)] <- 2  #people who didn't have cancer are coded into
```

```r
# 1 have hypertension ; 2 don't have hypertension
adult07$hypertension <- adult07$HYPEV
```

```r
# 01 White 02 Black/African American 03 AIAN 04 Asian 05 Unknow 06 other race
adult07$race <- adult07$RACERPI2
```

```r
adult07$age <-adult07$AGE_P
```

```r
# 1 Married 2 Widowed 3 Divorced 4 Separated 5 Never married 6 Living with partner 7 Unknow
adult07$marital_status <- adult07$R_MARITL
adult07$marital_status[adult07$marital_status %in% c(1,2,3)] <- 1
adult07$marital_status[adult07$marital_status == 4] <- 2
adult07$marital_status[adult07$marital_status == 5] <- 3
adult07$marital_status[adult07$marital_status == 6] <- 4
adult07$marital_status[adult07$marital_status == 7] <- 5
adult07$marital_status[adult07$marital_status == 8] <- 6
adult07$marital_status[adult07$marital_status == 9] <- 7
```

```r
adult07$BMI <- adult07$BMI
```

```r
adult07$sex <- adult07$SEX
```

```r
# 1 Current smokers 2 former smokers 3 never smokers 4 don't know
adult07$smoking_status <- adult07$SMKSTAT2
adult07$smoking_status[adult07$smoking_status %in% c(1,2)] <- 1
adult07$smoking_status[adult07$smoking_status==3] <- 2
adult07$smoking_status[adult07$smoking_status==4] <- 3
adult07$smoking_status[adult07$smoking_status %in% c(5,9)] <- 4
```

```r
adult2007 <- adult07[,c("breast_cancer","hypertension","race","age","marital_status","BMI","smoking_sta
```

clean the dataset adult08

```r
# 1 have breast cancer; 2 don't have breast cancer
adult08$breast_cancer <- adult08$CNKIND5
adult08$breast_cancer[is.na(adult08$breast_cancer)] <- 2  #people who didn't have cancer are coded into
```

```r
# 1 have hypertension ; 2 don't have hypertension
adult08$hypertension <- adult08$HYPEV
```

```r
# 01 White 02 Black/African American 03 AIAN 04 Asian 05 Unknow 06 other race
adult08$race <- adult08$RACERPI2
```

```r
adult08$age <-adult08$AGE_P
```

```r
# 1 Married 2 Widowed 3 Divorced 4 Separated 5 Never married 6 Living with partner 7 Unknow
adult08$marital_status <- adult08$R_MARITL
adult08$marital_status[adult08$marital_status %in% c(1,2,3)] <- 1
adult08$marital_status[adult08$marital_status == 4] <- 2
adult08$marital_status[adult08$marital_status == 5] <- 3
adult08$marital_status[adult08$marital_status == 6] <- 4
adult08$marital_status[adult08$marital_status == 7] <- 5
adult08$marital_status[adult08$marital_status == 8] <- 6
adult08$marital_status[adult08$marital_status == 9] <- 7
```

```r
adult08$BMI <- adult08$BMI
```

```r
adult08$sex <- adult08$SEX
```

```r
# 1 Current smokers 2 former smokers 3 never smokers 4 don't know
adult08$smoking_status <- adult08$SMKSTAT2
adult08$smoking_status[adult08$smoking_status %in% c(1,2)] <- 1
adult08$smoking_status[adult08$smoking_status==3] <- 2
adult08$smoking_status[adult08$smoking_status==4] <- 3
adult08$smoking_status[adult08$smoking_status %in% c(5,9)] <- 4
```

```r
adult2008 <- adult08[,c("breast_cancer","hypertension","race","age","marital_status","BMI","smoking_sta
```

clean the dataset adult09

```r
# 1 have breast cancer; 2 don't have breast cancer
adult09$breast_cancer <- adult09$CNKIND5
adult09$breast_cancer[is.na(adult09$breast_cancer)] <- 2  #people who didn't have cancer are coded into
```

```r
# 1 have hypertension ; 2 don't have hypertension
adult09$hypertension <- adult09$HYPEV
```

```r
# 01 White 02 Black/African American 03 AIAN 04 Asian 05 Unknow 06 other race
adult09$race <- adult09$RACERPI2
```

```
adult09$age <-adult09$AGE_P
```

```
# 1 Married 2 Widowed 3 Divorced 4 Separated 5 Never married 6 Living with partner 7 Unknow
adult09$marital_status <- adult09$R_MARITL
adult09$marital_status[adult09$marital_status %in% c(1,2,3)] <- 1
adult09$marital_status[adult09$marital_status == 4] <- 2
adult09$marital_status[adult09$marital_status == 5] <- 3
adult09$marital_status[adult09$marital_status == 6] <- 4
adult09$marital_status[adult09$marital_status == 7] <- 5
adult09$marital_status[adult09$marital_status == 8] <- 6
adult09$marital_status[adult09$marital_status == 9] <- 7
```

```
adult09$BMI <- adult09$BMI
```

```
adult09$sex <- adult09$SEX
```

```
# 1 Current smokers 2 former smokers 3 never smokers 4 don't know
adult09$smoking_status <- adult09$SMKSTAT2
adult09$smoking_status[adult09$smoking_status %in% c(1,2)] <- 1
adult09$smoking_status[adult09$smoking_status==3] <- 2
adult09$smoking_status[adult09$smoking_status==4] <- 3
adult09$smoking_status[adult09$smoking_status %in% c(5,9)] <- 4
```

```
adult2009 <- adult09[,c("breast_cancer","hypertension","race","age","marital_status","BMI","smoking_sta
```

clean the dataset adult10

```
# 1 have breast cancer; 2 don't have breast cancer
adult10$breast_cancer <- adult10$CNKIND5
adult10$breast_cancer[is.na(adult10$breast_cancer)] <- 2  #people who didn't have cancer are coded into
```

```
# 1 have hypertension ; 2 don't have hypertension
adult10$hypertension <- adult10$HYPEV
```

```
# 01 White 02 Black/African American 03 AIAN 04 Asian 05 Unknow 06 other race
adult10$race <- adult10$RACERPI2
```

```
adult10$age <-adult10$AGE_P
```

```
# 1 Married 2 Widowed 3 Divorced 4 Separated 5 Never married 6 Living with partner 7 Unknow
adult10$marital_status <- adult10$R_MARITL
adult10$marital_status[adult10$marital_status %in% c(1,2,3)] <- 1
adult10$marital_status[adult10$marital_status == 4] <- 2
adult10$marital_status[adult10$marital_status == 5] <- 3
adult10$marital_status[adult10$marital_status == 6] <- 4
adult10$marital_status[adult10$marital_status == 7] <- 5
adult10$marital_status[adult10$marital_status == 8] <- 6
adult10$marital_status[adult10$marital_status == 9] <- 7
```

```r
adult10$BMI <- adult10$BMI
```

```r
adult10$sex <- adult10$SEX
```

```r
# 1 Current smokers 2 former smokers 3 never smokers 4 don't know
adult10$smoking_status <- adult10$SMKSTAT2
adult10$smoking_status[adult10$smoking_status %in% c(1,2)] <- 1
adult10$smoking_status[adult10$smoking_status==3] <- 2
adult10$smoking_status[adult10$smoking_status==4] <- 3
adult10$smoking_status[adult10$smoking_status %in% c(5,9)] <- 4
```

```r
adult2010 <- adult10[,c("breast_cancer","hypertension","race","age","marital_status","BMI","smoking_sta
```

clean the dataset adult11

```r
# 1 have breast cancer; 2 don't have breast cancer
adult11$breast_cancer <- adult11$CNKIND5
adult11$breast_cancer[is.na(adult11$breast_cancer)] <- 2  #people who didn't have cancer are coded into
```

```r
# 1 have hypertension ; 2 don't have hypertension
adult11$hypertension <- adult11$HYPEV
```

```r
# 01 White 02 Black/African American 03 AIAN 04 Asian 05 Unknow 06 other race
adult11$race <- adult11$RACERPI2
```

```r
adult11$age <-adult11$AGE_P
```

```r
# 1 Married 2 Widowed 3 Divorced 4 Separated 5 Never married 6 Living with partner 7 Unknow
adult11$marital_status <- adult11$R_MARITL
adult11$marital_status[adult11$marital_status %in% c(1,2,3)] <- 1
adult11$marital_status[adult11$marital_status == 4] <- 2
adult11$marital_status[adult11$marital_status == 5] <- 3
adult11$marital_status[adult11$marital_status == 6] <- 4
adult11$marital_status[adult11$marital_status == 7] <- 5
adult11$marital_status[adult11$marital_status == 8] <- 6
adult11$marital_status[adult11$marital_status == 9] <- 7
```

