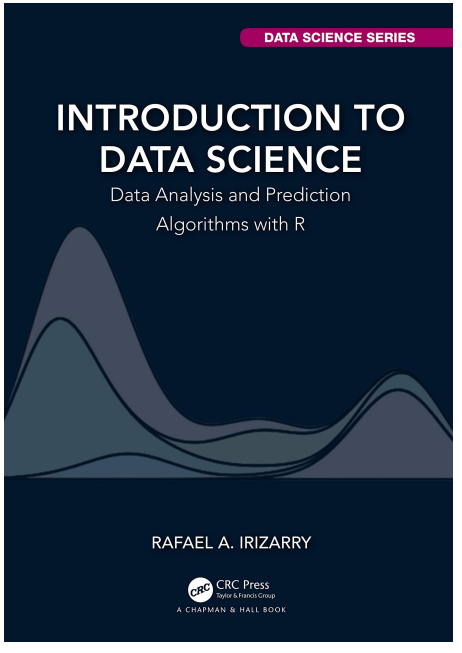
Introdução ao aprendizado de máquina: Métricas de Avaliação

Oscar J. O. Ayala

- Talvez Machine Learning (ML) seja a metodologia mais popular em **Data Science**. Seu diferencial é que suas decisões são **baseadas** em algoritmos *construídos com dados*.
- Este material é baseado no capítulo 27 do livro *Data Analysis and Prediction Algorithms with R* (Irizarry, 2022)[1].



[1] http://rafalab.dfci.harvard.edu/dsbook/introduction-to-machine-learning.html

Abordagem de ML

- Primeiro, treinar (**train**) o algoritmo usando um conjunto de recursos disponíveis para prever um conjunto de resultados observados. O que implica a optimização do algoritmo.
- Em segundo lugar, testar (**teste**) o algoritmo usando um conjunto de recursos disponíveis para prever um conjunto de resultados que se fingem que não ser observados. Para de fingir que não sabe o resultado para avaliar o algoritmo, mas só depois que terminarmos de construí-lo. Resultando em avaliação de algoritmo.
- Terceito, prever (**hat**) um conjunto de resultados não observados usando um conjunto de variáveis disponíveis.

Notação

Em *ML* os dados vem de dois tipos:

- Os **resultados** que se desejam prever.
- Os **recursos (variáveis)** disponíveis para prever os resultados.

Assim, os resultados são representados por Y, enquanto os recursos por $x_1,\ x_2, \ldots x_n, n \in Z_+$.

Tarefas de ML

- Quando o resultado é contínuo, diz-se que a tarefa de *ML* é de *previsão*. Portanto, pode-se obter um erro que nos diz quão próximas as previsões estão dos resultados reais. Uma maneira comum é fazer $y - \hat{y}$, $\hat{y} = f(x_1, x_2, \ldots x_n)$.

- Quando o resultado é categórico, a tarefa de *ML* é de *classificação*. A saída principal obedece a uma regra de decisão que prescreve qual das classes *K* deve ser prevista. A função de recursos ou preditores para cada classe *k* tomada como regra de decisão é dada por $f_k(x_1, x_2, \ldots x_n)$. Para o caso binário específico, temos,

$$\begin{cases} \text{Se, } f_1(x_1,\ x_2, \ldots x_n) > C, \text{ prever } 1 \\ \text{C.C., prever } 0 \end{cases}$$

Métricas de avaliação

- Definir o conceito de *melhor* entre um conjunto de abordagens.
- Descrever a forma como os algoritmos *ML* são avaliados.

Caso de estudo

Se quer **prever o sexo** de uma pessoa pela **altura**. É usado o conjunto de dados disponível no data frame `dslabs::heights`, que apresenta as alturas auto concedidas em polegadas para um total de `1050` homens e mulheres. São construidos dos algoritmos, sendo necessário avaliar e escolher o melhor.

Na verdade, o recurso **altura** não é suficiente para prever o sexo, mas como o objetivo é ilustrar os métodos de avaliação, suponha que não haja problemas com essa abordagem simplista.

