



Escola Politécnica de Pernambuco
Especialização em Ciência de Dados e Analytics

Estatística Computacional

Aula 3.1 – Modelos de Regressão – PARTE II

Prof. Dr. Rodrigo Lins Rodrigues

rodrigo.linsrodrigues@ufrpe.br

Regressão Logística

2

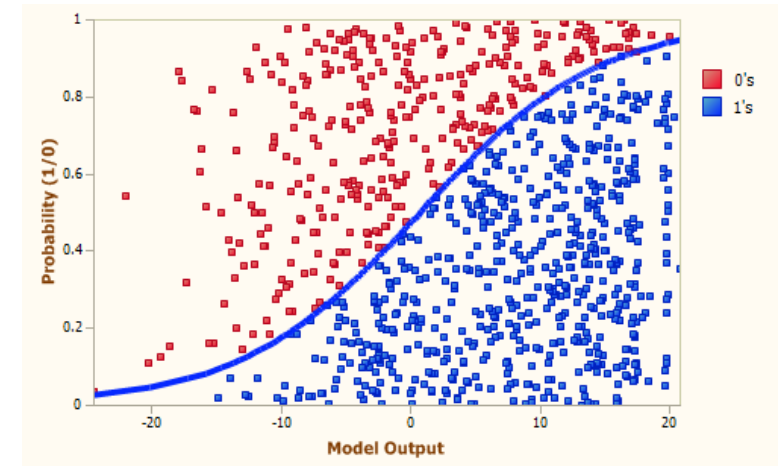
Regressão Logística

...O que seria a
Regressão **Logística**?



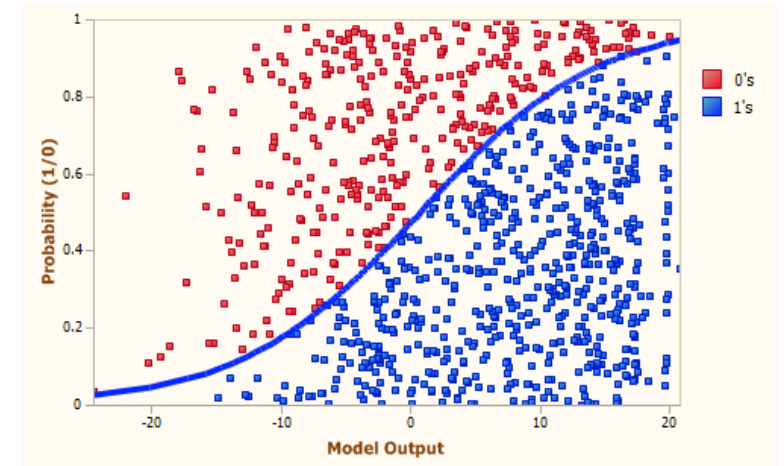
Regressão Logística

- É utilizada para **prever a probabilidade** de um **evento binário** ocorrer.
- Segue a mesma **lógica do modelo de regressão linear** com a particularidade da variável alvo ser binária;
- É uma técnica muito utilizada **quando não se tem bases de dados muito grandes**;



Regressão Logística

- Não tem muitas **exigências de pressupostos**;
- Há uma **infinidade de eventos** de interesse que podem ser modelados pela regressão logística;
- Uma das grandes vantagens é a **flexibilidade de seus pressupostos**, o que amplia sua aplicabilidade.



Regressão Logística

- Deriva seu nome da **transformação *logit*** usada como variável dependente;
- Um modelo é definido como logístico se a função segue a seguinte equação:

$$\text{logit}(p_i) = \ln \left(\frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

Regressão Logística

- Um modelo é definido como logístico se a função segue a seguinte equação:

$$\text{logit}(p_i) = \ln \left(\frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

- Em que p_i indica a probabilidade de ocorrência, x_1, \dots, x_n representa o vetor de variáveis explicativas (ou independentes) e β_0 e β_x indicam os coeficientes do modelo.