```r
adult11$BMI <- adult11$BMI
```

```r
adult11$sex <- adult11$SEX
```

```r
# 1 Current smokers 2 former smokers 3 never smokers 4 don't know
adult11$smoking_status <- adult11$SMKSTAT2
adult11$smoking_status[adult11$smoking_status %in% c(1,2)] <- 1
adult11$smoking_status[adult11$smoking_status==3] <- 2
adult11$smoking_status[adult11$smoking_status==4] <- 3
adult11$smoking_status[adult11$smoking_status %in% c(5,9)] <- 4
```

```r
adult2011 <- adult11[,c("breast_cancer","hypertension","race","age","marital_status","BMI","smoking_sta
```

clean the dataset adult12

```r
# 1 have breast cancer; 2 don't have breast cancer
adult12$breast_cancer <- adult12$CNKIND5
adult12$breast_cancer[is.na(adult12$breast_cancer)] <- 2  #people who didn't have cancer are coded into
```

```r
# 1 have hypertension ; 2 don't have hypertension
adult12$hypertension <- adult12$HYPEV
```

```r
# 01 White 02 Black/African American 03 AIAN 04 Asian 05 Unknow 06 other race
adult12$race <- adult12$RACERPI2
```

```r
adult12$age <-adult12$AGE_P
```

```r
# 1 Married 2 Widowed 3 Divorced 4 Separated 5 Never married 6 Living with partner 7 Unknow
adult12$marital_status <- adult12$R_MARITL
adult12$marital_status[adult12$marital_status %in% c(1,2,3)] <- 1
adult12$marital_status[adult12$marital_status == 4] <- 2
adult12$marital_status[adult12$marital_status == 5] <- 3
adult12$marital_status[adult12$marital_status == 6] <- 4
adult12$marital_status[adult12$marital_status == 7] <- 5
adult12$marital_status[adult12$marital_status == 8] <- 6
adult12$marital_status[adult12$marital_status == 9] <- 7
```

```r
adult12$BMI <- adult12$BMI
```

```r
# 1 Current smokers 2 former smokers 3 never smokers 4 don't know
adult12$smoking_status <- adult12$SMKSTAT2
adult12$smoking_status[adult12$smoking_status %in% c(1,2)] <- 1
adult12$smoking_status[adult12$smoking_status==3] <- 2
adult12$smoking_status[adult12$smoking_status==4] <- 3
adult12$smoking_status[adult12$smoking_status %in% c(5,9)] <- 4
```

```r
adult12$sex <- adult12$SEX
```

```r
adult2012 <- adult12[,c("breast_cancer","hypertension","race","age","marital_status","BMI","smoking_sta
```

clean the dataset adult13

```r
# 1 have breast cancer; 2 don't have breast cancer
adult13$breast_cancer <- adult13$CNKIND5
adult13$breast_cancer[is.na(adult13$breast_cancer)] <- 2  #people who didn't have cancer are coded into
```

```r
# 1 have hypertension ; 2 don't have hypertension
adult13$hypertension <- adult13$HYPEV
```

```r
# 01 White 02 Black/African American 03 AIAN 04 Asian 05 Unknow 06 other race
adult13$race <- adult13$RACERPI2
```

```r
adult13$age <-adult13$AGE_P
```

```r
# 1 Married 2 Widowed 3 Divorced 4 Separated 5 Never married 6 Living with partner 7 Unknow
adult13$marital_status <- adult13$R_MARITL
adult13$marital_status[adult13$marital_status %in% c(1,2,3)] <- 1
adult13$marital_status[adult13$marital_status == 4] <- 2
adult13$marital_status[adult13$marital_status == 5] <- 3
adult13$marital_status[adult13$marital_status == 6] <- 4
adult13$marital_status[adult13$marital_status == 7] <- 5
adult13$marital_status[adult13$marital_status == 8] <- 6
adult13$marital_status[adult13$marital_status == 9] <- 7
```

```r
adult13$BMI <- adult13$BMI
```

```r
adult13$sex <- adult13$SEX
```

```r
# 1 Current smokers 2 former smokers 3 never smokers 4 don't know
adult13$smoking_status <- adult13$SMKSTAT2
adult13$smoking_status[adult13$smoking_status %in% c(1,2)] <- 1
adult13$smoking_status[adult13$smoking_status==3] <- 2
adult13$smoking_status[adult13$smoking_status==4] <- 3
adult13$smoking_status[adult13$smoking_status %in% c(5,9)] <- 4
```

```r
adult2013 <- adult13[,c("breast_cancer","hypertension","race","age","marital_status","BMI","smoking_sta
```

clean the dataset adult14

```r
# 1 have breast cancer; 2 don't have breast cancer
adult14$breast_cancer <- adult14$CNKIND5
adult14$breast_cancer[is.na(adult14$breast_cancer)] <- 2  #people who didn't have cancer are coded into
```

```r
# 1 have hypertension ; 2 don't have hypertension
adult14$hypertension <- adult14$HYPEV
```

```r
# 01 White 02 Black/African American 03 AIAN 04 Asian 05 Unknow 06 other race
adult14$race <- adult14$RACERPI2
```

```r
adult14$age <-adult14$AGE_P
```

```r
# 1 Married 2 Widowed 3 Divorced 4 Separated 5 Never married 6 Living with partner 7 Unknow
adult14$marital_status <- adult14$R_MARITL
adult14$marital_status[adult14$marital_status %in% c(1,2,3)] <- 1
```

```
adult14$marital_status[adult14$marital_status == 4] <- 2
adult14$marital_status[adult14$marital_status == 5] <- 3
adult14$marital_status[adult14$marital_status == 6] <- 4
adult14$marital_status[adult14$marital_status == 7] <- 5
adult14$marital_status[adult14$marital_status == 8] <- 6
adult14$marital_status[adult14$marital_status == 9] <- 7
```

```
adult14$BMI <- adult14$BMI
```

```
adult14$sex <- adult14$SEX
```

```
# 1 Current smokers 2 former smokers 3 never smokers 4 don't know
adult14$smoking_status <- adult14$SMKSTAT2
adult14$smoking_status[adult14$smoking_status %in% c(1,2)] <- 1
adult14$smoking_status[adult14$smoking_status==3] <- 2
adult14$smoking_status[adult14$smoking_status==4] <- 3
adult14$smoking_status[adult14$smoking_status %in% c(5,9)] <- 4
```

```
adult2014 <- adult14[,c("breast_cancer","hypertension","race","age","marital_status","BMI","smoking_sta
```

clean the dataset adult15

```
# 1 have breast cancer; 2 don't have breast cancer
adult15$breast_cancer <- adult15$CNKIND5
adult15$breast_cancer[is.na(adult15$breast_cancer)] <- 2  #people who didn't have cancer are coded into
```

```
# 1 have hypertension ; 2 don't have hypertension
adult15$hypertension <- adult15$HYPEV
```

```
# 01 White 02 Black/African American 03 AIAN 04 Asian 05 Unknow 06 other race
adult15$race <- adult15$RACERPI2
```

```
adult15$age <-adult15$AGE_P
```

```
# 1 Married 2 Widowed 3 Divorced 4 Separated 5 Never married 6 Living with partner 7 Unknow
adult15$marital_status <- adult15$R_MARITL
adult15$marital_status[adult15$marital_status %in% c(1,2,3)] <- 1
adult15$marital_status[adult15$marital_status == 4] <- 2
adult15$marital_status[adult15$marital_status == 5] <- 3
adult15$marital_status[adult15$marital_status == 6] <- 4
adult15$marital_status[adult15$marital_status == 7] <- 5
adult15$marital_status[adult15$marital_status == 8] <- 6
adult15$marital_status[adult15$marital_status == 9] <- 7
```

```
adult15$BMI <- adult15$BMI*0.01
```

```
adult15$sex <- adult15$SEX
```

```r
# 1 Current smokers 2 former smokers 3 never smokers 4 don't know
adult15$smoking_status <- adult15$SMKSTAT2
adult15$smoking_status[adult15$smoking_status %in% c(1,2)] <- 1
adult15$smoking_status[adult15$smoking_status==3] <- 2
adult15$smoking_status[adult15$smoking_status==4] <- 3
adult15$smoking_status[adult15$smoking_status %in% c(5,9)] <- 4
```

```r
adult2015 <- adult15[,c("breast_cancer","hypertension","race","age","marital_status","BMI","smoking_sta
```

clean the dataset adult16

