APRENDIZAGEM AUTOMÁTICA

Mestrado em Engenharia Informática 2021

Universidade Do Minho

PREVENDO OCORRÊNCIA DO CANCRO MAMÁRIO

Raimundo Barros-PG42814

Pedro Ribeiro - PG42848

CANCRO DA MAMA

Em 2018, quase 2 milhões de novos casos de cancro da mama foram diagnosticados. Em2012, representou cerca de 12% de todos os novos casos de cancro e 25% de todos os cancro diagnosticados em pessoas do sexo feminino. Atualmente, o cancro da mama representa um em cada quatro de todos os cancros em mulheres. Desde 2008, a incidência mundial do cancro da mama aumentou em mais de 20% e a mortalidade aumentou 14%.

ANÁLISE EXPLORATÓRIA

A base de dados foi extraída do repositório de Machine Learning da UCI, esta contem um total de 569 observações explicadas num total de 32 variáveis, nomeadamente:

1. ID
2. Diagnóstico (M = maligno, B = benigno)
3. Dez atributos calculados para cada núcleo da célula:
    1. Raio
    2. Textura
    3. Perímetro
    4. Área
    5. Suavidade
    6. Compactidade
    7. Concavidade
    8. Pontos côncavos
    9. Simetria
    10. Aproximação do litoral

```
       id             diagnosis          radius_mean      texture_mean   
 Min.   :     8670   Length:569         Min.   : 6.981   Min.   : 9.71  
 1st Qu.:   869218   Class :character   1st Qu.:11.700   1st Qu.:16.17  
 Median :   906024   Mode  :character   Median :13.370   Median :18.84  
 Mean   : 30371831                      Mean   :14.127   Mean   :19.29  
 3rd Qu.:  8813129                      3rd Qu.:15.780   3rd Qu.:21.80  
 Max.   :911320502                      Max.   :28.110   Max.   :39.28  
 perimeter_mean     area_mean      smoothness_mean   compactness_mean  
 Min.   : 43.79   Min.   : 143.5   Min.   :0.05263   Min.   :0.01938  
 1st Qu.: 75.17   1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492  
 Median : 86.24   Median : 551.1   Median :0.09587   Median :0.09263  
 Mean   : 91.97   Mean   : 654.9   Mean   :0.09636   Mean   :0.10434  
 3rd Qu.:104.10   3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040  
 Max.   :188.50   Max.   :2501.0   Max.   :0.16340   Max.   :0.34540  
 concavity_mean     points_mean      symmetry_mean     dimension_mean   
 Min.   :0.00000   Min.   :0.00000   Min.   :0.1060   Min.   :0.04996  
 1st Qu.:0.02956   1st Qu.:0.02031   1st Qu.:0.1619   1st Qu.:0.05770  
 Median :0.06154   Median :0.03350   Median :0.1792   Median :0.06154  
 Mean   :0.08880   Mean   :0.04892   Mean   :0.1812   Mean   :0.06280  
 3rd Qu.:0.13070   3rd Qu.:0.07400   3rd Qu.:0.1957   3rd Qu.:0.06612  