Usa-se como recurso computacional o software **R** e os pacotes `caret`, `tidyverse`, `dslabs`, `ggrepel` `egridExtra`.

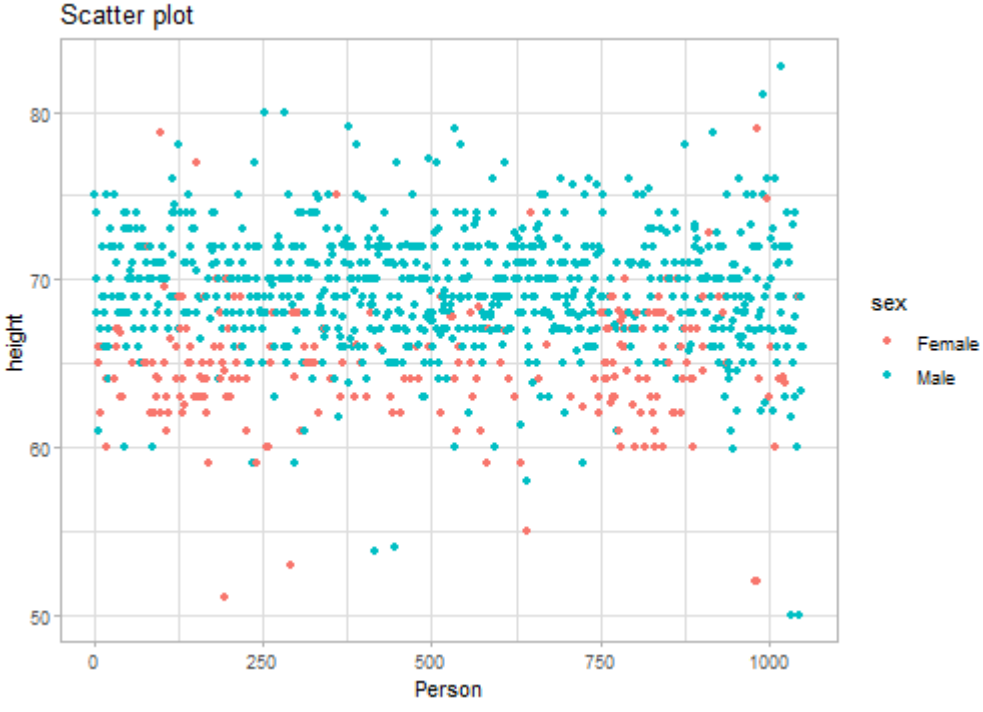
Analises descritiva

```r
# pacotes
require(caret)
require(tidyverse)
require(dslabs)
require(ggrepel)
require(gridExtra)

# dados
dat <- dslabs::heights

# Tabela descritiva
dat %>%
  dplyr::group_by(sex) %>%
  dplyr::summarise(n = n(),
                   mean = round(mean(height), 2),
                   sigma = round(sd(height), 2),
                   .groups = "drop") %>%
  knitr::kable(caption = "Descritive") -> tab1
```

Na Tabela e Figura a seguir, pode-se observar que a dispersão das *alturas* de homens e mulheres são semelhantes. No entanto, a média e a prevalência parecem ser maiores para os homens.

