Introdução à Aprendizagem Estatística

Projeto de Grupo
Perfil de Especialização em Ciência de Dados – MEI
Docente: Raquel Menezes 2018/2019

Ângela Barros, PG38407
Universidade do Minho

RELATÓRIO FINAL

DIABETES DATASET

2018/2019

Resumo Executivo

O *dataset* a ser trabalhado neste projeto é o de diabetes, um *dataset* que já vem incluído no R, por isso não foi necessário fazer nenhuma importação de ficheiros para o desenvolvimento deste projeto.
Inicialmente o *dataset* era um data.frame com 19 colunas.
Como a variável resposta vinha explicitamente explicada no *dataset* e era possível observá-la, este é claramente um problema supervisionado. Como y é quantitativa inicialmente fiz uma regressão linear e posteriormente adicionei mais uma coluna ao *dataset* para poder fazer regressão logística e abordar este problema como sendo um problema de classificação, pois esta coluna teria dois "níveis" – [0,1] – se fosse 0 o individuo não teria diabetes e se fosse 1 o individuo teria diabetes.
Contudo, inicialmente fiz uma análise exploratória dos dados. Pessoalmente, o que mais me intrigava era as idades dos indivíduos e a distribuição das medidas das suas respetivas cinturas. A meu ver, e dizendo isto de um ponto de vista desinformado sobre o tema, esses pareciam-me ser os preditores mais interessantes a ter em conta. Devido a essa razão, na minha fase exploratória decidi colocar uma nova coluna para poder associar cada individuo a uma categoria de idades.
O *dataset* final passaria a ter, então, 21 colunas.
Tanto ao fazer a regressão linear simples tal como a regressão logística, fui iterando várias vezes o modelo fazendo sempre ajustes de forma a poder encontrar o melhor modelo possível para o meu *dataset*. As minhas iterações estão documentadas no ficheiro em R que envio em anexo.
Depois de ter feito os vários modelos, testei a sua *accuracy* e documentei também no código do projeto. Aliás, durante o desenvolvimento do meu projeto fui tecendo vários comentários no meio do código para uma melhor futura compreensão por parte de quem esteja a ler e a tentar executar o código.

Descrição do problema

A Diabetes é uma doença crónica, onde a quantidade de glicose no sangue é muito elevada porque o pâncreas não produz qualquer insulina ou não a produz em quantidade suficiente. A Organização Mundial de Saúde (OMS) afirma que o problema tem uma dimensão de epidemia mundial. Segundo a OMS, o número de pessoas com diabetes continua a aumentar em todos os países do mundo. Em 2014, atingia 8,5% da população mundial, ou seja, 422 milhões de indivíduos.
O objetivo deste trabalho será analisar um conjunto de dados recolhidos a uma amostra de 403 indivíduos. Primeiramente, será feita uma análise relativamente à população em causa e, posteriormente, uma análise aos dados na tentativa de detetar padrões quanto ao surgimento da doença, analisar qual dos preditores terá maior peso para um diagnóstico positivo de diabetes e fazer predição relativamente a novos casos.

Descrição do conjunto de dados

Quatrocentos e três afro-americanos foram entrevistados num estudo para compreender a prevalência da obesidade, diabetes e outros fatores de risco de doenças cardiovasculares no estado da Virgínia. O conjunto de dados possui 403 registos e é composto por 19 componentes/variáveis.
O conjunto de dados é, portanto, um data.frame de dimensões 403x19.

ID	Subject ID	WEIGHT	Weight in pounds
CHOL	Total Cholesterol	FRAME	a factor with levels [small, medium, large]
STAB.GLU	Stabilized Glucose	BP.1S	First Systolic Blood Pressure
HDL	High Density Lipoprotein	BP.1D	First Diastolic Blood Pressure
RATIO	Cholesterol/HDL Ratio	BP.2S	Second Systolic Blood Pressure
GLYHB	Glycosolated Hemoglobin	BP.2D	Second Diastolic Blood Pressure
LOCATION	County - a factor with levels [Buckingham, Louisa]	WAIST	waist in inches
AGE	Age in years	HIP	Hip in inches
GENDER	a factor with levels [female, male]	TIME.PPN	Postprandial Time (in minutes) when Labs were Drawn
HEIGHT	Height in inches		

É importante referir que, nos detalhes do *dataset* utilizado, quando a *Glycosolated Hemoglobin* é superior a 7 é normalmente visto como um diagnóstico positivo de diabetes.

