Ex3

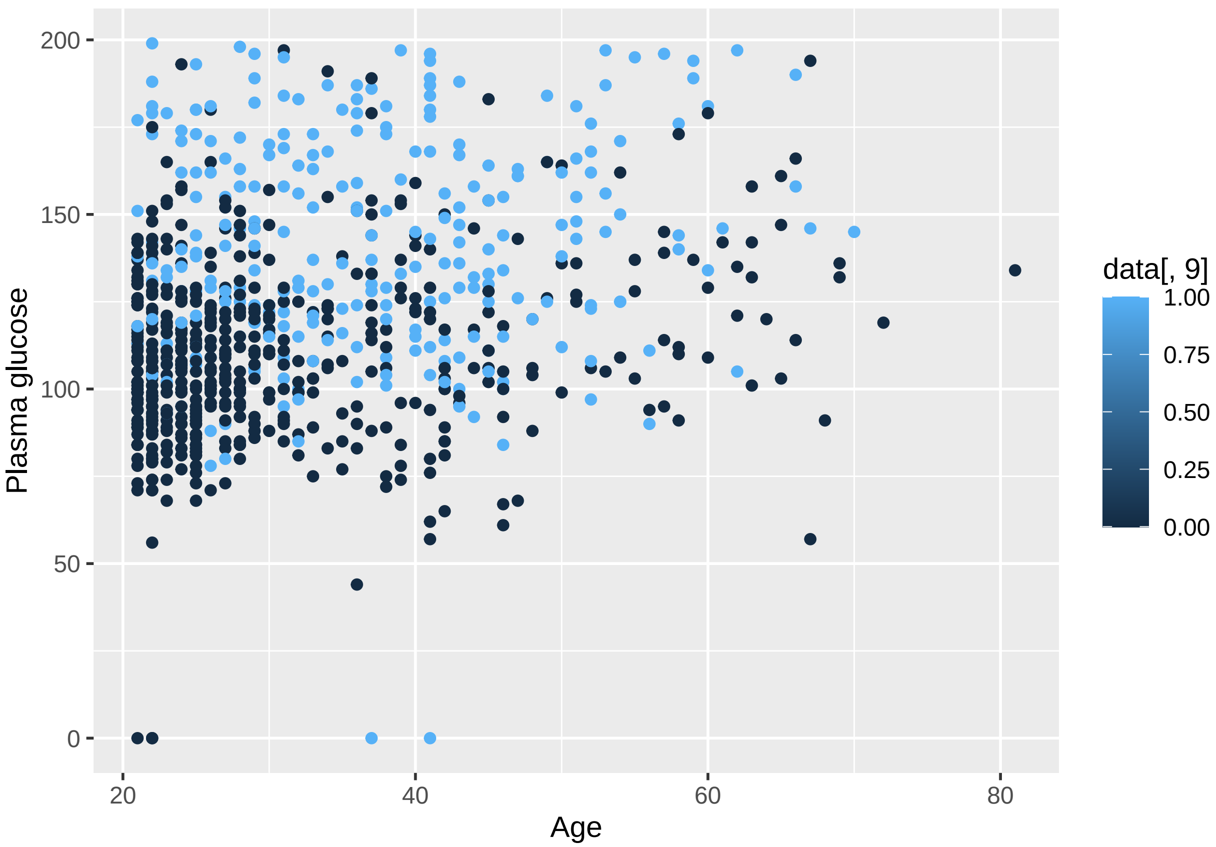
Daniele Bozzoli

2023-11-21

Ex 3

1)

```r
data <- read.csv("pima-indians-diabetes.csv")

age <- data[, 8]
plasma <- data[, 2]

names(data)[9] <- "diabetes"
names(data)[2] <- "plasma"
names(data)[8] <- "age"

x <- age
y <- plasma
df <- data.frame(x, y)

ggplot2::ggplot(df, ggplot2::aes(age, plasma)) +
  ggplot2::geom_point(ggplot2::aes(colour = data[, 9]))+
  ggplot2::labs(x = "Age", y="Plasma glucose")
```



We observe how a lot of people without diabete is highly concentrated in the lower part of the age axis, explaining how younger individuals tend to be non-diabetic, on the other hand, we notice how especially people with higher plasma glucose concentration tend to be diabetic, less concentrated in a specific age group, more spread across the x-age axis.

```
model <- glm(diabetes ~ plasma + age, fam = binomial(link="logit"), data = data)
summary(model)
```

2) and 3)