 Max.   :0.42680   Max.   :0.20120   Max.   :0.3040   Max.   :0.09744  
   radius_se        texture_se      perimeter_se       area_se        
 Min.   :0.1115   Min.   :0.3602   Min.   : 0.757   Min.   :  6.802  
 1st Qu.:0.2324   1st Qu.:0.8339   1st Qu.: 1.606   1st Qu.: 17.850  
 Median :0.3242   Median :1.1080   Median : 2.287   Median : 24.530  
 Mean   :0.4052   Mean   :1.2169   Mean   : 2.866   Mean   : 40.337  
 3rd Qu.:0.4789   3rd Qu.:1.4740   3rd Qu.: 3.357   3rd Qu.: 45.190  
 Max.   :2.8730   Max.   :4.8850   Max.   :21.980   Max.   :542.200  
 smoothness_se      compactness_se      concavity_se       points_se        
 Min.   :0.001713   Min.   :0.002252   Min.   :0.00000   Min.   :0.000000  
 1st Qu.:0.005169   1st Qu.:0.013080   1st Qu.:0.01509   1st Qu.:0.007638  
 Median :0.006380   Median :0.020450   Median :0.02589   Median :0.010930  
 Mean   :0.007041   Mean   :0.025478   Mean   :0.03189   Mean   :0.011796  
 3rd Qu.:0.008146   3rd Qu.:0.032450   3rd Qu.:0.04205   3rd Qu.:0.014710  
 Max.   :0.031130   Max.   :0.135400   Max.   :0.39600   Max.   :0.052790  
  symmetry_se        dimension_se        radius_worst   texture_worst   
 Min.   :0.007882   Min.   :0.0008948   Min.   : 7.93   Min.   :12.02  
 1st Qu.:0.015160   1st Qu.:0.0022480   1st Qu.:13.01   1st Qu.:21.08  
 Median :0.018730   Median :0.0031870   Median :14.97   Median :25.41  
 Mean   :0.020542   Mean   :0.0037949   Mean   :16.27   Mean   :25.68  
 3rd Qu.:0.023480   3rd Qu.:0.0045580   3rd Qu.:18.79   3rd Qu.:29.72  
 Max.   :0.078950   Max.   :0.0298400   Max.   :36.04   Max.   :49.54  
 perimeter_worst     area_worst     smoothness_worst  compactness_worst 
 Min.   : 50.41   Min.   : 185.2   Min.   :0.07117   Min.   :0.02729  
 1st Qu.: 84.11   1st Qu.: 515.3   1st Qu.:0.11660   1st Qu.:0.14720  
 Median : 97.66   Median : 686.5   Median :0.13130   Median :0.21190  
 Mean   :107.26   Mean   : 880.6   Mean   :0.13237   Mean   :0.25427  
 3rd Qu.:125.40   3rd Qu.:1084.0   3rd Qu.:0.14600   3rd Qu.:0.33910  
 Max.   :251.20   Max.   :4254.0   Max.   :0.22260   Max.   :1.05800  
 concavity_worst   points_worst     symmetry_worst   dimension_worst   
 Min.   :0.0000   Min.   :0.00000   Min.   :0.1565   Min.   :0.05504  
 1st Qu.:0.1145   1st Qu.:0.06493   1st Qu.:0.2504   1st Qu.:0.07146  
 Median :0.2267   Median :0.09993   Median :0.2822   Median :0.08004  
 Mean   :0.2722   Mean   :0.11461   Mean   :0.2901   Mean   :0.08395  
 3rd Qu.:0.3829   3rd Qu.:0.16140   3rd Qu.:0.3179   3rd Qu.:0.09208  
 Max.   :1.2520   Max.   :0.29100   Max.   :0.6638   Max.   :0.20750  
```