Análise Exploratória dos Dados

Inicialmente fiz uma análise exploratória dos dados para poder perceber que tipo de dados o dataset possuía.

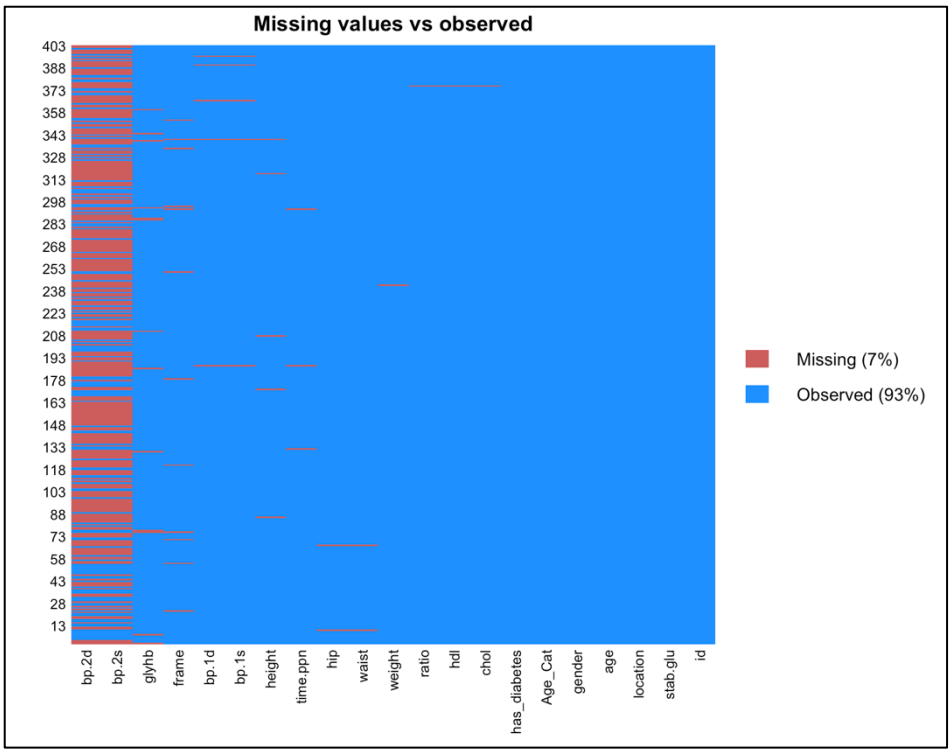

Figura 1 - Valores em falta VS Valores Observados

Como é possível observar na figura 1, o dataset não possui muitos valores em falta. Cerca de 93% do dataset está preenchido, porém, é facilmente observável que a maior incidência de valores em falta ocorro nas variáveis bp.2d e bp.2s.