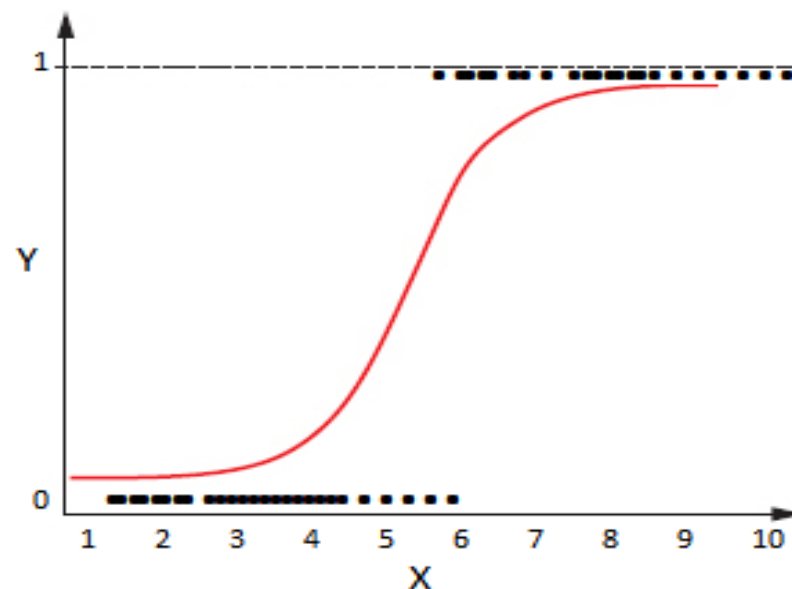
Regressão Logística

- Os coeficientes logísticos **são difíceis de interpretar** em sua forma original, pois são expressos em termos de logaritmos quando usamos a **função *logit***;
- É possível aplicar a **transformação de anti-logaritmo** por meio da **exponenciação dos coeficientes** originais, gerando a razão de desigualdades:

$$\text{Razão de Desigualdades}_i \text{ (odds)} = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n}$$


Regressão Logística

- Para cada observação, é previsto um valor de probabilidade entre 0 e 1;
- Os valores previstos para todos os valores da variável independente gera a curva logística:



Regressão Logística

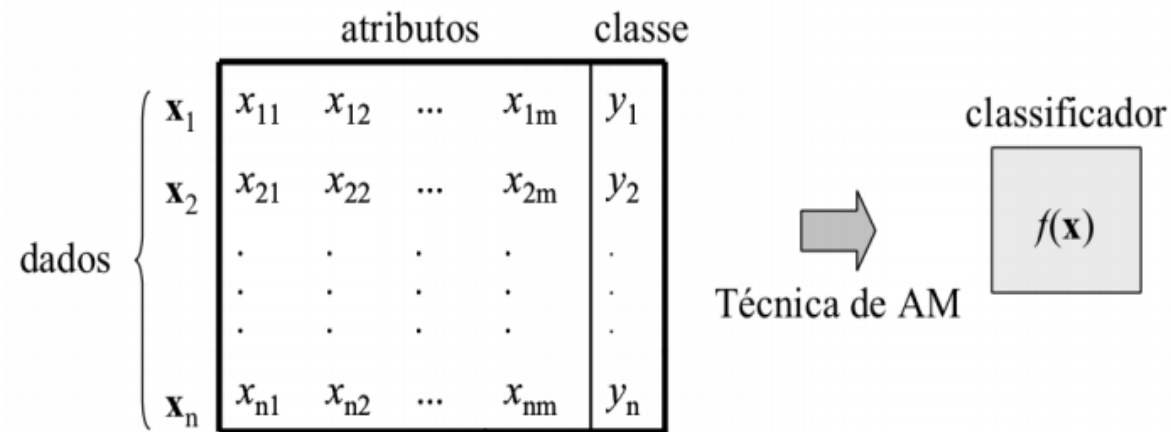
- Se a probabilidade prevista é maior do que 0,50, então a previsão é de que o resultado seja 1 (evento ocorreu);
- Caso a probabilidade prevista seja menor do que 0,50, então a previsão é de que o resultado seja 0 (evento não ocorreu);
- Esse **corte pode ser ajustado** e faz parte das configurações de parâmetros para melhorar o modelo.