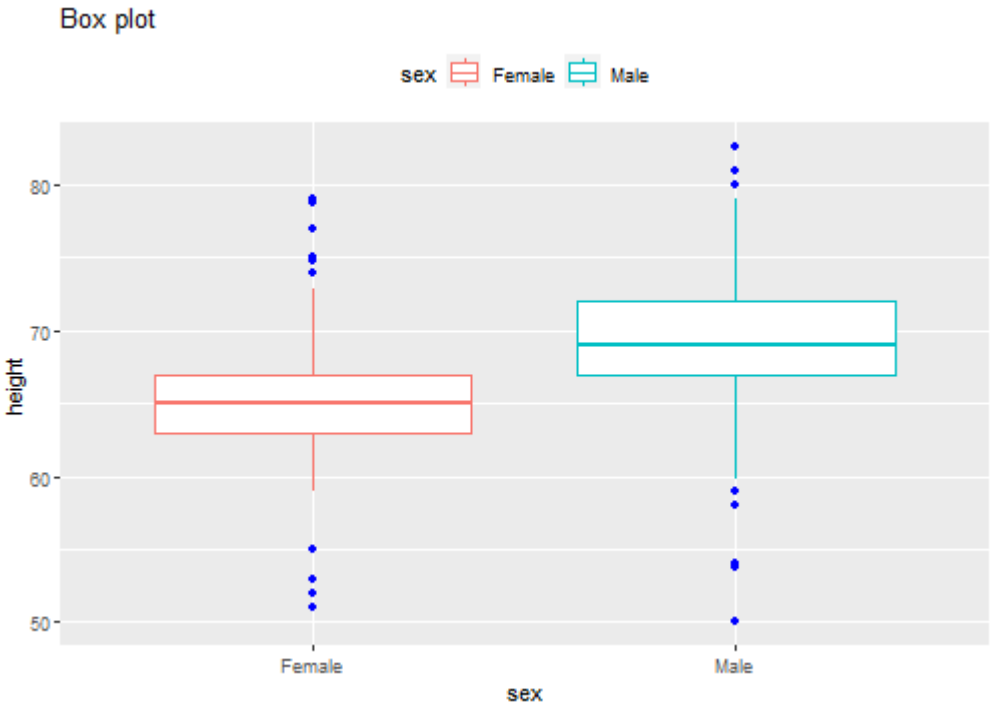
Table: Descritive

sex	n	mean	sigma
Female	238	64.94	3.76
Male	812	69.31	3.61



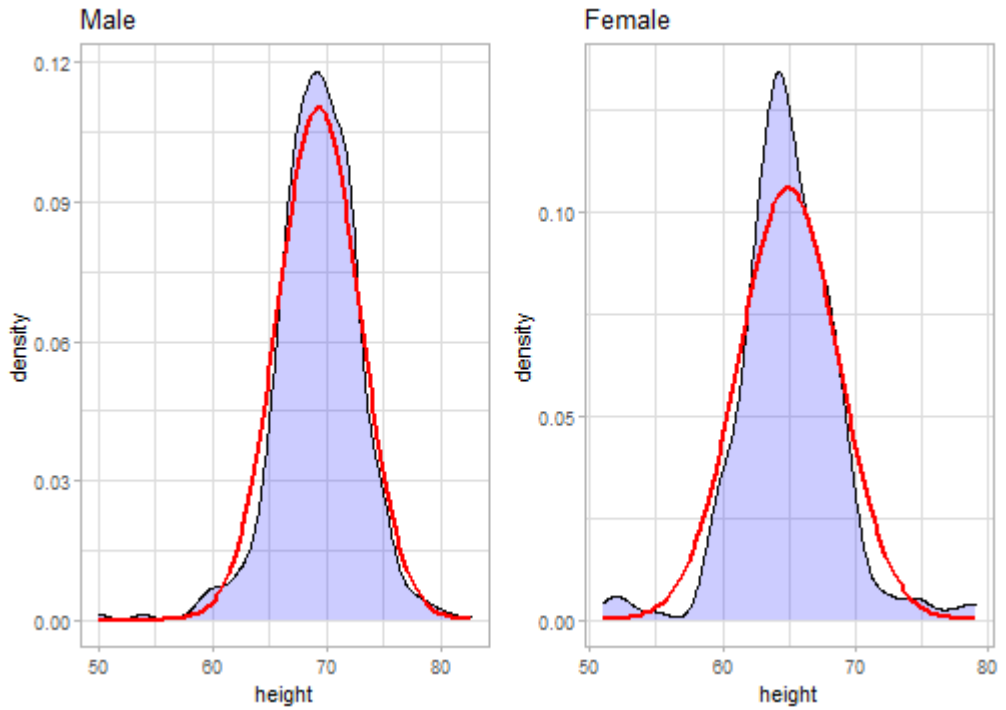
Box plot

Existem outliers em ambos os casos, mas como são consistentes, permanecem. Há indícios de que a estatura média geral dos homens seja maior que a das mulheres, mas a dispersão é semelhante.



