

Regressão logística - Estudo da diabetes

Análise de Clusters

ANA SOFIA FERREIRA PG38356

CAROLINA SILVA PG38335

CÉLIA FIGUEIREDO PG41022

MÁRCIA COSTA A67672

SAMUEL COSTA PG38352

Conteúdos

Análise do *dataset*

Tratamento dos dados

Análise Exploratória

Construção do modelo

Modelo final

- Interpretação dos coeficientes obtidos
- *Accuracy* do modelo
- Avaliação da capacidade preditiva do modelo

Conclusão

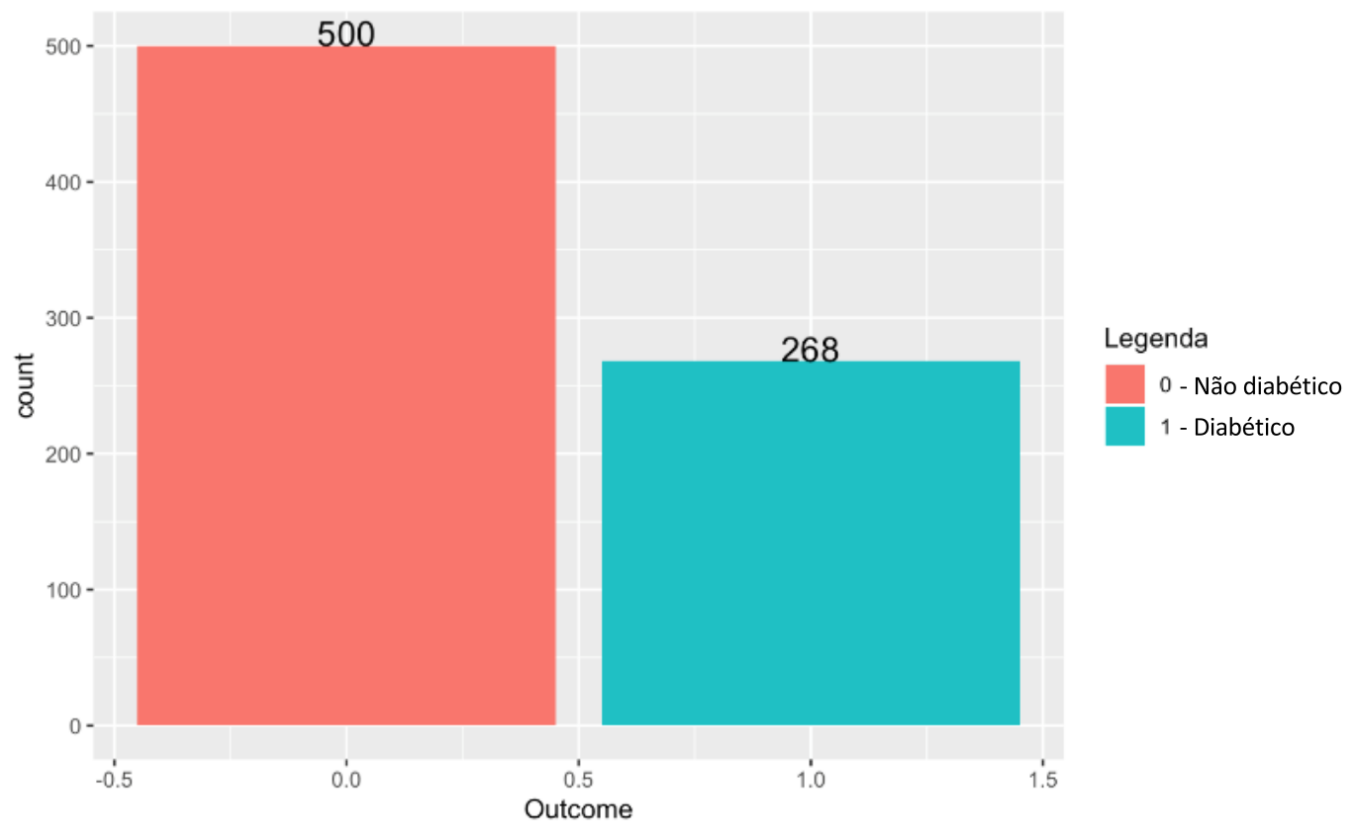


Prediction of Diabetes in PIMA Women

Dataset utilizado

Variáveis	Descrição
Pregnancies	Número de vezes que engravidou
Glucose	Concentração de glicose no plasma durante 2 horas presente num teste de tolerância à glicose por via oral
BloodPressure	Pressão sanguínea diastólica (mm Hg)
SkinThickness	Espessura da dobra da pele do tríceps (mm)
Insulin	Insulina sérica de 2 horas (mg/dl)
BMI	índice de massa corporal (kg/m ²)
DiabetesPedigreeFuction	Função de hereditariedade do diabetes (uma função que pontua a probabilidade de diabetes com base no histórico familiar)
Age	Idade (anos)
Outcome	Variável de classe (0 se não diabético, 1 se diabético)

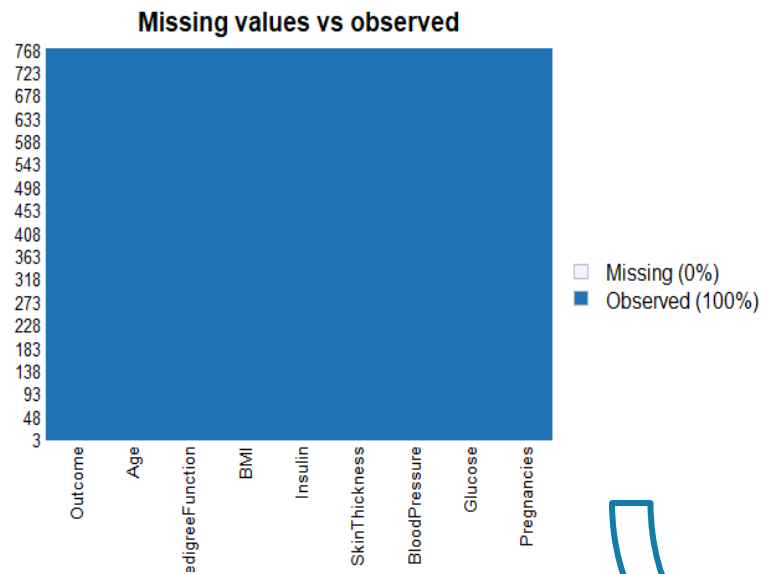
Distribuição da variável "Outcome"



