




# SHIFT

 FIAP





# PYTHON JOURNEY

MACHINE & DEEP LEARNING

# MODELOS CLASSIFICATÓRIOS E PREDITIVOS

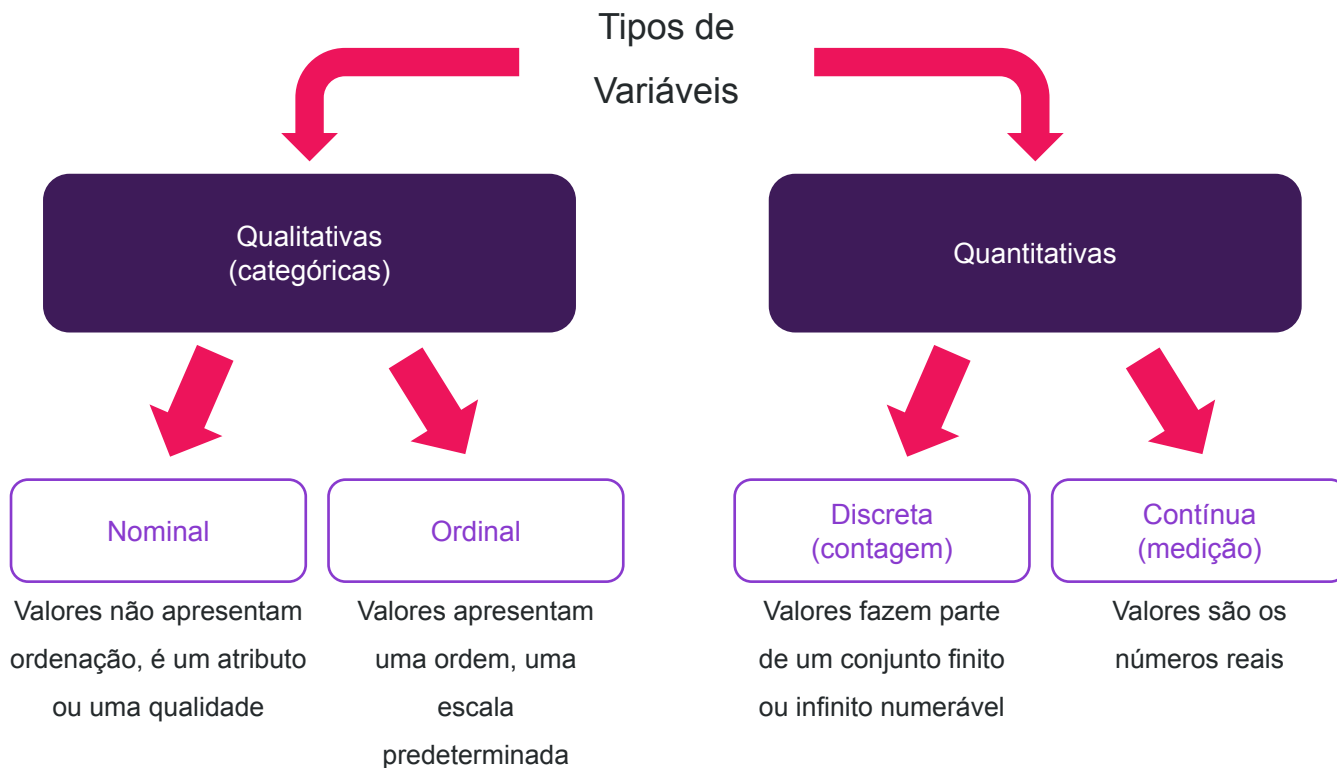
---



PARA RECORDARMOS



# ANÁLISE EXPLORATÓRIA DE DADOS



## PROBABILIDADE DE UM EVENTO OCORRER

---

$$P(A) = \frac{\text{número de elementos de } A}{\text{número de elementos de } \Omega} \Rightarrow P(A) = \frac{n(A)}{n(\Omega)}$$



## PROBABILIDADE DE UM EVENTO OCORRER

---

### Definição


Dados dois eventos A e B, com  $P(A) \neq 0$ , a probabilidade condicional de B, na certeza de A, é o número:

$$P(B | A) = \frac{P(A \cap B)}{P(A)}.$$

Se  $P(B) = 0$ , decretamos  $P(A | B) = 0$ .

É muito comum o uso dessa fórmula para o cálculo de  $P(A \cap B)$ .

Pois,


$$P(A \cap B) = P(A) \cdot P(B | A)$$



## PROBABILIDADE CONDICIONAL

---

Numa caixa, contendo 4 bolas vermelhas e 6 bolas brancas, retiram-se, sucessivamente e sem reposição, duas bolas dessas. Determine a probabilidade de a primeira bola ser vermelha, sabendo que a segunda bola é vermelha.

Solução: Sejam  $A = \{\text{a primeira bola é vermelha}\}$  e  $B = \{\text{a segunda bola é vermelha}\}$ , temos:

$$P(A | B) = \frac{P(A \cap B)}{P(B)}.$$





## TEOREMA DE BAYES

---

O teorema de Bayes é um corolário (consequência imediata de um teorema) do teorema da probabilidade total. E, com ele, é possível o cálculo da seguinte probabilidade:

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$

Em que:

- $P(A)$  e  $P(B)$  são as probabilidades a priori de A e B.
- $P(B|A)$  e  $P(A|B)$  são as probabilidades a posteriori de B condicional a A e de A condicional a B, respectivamente.



## EXEMPLO

A tabela abaixo, dá a distribuição dos funcionários de uma empresa, por sexo e por departamento em que está alocado.

Departamento	Homens (H)	Mulheres (M)	Total
Venda (V)	153	48	201
Operações (O)	161	153	314
Administrativo (A)	18	10	28
Demais (D)	45	25	70
Total	377	236	613

Escolhe-se, ao acaso, um funcionário. Defina os eventos:

H: o funcionário selecionado é do sexo masculino.

V: o funcionário selecionado é do departamento de Vendas.