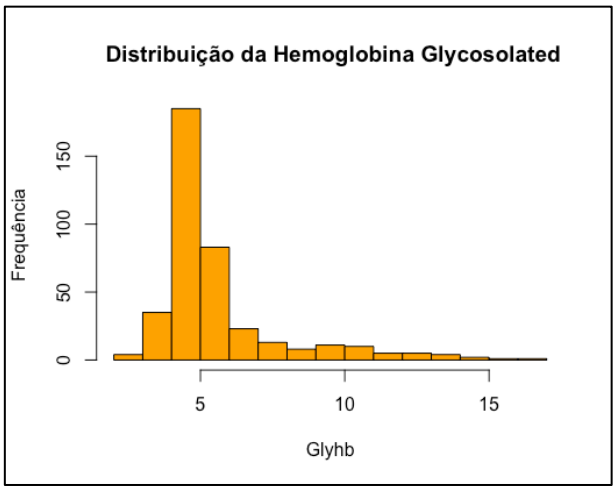

Figura 2 - Distribuição da Hemoglobina Glycosolated

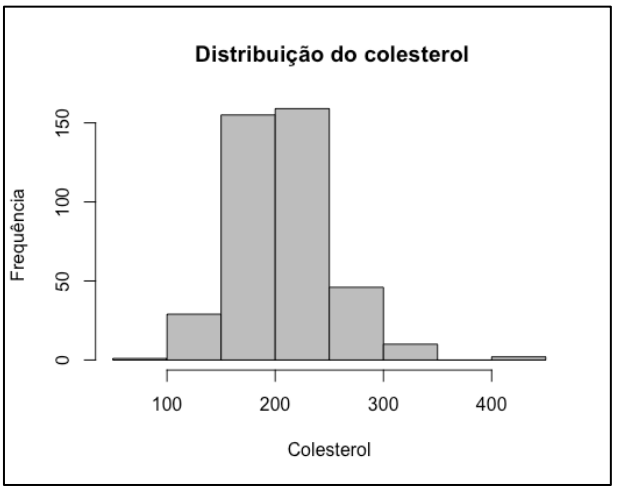

Figura 3 - Distribuição do colesterol

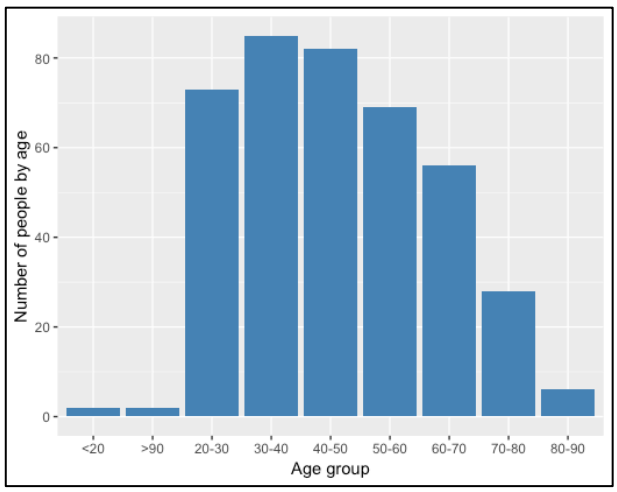

Figura 4 - Distribuição de idades por categorias

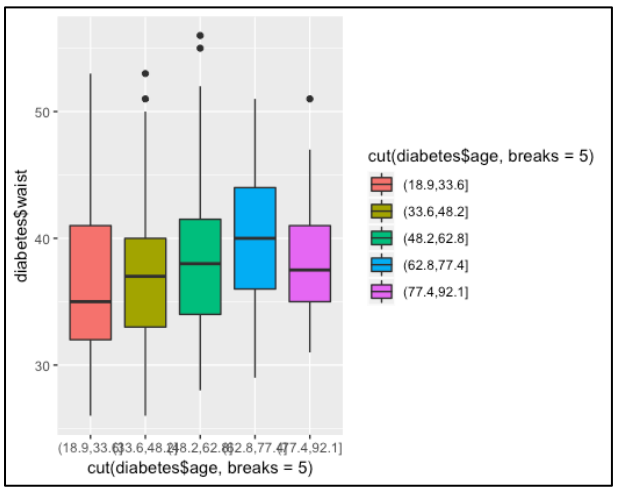

Figura 5 - Relação cintura e idade

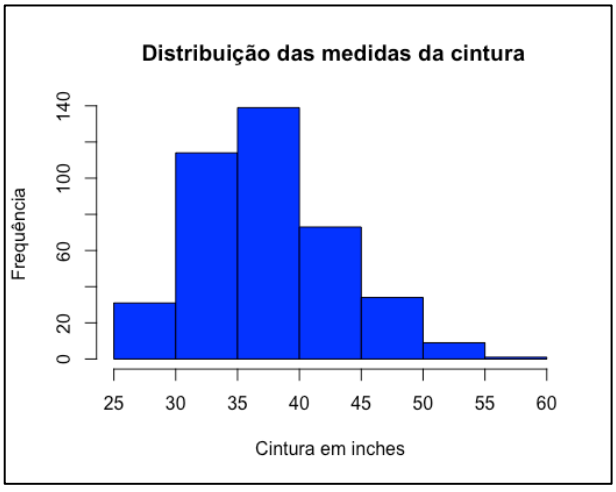

Figura 6 - Distribuição das medidas da cintura

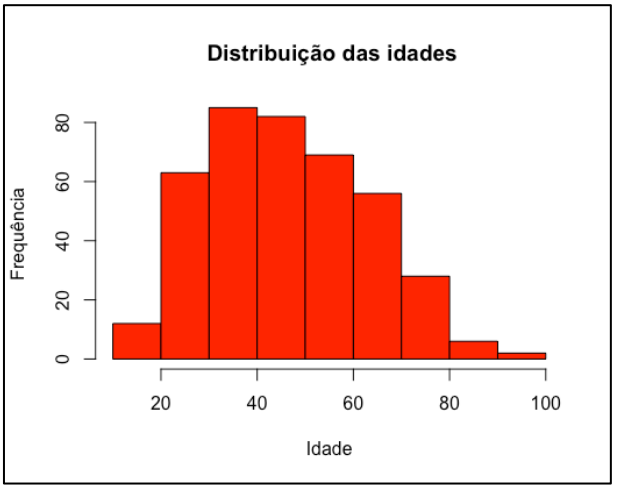

Figura 7 - Distribuição das idades

Após uma análise exploratória dos dados, consegui observar que a maioria dos indivíduos encontrava-se na faixa etária entre os 30 a 40 anos. Observei, também, a distribuição de alguns preditores que julguei mais relevantes e reparei que não possuíam uma distribuição gaussiana. Ao observar a figura 2, notei que havia uma grande prevalência de diagnósticos negativos de diabetes pois a distribuição da Glyhb é mais predominante do lado esquerdo do gráfico, abaixo do valor 7 no eixo do x.
Posteriormente, queria ver a correlação dos preditores para poder eliminar todos aqueles que tivessem correlação forte entre si. A figura abaixo é representativa da correlação existente no *dataset*.