```r
# 1 have breast cancer; 2 don't have breast cancer
adult16$breast_cancer <- adult16$CNKIND5
adult16$breast_cancer[is.na(adult16$breast_cancer)] <- 2  #people who didn't have cancer are coded into
```

```r
# 1 have hypertension ; 2 don't have hypertension
adult16$hypertension <- adult16$HYPEV
```

```r
# 01 White 02 Black/African American 03 AIAN 04 Asian 05 Unknow 06 other race
adult16$race <- adult16$RACERPI2
```

```r
adult16$age <-adult16$AGE_P
```

```r
# 1 Married 2 Widowed 3 Divorced 4 Separated 5 Never married 6 Living with partner 7 Unknow
adult16$marital_status <- adult16$R_MARITL
adult16$marital_status[adult16$marital_status %in% c(1,2,3)] <- 1
adult16$marital_status[adult16$marital_status == 4] <- 2
adult16$marital_status[adult16$marital_status == 5] <- 3
adult16$marital_status[adult16$marital_status == 6] <- 4
adult16$marital_status[adult16$marital_status == 7] <- 5
adult16$marital_status[adult16$marital_status == 8] <- 6
adult16$marital_status[adult16$marital_status == 9] <- 7
```

```r
adult16$BMI <- adult16$BMI*0.01
```

```r
adult16$sex <- adult16$SEX
```

```r
# 1 Current smokers 2 former smokers 3 never smokers 4 don't know
adult16$smoking_status <- adult16$SMKSTAT2
adult16$smoking_status[adult16$smoking_status %in% c(1,2)] <- 1
adult16$smoking_status[adult16$smoking_status==3] <- 2
adult16$smoking_status[adult16$smoking_status==4] <- 3
adult16$smoking_status[adult16$smoking_status %in% c(5,9)] <- 4
```

```r
adult2016 <- adult16[,c("breast_cancer","hypertension","race","age","marital_status","BMI","smoking_sta
```

clean the dataset adult17

```r
# 1 have breast cancer; 2 don't have breast cancer
adult17$breast_cancer <- adult17$CNKIND5
adult17$breast_cancer[is.na(adult17$breast_cancer)] <- 2  #people who didn't have cancer are coded into
```

```r
# 1 have hypertension ; 2 don't have hypertension
adult17$hypertension <- adult17$HYPEV
```

```r
# 01 White 02 Black/African American 03 AIAN 04 Asian 05 Unknow 06 other race
adult17$race <- adult17$RACERPI2
```

```r
adult17$age <-adult17$AGE_P
```

```r
# 1 Married 2 Widowed 3 Divorced 4 Separated 5 Never married 6 Living with partner 7 Unknow
adult17$marital_status <- adult17$R_MARITL
adult17$marital_status[adult17$marital_status %in% c(1,2,3)] <- 1
adult17$marital_status[adult17$marital_status == 4] <- 2
adult17$marital_status[adult17$marital_status == 5] <- 3
adult17$marital_status[adult17$marital_status == 6] <- 4
adult17$marital_status[adult17$marital_status == 7] <- 5
adult17$marital_status[adult17$marital_status == 8] <- 6
adult17$marital_status[adult17$marital_status == 9] <- 7
```

```r
adult17$BMI <- adult17$BMI*0.01
```

```r
adult17$sex <- adult17$SEX
```

```r
# 1 Current smokers 2 former smokers 3 never smokers 4 don't know
adult17$smoking_status <- adult17$SMKSTAT2
adult17$smoking_status[adult17$smoking_status %in% c(1,2)] <- 1
adult17$smoking_status[adult17$smoking_status==3] <- 2
adult17$smoking_status[adult17$smoking_status==4] <- 3
adult17$smoking_status[adult17$smoking_status %in% c(5,9)] <- 4
```

```r
adult2017 <- adult17[,c("breast_cancer","hypertension","race","age","marital_status","BMI","smoking_sta
```

clean the dataset adult18

```r
# 1 have breast cancer; 2 don't have breast cancer; 7 Refused; 8 Not ascertained; 9 Don't know
adult18$breast_cancer <- adult18$CNKIND5
adult18$breast_cancer[is.na(adult18$breast_cancer)] <- 2  #people who didn't have cancer are coded into
```

```r
# 1 have hypertension ; 2 don't have hypertension; 7 Refused; 8 Not ascertained; 9 Don't know
adult18$hypertension <- adult18$HYPEV
```

```r
# 01 White 02 Black/African American 03 AIAN 04 Asian 05 Unknow 06 other race
adult18$race <- adult18$RACERPI2
```

```r
adult18$age <-adult18$AGE_P
```

```r
# 1 Married 2 Widowed 3 Divorced 4 Separated 5 Never married 6 Living with partner 7 Unknow
adult18$marital_status <- adult18$R_MARITL
adult18$marital_status[adult18$marital_status %in% c(1,2,3)] <- 1
adult18$marital_status[adult18$marital_status == 4] <- 2
adult18$marital_status[adult18$marital_status == 5] <- 3
adult18$marital_status[adult18$marital_status == 6] <- 4
adult18$marital_status[adult18$marital_status == 7] <- 5
adult18$marital_status[adult18$marital_status == 8] <- 6
adult18$marital_status[adult18$marital_status == 9] <- 7
```

```r
adult18$BMI <- adult18$BMI*0.01
```

```r
adult18$sex <- adult18$SEX
```

```r
# 1 Current smokers 2 former smokers 3 never smokers 4 don't know
adult18$smoking_status <- adult18$SMKSTAT2
adult18$smoking_status[adult18$smoking_status %in% c(1,2)] <- 1
adult18$smoking_status[adult18$smoking_status==3] <- 2
adult18$smoking_status[adult18$smoking_status==4] <- 3
adult18$smoking_status[adult18$smoking_status %in% c(5,9)] <- 4
```

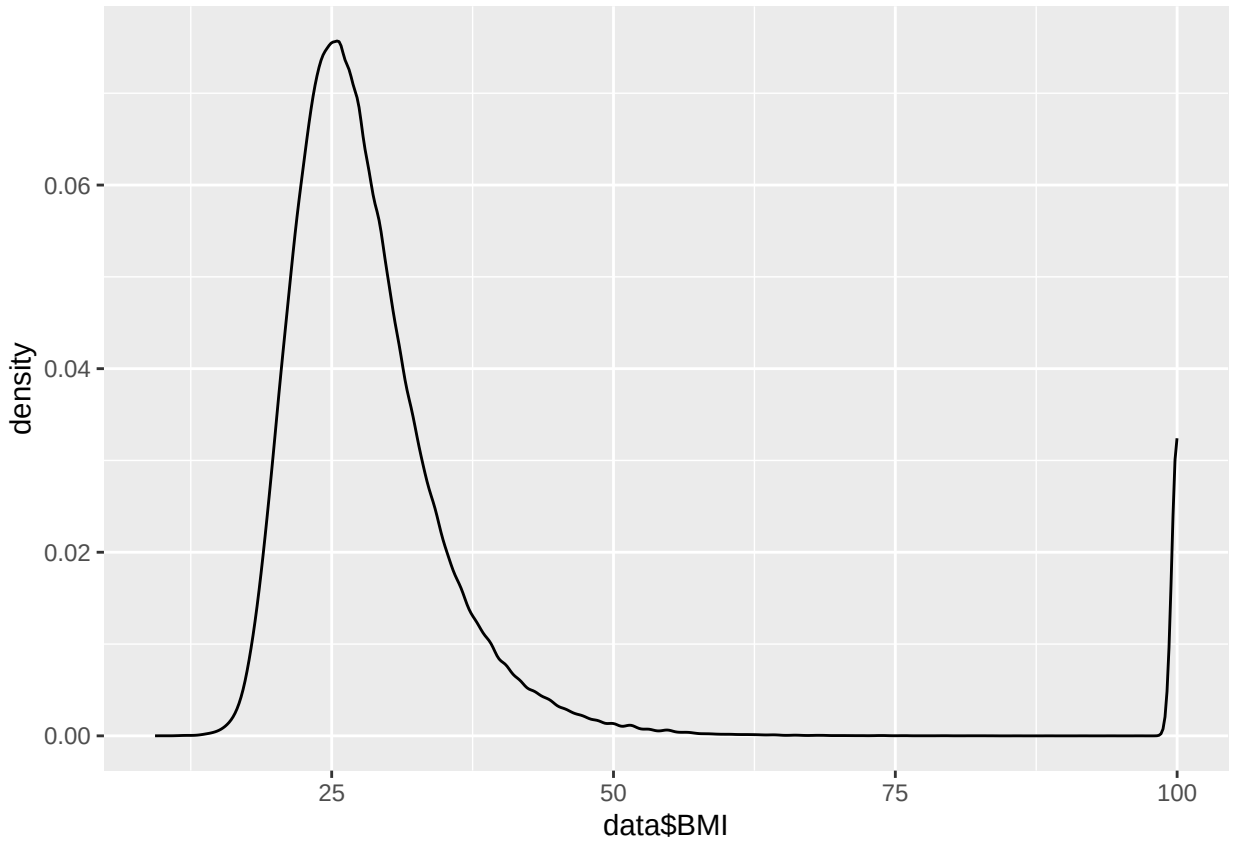
```r
adult2018 <- adult18[,c("breast_cancer","hypertension","race","age","marital_status","BMI","smoking_sta
```

the final dataset

