CENTRO UNIVERSITÁRIO DE BELO HORIZONTE
Graduação - Ciência da Computação

JOSUÉ FILIPE PONTES NERY - 119122537
MARCELA CRISTYNE - 11827313
THAIS CORDEIRO PEREIRA - 11821085

UC: ANÁLISE DE DADOS E BIG DATA

Prática 09 - Estatística Preditiva (Regressão Logística)

Belo Horizonte
2º Semestre de 2021

Objetivo: Analisar os dados de massas mamográficas de classificação em maligno ou benigno com base nos atributos BI-RADS e na idade da paciente. O BI-RADS se refere a *Breast Imaging Reporting and Data System* (Sistema de Relatório de Dados sobre Imagem da Mama), que estima qual a chance de determinada imagem da mamografia ser câncer. Variando de 0 a 6, o BI-RADS ajuda a orientar a conduta médica. Ele não estima o grau de crescimento ou o tipo do tumor, nem dá dicas do tratamento, apenas diz “a chance de haver câncer, por este exame, é X por cento” (Fonte: https://vidasaudavel.einstein.br/o-que-e-birads/).

No *dataset*, armazenado na pasta compartilhada, é possível observar 6 valores:

- Avaliação da imagem BI-RADS: 1 (sugestivo de benigno) a 5 (sugestivo de maligno);
- Idade: idade da paciente;
- Forma da massa: redondo=1, oval=2, lobular=3, irregular=4;
- Borda da massa: circunscrito = 1, microlobulado = 2, obscurecido = 3, mal definido = 4, espiculado = 5;
- Densidade da massa: alta=1, iso=2, baixa=3, contendo gordura=4;
- Gravidade: benigno=0 ou maligno=1, que representa o resultado da análise. Essa será a nossa variável de interesse (desfecho).

Um médico analisa a imagem BI-RADS juntamente com a avaliação dessa imagem, e além disso, analisa os atributos BI-RADS, que são: i) forma da massa; ii) borda da massa; iii) densidade da massa e iv) a idade do paciente. Após essa análise, há a indicação do médico em relação a gravidade do exame.
Usem o RStudio para os itens abaixo. Não esqueçam de interpretar os resultados.

Resultado -

O primeiro modelo foi montado utilizando as seguintes variáveis: i) forma da massa; ii) borda da massa; iii) densidade da massa e iv) a idade do paciente. Ao treinar o modelo pela primeira vez foi observado que o p-value encontrados nas variáveis de densidade e de forma2 e 3, estão acima do valor significativo requerido para a atividade (0.05), logo estas variáveis foram retiradas do modelo.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.166033   0.848170  -4.912 9.02e-07 ***
idade        0.054552   0.007812   6.983 2.89e-12 ***
forma2      -0.261392   0.319274  -0.819 0.412953
forma3       0.658661   0.375601   1.754 0.079496 .
forma4       1.382353   0.333191   4.149 3.34e-05 ***
borda2       1.636297   0.559479   2.925 0.003448 **
borda3       1.175188   0.352012   3.338 0.000842 ***
borda4       1.492385   0.302650   4.931 8.18e-07 ***
borda5       2.002598   0.374774   5.343 9.12e-08 ***
desindade2  -0.962804   0.797498  -1.207 0.227324
desindade3  -0.649215   0.718219  -0.904 0.366036
desindade4  -1.760597   1.062947  -1.656 0.097654 .
```

O Modelo resultante teve uma taxa de acerto médio de aproximadamente 82%, considerando um limiar de 0.5.

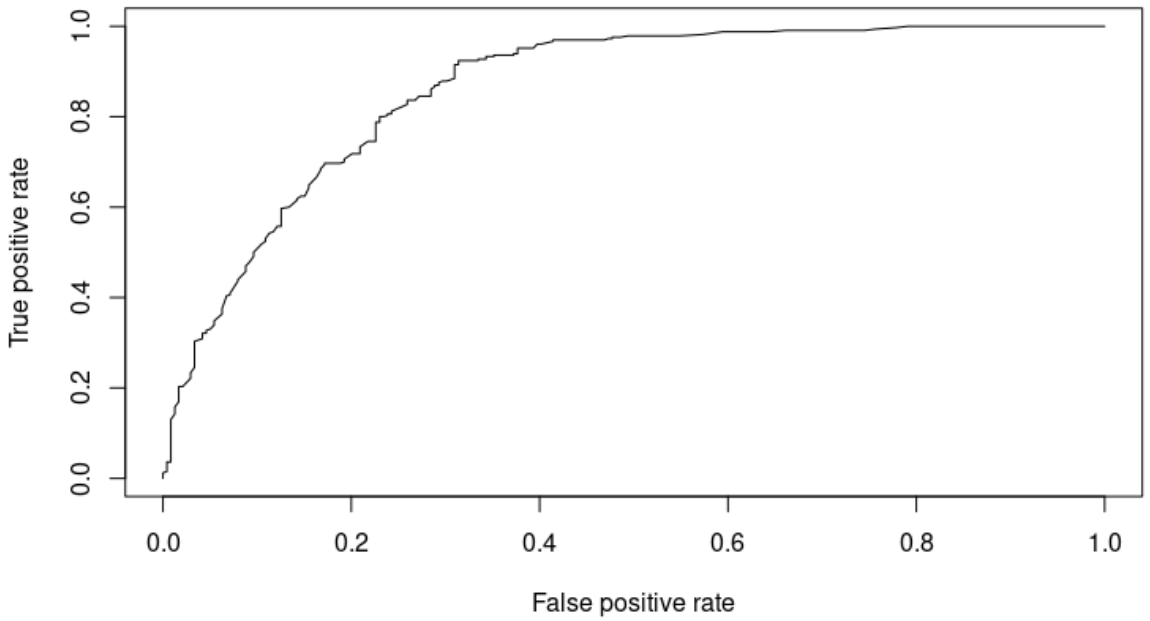
```
Call:
glm(formula = gravidade ~ idade + forma + borda, family = binomial,
    data = dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4761  -0.4911   0.4376   0.6772   2.3372

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.935339   0.579190  -8.521  < 2e-16 ***
idade        0.055400   0.009468   5.851 4.87e-09 ***
forma4       1.059599   0.437815   2.420  0.01551 *
borda2       1.812201   0.816246   2.220  0.02641 *
borda3       1.695570   0.535399   3.167  0.00154 **
borda4       1.864643   0.491821   3.791  0.00015 ***
borda5       2.239990   0.515471   4.346 1.39e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 774.19  on 568  degrees of freedom
Residual deviance: 494.46  on 562  degrees of freedom
AIC: 508.46

Number of Fisher Scoring iterations: 5
```

Neste novo modelo temos todas as variáveis com impacto significativo na variável resultado (p-value < 0.05). Sendo que a idade do paciente é a variável com maior intensidade do impacto seguida pelo tipo de borda de massa de número 5 (espiculado), são variáveis (borda e idade) de grande importância para o resultado da análise.

2) Apresente a Matriz de confusão, o limiar usado para a classificação, a taxa de acerto em média, a curva ROC e o valor AUC.

pred	0	1
benigno	164	25
maligno	75	305

Limiar	**0.5**
Média de acerto	**0.8242531**
AUC	**0.8613858**


True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0