Assignment 2

Pearline Nagabattula

2023-03-01

Task: Diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset, and to develop general guidelines for predicting diabetes in general.

```
library(readr)
Diabetes <- read.csv("diabetes.csv")
```

This dataset includes different variables that affect the occurrence of diabetes in an individual, and are as follows:

Response variable:

Outcome: whether or not a patient has diabetes. (yes = 1, no = 0) (nominal)

Predictor variables:

Glucose: Plasma glucose concentration after 2 hours in an oral glucose tolerance test

BloodPressure: Diastolic, mm Hg

SkinThickness: Triceps skin fold thickness in mm

Insulin: 2 hour serum insulin, μU/mL

Pregnancies: Number of times pregnant

BMI: weight in kg / (height in m)ˆ2

Age: years

DiabetesPedigreeFunction: standardized score of likelihood of diabetes based on family history X

We shall now see if there is any significant correlation between any of these factors to begin our model.

```
cor(Diabetes)
```

We find that:

Glucose and Outcome are highly correlated. (0.47).

BMI and Outcome are marginally correlated.(0.29)

Age and Outcome are marginally correlated (0.24)

Pregnancies and Outcome are marginally correlated.(0.22)

These 4 variables seem to affect the occurrence of diabetes the most. Glucose has the highest correlation as the very cause of diabetes is the increase of blood glucose levels to a level higher than normal.

Gestational diabetes is a type of diabetes that can develop during pregnancy in women who don't already have diabetes, and is not usually caused by lack of insulin, but by other hormones that make insulin lesss effective (CDC, 2022). Thus we see a correlation between the 2.

BMI affects the occurrence of Diabetes, as studies show that in obese individuals, the amount of nonesterified fatty acids, glycerol, hormones, cytokines, proinflammatory markers, and other substances that are involved in the development of insulin resistance, is increased (Al-Goblan et al. 2014).

Studies have also shown that the risk of developing diabetes increases with age. The CDC report that 4.0 percent of people aged 18 to 44 years are living with diabetes, 17 percent of those aged 45 to 64 years, and 25.2 percent of those aged over 65 years (Huizen, 2022).

Other related variables:

Pregnancy and Age are highly correlated (0.54)

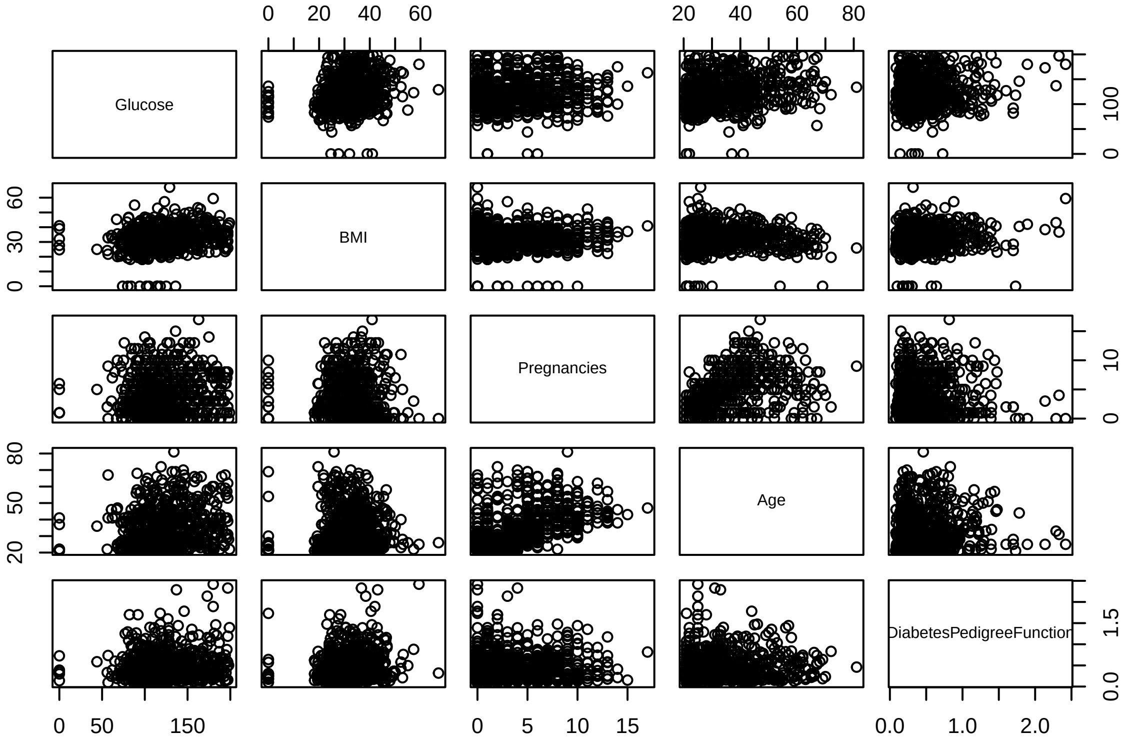
Glucose and Insulin are obviously correlated. (0.33)

Skin Thickness and BMI are also obviously correlated. (0.39)

We shall use only one of the 2 other variables as they are highly correlated. Thus we shall remove, SkinThickness, Insulin and Age and keep the other variables.