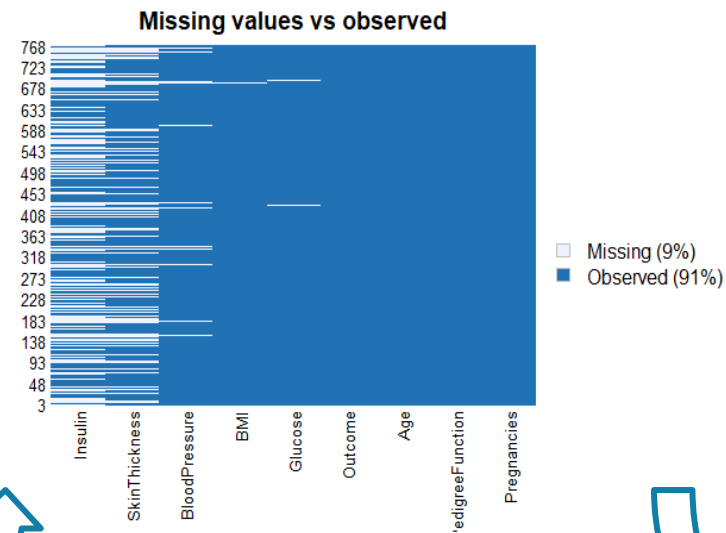
- Verificaram-se mais casos de ausência da doença
- Os casos de ocorrência da doença são, aproximadamente, metade dos casos de não ocorrência