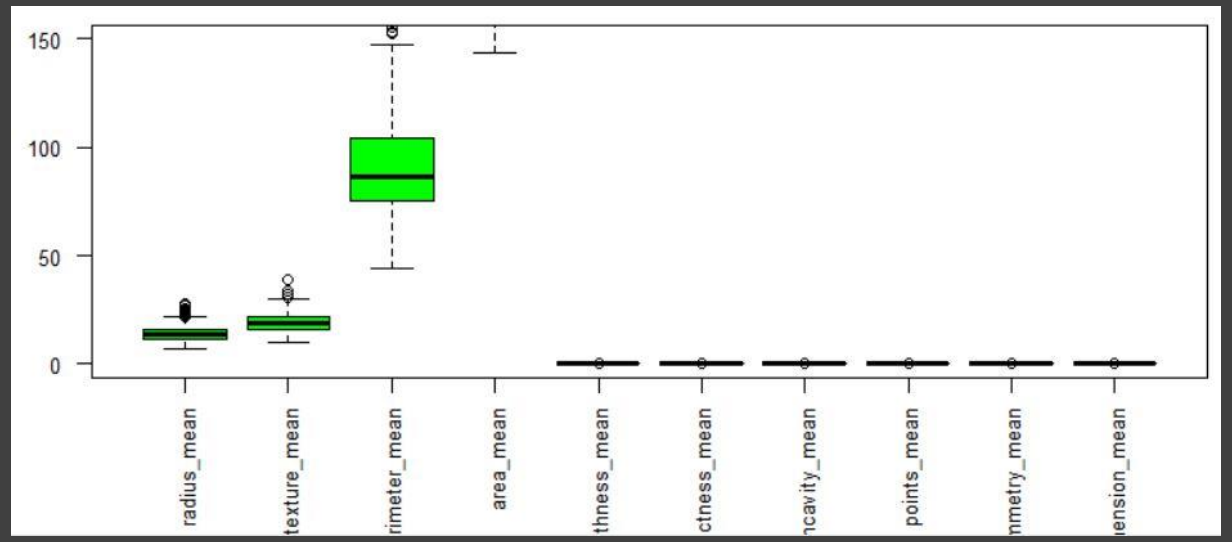
CORRELAÇÃO

Com base na matriz de correlação, observa-se que existem 13 preditores relacionados com a resposta "*diagnosis*", com r ≥ 0.6, portanto pode-se inferir que existe uma boa relação entre estes preditores e a variável ”alvo”.

NORMALIZAÇÃO

Existem algumas variáveis que estão em escala diferentes (eg: *texturemeanes*, *moothnessmean*). Para padronizar os valores das variáveis utilizou-se a função ”scale”, que é responsável por aplicar a função sobre os vetores do *dataset* para padronizar os valores numéricos, colocando-os na mesma escala.

Devido ao Teorema do Limite Central, sabe-se que as médias das amostras tendem a distribuir-se por uma distribuição normal.



MODELAÇÃO

1 Modelo

```
Deviance Residuals:
       Min          1Q      Median          3Q         Max
-1.595e-04  -2.100e-08  -2.100e-08   2.100e-08   1.461e-04

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)          1.942  41083.417   0.000    1.000
radius_mean      -1587.827 640777.917  -0.002    0.998
texture_mean       104.307  26569.258   0.004    0.997
perimeter_mean     874.416 650105.690   0.001    0.999
area_mean          657.045 252722.174   0.003    0.998
smoothness_mean    103.852  19090.879   0.005    0.996
compactness_mean  -501.788 119598.941  -0.004    0.997
concavity_mean     280.525  80196.914   0.003    0.997
points_mean         94.553  81987.260   0.001    0.999
symmetry_mean       -9.955  20902.749   0.000    1.000
dimension_mean      75.289  41398.123   0.002    0.999
radius_se           74.479  87488.042   0.001    0.999
texture_se           2.899  24040.631   0.000    1.000
perimeter_se        83.533  76946.504   0.001    0.999
area_se           -271.335 131353.186  -0.002    0.998
smoothness_se      -49.039  31875.388  -0.002    0.999
compactness_se     160.538  57058.368   0.003    0.998
concavity_se       -75.970  70258.679  -0.001    0.999
points_se           80.129  65409.593   0.001    0.999
symmetry_se        -58.103  33556.398  -0.002    0.999
dimension_se       -114.834  75442.836  -0.002    0.999
radius_worst       1152.857 323318.827   0.004    0.997
texture_worst       -39.629  32725.141  -0.001    0.999
perimeter_worst     -332.167 113655.842  -0.003    0.998
area_worst          -437.179 343529.338  -0.001    0.999
smoothness_worst      -2.621  36425.559   0.000    1.000
compactness_worst    104.522  50814.755   0.002    0.998
concavity_worst     -118.392  50493.620  -0.002    0.998
points_worst         -10.818  47320.066   0.000    1.000
symmetry_worst        48.952  33063.717   0.001    0.999
dimension_worst       59.513  47605.314   0.001    0.999

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5.2728e+02  on 398  degrees of freedom
Residual deviance: 2.1560e-07  on 368  degrees of freedom
AIC: 62
```