Though the correlation between DiabetesPedigreeFunction with Outcome is not that significant (0.17), we shall include it in our model as it can be an important factor.

The following plot shows the extent of the correlation between the above significant variables



All these variables are associated with each other, showing different patterns of associations, most of them positive, some neutral, some strong, some weak.

MODEL 1: Generalized Linear Model

```r
mod1 <- glm(Outcome ~ ., data = Diabetes, family = binomial)
summary(mod1)
```

After viewing the results of this model, we shall remove SkinThickness, Insulin, Age, as their p values are all greater than the significance level (α=0.05), which indicates that we cannot reject the null hypothesis that these variables do not have a significant effect on the occurrence of diabetes.

We can also remove blood pressure as it shows a negative estimate, which indicates that diastolic blood pressure gets lowered in diabetics compared to non-diabetics and we are looking at what factors positively correlate with diabetes. So we can disregard these variables, and frame our model again.

```r
mod2p <- glm(Outcome ~ . - SkinThickness - Age - Insulin - BloodPressure,
             data = Diabetes, family = binomial)
summary(mod2p)
```

```
##
## Call:
## glm(formula = Outcome ~ . - SkinThickness - Age - Insulin - BloodPressure,
##     family = binomial, data = Diabetes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7581  -0.7349  -0.4264   0.7580   2.9008
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -8.415851   0.656908 -12.811  < 2e-16 ***
## Pregnancies               0.141926   0.027105   5.236 1.64e-07 ***
## Glucose                   0.033826   0.003345  10.112  < 2e-16 ***
## BMI                       0.078097   0.013771   5.671 1.42e-08 ***
## DiabetesPedigreeFunction  0.901294   0.291696   3.090    0.002 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 734.31  on 763  degrees of freedom
## AIC: 744.31
##
## Number of Fisher Scoring iterations: 5
```

Checking our Performance

```r
pred1 <- predict(mod2p)
# pred1[1:10] (log odds)
pred2 <- predict.glm(mod2p, type = "response")
# pred2[1:10] (log odds)
```

Now, we shall make binary predictions. i.e. we shall predict a 1 if we think the patient has diabetes; and 0 if the patient does not.

```
pred2_cat <- ifelse(test = pred2 > 0.50, yes = 1, no = 0)
pred2[1:10]
```

```
##          1          2          3          4          5          6          7
## 0.65275605 0.04720267 0.79168475 0.05125442 0.83841714 0.16771338 0.06255627
##          8          9         10
## 0.44300837 0.74195506 0.05504190
```

```
pred2_cat[1:10]
```

```
##  1  2  3  4  5  6  7  8  9 10
##  1  0  1  0  1  0  0  0  1  0
```

We now make a confusion table.

```
table1 <- table(Diabetes$Outcome, pred2_cat)
table1
```

```
##    pred2_cat
##       0   1
##   0 442  58
##   1 117 151
```

```
errorrate1 <- (table1[1, 2] + table1[2, 1]) / sum(table1)
errorrate1 * 100
```

```
## [1] 22.78646
```

From this table we see that we have an error rate of 22% from 117 errors (false negatives) and 58 errors (false positives).

This means that 117 subjects show that they do not have diabetes from the methods used in the study, even though they actually have diabetes but the methods used in the study could not detect it, and 58 people show that they have diabetes even though they do not have diabetes.

False negatives are more concerning as we want to predict how many subject actually have diabetes, and if the results have more false negatives, it renders the study insignificant, and can cause problems for patients who actually have diabetes but believe they don't, and delay their treatment. To reduce this error rate, we shall try changing the threshold.

```
table(Diabetes$Outcome, rep(0, 768))
```

```
##
##       0
##   0 500
##   1 268
```

```
268/768*100
```

```
## [1] 34.89583
```

If we take a super naive approach and assume that everybody does not have diabetes, we have 268 errors and 12.3% difference. So let us change the threshold from 0.50 to 0.25 for diagnosing diabetes and see if there is a difference.