Tratamento dos dados

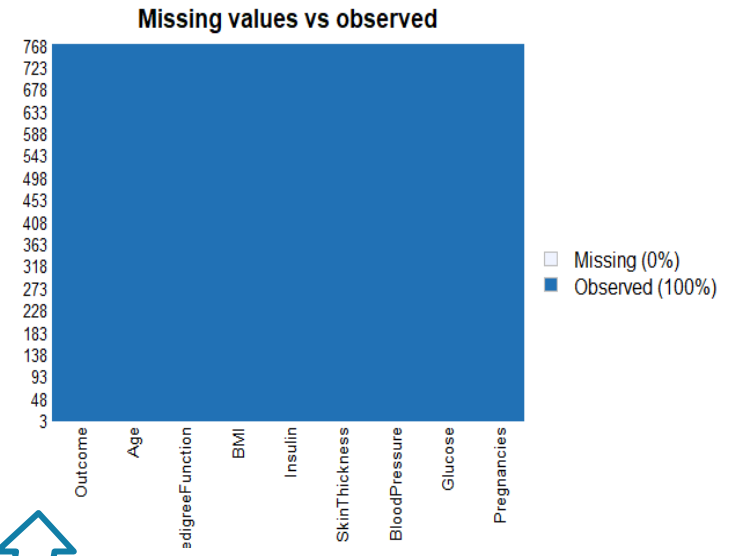
Antes do tratamento



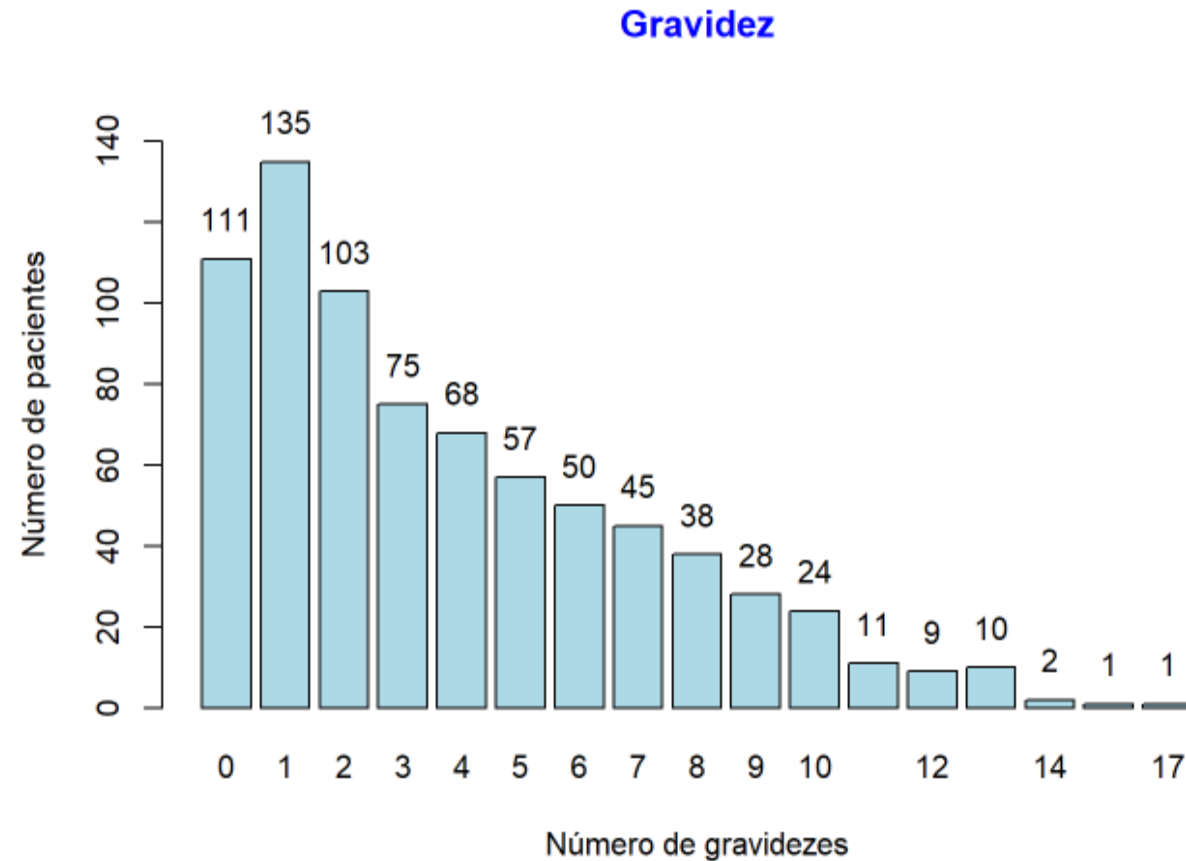
Durante o tratamento



Após o tratamento



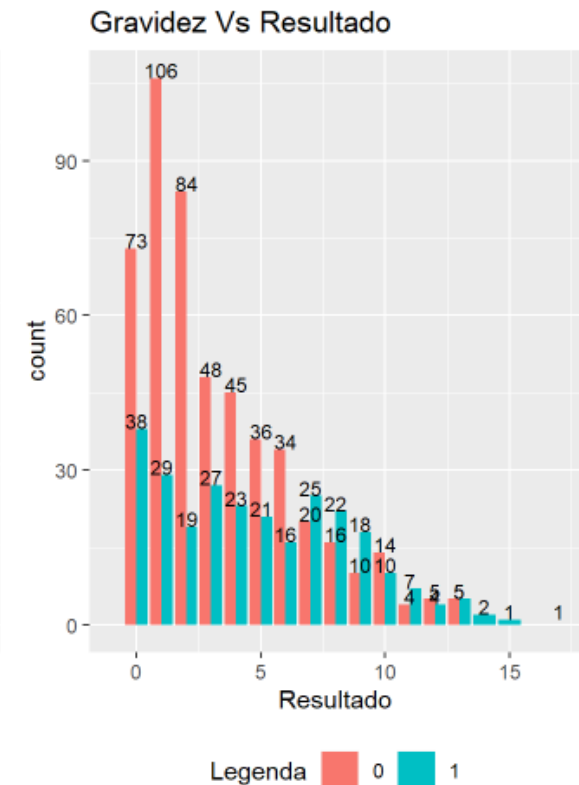
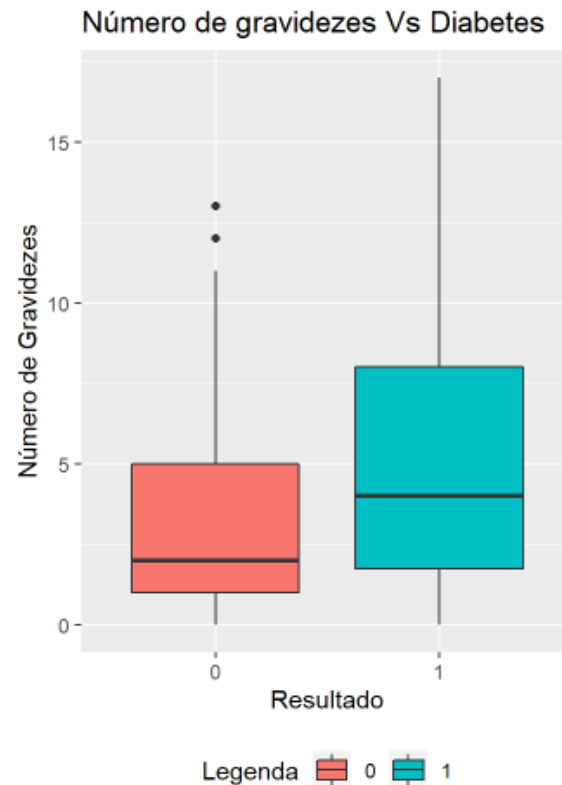
Análise da variável Gravidez



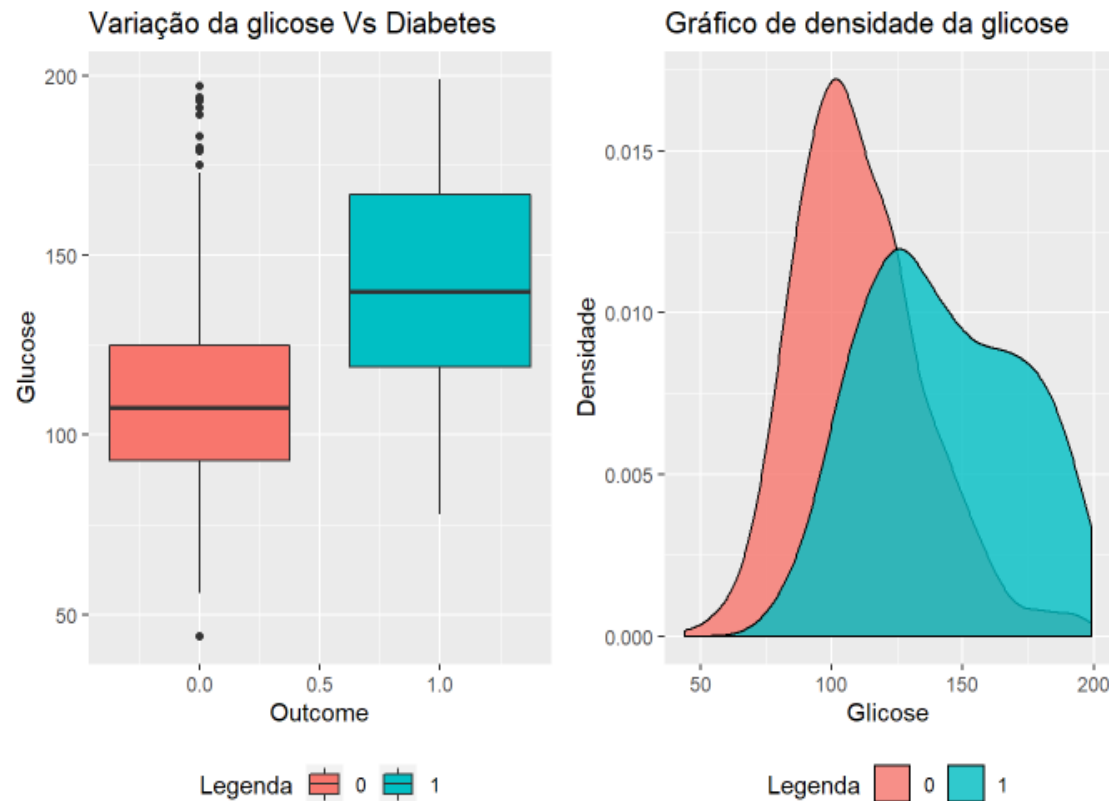
- Valor mais recorrente: 1 vez grávida
- Quanto mais vezes a ocorrência de gravidez menos pacientes verificados

Gravidez vs Diabetes

- Proporção de pacientes com diabetes aumenta à medida que o número de gravidezes aumenta.
- Existe um número muito elevado de pacientes sem diabetes que tiveram um número reduzido de gravidezes ou nenhuma.



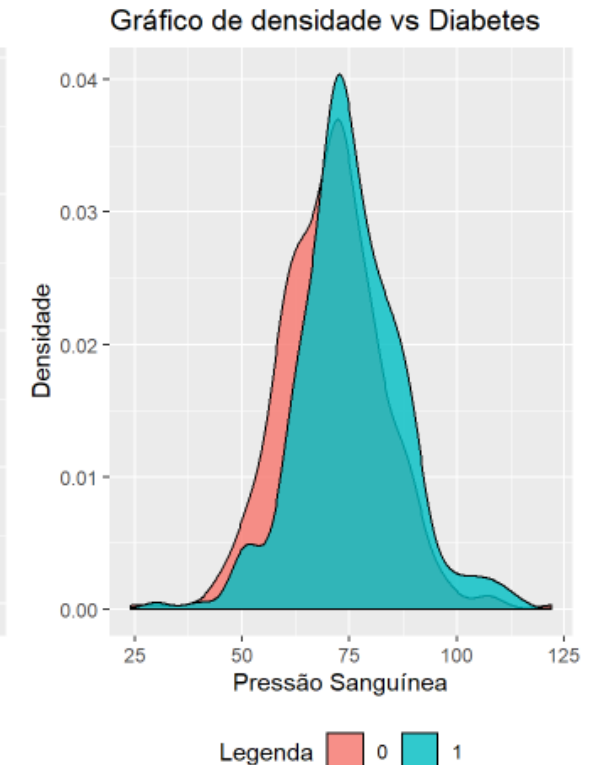
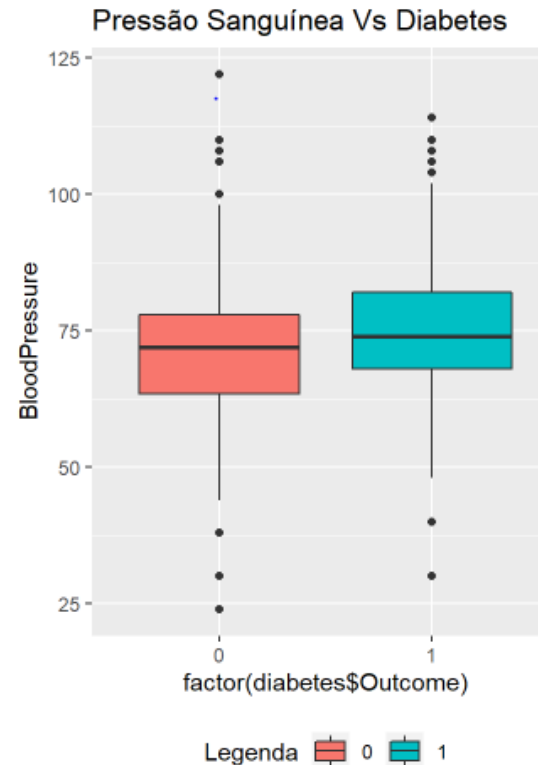
Glicose vs Diabetes