A young man with dark hair, wearing a red and white striped shirt, is holding a magnifying glass over a chalkboard. The chalkboard is filled with various mathematical diagrams, including bar charts, line graphs, and geometric shapes. The man has a focused expression, looking through the magnifying glass. A blue semi-transparent banner is overlaid on the bottom half of the image, containing the title in yellow text.

Construção do modelo de Regressão Logística

Construção do modelo

- A construção de um classificador por regressão logística pode ser representado de forma simplificada por:



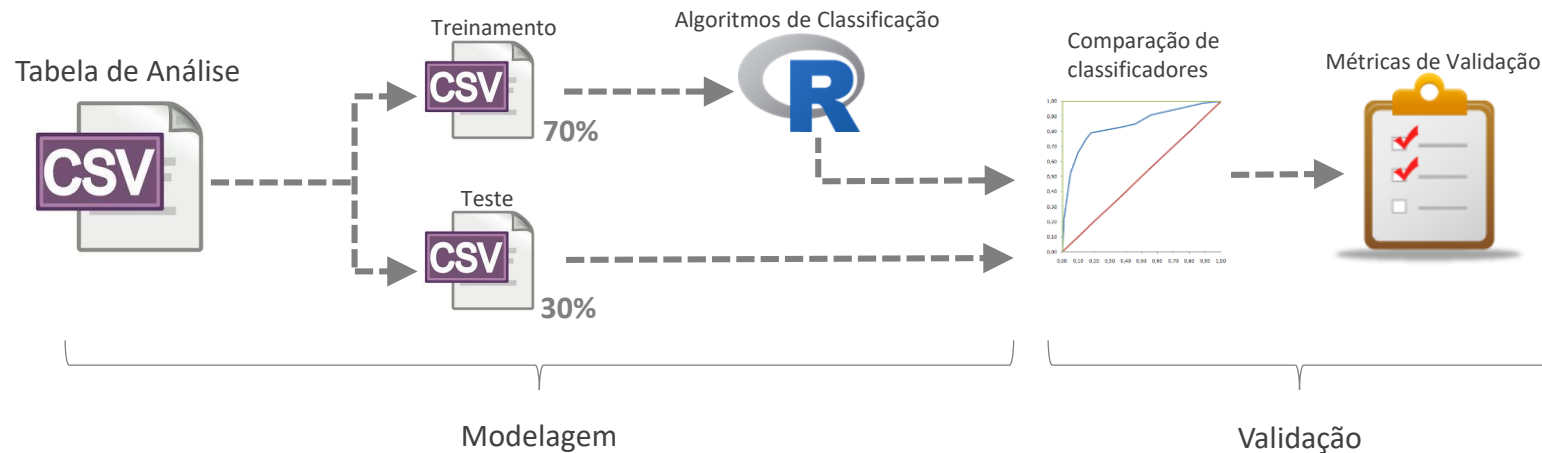
- Conjunto com n dados, onde cada observação x_i possui m atributos e as variáveis y_i representam as classes ou rótulos.

Construção do modelo

- A tarefa de classificação pode ser dividida em duas categorias, a classificação binária e a classificação multiclases.
 - ✓ Regressão Logística Binária;
 - ✓ Regressão Logística Multimodal;
- **Diversos algoritmos** vêm sendo desenvolvidos ao longo de pesquisas;
- A Regressão Logística é um algoritmo **criado pela estatística**.

Construção do modelo

- Construção de um classificador de forma computacional através do software R:



Avaliação do Modelo Logístico



Avaliação do Modelo Logístico

- ✓ Para entender os **erros gerados por um classificador** é possível visualizar por meio da construção de uma matriz de erros denominada **matriz de confusão**;
- ✓ A partir da matriz é possível **obter métricas de qualidade** para a avaliação do desempenho de um classificador;
- ✓ Resume o **número de instâncias previstas corretas ou incorretas** por um modelo de classificação.