Changing threshold for (diagnosing) classifying diabetes.

```
pred2_cat2 <- ifelse(test = pred2 > 0.25, yes = 1, no = 0)
table2 <- table(Diabetes$Outcome, pred2_cat2)
table2
```

```
##    pred2_cat2
##       0   1
##   0 317 183
##   1  42 226
```

```
errorrate2 <- (table2[1, 2] + table2[2, 1]) / sum(table2)
errorrate2 * 100
```

```
## [1] 29.29688
```

By changing the threshold to 0.25, we have a higher error rate of **29%**, but we have lower false negatives (only 42) compared to the false positives (183), which is better as patients with diabetes can begin treatment early, while patients without diabetes but got a false positive result can get retested.

Let us now use other algorithms to classify diabetes.

MODEL 2: Linear Discriminant Analysis

```
library(MASS)
set.seed(69)
mod3lda <- lda(Outcome ~ . - SkinThickness - Age - Insulin - BloodPressure,
               data = Diabetes)
mod3lda
```

```
## Call:
## lda(Outcome ~ . - SkinThickness - Age - Insulin - BloodPressure,
##     data = Diabetes)
##
## Prior probabilities of groups:
##         0         1
## 0.6510417 0.3489583
##
## Group means:
```

```
##   Pregnancies  Glucose      BMI DiabetesPedigreeFunction
## 0    3.298000 109.9800 30.30420                 0.429734
## 1    4.865672 141.2575 35.14254                 0.550500
##
## Coefficients of linear discriminants:
##                                  LD1
## Pregnancies              0.11186653
## Glucose                  0.02671716
## BMI                      0.05287374
## DiabetesPedigreeFunction 0.65998391
```

```
pred3lda <- predict(mod3lda)$posterior[, 2]
```

The **prior probabilities of groups** show that 65% of the population considered in the study does not have diabetes, while 35% of the population considered in the study has diabetes.

Group Means

DiabetesPedigreeFunction:

55% of those in the study with family history of diabetes are likely to have diabetes.

43% of those in the study without family history of diabetes are not likely to have diabetes.

Pregnancies:

Women with diabetes in the study were more likely to get pregnant 4-5 times.

Women without diabetes in the study were likely to get pregnant around 3 times.

Glucose:

Most people from our study with diabetes have an average blood glucose concentration of 141.25 which is higher than most people without diabetes who have an average blood glucose concentration of 109.9

We shall disregard **BMI**, as it is an innacurate measure of body fat content and does not take into account muscle mass, bone density, overall body composition, and racial and sex differences, according to researchers from the Perelman School of Medicine, University of Pennsylvania (Nordqvist, 2022) and thus cannot accurately signify the correlation between Diabetes and BMI.

We use the probabilities to predict:

```
pred2_lda <- ifelse(test = pred3lda > 0.25, yes = 1, no = 0)
table3 <- table(Diabetes$Outcome, pred2_lda)

errorrate3 <- (table3[1, 2] + table3[2, 1]) / sum(table3)
errorrate3 * 100
```

```
## [1] 28.51562
```

```
table(pred2_cat2, pred2_lda)
```

```
##           pred2_lda
## pred2_cat2   0   1
##          0 358   1
##          1  13 396
```

We see that the error rate is **29%** which is similar to the error rate we get using the generalized linear model, and just a slight change in numbers of the false positives and false negatives.

We see that there are 13+1=16 individiuals whose predicted diabetes status is different across LDA and logistic. We shall check the probabilities of those people.