```r
data <- rbind(adult2005,adult2006,adult2007,adult2008,adult2009,adult2010,adult2011,adult2012,adult2013
```

```r
library(ggplot2)
ggplot(data.frame(data$BMI),aes(data$BMI))+geom_density()
```

```
## Warning: Removed 9 rows containing non-finite values (‘stat_density()‘).
```




```
data$breast_cancer[data$breast_cancer==9 | data$breast_cancer==7] <- NA
data$hypertension[data$hypertension==9 | data$hypertension==7] <- NA
data$race[data$race==5] <- NA
data$race[data$race==6] <- 5
data$marital_status[data$marital_status==7] <- NA
data$BMI[data$BMI==99.99] <- NA
data$BMI[data$BMI<18.5] <- 1
data$BMI[data$BMI>=18.5 & data$BMI<25] <- 2
data$BMI[data$BMI>=25 & data$BMI<30] <- 3
data$BMI[data$BMI>=30] <- 4
data$smoking_status[data$smoking_status==4] <- NA
```

check the sample size

```
adult_female=data[data$sex==2,]
premenopausal_female=adult_female[adult_female$age>=18&adult_female$age<=60,]
premenopausal_female<-premenopausal_female[complete.cases(premenopausal_female),]
nrow(premenopausal_female)
```

```
## [1] 99944
```

```
table(data$BMI)
```

```
##
```

```
##     1     2     3     4
##  4638 89901 91135 81012
```

```
nrow(premenopausal_female)
```

```
## [1] 99944
```

```
pf_hypertension=premenopausal_female[premenopausal_female$hypertension==1,]
nrow(pf_hypertension)
```

```
## [1] 20424
```

```
pf_breast=premenopausal_female[premenopausal_female$breast_cancer==1,]
nrow(pf_breast)
```

```
## [1] 1304
```

```
pf_breast_h=pf_breast[pf_breast$hypertension==1,]
nrow(pf_breast_h)
```

```
## [1] 457
```

```
premenopausal_female$breast_cancer[premenopausal_female$breast_cancer==2]<-0
premenopausal_female$hypertension[premenopausal_female$hypertension==2]<-0
```

description tables

```
premenopausal_female$breast_cancer <- factor(premenopausal_female$breast_cancer, levels = c(0,1), labels
premenopausal_female$hypertension <- factor(premenopausal_female$hypertension, levels = c(0,1), labels =
premenopausal_female$race <- factor(premenopausal_female$race, levels = c(1, 2, 3, 4, 5), labels = c("W
premenopausal_female$marital_status <- factor(premenopausal_female$marital_status, levels = c(1, 2, 3, 4
premenopausal_female$BMI <- factor(premenopausal_female$BMI, levels = c(1, 2, 3, 4), labels = c("underwe
premenopausal_female$smoking_status <- factor(premenopausal_female$smoking_status, levels = c(1, 2, 3),
```

```
library(table1)
```

```
##
## Attaching package: 'table1'
```

```
## The following objects are masked from 'package:base':
##
##     units, units<-
```

```
table1(~breast_cancer+hypertension+race+age+marital_status+BMI+smoking_status,data = premenopausal_femal
```

```
## Get nicer `table1` LaTeX output by simply installing the `kableExtra` package
```

	Overall
	(N=99944)
breast_cancer	
don't have breast cancer	98640 (98.7%)
have breast cancer	1304 (1.3%)
hypertension	
don't have hypertension	79520 (79.6%)
have hypertension	20424 (20.4%)
race	
White	74116 (74.2%)
Black/African American	16715 (16.7%)
AIAN	1200 (1.2%)
Asian	5817 (5.8%)
Other race	2096 (2.1%)
age	
Mean (SD)	39.7 (12.1)
Median [Min, Max]	39.0 [18.0, 60.0]
marital_status	
married	44992 (45.0%)
widowed	2472 (2.5%)
divorced	13703 (13.7%)
separated	4174 (4.2%)
never married	27321 (27.3%)
living with partner	7282 (7.3%)
BMI	
underweight	2315 (2.3%)
healthy weight	39235 (39.3%)
overweight	26855 (26.9%)
obesity	31539 (31.6%)
smoking_status	
current smokers	18556 (18.6%)
former smokers	15068 (15.1%)
never smokers	66320 (66.4%)

```
pvalue <- function(x, ...) {
    y <- unlist(x)
    g <- factor(rep(1:length(x), times=sapply(x, length)))
    if (is.numeric(y)) {
        p <- t.test(y ~ g)$p.value
    } else {
        p <- chisq.test(table(y, g))$p.value
    }
    c("", sub("<", "&lt;", format.pval(p, digits=3, eps=0.001)))}
table1(~hypertension+race+age+marital_status+BMI+smoking_status | breast_cancer,data = premenopausal_fe
```

```
## Get nicer `table1` LaTeX output by simply installing the `kableExtra` package
```

	don't have breast cancer	have breast cancer	P-value
	(N=98640)	(N=1304)	
hypertension			

	don't have breast cancer	have breast cancer	P-value
don't have hypertension	78673 (79.8%)	847 (65.0%)	<0.001
have hypertension	19967 (20.2%)	457 (35.0%)	
race			
White	73080 (74.1%)	1036 (79.4%)	<0.001
Black/African American	16542 (16.8%)	173 (13.3%)	
AIAN	1184 (1.2%)	16 (1.2%)	
Asian	5763 (5.8%)	54 (4.1%)	
Other race	2071 (2.1%)	25 (1.9%)	
age			
Mean (SD)	39.5 (12.0)	52.1 (6.97)	<0.001
Median [Min, Max]	39.0 [18.0, 60.0]	54.0 [19.0, 60.0]	
marital_status			
married	44352 (45.0%)	640 (49.1%)	<0.001
widowed	2409 (2.4%)	63 (4.8%)	
divorced	13389 (13.6%)	314 (24.1%)	
separated	4108 (4.2%)	66 (5.1%)	
never married	27168 (27.5%)	153 (11.7%)	
living with partner	7214 (7.3%)	68 (5.2%)	
BMI			
underweight	2299 (2.3%)	16 (1.2%)	0.0103
healthy weight	38740 (39.3%)	495 (38.0%)	
overweight	26468 (26.8%)	387 (29.7%)	
obesity	31133 (31.6%)	406 (31.1%)	
smoking_status			
current smokers	18319 (18.6%)	237 (18.2%)	<0.001
former smokers	14740 (14.9%)	328 (25.2%)	
never smokers	65581 (66.5%)	739 (56.7%)	

```
table1(~breast_cancer+race+age+marital_status+BMI+smoking_status | hypertension,data = premenopausal_fe
```

```
## Get nicer 'table1' LaTeX output by simply installing the 'kableExtra' package
```

	don't have hypertension	have hypertension	P-value
	(N=79520)	(N=20424)	
breast_cancer			
don't have breast cancer	78673 (98.9%)	19967 (97.8%)	<0.001
have breast cancer	847 (1.1%)	457 (2.2%)	
race			
White	60538 (76.1%)	13578 (66.5%)	<0.001
Black/African American	11304 (14.2%)	5411 (26.5%)	
AIAN	923 (1.2%)	277 (1.4%)	
Asian	5097 (6.4%)	720 (3.5%)	
Other race	1658 (2.1%)	438 (2.1%)	
age			
Mean (SD)	37.7 (11.7)	47.2 (10.4)	<0.001
Median [Min, Max]	37.0 [18.0, 60.0]	50.0 [18.0, 60.0]	
marital_status			
married	36291 (45.6%)	8701 (42.6%)	<0.001
widowed	1420 (1.8%)	1052 (5.2%)	

	don't have hypertension	have hypertension	P-value
divorced	9583 (12.1%)	4120 (20.2%)	
separated	3004 (3.8%)	1170 (5.7%)	
never married	23139 (29.1%)	4182 (20.5%)	
living with partner	6083 (7.6%)	1199 (5.9%)	
BMI			
underweight	2140 (2.7%)	175 (0.9%)	<0.001
healthy weight	35360 (44.5%)	3875 (19.0%)	
overweight	21498 (27.0%)	5357 (26.2%)	
obesity	20522 (25.8%)	11017 (53.9%)	
smoking_status			
current smokers	13868 (17.4%)	4688 (23.0%)	<0.001
former smokers	11171 (14.0%)	3897 (19.1%)	
never smokers	54481 (68.5%)	11839 (58.0%)	

split the data