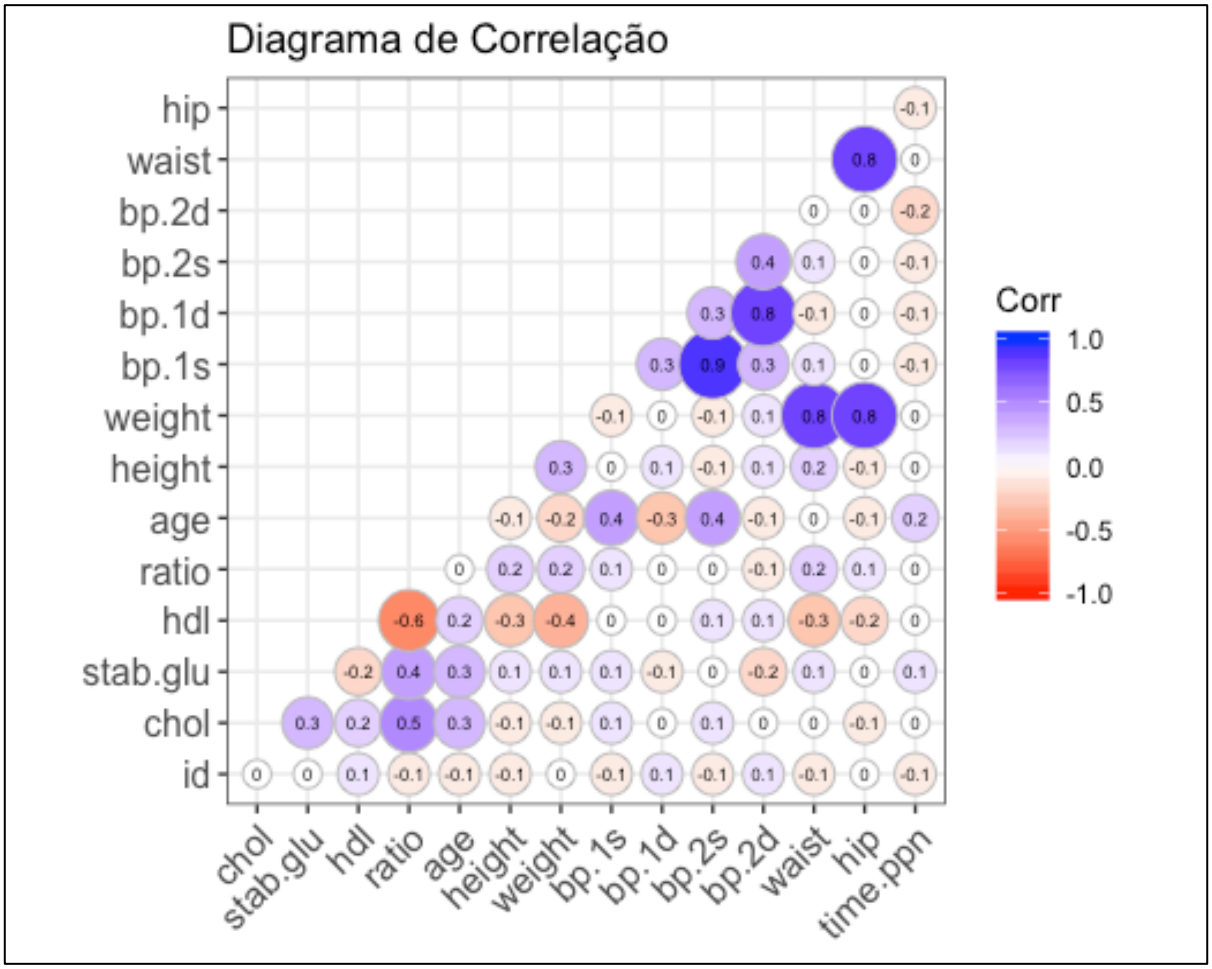

Figura 8 - Diagrama com a correlação dos preditores

Como é possível observar com a imagem anterior, existe algumas relações bastante fortes entre preditores e, para a construção de um melhor modelo é necessário remover estes preditores cujas correlações sejam muito fortes.

Regressão Linear

Como a variável resposta é uma variável quantitativa, o método mais apropriado seria aplicar uma regressão linear:

```
#BIG P-VALUES --> REMOVE PREDICTORS
m2<- lm(diabetes$glyhb ~ diabetes$weight + diabetes$height + diabetes$age + diabetes$chol +
     diabetes$frame + diabetes$hdl + diabetes$hip + diabetes$id + diabetes$location +
     diabetes$bp.1s + diabetes$bp.1d + diabetes$bp.2d + diabetes$bp.2s + diabetes$stab.glu + diabetes$rati
     diabetes$gender + diabetes$time.ppn)
summary(m2) #Adjusted R-squared: 0.7344

m3<- lm(diabetes$glyhb ~ diabetes$weight + diabetes$age + diabetes$chol +
     diabetes$frame + diabetes$hdl + diabetes$hip + diabetes$id + diabetes$location +
     diabetes$bp.1s + diabetes$bp.1d + diabetes$bp.2d + diabetes$bp.2s + diabetes$stab.glu + diabetes$rati
     diabetes$gender + diabetes$time.ppn)
summary(m3) #Adjusted R-squared: 0.7367

m4<- lm(diabetes$glyhb ~ diabetes$weight + diabetes$age + diabetes$chol +
     diabetes$frame + diabetes$hdl + diabetes$hip + diabetes$location +
     diabetes$bp.1s + diabetes$bp.1d + diabetes$bp.2d + diabetes$bp.2s + diabetes$stab.glu + diabetes$rati
     diabetes$gender + diabetes$time.ppn)
summary(m4) #Adjusted R-squared: 0.7391

m5<- lm(diabetes$glyhb ~ diabetes$weight + diabetes$age + diabetes$chol +
     diabetes$frame + diabetes$hip + diabetes$location +
     diabetes$bp.1s + diabetes$bp.1d + diabetes$bp.2d + diabetes$bp.2s + diabetes$stab.glu + diabetes$rati
     diabetes$gender + diabetes$time.ppn)
summary(m5) #Adjusted R-squared: 0.7408
```