Modelo não normalizado

Var1	Var2	Freq
radius_mean	perimeter_mean	0.9979
radius_worst	perimeter_worst	0.9937
radius_mean	area_mean	0.9874
perimeter_mean	area_mean	0.9865
radius_worst	area_worst	0.9840
perimeter_worst	area_worst	0.9776
radius_se	perimeter_se	0.9728
perimeter_mean	perimeter_worst	0.9704
radius_mean	radius_worst	0.9695
perimeter_mean	radius_worst	0.9695
radius_mean	perimeter_worst	0.9651
area_mean	radius_worst	0.9627
area_mean	area_worst	0.9592
area_mean	perimeter_worst	0.9591
radius_se	area_se	0.9518
perimeter_mean	area_worst	0.9415
radius_mean	area_worst	0.9411
perimeter_se	area_se	0.9377
concavity_mean	points_mean	0.9214
texture_mean	texture_worst	0.9120
points_mean	points_worst	0.9102

Correlação

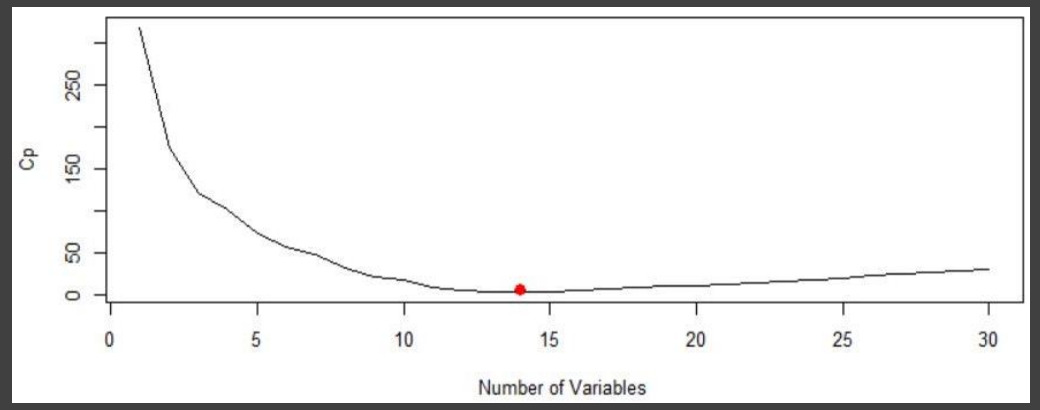
2 Modelo

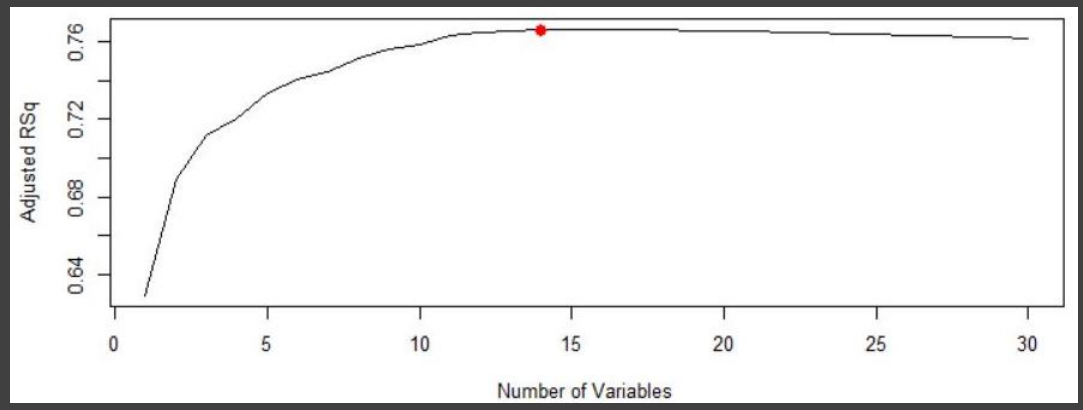
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1699  -0.1334  -0.0219   0.0538   3.4377