Note que  $P(H) = 377/613$ ,  $P(V) = 201/613$ , mas, dentre os funcionários do departamento de Vendas, temos que a probabilidade de ele ser do sexo masculino é:  $153/201$ . Isto é,  $P(H|V) = 153/201$ .



# ÁRVORE DE DECISÃO



# ÁRVORE DE DECISÃO

---

- Pontos positivos
  - Facilmente interpretável
  - Fácil de implementar
  - Lida bem com todo tipo de preditor (assimétrico, esparso, contínuo)
  - Realiza seleção de variáveis
- Pontos negativos
  - Alta variância / instabilidade
  - Baixa performance preditiva



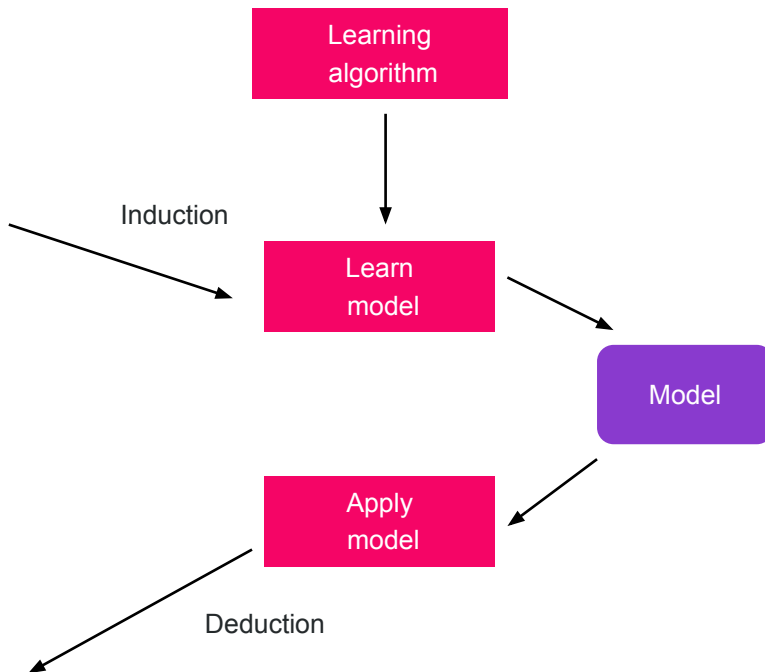
# ÁRVORE DE DECISÃO

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

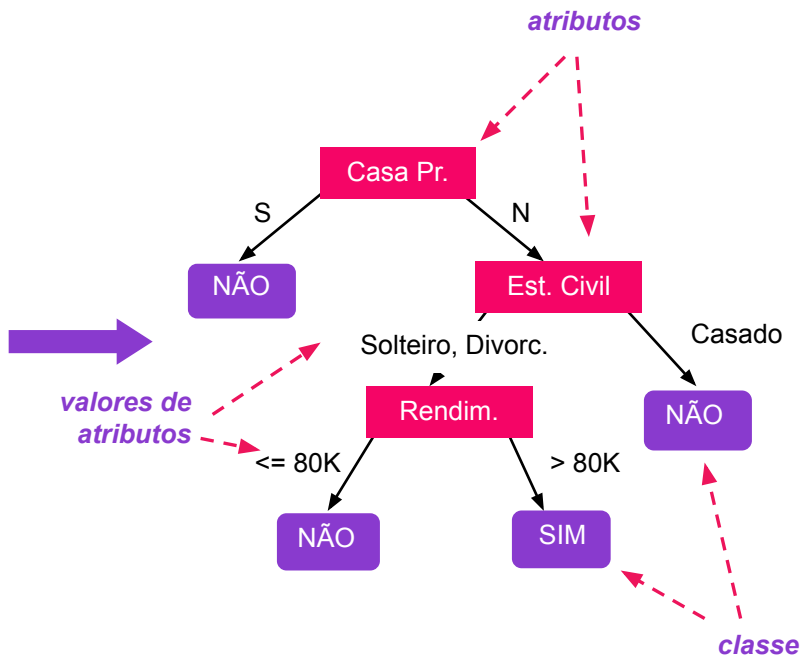


# EXEMPLO

Id	categórico		categórico		continuo	classe
	Casa própria	EstCivil	Rendim.	Mau Pagador		
1	S	Solteiro	125K	NÃO		
2	N	Casado	100K	NÃO		
3	N	Solteiro	70K	NÃO		
4	S	Casado	120K	NÃO		
5	N	Divorc.	95K	SIM		
6	N	Casado	60K	NÃO		
7	S	Divorc.	220K	NÃO		
8	N	Solteiro	85K	SIM		
9	N	Casado	75K	NÃO		
10	N	Solteiro	90K	SIM		

10

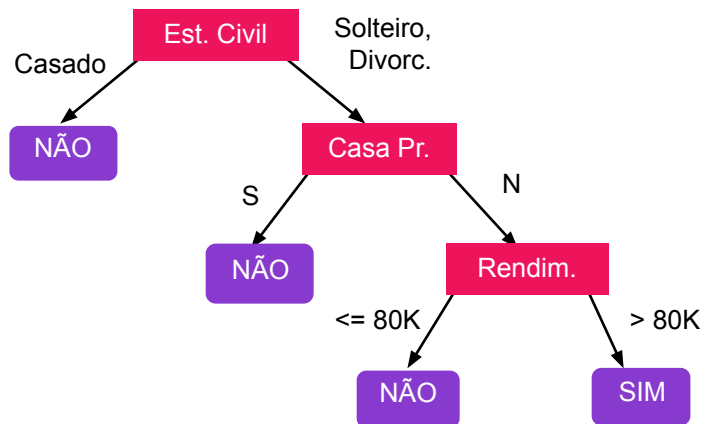
Dados de treinamento



Modelo: árvore de decisão

# EXEMPLO

Id	categórico		categórico		continuo	classe
	Casa própria	EstCivil	Rendim.	Mau Pagador		
1	S	Solteiro	125K	NÃO		
2	N	Casado	100K	NÃO		
3	N	Solteiro	70K	NÃO		
4	S	Casado	120K	NÃO		
5	N	Divorc.	95K	SIM		
6	N	Casado	60K	NÃO		
7	S	Divorc.	220K	NÃO		
8	N	Solteiro	85K	SIM		
9	N	Casado	75K	NÃO		
10	N	Solteiro	90K	SIM		



Pode haver mais de uma árvore para o mesmo conjunto de dados!