Figura 9 - Vários modelos de regressão linear simples

Fui sistematicamente removendo preditores, conforme os p-values fossem elevados. O meu objetivo era encontrar um modelo que tivesse um *Adjusted R-squared* alto. Após várias iterações, cheguei á conclusão que o meu modelo inicial era aquele que explicava melhor o comportamento do *dataset* pois foi o valor mais elevado que consegui obter no *Adjusted R-squared.* E foi o modelo que usava todas as variáveis do *dataset*.

```
m1<- lm(diabetes$glyhb~. , data = diabetes)
summary(m1) #Adjusted R-squared: 0.849
```

Figura 10 - Melhor modelo regressão linear

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.41852 -0.51747 -0.04607  0.55231  2.92563

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.979e+00  3.174e+00   0.938   0.3502
id             -1.028e-07  7.985e-06  -0.013   0.9898
chol           -3.230e-05  4.377e-03  -0.007   0.9941
stab.glu        1.644e-02  2.306e-03   7.127 1.45e-10 ***
hdl             9.297e-03  1.198e-02   0.776   0.4395
ratio           2.072e-01  1.221e-01   1.697   0.0928 .
locationLouisa -4.589e-01  2.295e-01  -2.000   0.0481 *
age            -4.033e-02  3.219e-02  -1.253   0.2130
genderfemale   -2.674e-01  3.160e-01  -0.846   0.3995
height         -1.309e-02  3.876e-02  -0.338   0.7363
weight          6.416e-03  6.304e-03   1.018   0.3112
framemedium     1.519e-01  2.659e-01   0.571   0.5691
framelarge     -5.038e-01  3.323e-01  -1.516   0.1326
bp.1s           1.982e-03  1.005e-02   0.197   0.8440
bp.1d          -7.621e-03  1.457e-02  -0.523   0.6019
bp.2s          -5.038e-03  9.890e-03  -0.509   0.6116
bp.2d           1.413e-02  1.481e-02   0.954   0.3423
waist          -8.855e-03  3.525e-02  -0.251   0.8022
hip            -3.401e-03  4.394e-02  -0.077   0.9385
time.ppn        3.535e-04  3.669e-04   0.963   0.3376
Age_Cat30-40    4.300e-01  5.012e-01   0.858   0.3929
Age_Cat40-50    1.193e+00  7.331e-01   1.627   0.1068
Age_Cat50-60    1.875e+00  1.036e+00   1.811   0.0730 .
Age_Cat60-70    2.256e+00  1.293e+00   1.745   0.0840 .
Age_Cat70-80    3.176e+00  1.717e+00   1.850   0.0672 .
Age_Cat80-90    2.986e+00  2.067e+00   1.445   0.1515
has_diabetes    2.901e+00  3.199e-01   9.068 8.71e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9857 on 103 degrees of freedom
  (273 observations deleted due to missingness)
Multiple R-squared:  0.8794,    Adjusted R-squared:  0.849
F-statistic: 28.89 on 26 and 103 DF,  p-value: < 2.2e-16
```

Figura 11 - Summary do modelo

Regressão Logística - Classificação

De forma a poder explorar a regressão logística, criei uma nova coluna "has_diabetes" e adicionei-a ao dataset. Ou seja, neste momento, o dataset possui 21 colunas.