```
set.seed(123)
n<-nrow(premenopausal_female)
train_size <- round(0.8*n)
train_indices <- sample(1:n,size=train_size)
train_set <- premenopausal_female[train_indices,]
test_set <- premenopausal_female[-train_indices,]
```

develop logistic models

```
single_model1<-glm(breast_cancer~hypertension, family = 'binomial',data=train_set)
single_model2<-glm(breast_cancer~race, family = 'binomial',data=train_set)
single_model3<-glm(breast_cancer~age, family = 'binomial',data=train_set)
single_model4<-glm(breast_cancer~marital_status, family = 'binomial',data=train_set)
single_model5<-glm(breast_cancer~BMI, family = 'binomial',data=train_set)
single_model6<-glm(breast_cancer~smoking_status, family = 'binomial',data=train_set)
summary(single_model1)
```

```
##
## Call:
## glm(formula = breast_cancer ~ hypertension, family = "binomial",
##     data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.2175  -0.1440  -0.1440  -0.1440   3.0247
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -4.56411    0.03925 -116.30   <2e-16 ***
## hypertensionhave hypertension  0.83194    0.06497   12.81   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11081  on 79954  degrees of freedom
## Residual deviance: 10931  on 79953  degrees of freedom
## AIC: 10935
##
## Number of Fisher Scoring iterations: 7
```

```
summary(single_model2)
```

```
##
## Call:
## glm(formula = breast_cancer ~ race, family = "binomial", data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.1669  -0.1669  -0.1669  -0.1613   3.0466
##
## Coefficients:
##                             Estimate Std. Error  z value Pr(>|z|)
## (Intercept)                 -4.26695    0.03519 -121.265  < 2e-16 ***
## raceBlack/African American -0.27725    0.09168   -3.024  0.00249 **
## raceAIAN                    -0.19686    0.30528   -0.645  0.51901
## raceAsian                   -0.36410    0.15387   -2.366  0.01797 *
## raceOther race              -0.06855    0.21746   -0.315  0.75257
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11081  on 79954  degrees of freedom
## Residual deviance: 11066  on 79950  degrees of freedom
## AIC: 11076
##
## Number of Fisher Scoring iterations: 7
```

```
summary(single_model3)
```

```
##
## Call:
## glm(formula = breast_cancer ~ age, family = "binomial", data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3437  -0.1950  -0.1038  -0.0583   3.8807
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.721254   0.210336  -46.22   <2e-16 ***
## age          0.115366   0.003999   28.85   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11081.0  on 79954  degrees of freedom
## Residual deviance:  9805.6  on 79953  degrees of freedom
## AIC: 9809.6
##
## Number of Fisher Scoring iterations: 8
```

```
summary(single_model4)
```

```
##
## Call:
## glm(formula = breast_cancer ~ marital_status, family = "binomial",
##     data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.2277  -0.1701  -0.1701  -0.1040   3.2317
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -4.22882    0.04430 -95.461  < 2e-16 ***
## marital_statuswidowed             0.58917    0.14861   3.965 7.35e-05 ***
## marital_statusdivorced            0.45131    0.07823   5.769 7.99e-09 ***
## marital_statusseparated           0.15375    0.14301   1.075 0.282335
## marital_statusnever married      -0.98782    0.10239  -9.648  < 2e-16 ***
## marital_statusliving with partner -0.49665    0.14746  -3.368 0.000757 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11081  on 79954  degrees of freedom
## Residual deviance: 10856  on 79949  degrees of freedom
## AIC: 10868
##
## Number of Fisher Scoring iterations: 8
```

```
summary(single_model5)
```

```
##
## Call:
## glm(formula = breast_cancer ~ BMI, family = "binomial", data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.1739  -0.1739  -0.1611  -0.1555   3.1443
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.46703    0.07349 -60.786   <2e-16 ***
```

```
## BMI.L        0.45140    0.19118   2.361   0.0182 *
## BMI.Q       -0.34052    0.14698  -2.317   0.0205 *
## BMI.C       -0.01772    0.08156  -0.217   0.8280
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11081  on 79954  degrees of freedom
## Residual deviance: 11066  on 79951  degrees of freedom
## AIC: 11074
##
## Number of Fisher Scoring iterations: 7
```

```
summary(single_model6)
```

```
##
## Call:
## glm(formula = breast_cancer ~ smoking_status, family = "binomial",
##     data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.2060  -0.1594  -0.1505  -0.1505   2.9956
##
## Coefficients:
##                  Estimate Std. Error  z value Pr(>|z|)
## (Intercept)      -4.22586    0.03517 -120.160  < 2e-16 ***
## smoking_status.L -0.08185    0.05962   -1.373     0.17
## smoking_status.Q -0.46983    0.06218   -7.556 4.14e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11081  on 79954  degrees of freedom
## Residual deviance: 11017  on 79952  degrees of freedom
## AIC: 11023
##
## Number of Fisher Scoring iterations: 7
```

```
full_model <- glm(breast_cancer~hypertension+race+age+marital_status+BMI+smoking_status, family = 'bino
summary(full_model)
```

```
##
## Call:
## glm(formula = breast_cancer ~ hypertension + race + age + marital_status +
##     BMI + smoking_status, family = "binomial", data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4597  -0.1909  -0.1026  -0.0577   3.9173
##
```

```
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -9.662401   0.233679 -41.349  < 2e-16 ***
## hypertensionhave hypertension    0.171466   0.071196   2.408  0.01602 *
## raceBlack/African American      -0.158901   0.097009  -1.638  0.10142
## raceAIAN                        -0.052554   0.308524  -0.170  0.86474
## raceAsian                       -0.162068   0.157495  -1.029  0.30346
## raceOther race                   0.168062   0.220656   0.762  0.44627
## age                              0.112268   0.004242  26.468  < 2e-16 ***
## marital_statuswidowed           -0.152999   0.151179  -1.012  0.31152
## marital_statusdivorced           0.111528   0.080366   1.388  0.16521
## marital_statusseparated          0.264021   0.146189   1.806  0.07091 .
## marital_statusnever married     -0.115050   0.107059  -1.075  0.28254
## marital_statusliving with partner  0.158613   0.150631   1.053  0.29235
## BMI.L                            0.030985   0.194620   0.159  0.87351
## BMI.Q                           -0.284547   0.148565  -1.915  0.05545 .
## BMI.C                            0.037027   0.082413   0.449  0.65322
## smoking_status.L                 0.057720   0.061495   0.939  0.34793
## smoking_status.Q                -0.171480   0.063373  -2.706  0.00681 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11081.0  on 79954  degrees of freedom
## Residual deviance:  9766.7  on 79938  degrees of freedom
## AIC: 9800.7
##
## Number of Fisher Scoring iterations: 8
```

```
ci <- confint(full_model)
```

```
## Waiting for profiling to be done...
```

```
exp_ci <- exp(ci)
print(exp_ci)
```

```
##                                       2.5 %       97.5 %
## (Intercept)                     3.994109e-05 9.985968e-05
## hypertensionhave hypertension   1.031818e+00 1.364079e+00
## raceBlack/African American      7.029332e-01 1.028446e+00
## raceAIAN                        4.871329e-01 1.651855e+00
## raceAsian                       6.161902e-01 1.144112e+00
## raceOther race                  7.454316e-01 1.777783e+00
## age                             1.109660e+00 1.128271e+00
## marital_statuswidowed           6.309019e-01 1.142496e+00
## marital_statusdivorced          9.537612e-01 1.307128e+00
## marital_statusseparated         9.676643e-01 1.718190e+00
## marital_statusnever married     7.196885e-01 1.095382e+00
## marital_statusliving with partner 8.624438e-01 1.558485e+00
## BMI.L                           7.238441e-01 1.561966e+00
## BMI.Q                           5.489871e-01 9.873934e-01
## BMI.C                           8.878874e-01 1.228455e+00
```

```
## smoking_status.L                  9.405184e-01 1.197062e+00
## smoking_status.Q                  7.447221e-01 9.548537e-01
```