Função de densidade por sexo vs N(0, 1)

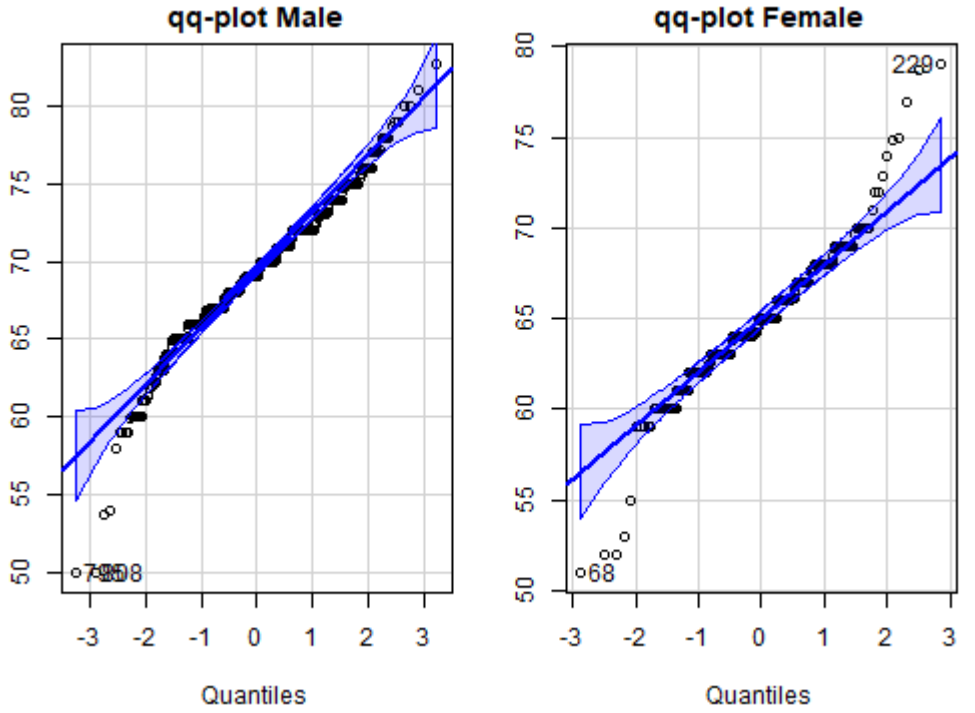
Observa-se que a distribuição das alturas das pessoas, independentemente do sexo, não apresenta assimetria acentuada em relação a $N(0, 1)$. No entanto, devido a possíveis outlier, parece que a densidade das alturas das mulheres se desviam marcadamente da normal padrão.



QQ - plot

Há indícios de que as alturas de homens e mulheres se desviam do padrão normal, sendo o problema mais *pronunciado* para as últimas.

```
## [1] 795 808
```




```
## [1] 229  68
```

Counjunto Treino e Teste

Define-se o conjunto de recursos disponíveis e resultados a prever, para os passos de treino e teste.

```
# indice de selecao aleatorio
set.seed(2022)
index <- caret::createDataPartition(y = dat$sex,
                                    times = 1,
                                    p = 0.5,
                                    list = FALSE) %>%
  as.vector()

# conjunto de dados traino e teste
trainSet <- dat[-index, ]
testSet  <- dat[index, ]
```

Métodos Sample vs Cutoff: Trainamento

- Sample: Consiste em criar uma amostra construída com dados, para diferentes probabilidades, $1 - p$, de escolha da classe *Mulher*.