```
glm.fit=glm(diabetes$has_diabetes~chol+stab.glu+hdl+age+height+weight+frame+
              bp.1s+bp.1d+bp.2s+bp.2d+waist+hip, data=diabetes, family=binomial)
summary(glm.fit) #AIC: 79.981

glm.fit=glm(diabetes$has_diabetes~chol+stab.glu+hdl+age+gender+height+weight+
              frame+bp.1s+bp.1d+bp.2s+bp.2d+waist+hip, data=diabetes, family=binomial)
summary(glm.fit) #AIC: 79.67

glm.fit=glm(diabetes$has_diabetes~chol+stab.glu+hdl+height+weight+frame+bp.1s+
              bp.1d+bp.2s+bp.2d+waist+hip, data=diabetes, family=binomial)
summary(glm.fit) #AIC: 79.614

glm.fit=glm(diabetes$has_diabetes~chol+stab.glu+hdl+height+weight+bp.1s+
              bp.1d+bp.2s+bp.2d+waist+hip, data=diabetes, family=binomial)
summary(glm.fit) #AIC: 77.964

glm.fit=glm(diabetes$has_diabetes~chol+stab.glu+height+weight+bp.1s+
              bp.1d+bp.2s+bp.2d+hip, data=diabetes, family=binomial)
summary(glm.fit) #AIC 74.084

glm.fit=glm(diabetes$has_diabetes~chol+stab.glu+height+weight+bp.1s+bp.1d+
              bp.2d+hip, data=diabetes, family=binomial)
summary(glm.fit) #AIC 72.468

glm.fit=glm(diabetes$has_diabetes~chol+stab.glu+height+weight+bp.1d+
              bp.2d+hip, data=diabetes, family=binomial)
summary(glm.fit) #AIC 70.884    MELHOR
```

Figura 12 - Iterações do modelo de regressão logística

O objetivo seria encontrar um modelo que obtivesse um AIC mais baixo possível. Para isso, eu repeti várias vezes o processo de selecionar/remover preditores e verificar qual o AIC que me era retornado. Após várias iterações o melhor AIC que consegui foi 70.884 e nessa altura eu decidi não continuar a remover mais preditores pois o meu modelo já só possuía 7 preditores.

```
Call:
glm(formula = diabetes$has_diabetes ~ chol + stab.glu + height +
    weight + bp.1d + bp.2d + hip, family = binomial, data = diabetes)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.45396  -0.33196  -0.15159  -0.05556   2.48662

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.87507    9.57299   0.091   0.9272
chol         0.02225    0.01015   2.192   0.0284 *
stab.glu     0.08496    0.02148   3.955 7.65e-05 ***
height      -0.21524    0.13153  -1.636   0.1017
weight       0.03392    0.02139   1.586   0.1127
bp.1d       -0.07585    0.04989  -1.520   0.1285
bp.2d        0.08002    0.05608   1.427   0.1536
hip         -0.21197    0.13114  -1.616   0.1060
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 148.039  on 139  degrees of freedom
Residual deviance:  54.884  on 132  degrees of freedom
  (263 observations deleted due to missingness)
AIC: 70.884

Number of Fisher Scoring iterations: 8
```

Figura 13 - Summary do melhor modelo

Pude analisar, que a altura tem uma relação negativa com a probabilidade de o individuo ter diabetes. Outro preditor com igual relação negativa é, surpreendentemente, a medida das ancas. Ou seja, por cada aumento unitário das ancas a probabilidade de o paciente ter diabetes diminui cerca de 20%.

Outro facto curioso com o qual me deparei foi constatar que neste modelo, existe apenas 139 graus de liberdade, que o modelo "rejeitou" 263 registos. Ou seja, n=140.

```
#DIVISÃO DADOS DE TREINO -- DADOS DE TESTE
n = 201
nr = nrow(diabetes)
train_data_aux = split(diabetes, rep(1:ceiling(nr/n), each=n, length.out=nr)) #DIVISÃO DE METADE DOS DADOS
train_data = train_data_aux$`1` #PRIMEIROS 201 REGISTOS
test_data = train_data_aux$`2` #RESTANTES
```

Figura 14 - Divisão dos dados

Um dos passos seguintes foi dividir o *dataset* em duas partes: dados de treino e dados de teste. Como o *dataset* é impar e constituído por 403 registos, dividi em duas pequenas data.frames em 201 registos cada.