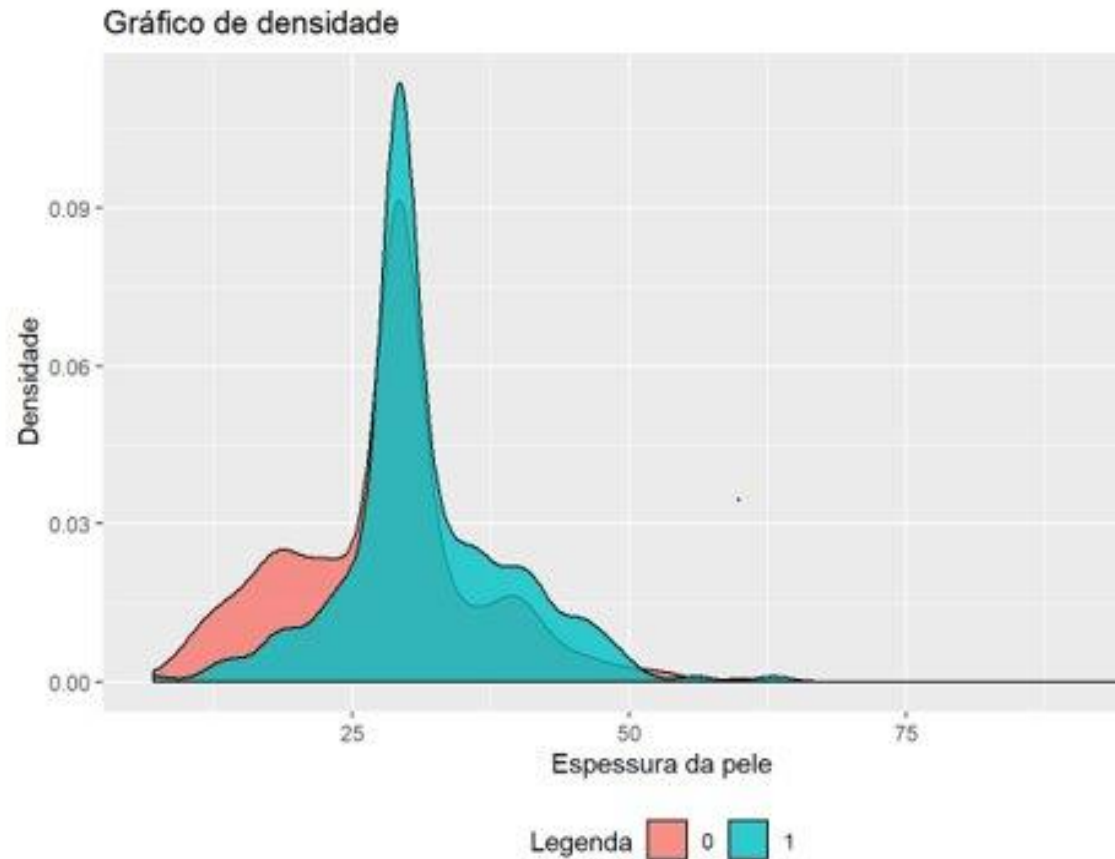
- Pessoas com diabetes apresentam no sangue uma maior quantidade de glicose.

Pressão sanguínea vs Diabetes

- Em média, a pressão sanguínea de uma mulher com diabetes é muito similar à de uma mulher sem diabetes.
- Como tal, suspeita-se que a pressão sanguínea poderá ser uma variável a retirar do modelo.