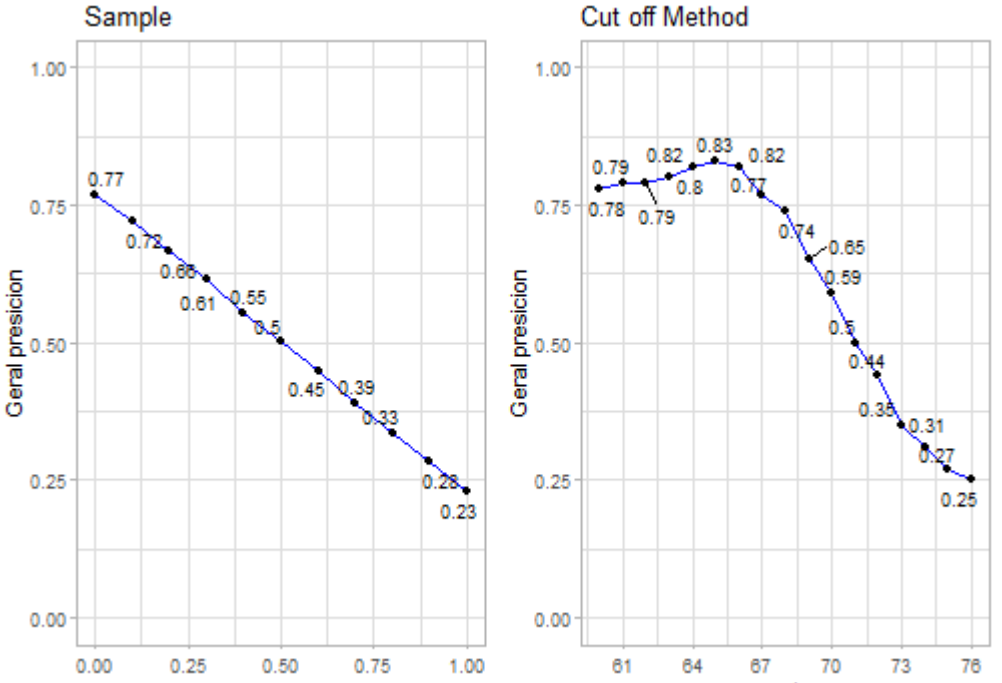
```
## probabilidades de escolha classe homem
p <- seq(0, 1, 0.1)

## Treinamento
methodSample <- purrr::map_dbl(p, function(x){

  # prever
  preGeral <- numeric()

  for(i in 1:100){

      yHat <- base::sample(x = base::levels(trainSet$sex),
                           size = base::length(trainSet$sex),
                           replace = TRUE,
                           prob = c(x, 1 - x)) %>%
        base::factor(levels =  base::levels(trainSet$sex))

  # metricas de avaliacao: precisao geral

  preGeral[i] <- round(unique(caret::confusionMatrix(dat = yHat,
                                                     reference = tra
```

- cutoff: Consiste na criação de um algoritmo que faz diferentes corte (>=) na altura das pessoas e lhes atribui um sexo.

```
## cortes
cut <- seq(round(mean(dat$height) - 2 * sd(dat$height), 0),
           round(mean(dat$height) + 2 * sd(dat$height), 0))

## Treinamento
methodCutoff <- purrr::map_dbl(cut, function(x){

  # prever
yHat <- dplyr::case_when(trainSet$height >= x ~ "Male",
                                TRUE ~ "Female") %>%
    base::factor(levels = base::levels(trainSet$sex))

  # metricas de avaliacao: presicao geral
  preGeral <- round(unique(caret::confusionMatrix(dat = yHat,
                                                  reference = tra

  return(preGeral) })
```

Treinamento: Presição Geral

- A presição geral é a proporção geral prevista corretamente.
- Na Figura pode ser visto que a relação entre a precisão geral e o parâmetro p no método *Sample* parece ser fortemente negativa. Isso pode ser devido à prevalência de homens no conjunto de dados. Observe que quando $p = 0,5$, a previsão geral é de $0,5$, ou seja, se está adivinhando. A melhor probabilidade é $p = 0$.



A partir da figura acima, tem-se que a relação entre os cortes (*cutoff*) e a presição geral parece ser representada por uma curva de segundo grau. Atingindo seu máximo (`0.83`) quando `cut = 65`.

Sample vs Cutoff: Teste

Na tabela abaixo, o método de corte de altura funciona melhor. No entanto, deve-se levar em conta que há uma alta predominância de homens, e a pressão geral é sensível à prevalência. Portanto, é importante aplicar outras métricas de avaliação.