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.52739    0.34669  -4.406 1.06e-05 ***
smoothness_mean   -0.06240    0.77347  -0.081 0.935699
compactness_mean   5.61340    1.36673   4.107 4.01e-05 ***
symmetry_mean     -0.82233    0.56370  -1.459 0.144622
dimension_mean    -4.89155    1.05443  -4.639 3.50e-06 ***
texture_se         1.21057    0.34746   3.484 0.000494 ***
smoothness_se      0.76958    0.66114   1.164 0.244415
compactness_se    -1.00976    1.29708  -0.778 0.436283
concavity_se      -2.48119    1.44180  -1.721 0.085269 .
points_se          1.86603    0.82504   2.262 0.023713 *
symmetry_se       -0.83034    0.59059  -1.406 0.159737
dimension_se      -0.32048    1.44962  -0.221 0.825031
smoothness_worst  -0.08558    0.82570  -0.104 0.917453
compactness_worst -3.84131    1.93802  -1.982 0.047470 *
concavity_worst    4.53524    1.42838   3.175 0.001498 **
symmetry_worst     2.02840    0.75159   2.699 0.006958 **
dimension_worst    2.68691    1.51713   1.771 0.076552 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 527.28  on 398  degrees of freedom
Residual deviance: 110.15  on 382  degrees of freedom
AIC: 144.15
```

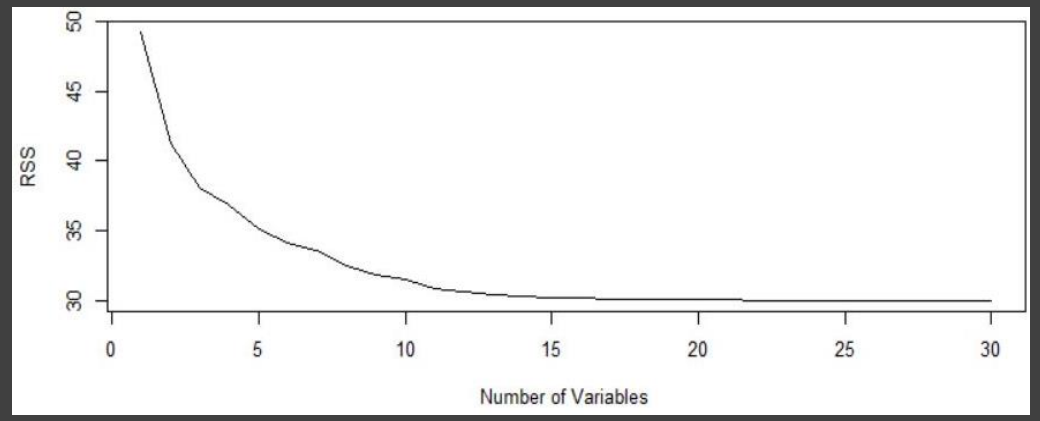
```
          test2_class
conv_log2   0   1
        0 103   5
        1   4  58

               Accuracy : 0.9471
                 95% CI : (0.9019, 0.9755)
    No Information Rate : 0.6294
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8861