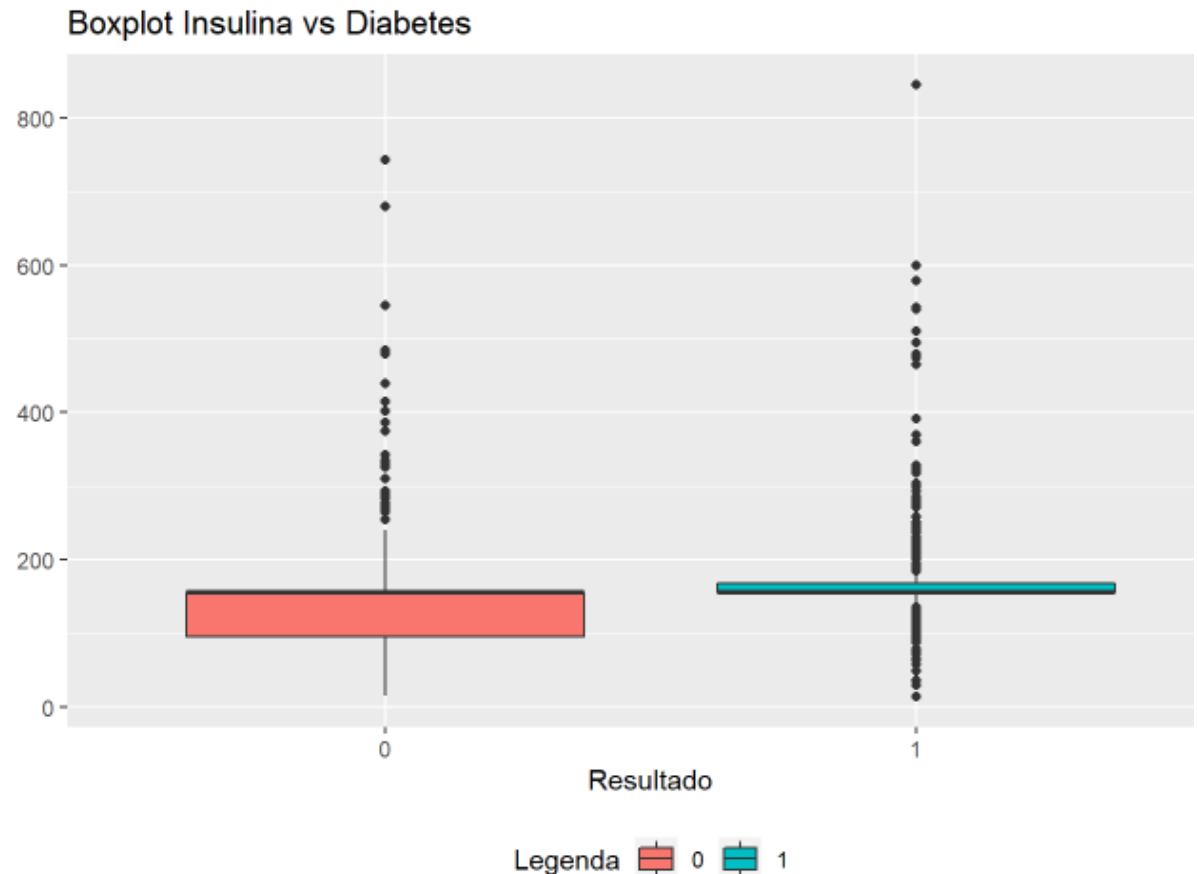
Espessura da pele vs Diabetes



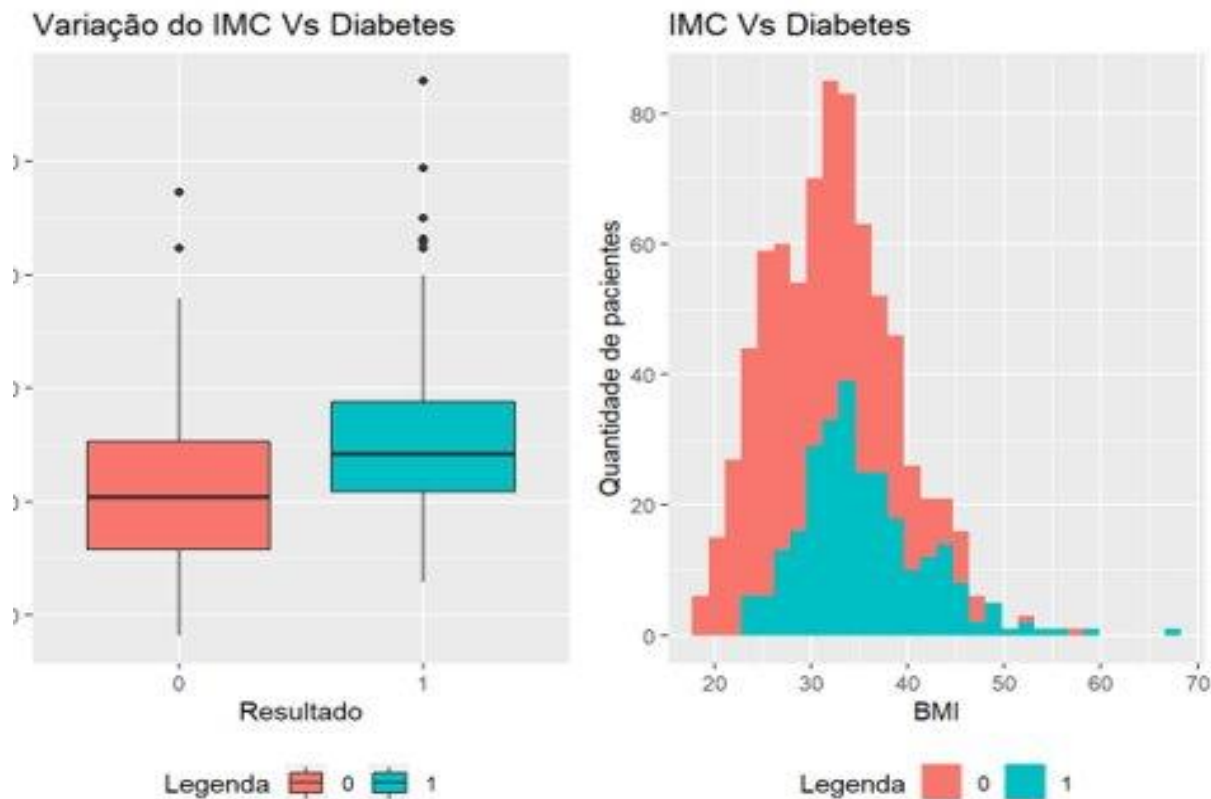
- Pacientes com maiores valores de espessura da pele possuem uma maior probabilidade de ter diabetes.
- Em contrapartida, menores valores de espessura da pele possuem uma densidade superior de mulheres sem diabetes, comparativamente às mulheres que possuem.

Insulina vs Diabetes

- Percebe-se que em mulheres sem diabetes existe uma maior concentração de registos com quantidade de insulina verificada abaixo de, aproximadamente, 155.
- No caso de mulheres com diabetes a concentração é maior acima desse mesmo número.



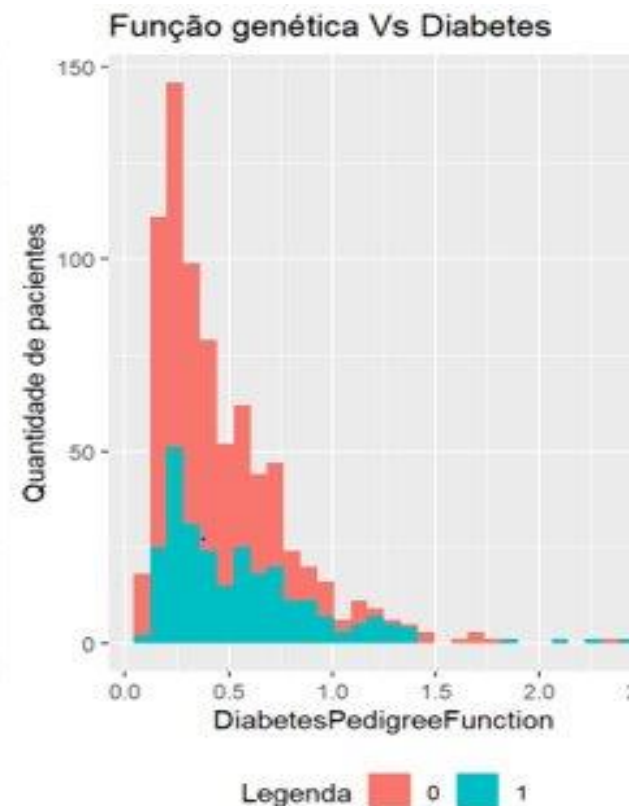
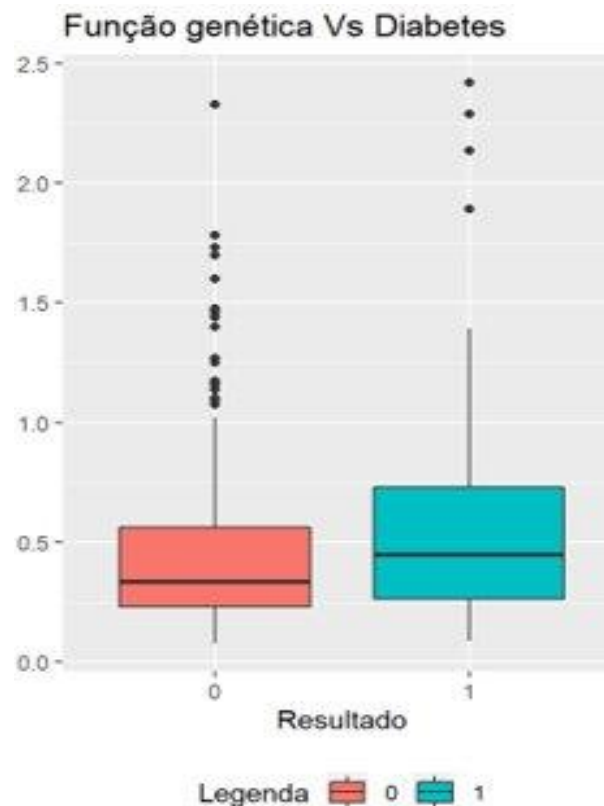
IMC vs Diabetes