Avaliação do Modelo Logístico

✓ Representação da matriz de confusão:

Matriz de Confusão		Classe Atual	
		Negativa (-)	Positiva (+)
Classe Prevista	Negativa (-)	$f -- (TN)$	$f +- (FN)$
	Positiva (+)	$f - + (FP)$	$f ++ (TP)$

Avaliação do Modelo Logístico

- As seguinte **terminologias** são usadas para o entendimento da matriz de confusão:
 - ✓ **Positivo verdadeiro (TP)**: é relacionado ao número de instâncias positivas previstas corretamente pelo classificador;
 - ✓ **Negativo falso (FN)**: é o número de instâncias previstas erroneamente como negativos pelo classificador;
 - ✓ **Positivo falso (FP)**: é o número de exemplos negativos previstos erroneamente como positivos pelo classificador;
 - ✓ **Falso verdadeiro (TN)**: é o número de exemplos negativos previstos corretamente pelo classificador.

Avaliação do Modelo Logístico

- Uma das maneiras mais comuns de **avaliar modelos** é por meio da **derivação de medidas** que, tentam medir a “qualidade” do modelo;
- Essas medidas geralmente podem ser obtidas **a partir da matriz** de confusão:
 - ✓ A **acurácia** é definida como sendo o número de instâncias corretas divididas pelo número total de instâncias;

$$Acurácia = \frac{TP + TN}{(TP + TN + FN + FP)}$$

Avaliação do Modelo Logístico

- A **precisão** determina o percentual de registros que são positivos no grupo que o classificador previu como classe positiva.

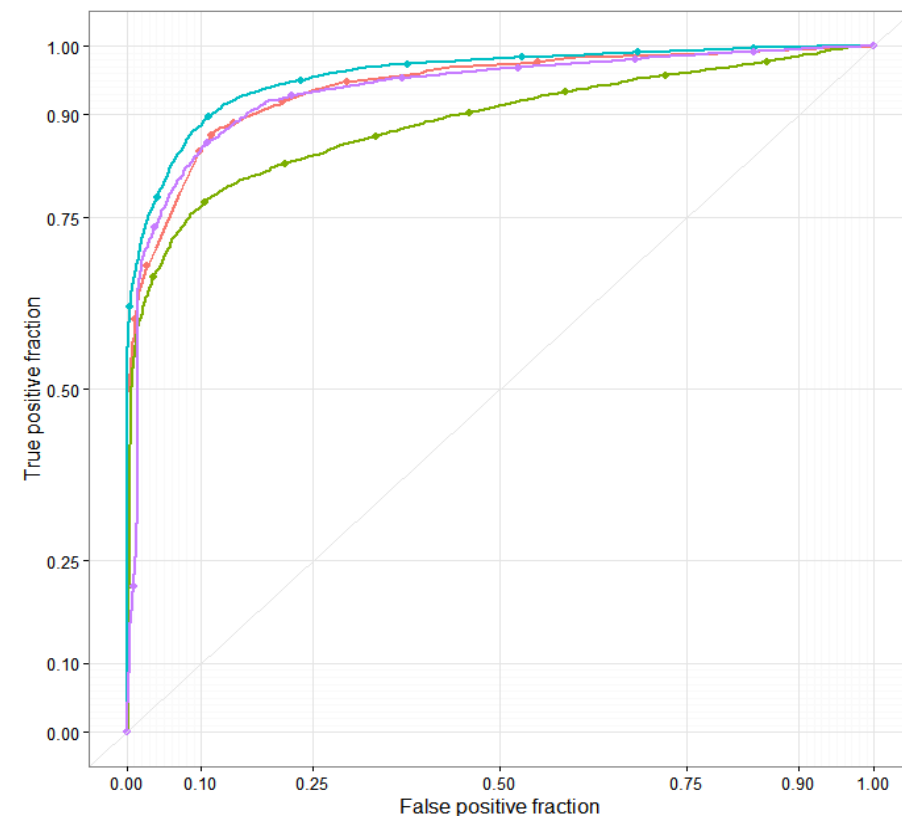
$$\textit{Precisão} = \frac{TP}{TP + FP}$$

- A **lembrança** (*Recall*) mede o percentual de instâncias positivas previstas corretamente pelo classificador.