## PASSOS PARA A CONSTRUÇÃO DA ÁRVORE DE DECISÃO

---

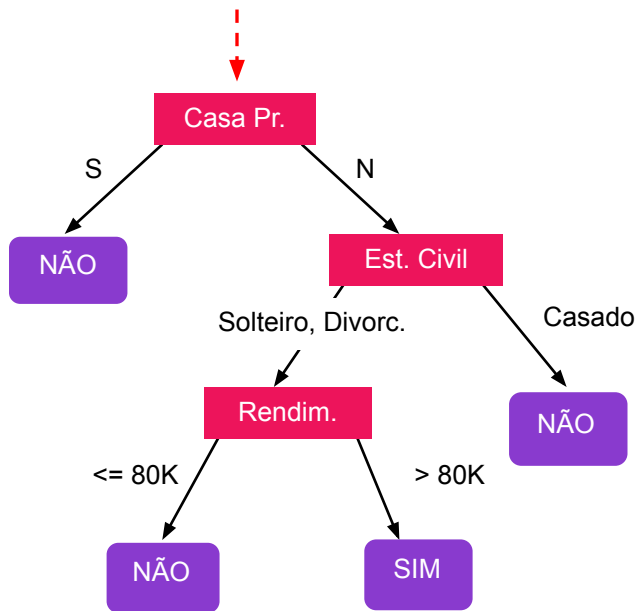
1. Seleciona um atributo como sendo o nodo raiz.
2. Arcos são criados para todos os diferentes valores do atributo selecionado no passo 1.
3. Se todos os exemplos de treinamento (registros) sobre uma folha pertencerem a uma mesma classe, essa folha recebe o nome da classe. Se todas as folhas possuem uma classe, o algoritmo termina.
4. Senão, o nodo é determinado com um atributo que não ocorra no trajeto da raiz, e arcos são criados para todos os valores. O algoritmo retorna ao passo 3.





# PASSO A PASSO

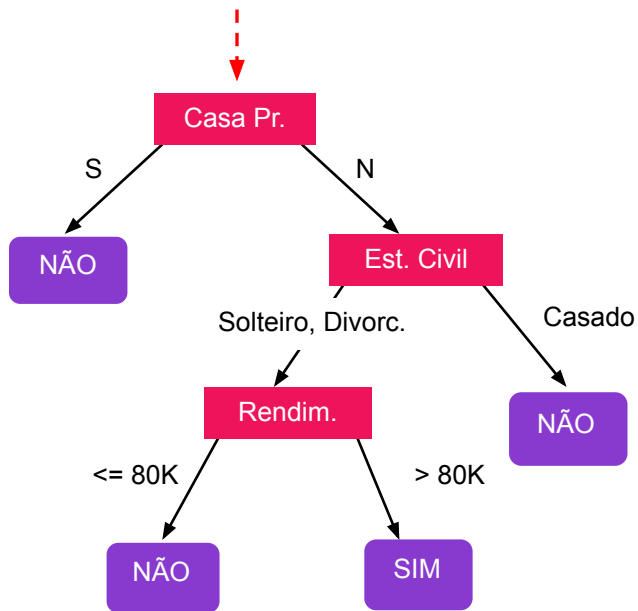
Comece pela raiz da árvore.



Casa Própria	Estado Civil	Rendim.	Mau pagador
N	Casado	80K	?

# PASSO A PASSO

Comece pela raiz da árvore.



Dado para teste

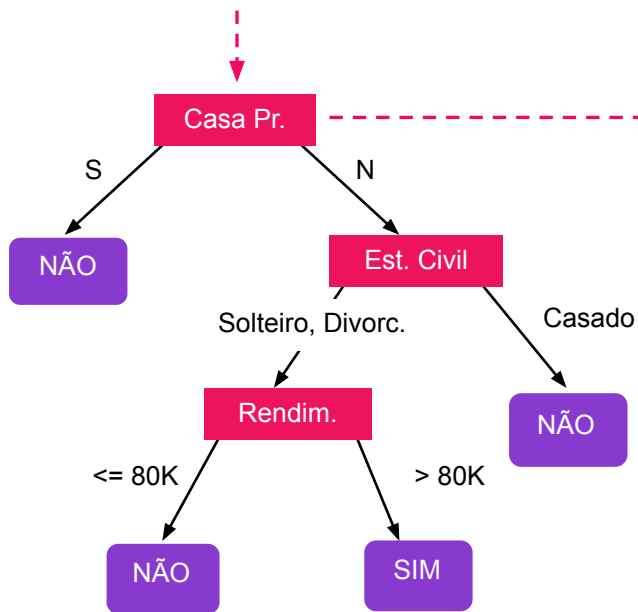
Casa Própria	Estado Civil	Rendim.	Mau pagador
N	Casado	80K	?

# PASSO A PASSO

Comece pela raiz da árvore.

Dado para teste

Casa Própria	Estado Civil	Rendim.	Mau pagador
N	Casado	80K	?

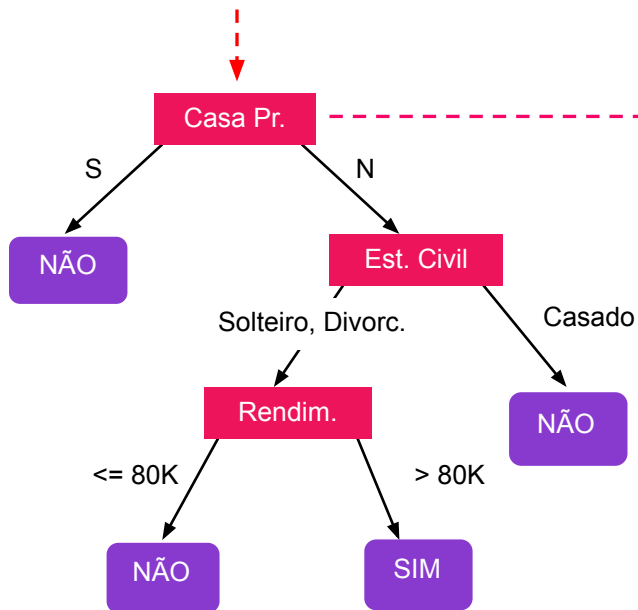


# PASSO A PASSO

Comece pela raiz da árvore.