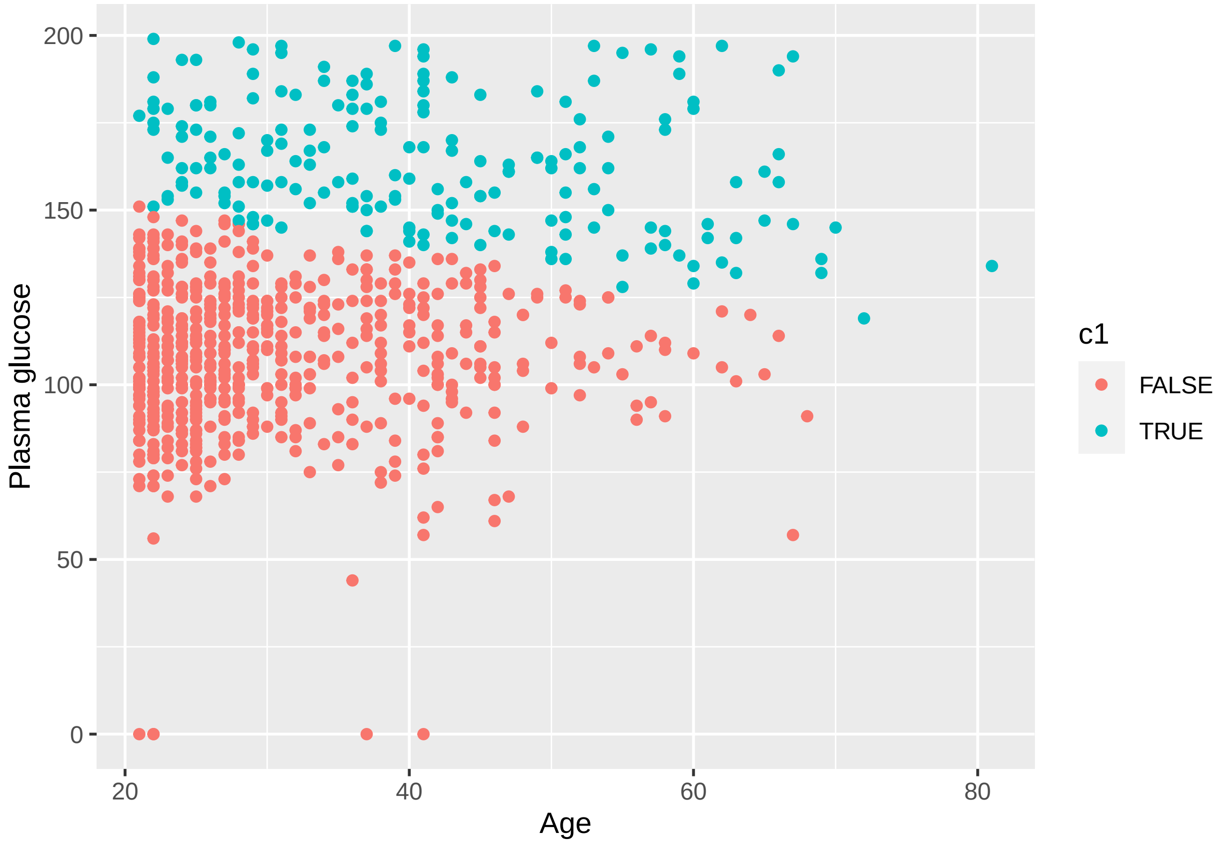
```
##
## Call:
## glm(formula = diabetes ~ plasma + age, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.897858   0.462450  -12.75  < 2e-16 ***
## plasma       0.035582   0.003288   10.82  < 2e-16 ***
## age          0.024502   0.007379    3.32 0.000899 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 991.38  on 766  degrees of freedom
## Residual deviance: 796.49  on 764  degrees of freedom
## AIC: 802.49
##
## Number of Fisher Scoring iterations: 4
```

```
# r = 0.5

c1 <- fitted(model) >= .5

missC_error <- sum(data$diabetes != c1) / nrow(data)
cat("Misclassification error:", missC_error, "\n")
```

```
## Misclassification error: 0.2659713
```