 Mcnemar's Test P-Value : 1
```

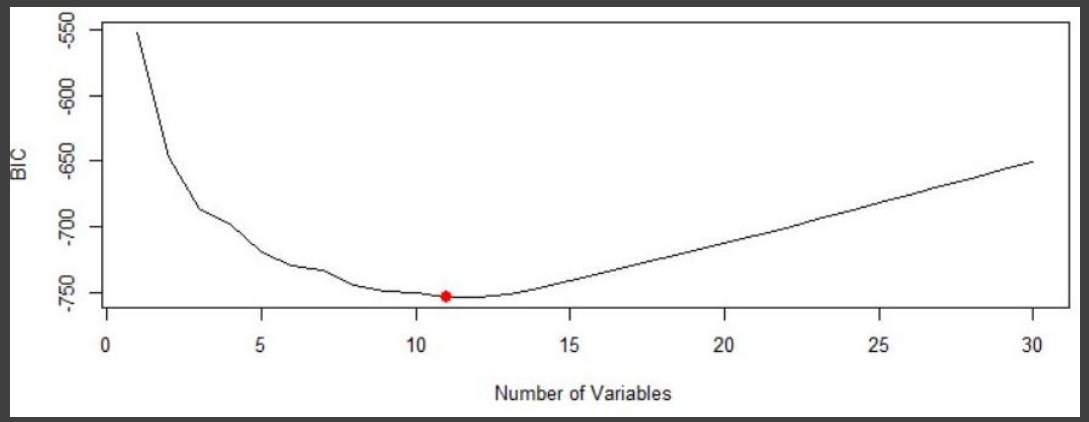
Primeiro Modelo após a analise da correlação

Nº DE PREDITORES POR TÉCNICA DE AJUSTAMENTO