Dado para teste

Casa Própria	Estado Civil	Rendim.	Mau pagador
N	Casado	80K	?

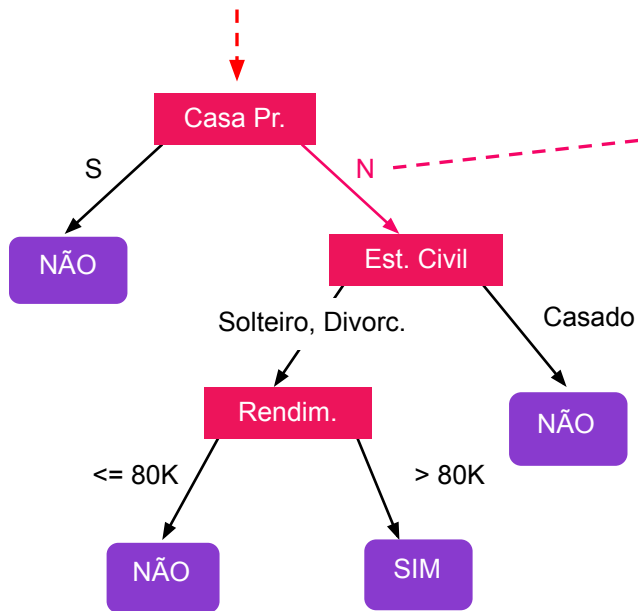


# PASSO A PASSO

Comece pela raiz da árvore.

Dado para teste

Casa Própria	Estado Civil	Rendim.	Mau pagador
N	Casado	80K	?

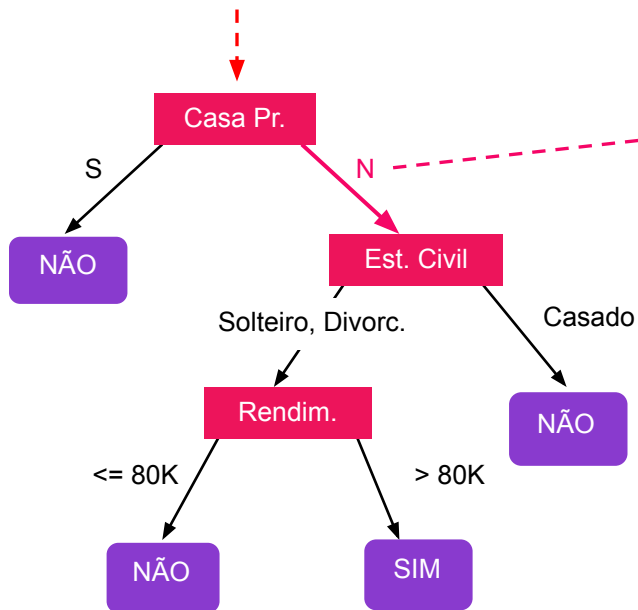


# PASSO A PASSO

Comece pela raiz da árvore.

Dado para teste

Casa Própria	Estado Civil	Rendim.	Mau pagador
N	Casado	80K	?

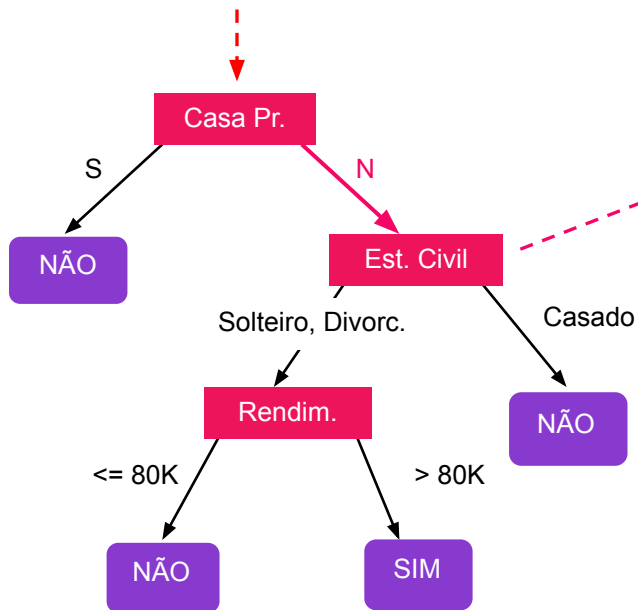


# PASSO A PASSO

Comece pela raiz da árvore.

Dado para teste

Casa Própria	Estado Civil	Rendim.	Mau pagador
N	Casado	80K	?

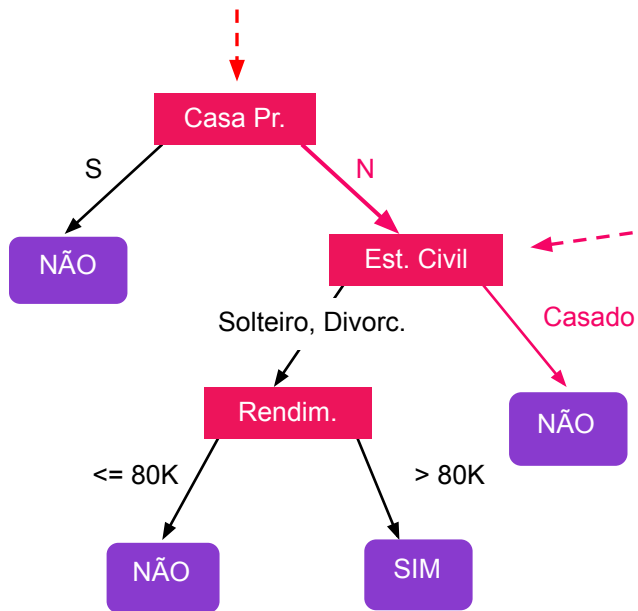


# PASSO A PASSO

Comece pela raiz da árvore.

