Breast Cancer Prediction Model

Gabriel Valverde Zanata da Silva

2023-01-20

Introdução e Objetivos

Este projeto tem como objetivo utilizar os dados do Database sobre Câncer de Mama da Universidade de Wisconsin Hospitals, Madison de Dr. William H. Wolberg, para desenvolver um modelo preditivo capaz de prever se o câncer é benígno ou malígno com base nas informações disponíveis.

Os atributos da base de dados são:

Attribute	Domain
1.	Sample code number id number
2.	Clump Thickness 1 - 10
3.	Uniformity of Cell Size 1 - 10
4.	Uniformity of Cell Shape 1 - 10
5.	Marginal Adhesion 1 - 10
6.	Single Epithelial Cell Size 1 - 10
7.	Bare Nuclei 1 - 10
8.	Bland Chromatin 1 - 10
9.	Normal Nucleoli 1 - 10
10.	Mitoses 1 - 10
11.	Class: (2 for benign, 4 for malignant)

Etapa 1 - Pacotes e Bibliotecas

Para este projeto, foram utilizados apenas 3 pacotes:

```
# Tratamento de dados
library(dplyr)

# Modelo de classificação knn
library(class)

# Avaliação do modelo
library(gmodels)
```

Etapa 2 - Carregando os Dados

O arquivo do database foi obtido em formato texto(.txt), sem nomes de colunas, e com observações separadas por vírgula. Portanto, foi necessário criar um vetor para nomear variáveis corretamente conforme o dicionário de dados.

```r
#Carregando dataset

col_names <- c("ID","clump_thickness","uniformity_cell_size",
               "uniformity_cell_shape", "marginal_adhesion",
               "single_epithelial_cell_size", "bare_nuclei",
               "bland_chromatin","normal_nucleoli", "mitoses",
               "class")

# Backup dos dados não tratados
dados_bruto <- read.table("breast-cancer-wisconsin.data", header = FALSE,
                     sep = ",", dec = ".", col.names = col_names)

# Df para tratamento dos dados
dados <- dados_bruto
```