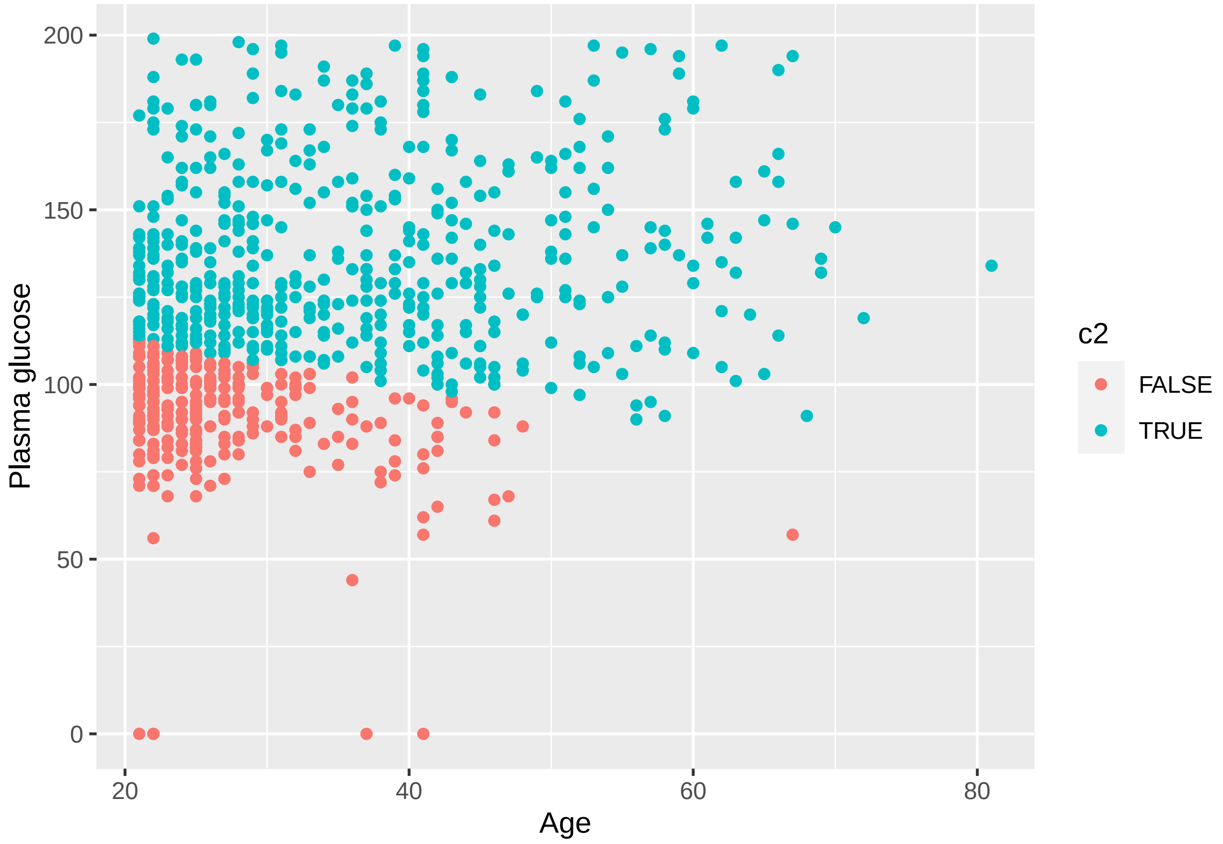
```
ggplot2::ggplot(df, ggplot2::aes(age, plasma)) +
  ggplot2::geom_point(ggplot2::aes(colour = c1))+
  ggplot2::labs(x = "Age", y="Plasma glucose")      # T = Diabetic, F = Non-diabetic
```