Dado para teste

Casa Própria	Estado Civil	Rendim.	Mau pagador
N	Casado	80K	?



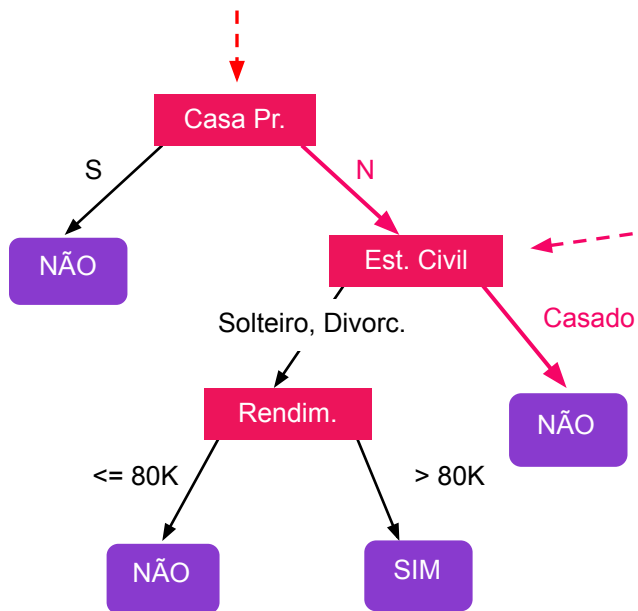


# PASSO A PASSO

Comece pela raiz da árvore.

Dado para teste

Casa Própria	Estado Civil	Rendim.	Mau pagador
N	Casado	80K	?

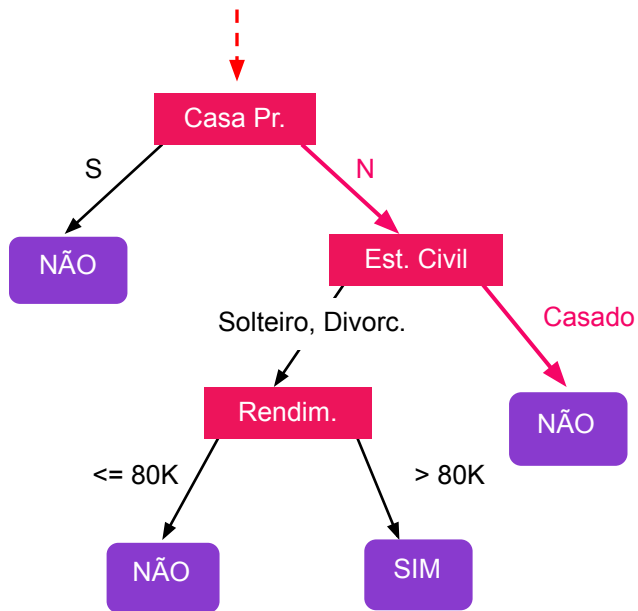


# PASSO A PASSO

Comece pela raiz da árvore.

Dado para teste

Casa Própria	Estado Civil	Rendim.	Mau pagador
N	Casado	80K	?



Atribua à classe (Mau Pagador) o valor NÃO

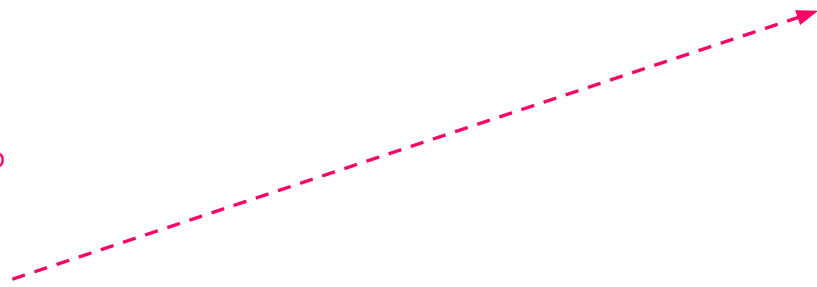
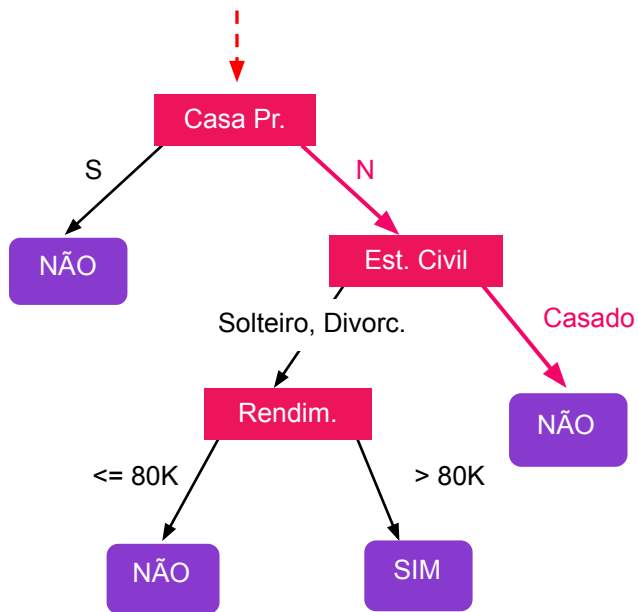


# PASSO A PASSO

Comece pela raiz da árvore.