```
print(exp(full_model$coefficients))
```

```
##                         (Intercept)     hypertensionhave hypertension
##                        6.363153e-05                        1.187044e+00
##         raceBlack/African American                            raceAIAN
##                        8.530812e-01                        9.488030e-01
##                           raceAsian                      raceOther race
##                        8.503832e-01                        1.183010e+00
##                                 age              marital_statuswidowed
##                        1.118813e+00                        8.581305e-01
##             marital_statusdivorced           marital_statusseparated
##                        1.117985e+00                        1.302156e+00
##       marital_statusnever married marital_statusliving with partner
##                        8.913218e-01                        1.171884e+00
##                               BMI.L                               BMI.Q
##                        1.031470e+00                        7.523549e-01
##                               BMI.C                   smoking_status.L
##                        1.037721e+00                        1.059418e+00
##                    smoking_status.Q
##                        8.424170e-01
```

```
library(car)
```

```
## Loading required package: carData
```

```
vif(full_model)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## hypertension   1.180921  1        1.086702
## race           1.142009  4        1.016737
## age            1.121582  1        1.059048
## marital_status 1.161760  5        1.015107
## BMI            1.137963  3        1.021774
## smoking_status 1.068992  2        1.016819
```

```
adjust_full_model <- glm(breast_cancer~hypertension+race+age+marital_status+BMI+smoking_status+hyperten
summary(adjust_full_model)
```

```
##
## Call:
## glm(formula = breast_cancer ~ hypertension + race + age + marital_status +
##     BMI + smoking_status + hypertension * age, family = "binomial",
##     data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4711  -0.1944  -0.1027  -0.0553   3.9642
##
```

```
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -9.966390   0.265755 -37.502  < 2e-16 ***
## hypertensionhave hypertension    1.530917   0.503331   3.042  0.00235 **
## raceBlack/African American      -0.160429   0.097032  -1.653  0.09826 .
## raceAIAN                        -0.055783   0.308415  -0.181  0.85647
## raceAsian                       -0.158988   0.157504  -1.009  0.31277
## raceOther race                   0.166557   0.220626   0.755  0.45029
## age                              0.118224   0.004883  24.211  < 2e-16 ***
## marital_statuswidowed           -0.143721   0.151134  -0.951  0.34163
## marital_statusdivorced           0.111862   0.080319   1.393  0.16370
## marital_statusseparated          0.262210   0.146146   1.794  0.07279 .
## marital_statusnever married     -0.110069   0.107041  -1.028  0.30382
## marital_statusliving with partner  0.158322   0.150615   1.051  0.29318
## BMI.L                            0.026909   0.194658   0.138  0.89005
## BMI.Q                           -0.282337   0.148583  -1.900  0.05741 .
## BMI.C                            0.037144   0.082417   0.451  0.65221
## smoking_status.L                 0.062534   0.061507   1.017  0.30930
## smoking_status.Q                -0.173528   0.063361  -2.739  0.00617 **
## hypertensionhave hypertension:age -0.025517   0.009407  -2.713  0.00668 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11081.0  on 79954  degrees of freedom
## Residual deviance:  9759.7  on 79937  degrees of freedom
## AIC: 9795.7
##
## Number of Fisher Scoring iterations: 8
```

```
print(exp(adjust_full_model$coefficients))
```

```
##                       (Intercept)     hypertensionhave hypertension
##                      4.695176e-05                      4.622412e+00
##        raceBlack/African American                          raceAIAN
##                      8.517779e-01                      9.457446e-01
##                         raceAsian                     raceOther race
##                      8.530064e-01                      1.181231e+00
##                               age              marital_statuswidowed
##                      1.125496e+00                      8.661295e-01
##            marital_statusdivorced           marital_statusseparated
##                      1.118358e+00                      1.299799e+00
##       marital_statusnever married marital_statusliving with partner
##                      8.957722e-01                      1.171544e+00
##                             BMI.L                             BMI.Q
##                      1.027274e+00                      7.540194e-01
##                             BMI.C                  smoking_status.L
##                      1.037843e+00                      1.064531e+00
##                  smoking_status.Q hypertensionhave hypertension:age
##                      8.406935e-01                      9.748061e-01
```

```r
ci <- confint(adjust_full_model)
```

```
## Waiting for profiling to be done...
```

```r
exp_ci <- exp(ci)
print(exp_ci)
```

```
##                                         2.5 %       97.5 %
## (Intercept)                       2.760567e-05 7.829104e-05
## hypertensionhave hypertension     1.689246e+00 1.217376e+01
## raceBlack/African American        7.018264e-01 1.026930e+00
## raceAIAN                          4.856431e-01 1.646158e+00
## raceAsian                         6.180752e-01 1.147676e+00
## raceOther race                    7.443402e-01 1.775024e+00
## age                               1.114928e+00 1.136489e+00
## marital_statuswidowed             6.368341e-01 1.153038e+00
## marital_statusdivorced            9.541654e-01 1.307446e+00
## marital_statusseparated           9.659853e-01 1.714944e+00
## marital_statusnever married       7.232970e-01 1.100825e+00
## marital_statusliving with partner 8.622081e-01 1.558003e+00
## BMI.L                             7.208241e-01 1.555738e+00
## BMI.Q                             5.501802e-01 9.896356e-01
## BMI.C                             8.879764e-01 1.228615e+00
## smoking_status.L                  9.450291e-01 1.202870e+00
## smoking_status.Q                  7.432148e-01 9.528797e-01
## hypertensionhave hypertension:age 9.572636e-01 9.932506e-01
```

```r
null_model <-glm(breast_cancer~1, family = 'binomial',data=premenopausal_female)
```

```r
adjust_full_model_2 <- glm(breast_cancer~hypertension+race+age+marital_status+BMI+smoking_status+hyperte
summary(adjust_full_model_2)
```

```
##
## Call:
## glm(formula = breast_cancer ~ hypertension + race + age + marital_status +
##     BMI + smoking_status + hypertension * age + age * BMI, family = "binomial",
##     data = premenopausal_female)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4716  -0.1933  -0.1039  -0.0564   3.9857
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -9.9417792  0.4250686 -23.389  < 2e-16 ***
## hypertensionhave hypertension      1.2781758  0.4697986   2.721 0.006515 **
## raceBlack/African American        -0.1771876  0.0873623  -2.028 0.042541 *
## raceAIAN                           0.1093325  0.2566484   0.426 0.670107
## raceAsian                         -0.1979361  0.1434714  -1.380 0.167703
## raceOther race                     0.0834319  0.2064126   0.404 0.686066
## age                                0.1181127  0.0080969  14.587  < 2e-16 ***
```

```
## marital_statuswidowed             -0.1438144  0.1359943  -1.058 0.290282
## marital_statusdivorced             0.1433713  0.0714513   2.007 0.044797 *
## marital_statusseparated            0.2210824  0.1330553   1.662 0.096596 .
## marital_statusnever married       -0.0732435  0.0944462  -0.776 0.438041
## marital_statusliving with partner  0.2240448  0.1311468   1.708 0.087571 .
## BMI.L                              0.5512508  1.0564467   0.522 0.601812
## BMI.Q                             -0.0261612  0.8155320  -0.032 0.974409
## BMI.C                              0.1184315  0.4708540   0.252 0.801408
## smoking_status.L                   0.0458399  0.0548002   0.836 0.402878
## smoking_status.Q                  -0.2032928  0.0559819  -3.631 0.000282 ***
## hypertensionhave hypertension:age -0.0220705  0.0087975  -2.509 0.012117 *
## age:BMI.L                         -0.0098077  0.0201744  -0.486 0.626864
## age:BMI.Q                         -0.0052360  0.0155657  -0.336 0.736585
## age:BMI.C                         -0.0003443  0.0089702  -0.038 0.969387
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13907  on 99943  degrees of freedom
## Residual deviance: 12253  on 99923  degrees of freedom
## AIC: 12295
##
## Number of Fisher Scoring iterations: 9
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