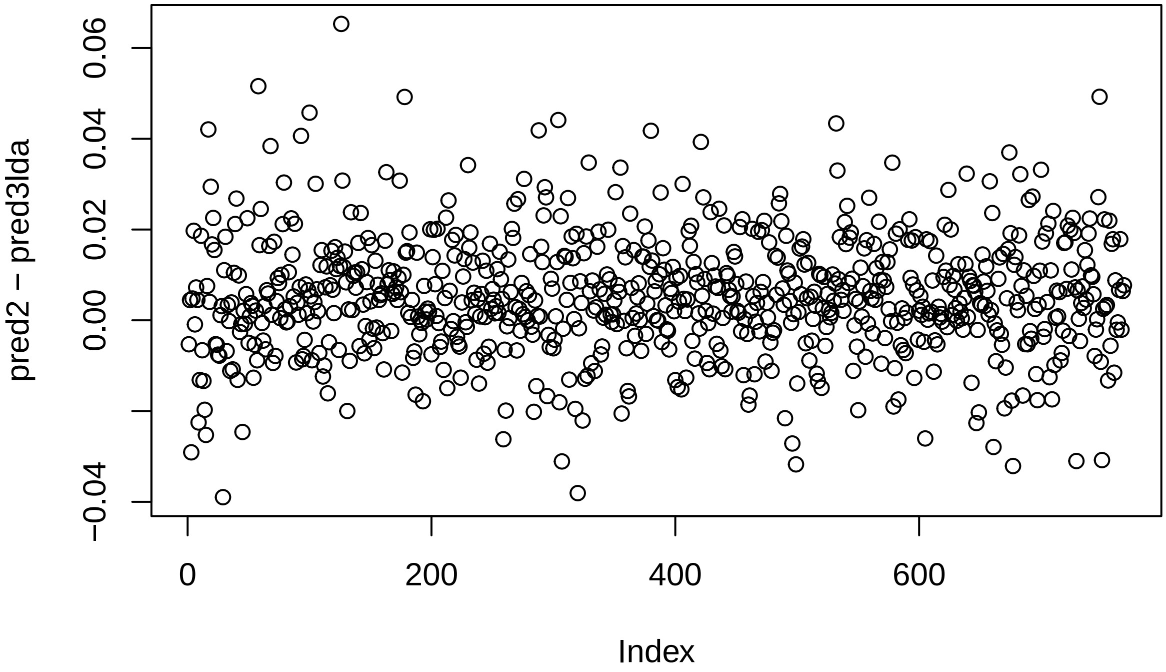
```
mismatchP <- which(pred2_cat2 + pred2_lda == 1)
pred2[mismatchP]
```

```
##        11       134       163       166       182       276       329       338 
## 0.2652896 0.2619182 0.2774300 0.2572102 0.2666877 0.2736343 0.2662300 0.2553232 
##       380       443       520       628       727       748 
## 0.2768037 0.2515531 0.2413229 0.2561221 0.2528015 0.2828672
```

```
pred3lda[mismatchP]
```

```
##        11       134       163       166       182       276       329       338 
## 0.2466671 0.2381090 0.2448027 0.2483609 0.2473335 0.2424830 0.2314820 0.2485178 
##       380       443       520       628       727       748 
## 0.2350470 0.2417254 0.2562174 0.2463382 0.2335713 0.2336199
```

```
plot(pred2 - pred3lda)
```



These probabilities differ mostly by +/-0.02, and we see the confusion matrix differences are almost entirely due to flip-flopping on the threshold line.

MODEL 3: Naive Bayes

We shall finally perform a Naive Bayes and see how it classifies diabetes with very few assumptions.

```
library(e1071)
set.seed(69)
mod_nbs <- naiveBayes(Outcome ~ . - SkinThickness - Age - Insulin, data = Diabetes)
mod_nbs
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##         0         1
## 0.6510417 0.3489583
##
## Conditional probabilities:
##    Pregnancies
## Y       [,1]     [,2]
##   0 3.298000 3.017185
##   1 4.865672 3.741239
##
##    Glucose
## Y       [,1]     [,2]
##   0 109.9800 26.14120
##   1 141.2575 31.93962
##
##    BloodPressure
## Y       [,1]     [,2]
##   0 68.18400 18.06308
##   1 70.82463 21.49181
##
##    BMI
## Y       [,1]     [,2]
##   0 30.30420 7.689855
##   1 35.14254 7.262967
##
##    DiabetesPedigreeFunction
## Y       [,1]      [,2]
##   0 0.429734 0.2990853
##   1 0.550500 0.3723545
```

```
pred_nbs <- predict(mod_nbs, newdata = Diabetes, type = "raw")[, 2]
str(pred_nbs)
```

```
##  num [1:768] 0.5749 0.0318 0.8667 0.0382 0.9997 ...
```

As we can see, we get the same results as our linear discriminant analysis model. This concludes our analysis of the study. We can use the Naive Bayes approach in the same way we used the previous 2 approaches to check our prediction probabilities, if we want to.

References:

Centre for Disease Control and Prevention. (2022). Gestational diabetes. https://www.cdc.gov/diabetes/basics/gestational.html (Accessed: 01/03/2023)

Al-Goblan, A. S., Al-Alfi, M. A., & Khan, M. Z. (2014). Mechanism linking diabetes mellitus and obesity. Diabetes, metabolic syndrome and obesity: Targets and therapy, 2014(7):587–591. https://doi.org/10.2147/DMSO.S67400

Huizen, J. (updated September 26th, 2022). The average age of onset for type 2 diabetes. Medical News Today. https://www.medicalnewstoday.com/articles/317375 (Accessed: 01/03/2023)

Nordqvist, C. (updated January 20th, 2022). Why BMI is inaccurate and misleading. Medical News Today. https://www.medicalnewstoday.com/articles/265215 (Accessed: 01/03/2023)