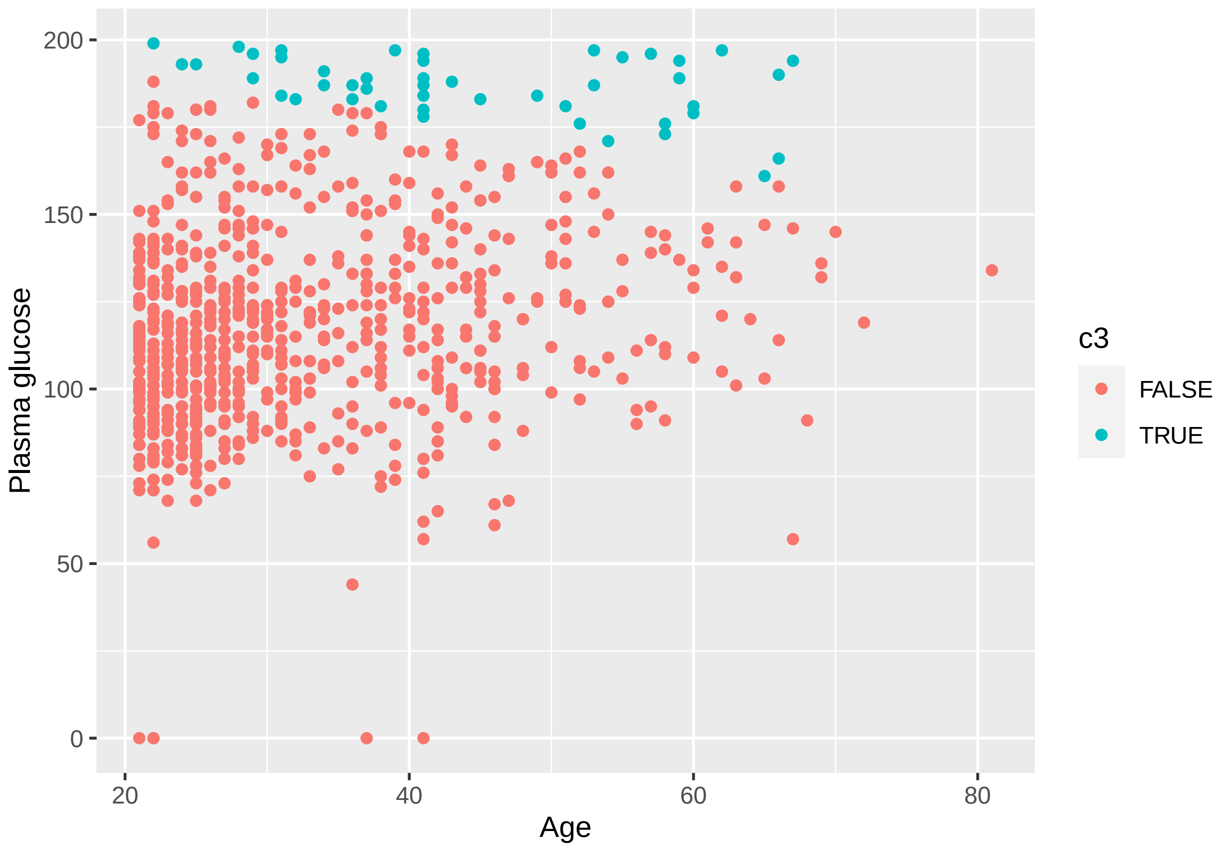
Looking at the graph, we notice how the classification is not very accurate. Though, it catches the fact that people with higher plasma glucose concentration are more commonly diabetic. We obtain a misclassification error = 0.266.

4)

```
# r = 0.2

c2 <- fitted(model) >= .2

ggplot2::ggplot(df, ggplot2::aes(age, plasma)) +
  ggplot2::geom_point(ggplot2::aes(colour = c2))+
  ggplot2::labs(x = "Age", y="Plasma glucose")
```



```
# r = 0.8

c3 <- fitted(model) >= .8

ggplot2::ggplot(df, ggplot2::aes(age, plasma)) +
  ggplot2::geom_point(ggplot2::aes(colour = c3))+
  ggplot2::labs(x = "Age", y="Plasma glucose")
```



It looks like the line that splits the predicted observations into diabetic and non diabetic into the two cases (r= 0.2 and r=0.8) has the same slope, but different intercept. Both cases are pretty bad predictors as 0.2 and 0.8 are respectively too low and too high to get a good classification for our case.

5)

```
# Adding new variables, logit function with r= 0.5

z1 <- data$plasma^4
z2 <- data$plasma^3 * data$age
z3 <- data$plasma^2 * data$age^2
z4 <- data$plasma^1 * data$age^3
z5 <- data$age^4

newdata <- cbind(data, z1, z2, z3, z4, z5)

newmodel <- glm(diabetes ~ plasma + age + z1 + z2 + z3 + z4 + z5 , family = binomial(link = "logit") ,

summary(newmodel)
```