```
#TODOS OS DADOS
glm_all.probs_all_data=predict(glm.fit, type="response")
glm_all.probs_all_data
glm_all.pred=rep(0,403)
glm_all.pred[glm_all.probs_all_data>0.5]=1
table(glm_all.pred, diabetes$has_diabetes)
```

Figura 15 - Prediction com todos os dados

A seguir, peguei em todos os dados, e adicionei uma coluna extra "*pred*" que inicialmente era apenas uma coluna preenchida com zeros. Posteriormente, inseri um *threshold* em que se algum valor fosse superior a 0.5, então esse registo seria 1.

```
> table(glm_all.pred, diabetes$has_diabetes)

glm_all.pred   0   1
           0 288  45
           1  55  15
>
```

Figura 16 - Tabela de confusão

A tabela de confusão que mostra 288 casos negativos de diabetes estavam corretos e que 45 estavam errados. Ou seja, 45 casos de falsos negativos.
Por sua vez, também mostra que 55 casos foram falsos positivos e que 15 casos estavam corretos.

```
#DADOS DE TESTE
glm.probs_teste=predict(glm.fit, test_data, type="response")
glm.probs_teste
glm.pred=rep(0,201)
glm.pred[glm.probs_teste>0.5]=1
table(glm.pred, test_data$has_diabetes)
```

Figura 17 - Dados de teste

A seguir, peguei no dados de teste que tinham sido previamente divididos, e repeti o processo. Adicionei uma variável "*pred*", defini um *threshold* e coloquei todos os registos que estavam a cima desse *threshold* como sendo 1 – ou seja, diagnóstico positivo de diabetes.

```
> table(glm.pred, test_data$has_diabetes)

glm.pred   0   1
       0 173  14
       1   2  12
```

Figura 18 - Tabela de confusão dos dados de teste

A tabela de confusão que mostra 173 casos negativos de diabetes estavam corretos e que 14 estavam errados. Ou seja, 14 casos de falsos negativos.
Por sua vez, também mostra que 2 casos foram falsos positivos e que 12 casos estavam corretos.

Posteriormente, repeti os passos acima descritos nas figuras 17 e 18 para os dados de treino.

Por fim, por sugestão da docente Raquel, "limpei" os dados certificando-me que não existia nenhum NA nos valores usados para o modelo. Essa parte encontra-se na parte final do código, onde também vejo a percentagem de *accuracy* tendo em conta os dados "limpos" que estão a ser usados.

Conclusões e trabalho futuro

Apesar de o conjunto de dados parecer ser simples, julgo que consegui trabalhar o problema de forma a explorar bem grande parte dos conhecimentos adquiridos na unidade curricular. Sem tentar cair no cliché, confesso que gostei imenso do trabalho, sinto que assimilei vários conhecimentos e este trabalho permitiu-me explorar e apreciar o mundo da Estatística e Data Science. Talvez pelo facto de ter sido só um elemento a fazer o trabalho de grupo, não consegui explorar a parte final da matéria lecionada por grande pena minha. Gostava de ter abordado o KNN, Bootstrap e Cross-validation. Julgo que se tivesse um bocado mais de tempo ou quiçá um elemento extra no grupo que talvez tivesse conseguido chegar até essa parte e, assim dessa forma, cobrir realmente toda a matéria lecionada na aula. Portanto, como trabalho futuro julgo que fosse pertinente explorar o KNN e por fim, fazer cross-validation para ver qual modelo se aplicaria melhor ao problema. Também seria interessante abordar o LDA e/ou QDA, aliás, cheguei a abordar essa questão com a docente, mas ambas chegamos á conclusão que para esta versão inicial do trabalho e com o prazo pré-estabelecido que não seria necessário nem oportuno explorar LDA. Para isso, exigiria reformular o meu problema e meter uma outra variável categórica, mas desta vez com mais do que dois "níveis", mas revelou ser algo supérfluo para esta altura. Contudo, gostaria de poder ter explorado isso juntamente com os temas referidos previamente.