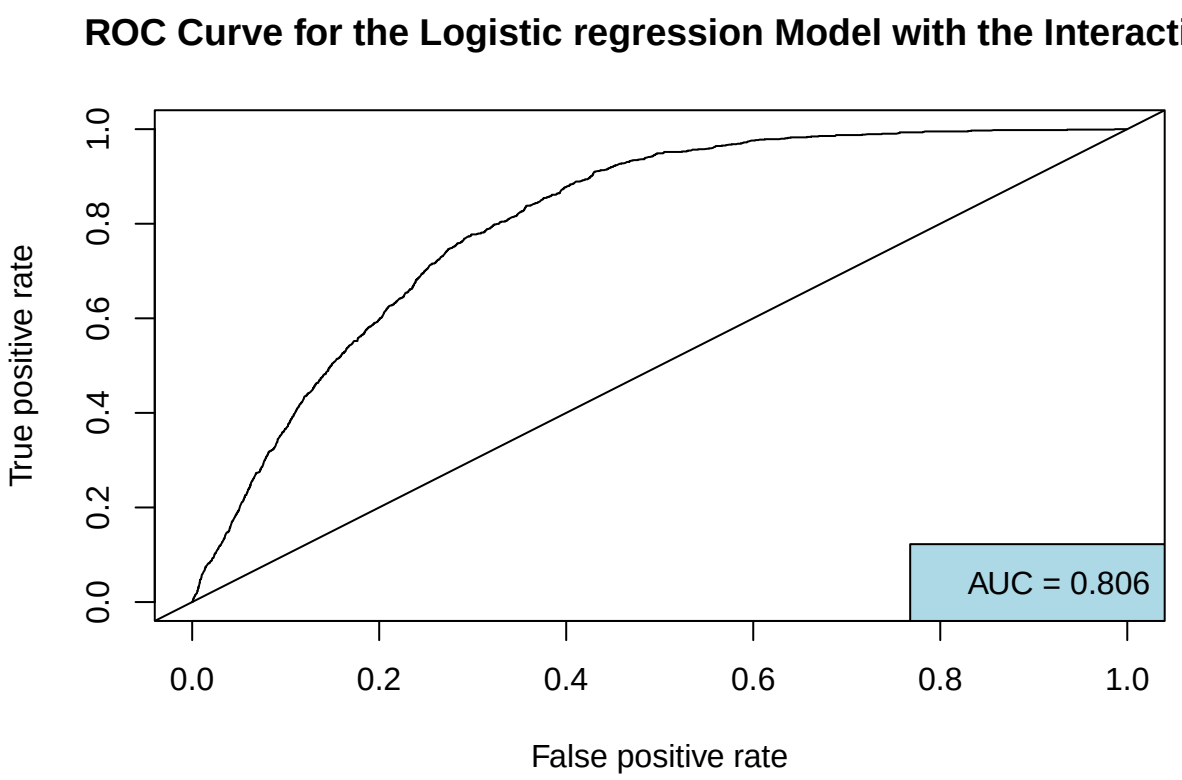
```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
lrtest(adjust_full_model,full_model)
```

```
## Likelihood ratio test
##
## Model 1: breast_cancer ~ hypertension + race + age + marital_status +
##     BMI + smoking_status + hypertension * age
## Model 2: breast_cancer ~ hypertension + race + age + marital_status +
##     BMI + smoking_status
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  18 -4879.9
## 2  17 -4883.4 -1 7.0331   0.008002 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
library(ROCR)
prob <- predict(adjust_full_model,data=train_set,type = "response")
pred <- prediction(prob,train_set$breast_cancer)
perf <- performance(pred, measure="tpr",x.measure="fpr")
plot(perf,xlim=c(0,1),ylim=c(0,1),main = "ROC Curve for the Logistic regression Model with the Interacti
auc_pred <- performance(pred,"auc")
auc_value <- auc_pred@y.values[[1]]
legend("bottomright", legend = paste("AUC =", round(auc_value, 3)), box.lty = 1, box.col = "black", bg =
abline(0,1)
```




```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

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

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

```r
train_set$model_prob <- predict(adjust_full_model,train_set,type="response")
train_set <- train_set %>% mutate(model_pred = 1*(model_prob>0.53)+0,
                                  breast_binary=1*(breast_cancer=="have breast cancer")+0)
train_set <- train_set %>% mutate(accurate = 1*(model_pred==breast_binary))
sum(train_set$accurate)/nrow(train_set)
```

```
## [1] 0.9870177
```

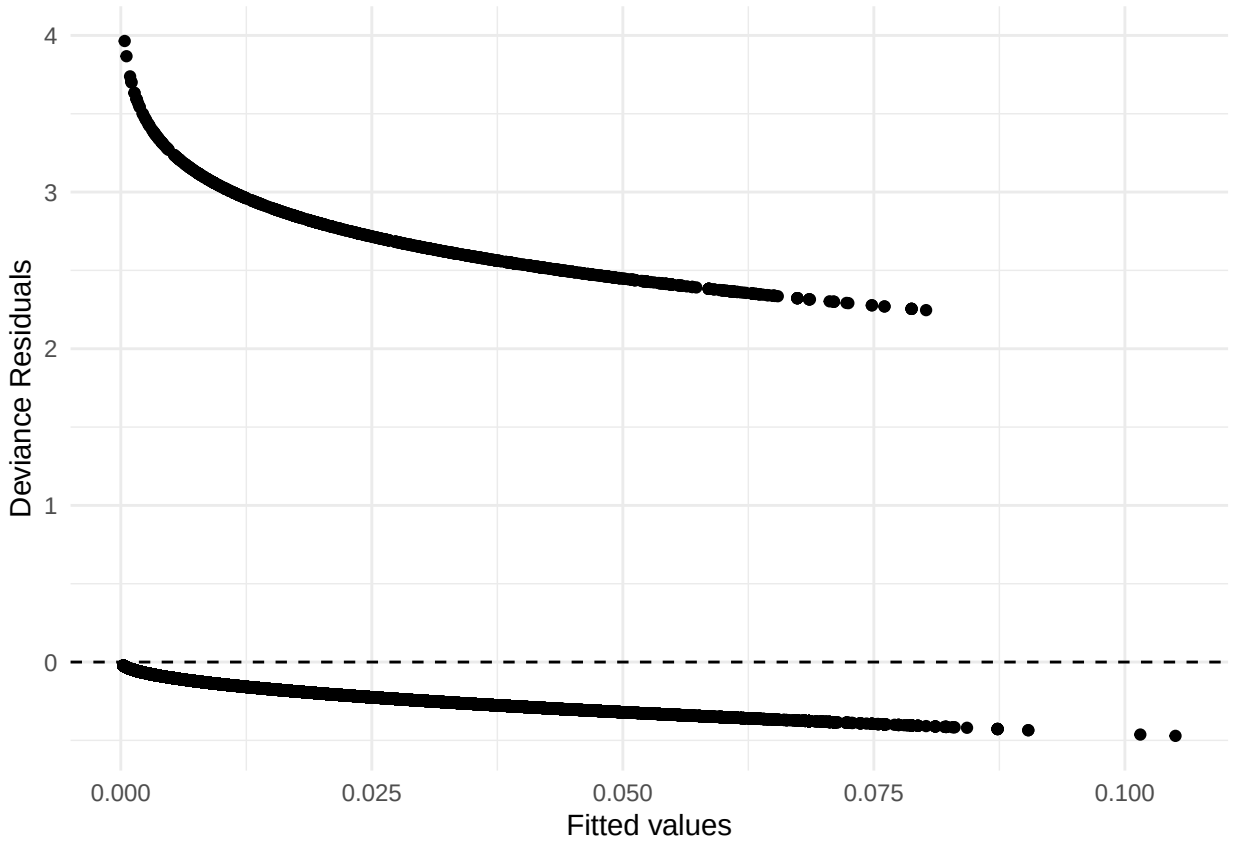
```r
test_set$model_prob <- predict(adjust_full_model,test_set,type="response")
test_set <- test_set %>% mutate(model_pred = 1*(model_prob>0.53)+0,
                                  breast_binary=1*(breast_cancer=="have breast cancer")+0)
test_set <- test_set %>% mutate(accurate = 1*(model_pred==breast_binary))
sum(test_set$accurate)/nrow(test_set)
```

```
## [1] 0.9866927
```

```r
# Assuming adjust_full_model is your logistic regression model
predicted_probabilities <- predict(adjust_full_model, newdata = train_set, type = "response")
predicted_classes <- ifelse(predicted_probabilities > 0.5, 1, 0)
actual_classes <- train_set$breast_cancer
accuracy <- mean(predicted_classes == actual_classes)
summary(predicted_classes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0       0       0       0
```

```r
library(ggplot2)
ggplot(train_set,aes(x = fitted(adjust_full_model), y = residuals(adjust_full_model,type="deviance"))) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  theme_minimal() +
  labs(x = "Fitted values", y = "Deviance Residuals")
```




```
library(ResourceSelection)
```

```
## ResourceSelection 0.3-6   2023-06-27
```

```
hoslem.test(adjust_full_model$y, fitted.values(adjust_full_model))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  adjust_full_model$y, fitted.values(adjust_full_model)
## X-squared = 26.38, df = 8, p-value = 0.000904
```