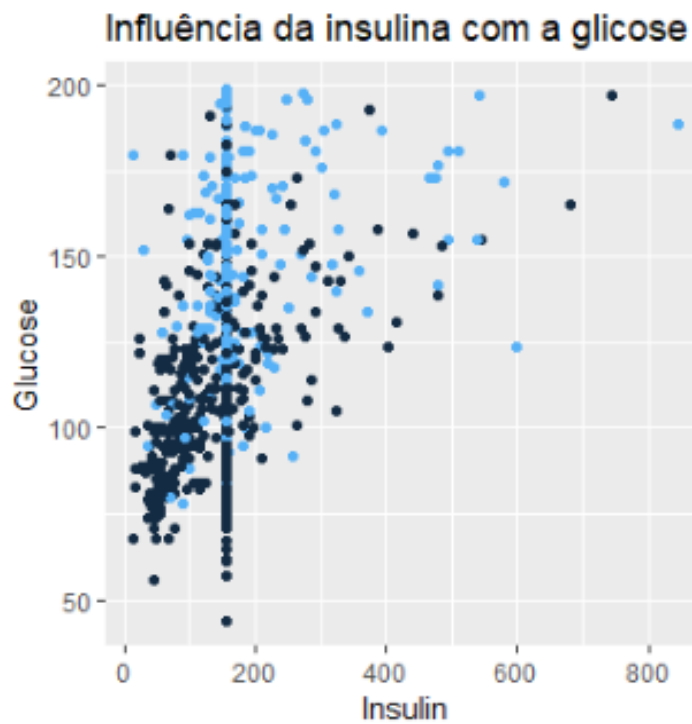
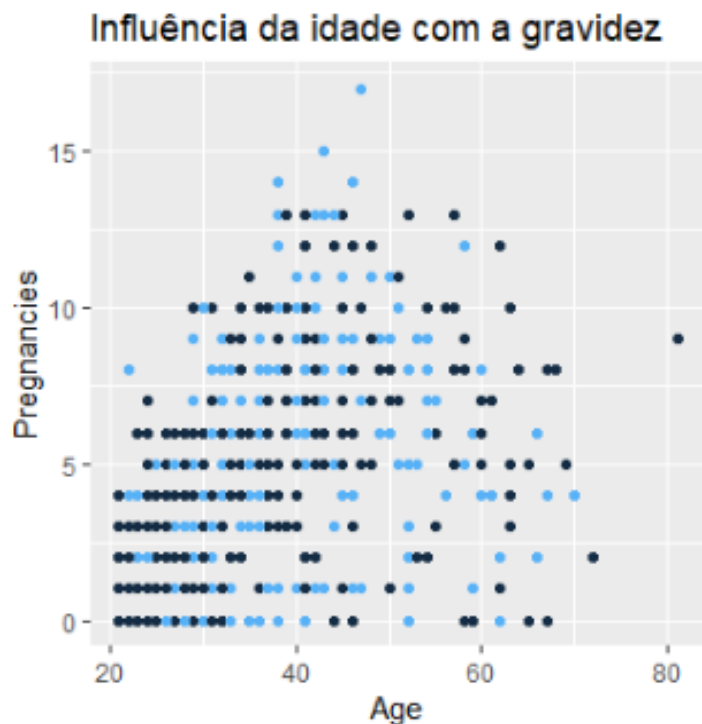
- Mulheres com diabetes, em média, têm um maior índice de massa corporal.

Função genética vs Diabetes

- A possibilidade de uma pessoa ter diabetes, transmitida hereditariamente, é maior nas mulheres com diabetes, efetivamente.