Dado para teste

Casa Própria	Estado Civil	Rendim.	Mau pagador
N	Casado	80K	Não



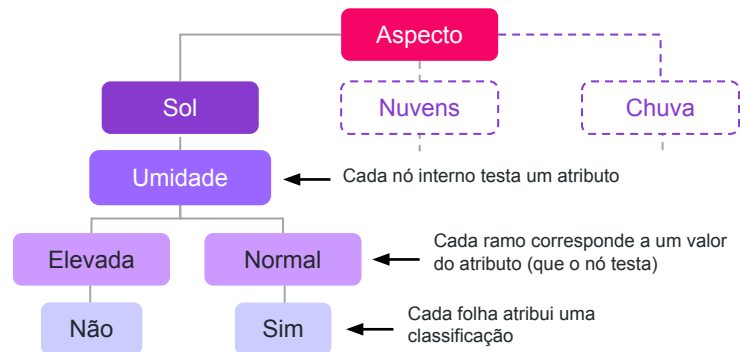
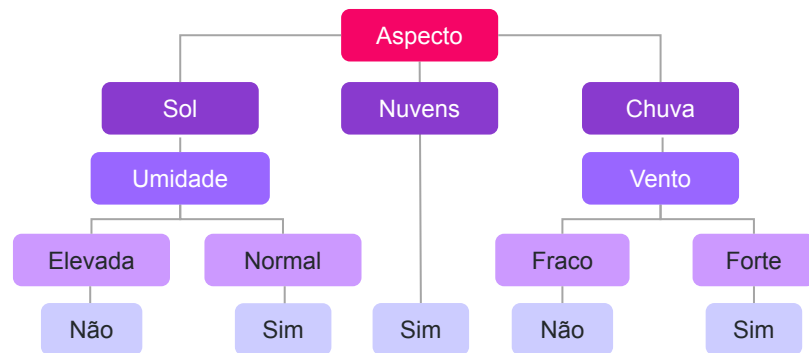
## OUTRO EXEMPLO - EXEMPLOS DE TREINO

Dia	Aspecto	Temperatura	Umidade	Vento	Jogar Tênis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não



## OUTRO EXEMPLO - EXEMPLOS DE TREINO

Dia	Aspecto	Temp.	Umidade	Vento	Jogar Tênis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não



## ESPERAR POR UMA MESA EM UM RESTAURANTE

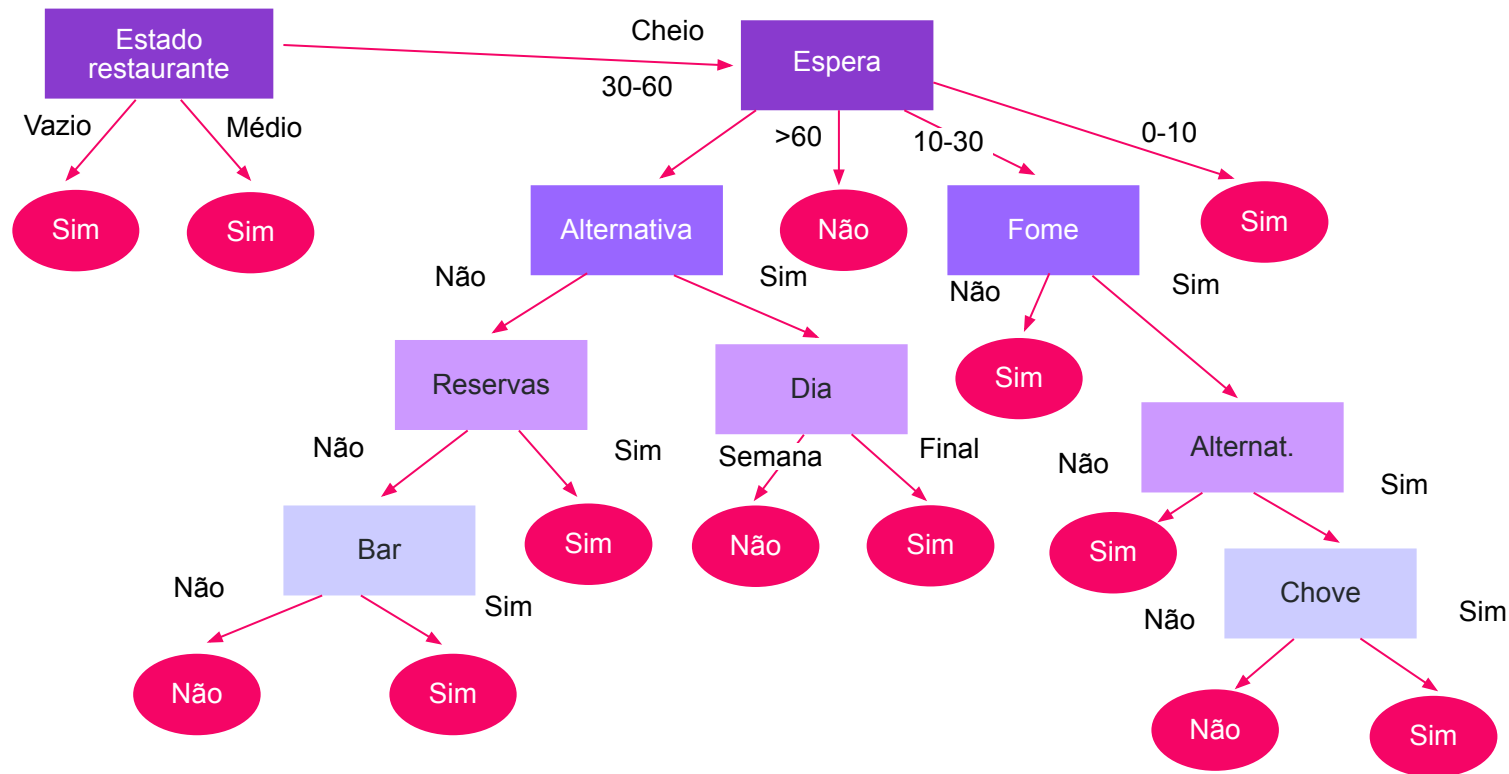
---

Decidir que propriedades ou atributos estão disponíveis para descrever os exemplos do domínio.

Existem alternativas?, Existe um bar no local?, dia da semana, estado da fome, estado do restaurante, preço, chuva, reserva, tipo de comida, tempo de espera...