Etapa 3 - Exploração e Tratamento dos Dados

Uma vez carregados os dados, é necessário realizar a adequação dos mesmos ao modelo preditivo que será utilizado. Para isso, elimina-se as variáveis irrelevantes, como o ID e os valores NA, neste caso representados pelo caracter "?" e presentes apenas em uma das colunas, conforme o dicionário de dados. Além disso é necessário adequar os tipos das variáveis, fatorizando a variável alvo "Class", por exemplo.

```r
# Estudando NAs
any(is.na(dados))
```

```
## [1] FALSE
```

```r
qtd_vazios <-  100 * unlist(dados %>% filter(bare_nuclei == "?") %>% count()) / length(dados$bare_nuclei
qtd_vazios # 2.3% de NAs (16 observações de 699) -> dropar
```

```
##        n
## 2.288984
```

```r
# Excluindo linhas com NA
dados <- dados[dados$bare_nuclei != "?",]

str(dados)
```

```
## 'data.frame':    683 obs. of  11 variables:
##  $ ID                         : int  1000025 1002945 1015425 1016277 1017023 1017122 1018099 1018561
##  $ clump_thickness            : int  5 5 3 6 4 8 1 2 2 4 ...
##  $ uniformity_cell_size       : int  1 4 1 8 1 10 1 1 1 2 ...
##  $ uniformity_cell_shape      : int  1 4 1 8 1 10 1 2 1 1 ...
##  $ marginal_adhesion          : int  1 5 1 1 3 8 1 1 1 1 ...
##  $ single_epithelial_cell_size: int  2 7 2 3 2 7 2 2 2 2 ...
##  $ bare_nuclei                : chr  "1" "10" "2" "4" ...
##  $ bland_chromatin            : int  3 3 3 3 3 9 3 3 1 2 ...
##  $ normal_nucleoli            : int  1 2 1 7 1 7 1 1 1 1 ...
##  $ mitoses                    : int  1 1 1 1 1 1 1 1 5 1 ...
##  $ class                      : int  2 2 2 2 2 4 2 2 2 2 ...
```

```r
# Removendo a coluna ID (irelevante/prejudicial para o modelo preditivo)
dados <- dados [-1]
str(dados)
```

```
## 'data.frame':    683 obs. of  10 variables:
##  $ clump_thickness            : int  5 5 3 6 4 8 1 2 2 4 ...
##  $ uniformity_cell_size       : int  1 4 1 8 1 10 1 1 1 2 ...
##  $ uniformity_cell_shape      : int  1 4 1 8 1 10 1 2 1 1 ...
##  $ marginal_adhesion          : int  1 5 1 1 3 8 1 1 1 1 ...
##  $ single_epithelial_cell_size: int  2 7 2 3 2 7 2 2 2 2 ...
##  $ bare_nuclei                : chr  "1" "10" "2" "4" ...
##  $ bland_chromatin            : int  3 3 3 3 3 9 3 3 1 2 ...
##  $ normal_nucleoli            : int  1 2 1 7 1 7 1 1 1 1 ...
##  $ mitoses                    : int  1 1 1 1 1 1 1 1 5 1 ...
##  $ class                      : int  2 2 2 2 2 4 2 2 2 2 ...
```

```r
# Tranformando os tipos de dados para Numérico
dados[c(1:9)] <- sapply(dados[c(1:9)], as.numeric)

# Fatorizando a coluna Class
table(dados$class)
```

```
##
##   2   4
## 444 239
```

```r
dados$class <- factor(dados$class, levels = c(2,4),
                      labels = c("Benigno", "Maligno"))
str(dados)
```

```
## 'data.frame':    683 obs. of  10 variables:
##  $ clump_thickness            : num  5 5 3 6 4 8 1 2 2 4 ...
##  $ uniformity_cell_size       : num  1 4 1 8 1 10 1 1 1 2 ...
##  $ uniformity_cell_shape      : num  1 4 1 8 1 10 1 2 1 1 ...
##  $ marginal_adhesion          : num  1 5 1 1 3 8 1 1 1 1 ...
##  $ single_epithelial_cell_size: num  2 7 2 3 2 7 2 2 2 2 ...
##  $ bare_nuclei                : num  1 10 2 4 1 10 10 1 1 1 ...
##  $ bland_chromatin            : num  3 3 3 3 3 9 3 3 1 2 ...
##  $ normal_nucleoli            : num  1 2 1 7 1 7 1 1 1 1 ...
##  $ mitoses                    : num  1 1 1 1 1 1 1 1 5 1 ...
##  $ class                      : Factor w/ 2 levels "Benigno","Maligno": 1 1 1 1 1 2 1 1 1 1 ...
```

```r
# Verificando a proporção de Class
round(prop.table(table(dados$class)) * 100, digits = 1)
```

```
##
## Benigno Maligno
##      65      35
```

Etapa 4 - Modelo Preditivo

Com os dados tratados, é possível partir para a criação, treinamento e avaliação do modelo preditivo. O modelo escolhido foi o modelo de classificação KNN - K Nearest Neighbour.

```
# Divisão dados treino e teste
d_treino <- dados[1:480,1:9] # ~ 70%
d_teste <- dados [481:683,1:9] # ~ 30%

# Labels
d_treino_labels <- dados[1:480,10]
d_teste_labels <- dados [481:683,10]

# Criando e treinando o modelo
modelo <- knn(train = d_treino, test = d_teste, cl = d_treino_labels, k=21)

# Avaliando modelo
CrossTable(x = d_teste_labels, y = modelo, prop.chisq = FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  203
##
##
##                | modelo
## d_teste_labels |   Benigno |   Maligno | Row Total |
## ---------------|-----------|-----------|-----------|
##        Benigno |       159 |         0 |       159 |
##                |     1.000 |     0.000 |     0.783 |
##                |     1.000 |     0.000 |           |
##                |     0.783 |     0.000 |           |
## ---------------|-----------|-----------|-----------|
##        Maligno |         0 |        44 |        44 |
##                |     0.000 |     1.000 |     0.217 |
##                |     0.000 |     1.000 |           |
##                |     0.000 |     0.217 |           |
## ---------------|-----------|-----------|-----------|
##   Column Total |       159 |        44 |       203 |
##                |     0.783 |     0.217 |           |
## ---------------|-----------|-----------|-----------|
##
##
```

Nota-se que o modelo pôde atingir uma taxa de ~100% de acertos para os dados de teste fornecidos, indicando que é um modelo altamente eficiente para a previsão da classe do câncer de mama através dos critérios e

variáveis utilizadas. Por se tratar de um tema de saúde e que envolve vidas, é importante que as taxas de eficiência de modelos preditivos nesta área sejam tão altas quanto possível.