```
library(car)
vif_values <- vif(full_model)
print(vif_values)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## hypertension    1.180921  1        1.086702
## race            1.142009  4        1.016737
## age             1.121582  1        1.059048
## marital_status 1.161760  5        1.015107
## BMI             1.137963  3        1.021774
## smoking_status 1.068992  2        1.016819
```

```
library(randomForest)
```

```
## randomForest 4.7-1.1
```

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

```
##
## Attaching package: 'randomForest'
```

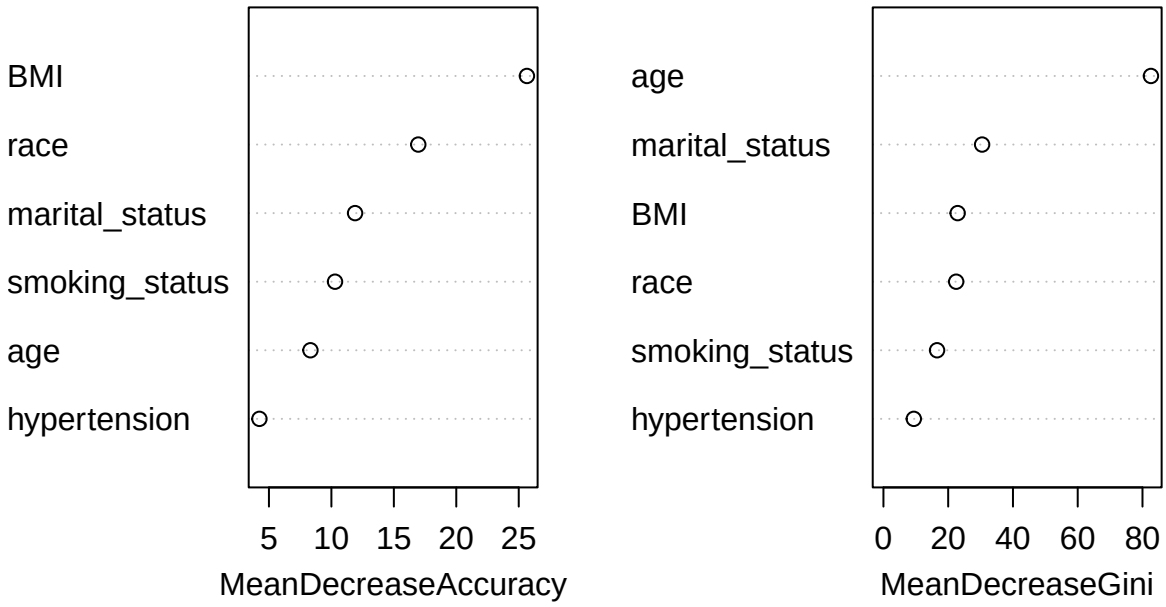
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

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

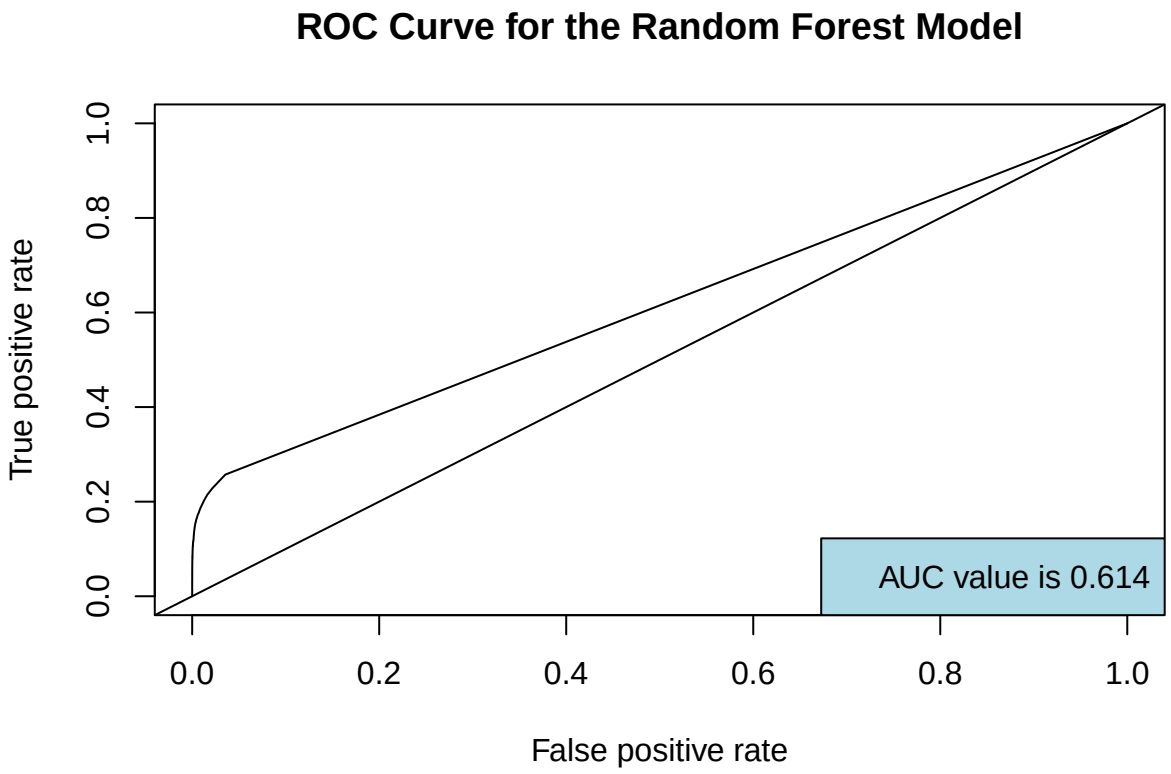
```
rf_model <- randomForest(breast_cancer~hypertension+race+age+marital_status+BMI+smoking_status, importa
importance(rf_model)
```

```
##                don't have breast cancer have breast cancer MeanDecreaseAccuracy
## hypertension                   2.859414         7.54694139             4.236089
## race                          16.994386        -0.36945290            16.957154
## age                            6.695751        10.71619934             8.322360
## marital_status                11.858020        -0.01964496            11.895507
## BMI                           25.474820         1.21595654            25.655250
## smoking_status                 9.231166         7.28931437            10.291012
##                MeanDecreaseGini
## hypertension           9.382257
## race                  22.471101
## age                   82.612277
## marital_status        30.472521
## BMI                   22.926211
## smoking_status        16.565018
```

```
varImpPlot(rf_model)
```




```r
library(randomForest)  # Assuming you've loaded randomForest for the rf_model
library(ROCR)

# Assuming rf_model is already trained and train_set is your dataset
# Predict probabilities
predictions <- predict(rf_model, newdata = train_set, type = "prob")

# Extract probabilities for the positive class
positive_probs <- predictions[, 2]  # Assuming the second column corresponds to the '1' class

# Create the prediction object needed for ROCR
pred <- prediction(predictions[,2], train_set$breast_cancer)

# Calculate performance measures
perf <- performance(pred, measure = "tpr", x.measure = "fpr")

# Calculate AUC
auc_perf <- performance(pred, measure = "auc")
auc_value <- auc_perf@y.values[[1]]  # Extract the AUC value

# Plot the ROC curve
plot(perf, main = "ROC Curve for the Random Forest Model")
abline(0, 1)  # Adding the diagonal line

# Add AUC value to the plot
legend("bottomright", legend = paste("AUC value is", round(auc_value, 3)), box.lty = 1, box.col = "blac
```




```
# Prediction
predicted_classes <- predict(rf_model, newdata = train_set)

# Actual classes from the data
actual_classes <- train_set$breast_cancer

# Generating the confusion matrix
conf_matrix <- table(Predicted = predicted_classes, Actual = actual_classes)
print(conf_matrix)
```

```
##                            Actual
## Predicted                   don't have breast cancer have breast cancer
##   don't have breast cancer                    78917               1037
##   have breast cancer                              0                  1
```

```
78917/(78917+1038)
```

```
## [1] 0.9870177
```

```
# Prediction
predicted_classes <- predict(rf_model, newdata = test_set)

# Actual classes from the data
actual_classes <- test_set$breast_cancer

# Generating the confusion matrix
conf_matrix <- table(Predicted = predicted_classes, Actual = actual_classes)
print(conf_matrix)
```

```
##                            Actual
## Predicted                   don't have breast cancer have breast cancer
##   don't have breast cancer                    19723                266
##   have breast cancer                              0                  0
```

```r
19723/(19723+266)
```

```
## [1] 0.9866927
```