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

Sample Method	Cutoff Method
0.77	0.84

Matriz de confusão

- A **matriz de confusão** permite definições precisas de outras métricas de precisão, normalmente possui quatro entradas.

	realmente positivo = 1	**realmente negativo = 0**
Previsão positiva = 1	Verdadeiros positivos (TP)	Falsos positivos (FP)
Negativo previsto = 0	Falsos negativos (FN)	Verdadeiros negativos (TN)

- Logo, as formas de médidas de avaliação mais importantes são dadas por:

Medida de	**Nome**	**Nome 2**	**Definição**	**Representação de probabilidade**
Sensibilidade	TPR	recall	PT/(PT+FN)	Pr(hat(Y)=1 | Y=1)
Especificidade	TNR	1-FPR	TN/(TN+PF)	Pr(hat(Y)=0 | Y=0)
Especificidade	PPV	Precisão	PT/(PT+PF)	Pr(Y=1 | hat(Y)=1)

Estrutura da matriz de confusão: estudo de caso

- Tomando como referência a classe *mulher*, a estrutura da matriz de confusão é dada por:

	realmente mulher = 1	**realmente homem = 0**
Previsão mulher = 1	Verdadeiros positivos (TP)	Falsos positivos (FP)
Previsão homem = 0	Falsos negativos (FN)	Verdadeiros negativos (TN)

- Observe que os valores dentro da matriz são preenchidos de acordo com o algoritmo construído e seus argumentos.

Precisão equilibrada e pontuação

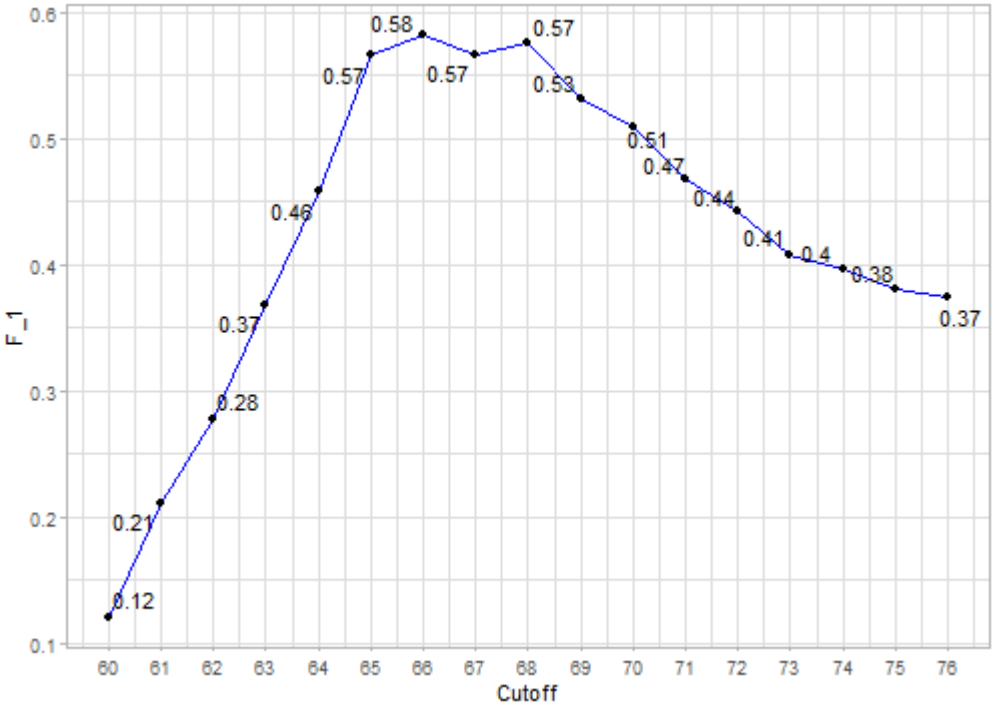
- Esta é uma métrica preferida em *relação* a precisão geral, sendo a média harmônica de sensibilidade e especificidade [1].
- O conceito por trás disso se refere ao fato de que muitas vezes há erros que custam mais do que outros. Por exemplo, em um caso de homicídio criminal, tomar uma decisão com base em um *falso positivo* levaria à execução de uma pessoa inocente. Portanto, é mais importante maximizar a sensibilidade do que a especificidade.
- A função F_1 Score foi adaptada definindo um β para representar a importância da sensibilidade (recall) sobre a especificidade, dada por:

$$\frac{1}{\frac{\beta^2}{1+\beta^2}\frac{1}{\text{recall}} + \frac{1}{1+\beta^2}\frac{1}{\text{precision}}}$$

[1] porque a TPR e TNR são probabilidades.