$$\textit{Recall} = \frac{TP}{TP + FN}$$

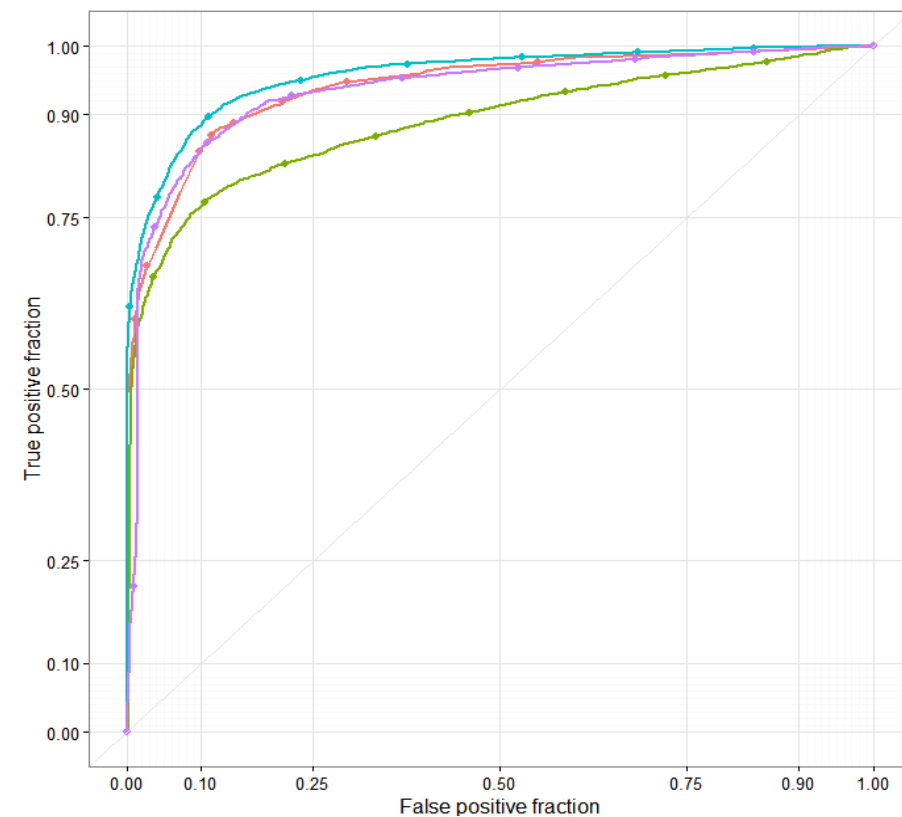
Avaliação do Modelo Logístico

- Área da curva ROC:
 - ✓ Representação gráfica para descrever o desempenho de um sistema **classificador binário**;
 - ✓ É baseada na taxa de verdadeiros positivos TPR , e na taxa de falsos positivos FPR ;



Avaliação do Modelo Logístico

- Área da curva ROC:
 - ✓ É muito utilizada quando queremos **comprar a performance** entre diversos classificadores;
 - ✓ Seus valores variam entre **zero e um**;
 - ✓ Muito utilizada quando se tem **classe desbalanceada**.



Avaliação do Modelo Logístico

- **Utilização de métricas conjuntamente:**

- ✓ É comum encontrar situações em que um modelo aparenta ser melhor que outro **para algumas das métricas**, mas **pior com relação a outras**;
- ✓ Nesses casos, utilizar **uma única medida pode dar a falsa impressão** de que o desempenho pode ser avaliado utilizando-se apenas essa medida;
- ✓ Para uma avaliação mais precisa é ideal que seja utilizado um **conjunto de métricas** levando em consideração o objetivo da pesquisa.



Regressão Logística em R

Regressão Logística em R

- **Exemplo:**

- ✓ Vamos continuar com exemplo relacionado a deslocamento de alunos;
- ✓ Agora temos o interesse em investigar se as variáveis explicativas influenciam a probabilidade de um aluno chegar atrasado na aula;
- ✓ O fenômeno em estudo agora apresenta somente duas categorias (atrasado e não atrasado);
- ✓ O evento de interesse refere-se a chegar atrasado.