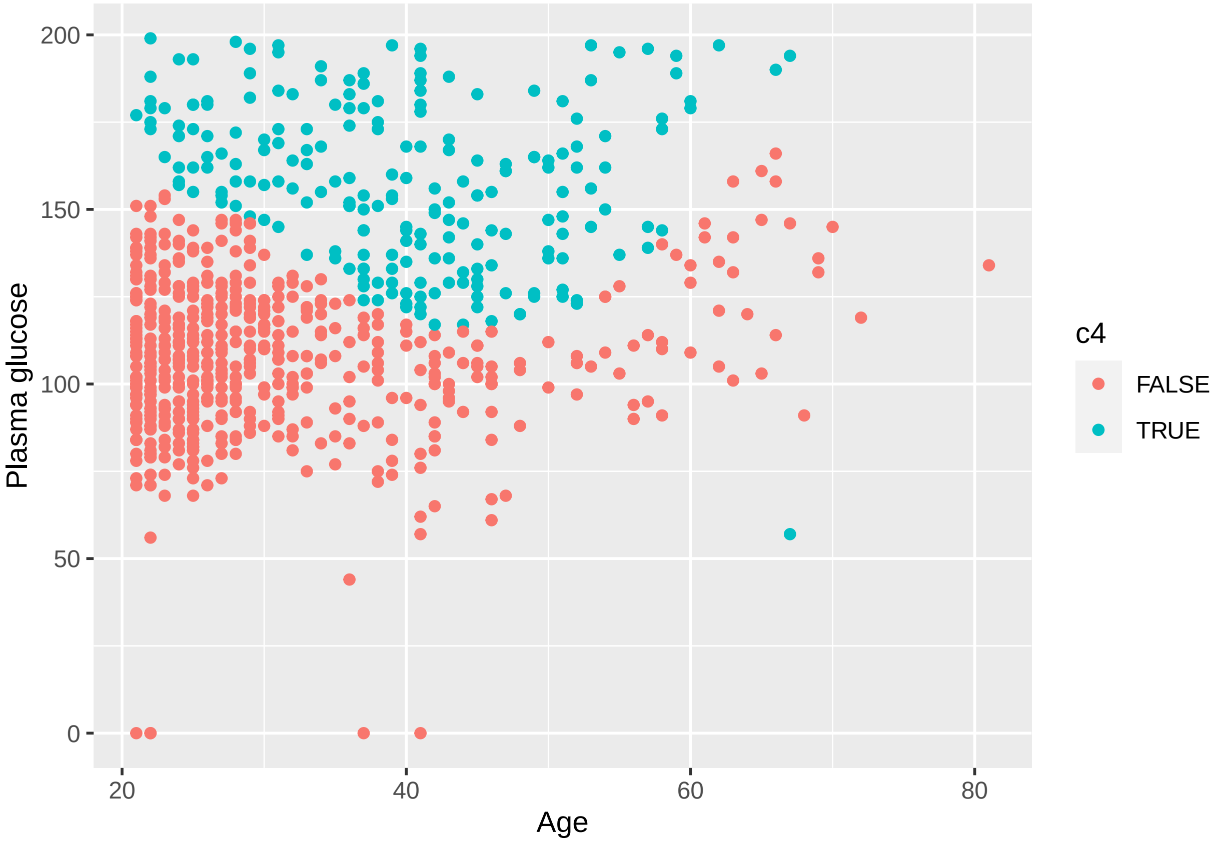
```
##
## Call:
## glm(formula = diabetes ~ plasma + age + z1 + z2 + z3 + z4 + z5,
##     family = binomial(link = "logit"), data = newdata)
##
```

```
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.279e+00  1.129e+00  -8.217  < 2e-16 ***
## plasma       3.772e-02  9.473e-03   3.981 6.85e-05 ***
## age          1.453e-01  2.072e-02   7.014 2.32e-12 ***
## z1           1.266e-08  5.610e-09   2.257  0.02402 *
## z2          -1.760e-07  7.638e-08  -2.304  0.02122 *
## z3           8.424e-07  3.439e-07   2.450  0.01430 *
## z4          -1.682e-06  6.317e-07  -2.662  0.00776 **
## z5           8.045e-07  4.056e-07   1.983  0.04732 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 991.38  on 766  degrees of freedom
## Residual deviance: 741.07  on 759  degrees of freedom
## AIC: 757.07
##
## Number of Fisher Scoring iterations: 5
```

```
c4 <- fitted(newmodel) >= .5

new_missC_error <- sum(data$diabetes != c4) / nrow(data)
cat("Misclassification error:", new_missC_error, "\n")
```

```
## Misclassification error: 0.2464146
```

```
ggplot2::ggplot(df, ggplot2::aes(age, plasma)) +
  ggplot2::geom_point(ggplot2::aes(colour = c4))+
  ggplot2::labs(x = "Age", y="Plasma glucose")
```



The outcome of this analysis is better than the first case, as we obtain a lower misclassification error than before, and also graphically we notice how this version captures what we were saying in the first part of the analysis, how people with higher plasma glucose concentration are more exposed to the risk of having diabete. We also notice how we strangely get a misclassificated predicted value in the bottom right corner of the graph.