F1 vs Cutoff: Trainamento

- Considera-se $\beta = 1$ o que leva a encontrar um corte (cut) que equilibre as duas taxas. O resultado indica que o melhor corte na altura para a classe masculina esta dado por `>= 66`.



F1 vs Cutoff: Teste

- A tabela a seguir mostra que o Teste Algorítmico possui sensibilidade (`0,63`) e especificidade (`0,89`) mais equilibradas, sendo ambas relativamente altas. Observe que uma altura maior ou igual a `66` parece mais lógica para prever a classe `Masculino` do que `>= 65`, pois a altura média da amostra em homens é maior.
- Até agora, o método *Cutoff* parece melhor do que o método *Sample*. No entanto, dada a prevalência da classe *Masculino* nos dados, outras métricas devem ser avaliadas antes de indicar a melhor abordagem.

cut	sensibilidade	especificidade
66	0.63	0.89

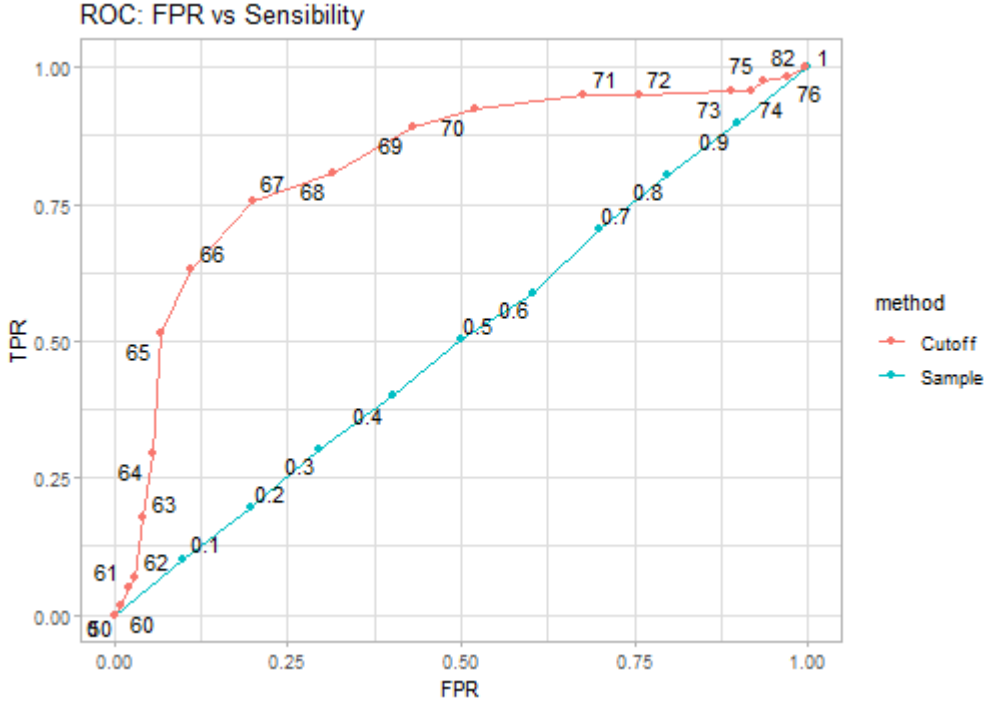
ROC: FPR vs Sensibilidade

- No caso em estudo, a classe positiva (`Y = 1`) refere-se a `Female`, e a negativa (`Y = 0`) a `Male`.
- Neste estudo, a prevalência pode levar a considerar um modelo com menor sensibilidade. Uma vez que na presição geral as poucas previsões positivas corretas são compensadas pelas previsões negativas corretas, ou os altos erros das previsões positivas são compensados pelas previsões negativas corretas.
- Observe que o método *Sample* tem o melhor parâmetro $p = 0$, o que dá uma presição geral mais alta, mas um custo de sensibilidade menor. O que pode estar acontecendo com o método `cutoff`. Assim, resulta conveniente aplicar outras diferentes métricas de avaliação.
- O gráfico ROC[1] permite analisar a perda de sensibilidade em relação à especificidade. Este gráfico é feito abaixo.