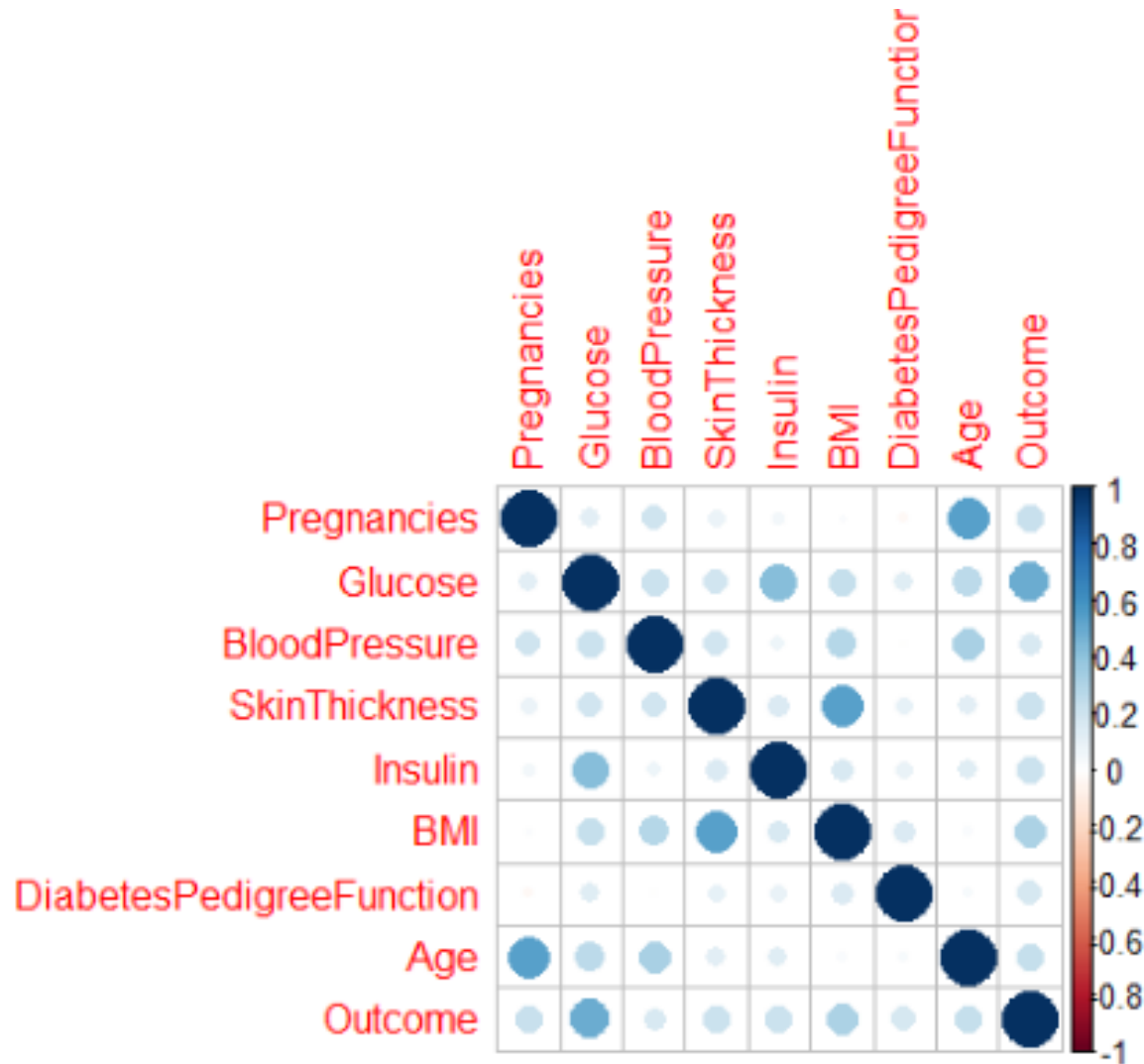
Idade vs Gravidez | Insulina vs Glicose



● Não ter diabetes ● Ter diabetes

- Há muitas ocorrências de aproximadamente 155. Deve-se ao facto de os *missing values* terem sido substituídos pela média.
- Níveis de glicose mais baixos dizem respeito a ocorrências menos frequentes de diabetes.

Matriz de correlação



Variáveis mais correlacionadas:

- Pregnancies e Age
- BMI e SkinThickness
- Insulin e Glucose

Variáveis que mais se relacionam com a resposta:

- Glucose
- BMI

Construção do modelo

Dados para treino: 70%
Dados para teste: 30%

```
Call:
glm(formula = Outcome ~ ., family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5957  -0.7006  -0.3841   0.6853   2.3762

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.6526443   0.9893468  -9.757  < 2e-16 ***
Pregnancies    0.1185184   0.0387091   3.062  0.0022 **
Glucose        0.0375306   0.0047726   7.864 3.73e-15 ***
BloodPressure -0.0082741   0.0106001  -0.781  0.4351
SkinThickness  0.0180122   0.0161984   1.112  0.2661
Insulin        0.0005201   0.0015363   0.339  0.7350
BMI            0.0921209   0.0218500   4.216 2.49e-05 ***
DiabetesPedigreeFunction 0.8526290   0.3587499   2.377  0.0175 *
Age           0.0127746   0.0114296   1.118  0.2637
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 696.28  on 537  degrees of freedom
Residual deviance: 487.84  on 529  degrees of freedom
AIC: 505.84


Number of Fisher Scoring iterations: 5
```

Regressão gradual

- Adicionar e remover iterativamente preditores ao modelo.
- Através da função `stepAIC` testamos os 3 métodos: backward, forward e both

AIC "both"	501.0397
AIC "backward"	501.0397
AIC "forward"	505.8351

Método *backward*
escolhido



Número de variáveis

`nvmax`
`<int>`

4

Modelo Final

(Intercept)	Pregnancies	Glucose	BMI	DiabetesPedigreeFunction
-9.62124129	0.13877111	0.03852866	0.09838685	0.87090498

- Coeficientes positivos aumentam a probabilidade de ter a diabetes
- Coeficientes negativos diminuem a probabilidade de ter a diabetes

Accuracy = 0.75

Modelo de Regressão Logística:

$$y = -9.62124129 + 0.13877111\text{pregnancies} + 0.03852866\text{Glucose} + 0.09838685\text{BMI} + 0.87090498\text{DiabetesPedigreeFunction}$$

Tabela da análise do desvio

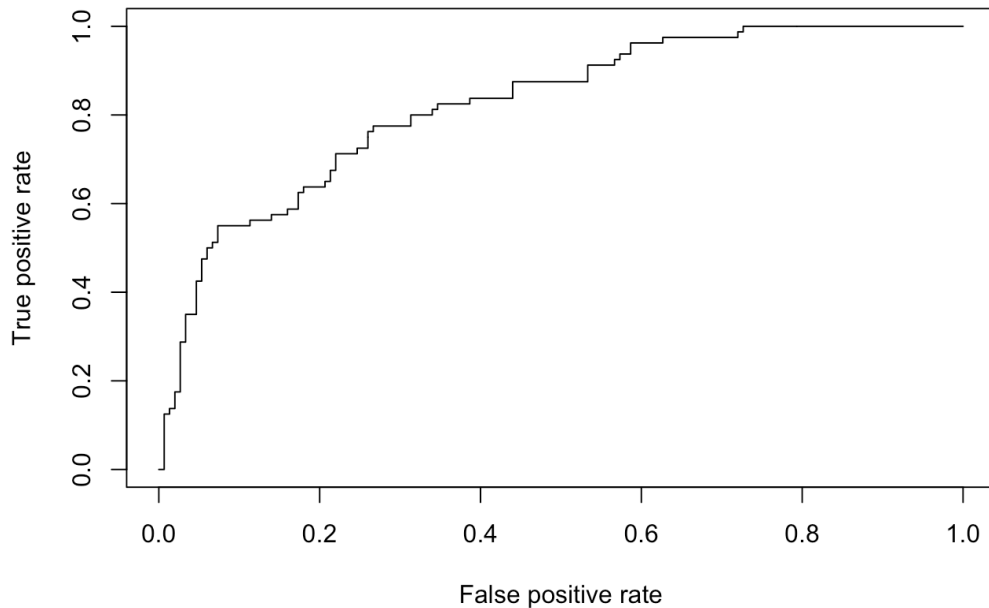
	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			537	696.28		
Pregnancies	1	29.047	536	667.23	7.065e-08	***
Glucose	1	134.568	535	532.67	< 2.2e-16	***
BMI	1	35.684	534	496.98	2.321e-09	***
DiabetesPedigreeFunction	1	5.943	533	491.04	0.01478	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