Regressão Logística em R

- **Exemplo:**

- ✓ Para realizar esse experimento fizemos uma pesquisa com 100 de uma escola específica;
- ✓ O aluno deveria informar se no dia da pesquisa chegou ou não atrasado;
- ✓ Outras variáveis coletadas foram:
 - Distância percorrida no trajeto (em quilômetros);
 - Número de semáforos pelos quais o aluno passou;
 - Período de realização do trajeto (manhã ou tarde);
 - E como cada aluno considera-se no volante (calmo, moderado ou agressivo);

Regressão Logística em R

- **Exemplo:**

- ✓ Tabela com as respostas de 100 alunos;

Estudante	Chegou atrasado à escola (Y_i)	Distância percorrida (X_{1i})	Quantidade de semáforos (X_{2i})	Período do dia (X_{3i})	Perfil ao volante (X_{4i})
Gabriela	Não	12.5	7	manhã	calmo
Patrícia	Não	13.3	10	manhã	calmo
Gustavo	Não	13.4	8	manhã	moderado
Letícia	Não	23.5	7	manhã	calmo
Luiz Ovídio	Não	9.5	8	manhã	calmo
Leonor	Não	13.5	10	manhã	calmo
Dalila	Não	13.5	10	manhã	calmo
Antônio	Não	15.4	10	manhã	calmo
Júlia	Não	14.7	10	manhã	calmo
...

Regressão Logística em R

- **Exemplo:**

- ✓ Foi necessário aplicar algumas transformações nas variáveis;
- ✓ A variável alvo (Chegou atrasado) recebeu valores de 0 para não chegou e 1 para chegou atrasado;
- ✓ Período do dia ficou: 0 para manhã, e 1 para tarde;
- ✓ A variável (Perfil ao volante) foi transformada em variável binária: 0 calmo e 1 moderado;
- ✓ Criou-se uma nova variável binária informando se o condutor se considera agressivo ou não (1 ou 0);
- ✓ Foi retirado espaços (renomeado) os nomes das variáveis.

Regressão Logística em R

- **Exemplo:**

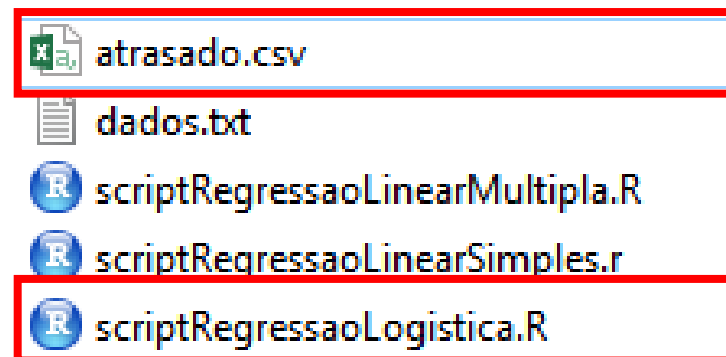
- ✓ A nova tabela de dados ficou da seguinte forma;

Estudante	chegou_atrasado	distancia	quantidade_semaforos	periodo	perfil_volante2
Gabriela	0	12.5	7	1	0
Patrícia	0	13.3	10	1	0
Gustavo	0	13.4	8	1	1
Letícia	0	23.5	7	1	0
Luiz Ovídio	0	9.5	8	1	0
Leonor	0	13.5	10	1	0
Dalila	0	13.5	10	1	0
Antônio	0	15.4	10	1	0
Júlia	0	14.7	10	1	0
...

Regressão Logística em R

- **Exemplo:**

- ✓ Vamos abrir o script “scriptRegressaoLogistica.R”;
- ✓ Abrir a Base “atrasado.CSV”



Regressão Logística em R

- Exemplo:

```
4 # Pacotes necessários
5 library(caret) # Para a Matriz de confusao, acuracia, sensibilidade e especificidade
6 library(ROCR)  # Para a Curva ROC
7
8 # Importando base de dados
9 dados <- read.csv2("atrasado.csv", header = T)
```

```
11 # Separando os dados de treinamento e de teste com o pacote caret
12 split <- createDataPartition(y = dados$chegou_atrasado, p = 0.7, list = FALSE)
13 treinamento <- dados[split,]
14 teste <- dados[-split,]
```

Regressão Logística em R

- Exemplo:

```
16 # Criando o Modelo a partir dos dados de treinamento
17 modelo <- glm(chegou_atrasado~distancia+quantidade_semaforos
18               +periodo+perfil_volante2+perfil_volante3,
19               family=binomial(link="logit"),data=treinamento)
```