[1] *Receive Operating Characteristic*

ROC: FPR vs Sensibilidade

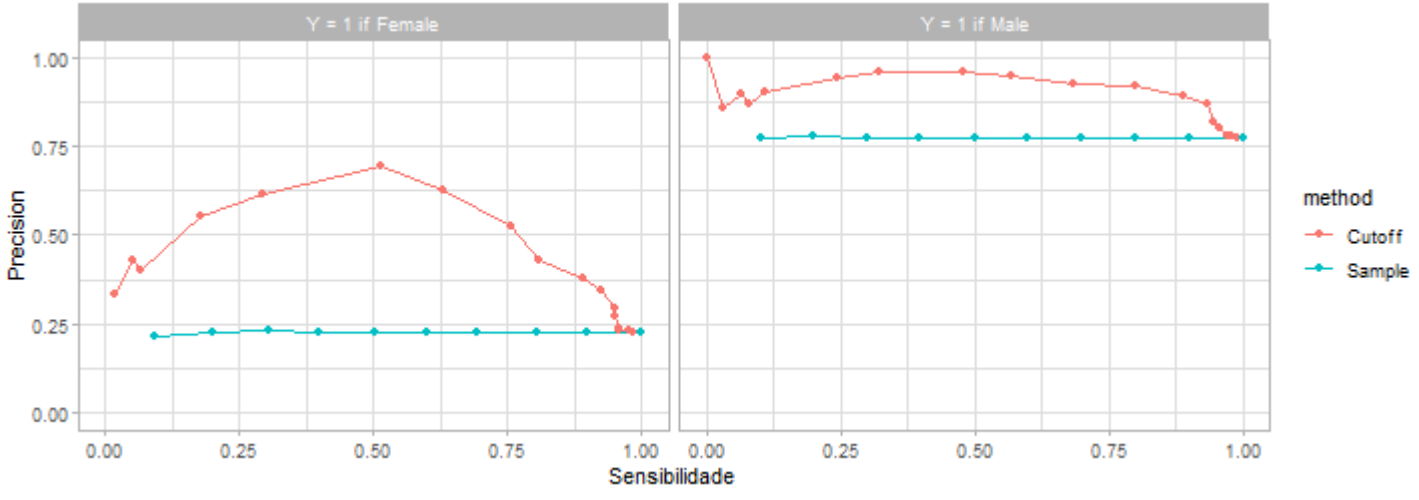
O gráfico mostra dos tipos de relação positiva entre FPR vs. Sensibilidade. O método *Sample* apresenta um crescimento linear próximo à linha identidade, enquanto ao método *Cutoff* parece que uma curva de segundo grau representa a relação entre as variáveis. Note, que o método *Cutoff* tem uma sensibilidade maior para todos os valores de FPR em comparação com o método *Sample*. Logo, o corte de alturas parece uma abordagem melhor.



ROC: Sensibilidade vs Previsão

Agora, como há prevalência de uma classe, se pode usar as métricas de sensibilidade e previsão. Sem importar a classe de referência positiva, o método *Sample* tem indícios de manter uma presição aproximadamente constante para todos os valores de sensibilidade. A presição parece ser maior para o método *Cutoff* para todos os valores da sensibilidade. Pela prevalência da classe *Masculino*.

Quando a designação positiva *Feminino* é alterada para *Masculino*, a sensibilidade é mantida, mas o presição parece aumentar substancialmente, devido ao domínio da classe masculina. No entanto, suspeita-se que a abordagem *Cutoff* seja a melhor.



Referência

- Irizarry, A. R. (2022). *Data Analysis and Prediction Algorithms with R*. Disponível no link
- Wickham H. and Grolemund G. *R for Data Science*. Disponível no link