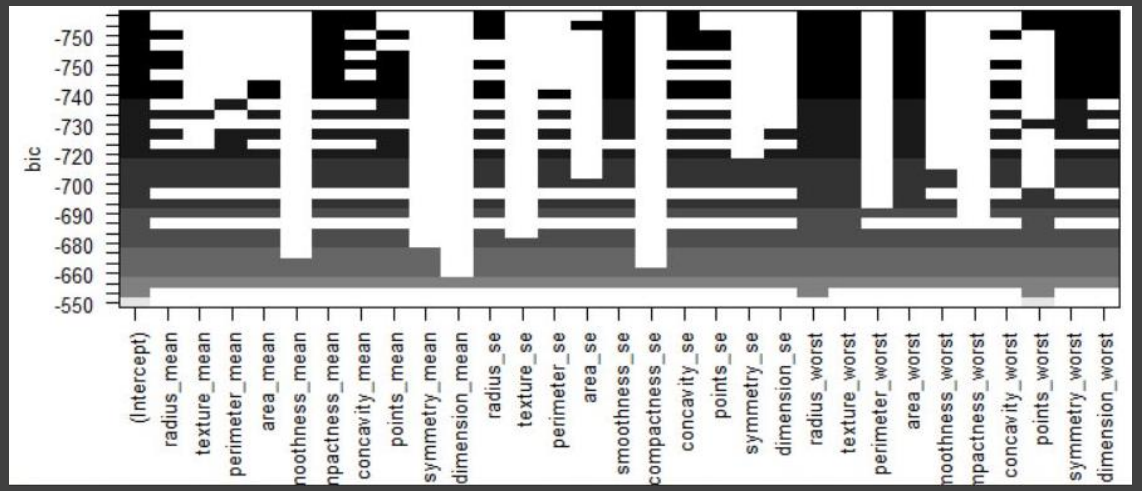


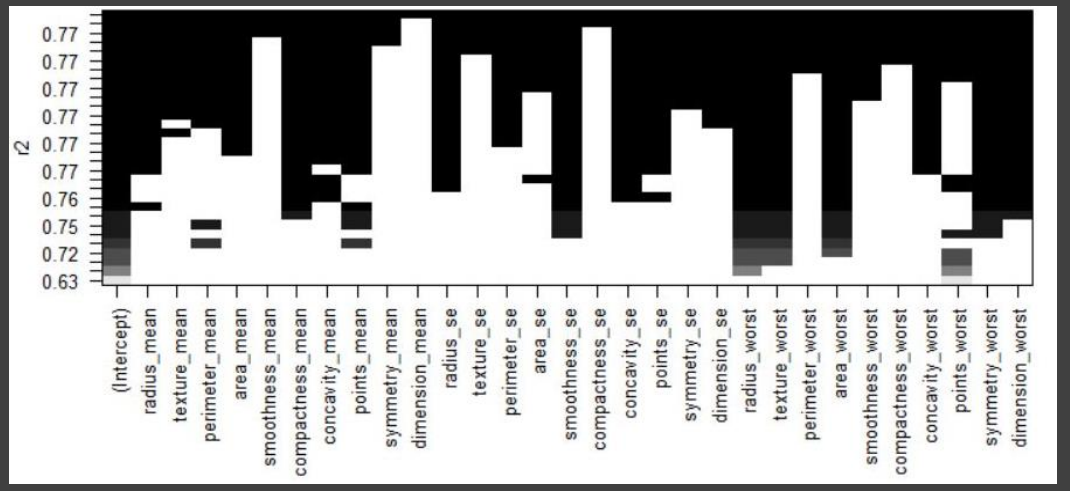


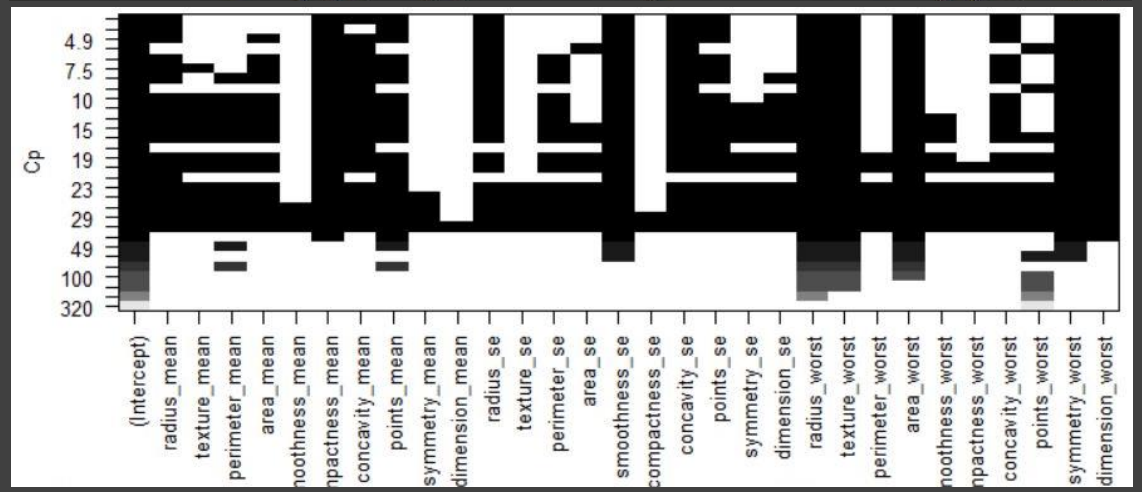




GRÁFICOS DOS PREDITORES MAIS RELEVANTES







MODELO LOGÍSTICO ESCOLHIDO