# ESPERAR POR UMA MESA EM UM RESTAURANTE



# REGRESSÃO LOGÍSTICA





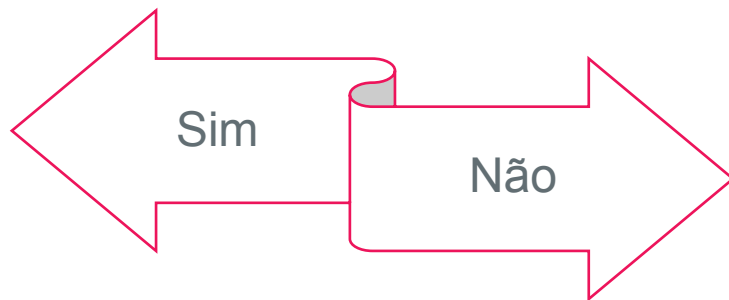
# EXEMPLOS



# ANÁLISE INFERENCIAL BANCO - BACEN

---

Manifestação?



- O que diferencia aqueles que abriram manifestação dos que não abriram?
- Quais são os ofensores?
- Qual a probabilidade de um cliente abrir manifestação dados os ofensores?



# POSSÍVEIS OFENSORES

---

Variáveis avaliadas:

## Empresa

- B – Relacionamento
- B - SAC

## Manifestação do BMG

- Não
- Sim

## Tipo de Contato

- Acompanhamento de proposta
- Atendimento
- Back Office
- Canais de Atendimento
- Características
- Contrato
- Fatura
- Ligação interrompida pelo cliente
- Pesquisa Satisfação
- Prêmio Época Reclame Aqui 2016
- Produtos
- Outros



## POSSÍVEIS OFENSORES

1.487 clientes que abriram manifestação no BACEN entraram em contato com o B antes da abertura.

Quantidade de chamadas realizadas		
Manifestação no BACEN	Média	Desvio-padrão
Não	2	5
Sim	6	14

10.461 chamadas realizadas antes da abertura da manifestação.	63% no mês da abertura
	29% no mês anterior à abertura
	4% nos dois meses anteriores à abertura
	4% nos três ou mais meses anteriores à abertura

- Aparentemente, a quantidade de chamadas realizadas nos dois meses anteriores pode diferenciar os grupos de clientes em relação à abertura de manifestação.



# RESULTADOS DO MODELO:

## CHANCE DE ABERTURA DE MANIFESTAÇÃO NO BACEN

---



### Empresa

Chance 13% menor no  
B – SAC



### Manifestação no B

Chance 600% maior para aqueles que  
abriram manifestação no B



### Quantidade de chamadas nos últimos 2 meses

Chance 2% maior para cada chamada  
a mais



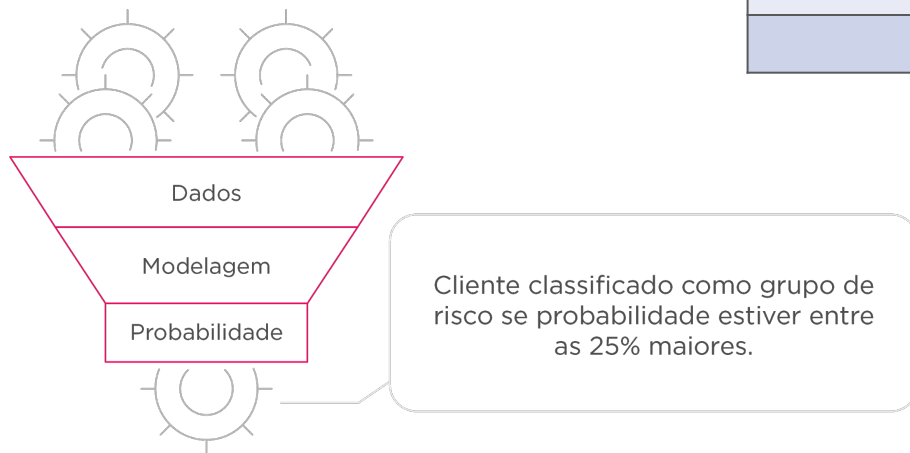
### Tipo de contato

- 147% maior para Back Office
- 82% maior para Ligação Interrompida pelo Cliente
- 79% maior para Contrato
- 73 % maior para Atendimento
- 52% maior para outros tipos
- 32% menor para Canais de Atendimento
- 33% menor para Características
- 75% menor para Fatura
- 86% menor para Produto



# CLASSIFICAÇÃO DOS CLIENTES E ASSERTIVIDADE

Para cada cliente, uma probabilidade  
de abertura de manifestação:



	Grupo de Risco	
Manifestação no BACEN	Não	Sim
Não	43.091	17.266
Sim	79	174

Assertividade:

- Dentro de BACEN “Não”: 71%
- Dentro de BACEN “Sim”: 69%
- Total: 71%



# CONCLUSÃO E MELHORIAS

---

Conclusão	Melhorias
<p>Maiores ofensores da abertura de manifestação no BACEN:</p> <ul style="list-style-type: none"><li>• Abertura de manifestação no B.</li><li>• Tipo de contato: Back Office.</li></ul> <p>Ofensores em menor escala:</p> <ul style="list-style-type: none"><li>• Tipos de contato: Ligação Interrompida pelo Cliente, Contrato e Atendimento.</li><li>• Quantidade de chamadas realizadas no período de dois meses.</li></ul>	<p>Ações que podem levar à melhoria da assertividade do modelo:</p> <ul style="list-style-type: none"><li>• Inclusão de variáveis (exemplo: gênero, idade, escolaridade, estado civil).</li><li>• Diferenciação das chamadas, como Informação, Solicitação e Reclamação.</li><li>• Ligação direta entre as chamadas realizadas no B e as chamadas realizadas no BACEN.</li></ul>