Call: glm(formula = chegou_atrasado ~ distancia + quantidade_semaforos +
periodo + perfil_volante2 + perfil_volante3, family = binomial(link = "logit"),
data = treinamento)

Coefficients:

(Intercept)	distancia	quantidade_semaforos
-202.4349	0.1145	22.6367
periodo	perfil_volante2	perfil_volante3
-2.3565	-2.0163	1.9273

Regressão Logística em R

- Exemplo:

```
25 # Aplicando os dados de teste no modelo construido
26 classificacaoProb <- predict(modelo,newdata=teste,type="response")
27 classificacaoBinaria <- ifelse(classificacaoProb > 0.5,1,0)
```

```
29 # Gerando a Matriz de confusao e metricas para a analise do modelo
30 MatrizDeConfusao<- confusionMatrix(data=classificacaoBinaria,
31 reference=teste$chegou_atrasado,positive = "1")
32 print(MatrizDeConfusao)
```

Regressão Logística em R

- Exemplo:

	Reference	
Prediction	0	1
0	12	0
1	1	17

```
Accuracy : 0.9667
95% CI : (0.8278, 0.9992)
No Information Rate : 0.5667
P-Value [Acc > NIR] : 9.527e-07

Kappa : 0.9315
McNemar's Test P-Value : 1

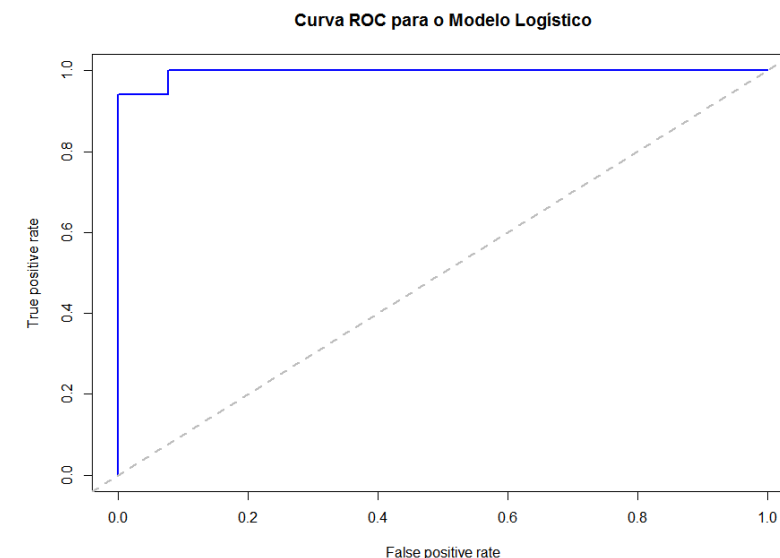
Sensitivity : 1.0000
Specificity : 0.9231
Pos Pred Value : 0.9444
Neg Pred Value : 1.0000
Prevalence : 0.5667
Detection Rate : 0.5667
Detection Prevalence : 0.6000
Balanced Accuracy : 0.9615

'Positive' Class : 1
```

Regressão Logística em R

- Exemplo:

```
34 # Curva ROC
35 predicao <- prediction(classificacaoProb, teste$chegou_atrasado)
36 perform <- performance(predicao, measure = "tpr", x.measure = "fpr")
37 plot(perform, col="blue", lwd=2, main="Curva ROC para o Modelo Logístico")
38 abline(a=0, b=1, lwd=2, lty=2, col="gray")
39 aucFG <- performance(FG, measure = "auc")
40 aucFG <- aucFG@y.values[[1]]
41 aucFG
```



Agora é com vocês !

- Qual a diferença entre a Regressão Linear Múltipla e a Regressão Logística ?
- Quais as formas de validar um modelo logístico ?
- Diga alguns fenômenos que podemos modelar através da regressão logística ?



Dúvidas



- **Contatos:**

- ✓ Email: rodrigo.linsrodrigues@ufrpe.br

- ✓ Facebook: [/rodrigomuribec](https://www.facebook.com/rodrigomuribec)