```
          test6_class
conv_log6   0   1
        0 105   4
        1   2  59

               Accuracy : 0.9647
                 95% CI : (0.9248, 0.9869)
    No Information Rate : 0.6294
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9238

 Mcnemar's Test P-Value : 0.6831

            Sensitivity : 0.9365
            Specificity : 0.9813
         Pos Pred Value : 0.9672
         Neg Pred Value : 0.9633
             Prevalence : 0.3706
         Detection Rate : 0.3471
   Detection Prevalence : 0.3588
      Balanced Accuracy : 0.9589

       'Positive' Class : 1
```

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.95634  -0.02183  -0.00199   0.00003   2.86349

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.3936     2.0236   0.195  0.84576
compactness_mean   -3.4024     2.7805  -1.224  0.22107
points_mean         4.2922     2.9816   1.440  0.14999
radius_se           4.8680     9.7939   0.497  0.61916
area_se            -1.6011    19.1713  -0.084  0.93344
smoothness_se       0.9527     0.7152   1.332  0.18286
compactness_se     -1.5693     1.4112  -1.112  0.26613
concavity_se       -0.9775     1.2339  -0.792  0.42822
radius_worst        4.2624    16.7631   0.254  0.79928
texture_worst       2.0257     0.6500   3.116  0.00183 **
area_worst          2.2841    21.4729   0.106  0.91529
points_worst        1.8598     2.2681   0.820  0.41221
symmetry_worst      0.6588     0.7388   0.892  0.37251
concavity_worst     2.6892     1.9816   1.357  0.17474
dimension_worst     2.1129     1.6491   1.281  0.20011
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 527.285  on 398  degrees of freedom
Residual deviance:  35.596  on 384  degrees of freedom
AIC: 65.596

Number of Fisher Scoring iterations: 13
```

MODELO K-NEAREST NEIGHBORS (KNN)

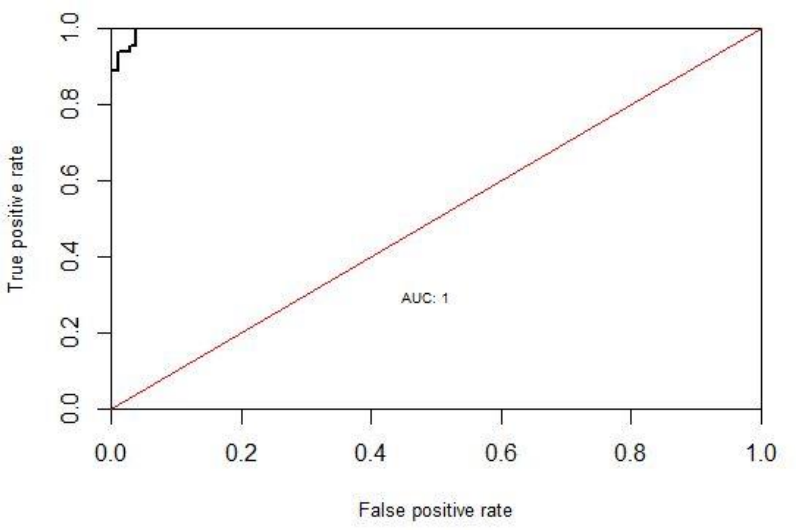
Antes de gerar o modelo KNN é importante determinar o valor de K e para isso utiliza-se o Cross Validation para treinar diferentes subconjuntos de dados com diferentes valores de K de forma a identificar o K que possui uma melhor acurácia.

Após a preparação do dataset e a separação deste em dataset de treino e dados de teste, aplica-se a função KNN, obtendo a seguinte ”Confusion Matrix”.

```
Confusion Matrix and Statistics

               dados_teste_labels
modelo_knn_v1   0   1
            0 105   3
            1   2  60

               Accuracy : 0.9706
                 95% CI : (0.9327, 0.9904)
    No Information Rate : 0.6294
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9367

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9524
            Specificity : 0.9813
         Pos Pred Value : 0.9677
         Neg Pred Value : 0.9722
             Prevalence : 0.3706
         Detection Rate : 0.3529
   Detection Prevalence : 0.3647
      Balanced Accuracy : 0.9668

       'Positive' Class : 1
```

”Confusion Matrix”

TABELA DE COMPARAÇÃO



	Accuracy	auc	aic
Logistic - M1	0.9176471	0.9117342	62
Logistic - M2	0.9470588	0.9416259	144.14862
Logistic - M3	0.9352941	0.9355437	168.65441
Logistic - M4	0.9588235	0.9477081	58.06722
Logistic - M5	0.9588235	0.9477081	62.04357
Logistic - M6	0.9647059	0.9589082	65.59571
KNN	0.9705882	----	---

CONCLUSÃO

- O modelo de Regressão Logística e o modelo KNN possuem desempenhos semelhantes.
- Modelo logístico pode-se visualizar os coeficientes de todos os preditores o que permite uma análise mais completa.
- KNN apenas é possível identificar o Valor de k para o modelo


OBRIGADO