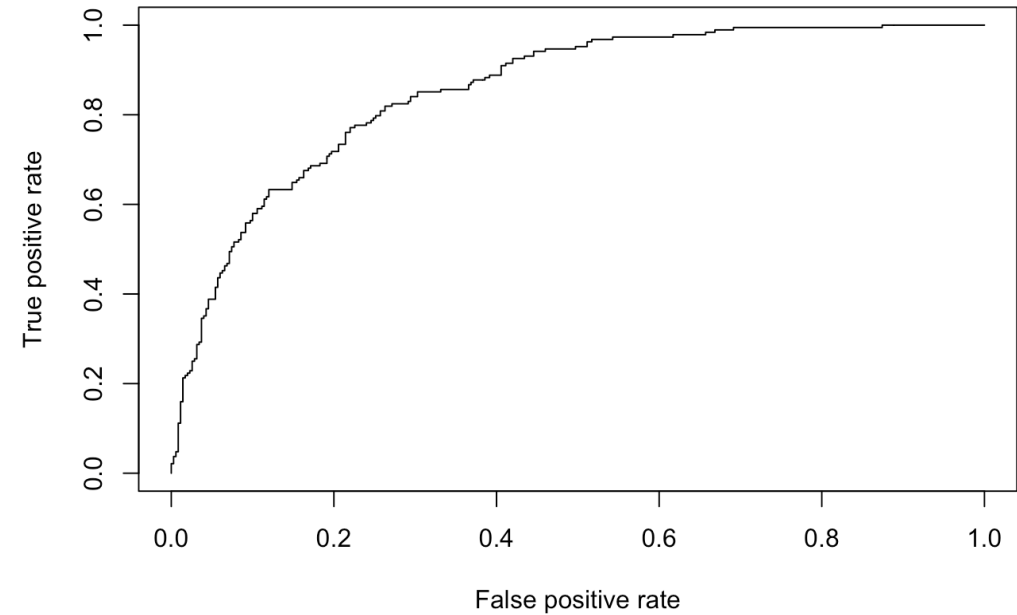
- Quanto maior a diferença entre o desvio nulo e o desvio residual, melhor o modelo.
- Sempre que se adiciona uma variável há um decréscimo no desvio residual.

Avaliação da capacidade preditiva do modelo

Curvas ROC



- Área abaixo da curva ROC - auc
- 0.8230833



- Validação interna - auc_i
- 0.8537538

Significado do modelo

- **Glucose**

`exp(0.03852866*10)`



OR = 1.470036

Um aumento de 10 mg de glicose no sangue faz com que uma mulher tenha cerca de 1.47 vezes mais possibilidade de ter diabetes.

- **BMI**

`exp(0.09838685*5)`



OR = 1.635477

Um aumento de 5 unidades no IMC faz com que uma mulher tenha cerca de 1.64 vezes mais possibilidade de ter diabetes.

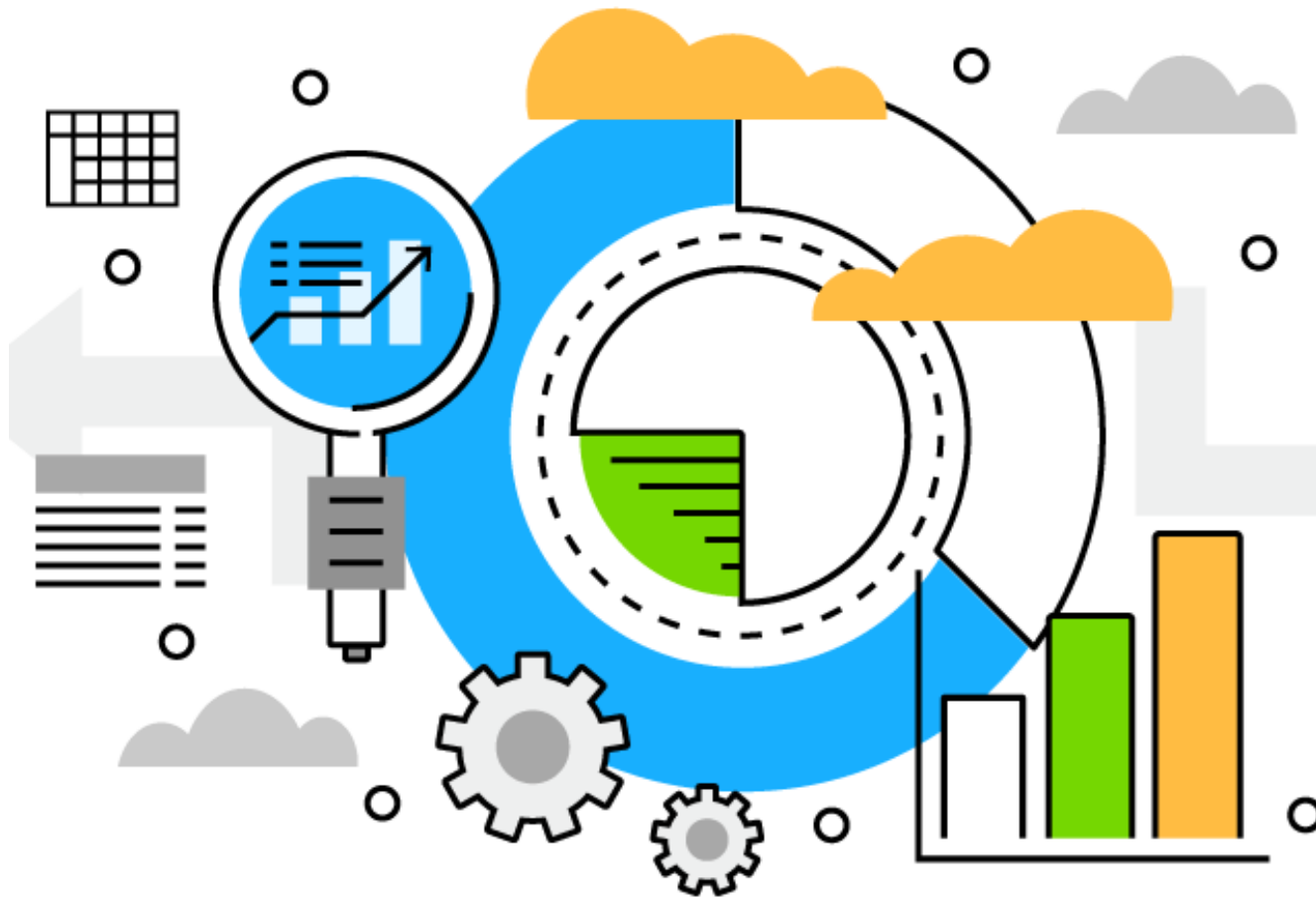
- **Predict**

```
t=(1+exp(-(predict(model2,list(Pregnancies=5,BMI=30,DiabetesPedigreeFunction=0.25,Glucose=100)))))^-1
```

t=0.1295224

Conclusões

- A regressão logística tem muitas vantagens perante a regressão linear (resíduos não precisam estar normalizados);
- A análise exploratória inicial das variáveis está coerente com os resultados obtidos;
- Todas as variáveis que compõem o modelo influenciam positivamente o resultado final;
- A precisão de 0.75 no conjunto de testes é um bom resultado. No entanto, este resultado depende um pouco da divisão dos dados para treino/teste;
- A área abaixo da curva ROC com valor de 0.85 é considerado um valor bastante bom.



Regressão logística - Estudo da diabetes

Análise de Clusters

ANA SOFIA FERREIRA PG38356

CAROLINA SILVA PG38335

CÉLIA FIGUEIREDO PG41022

MÁRCIA COSTA A67672

SAMUEL COSTA PG38352