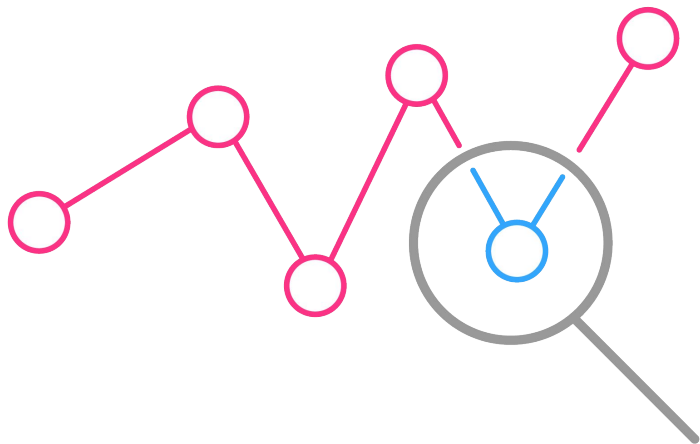
# DISCRIMINAR OS OFENSORES RELACIONADOS AO TURNOVER





# CLASSIFICAÇÃO DE **TURNOVER**

---



- Entender melhor as diferentes distribuições de perfis entre os operadores da empresa.
- Primeiro passo para o desenvolvimento de um modelo preditivo.

Escopo: Buscar entendimento no perfil dos operadores e quais variáveis influenciam seu possível desligamento da operação.



# DISTRIBUIÇÃO DE CARGOS

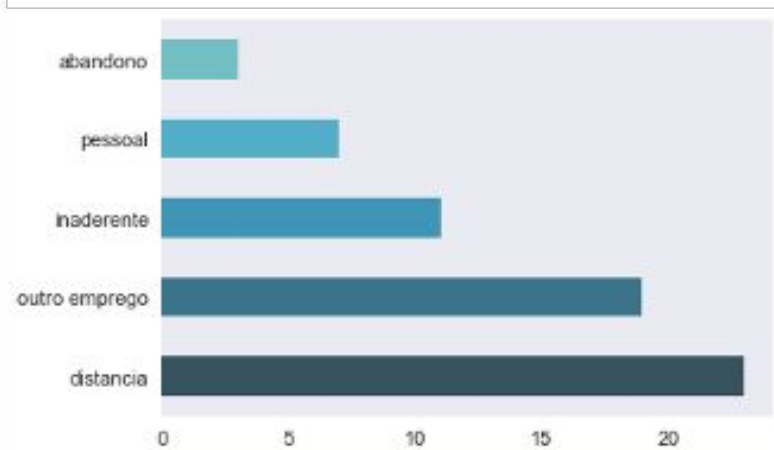
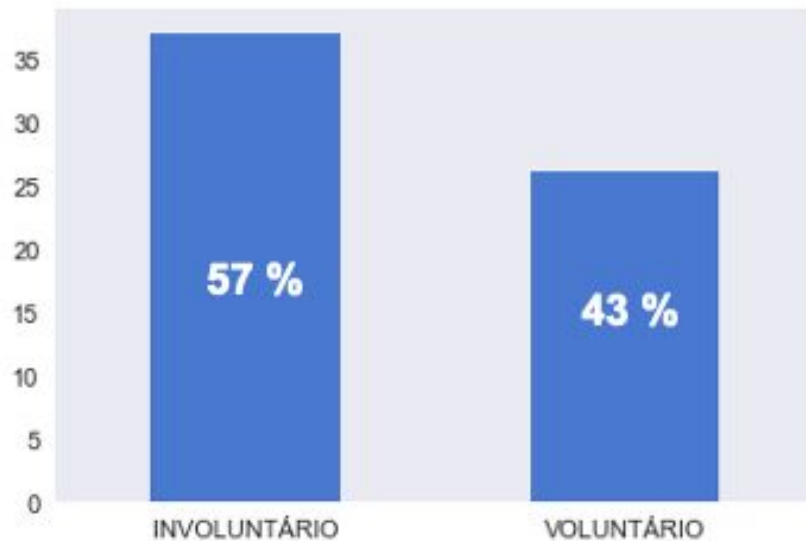
Atual



Desligada



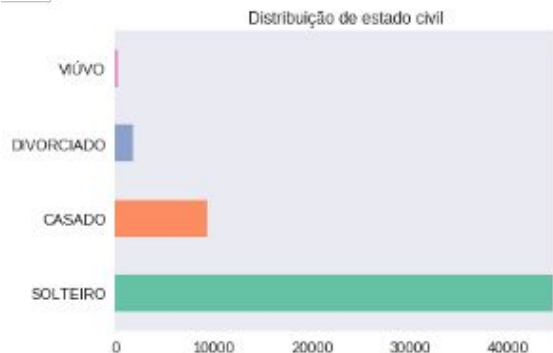
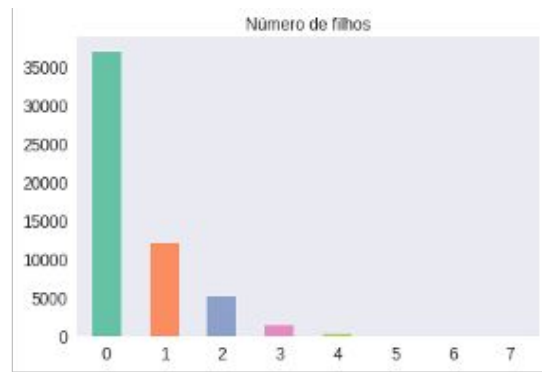
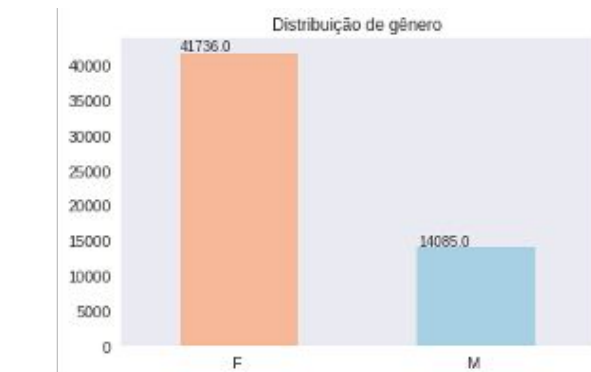
## DISTRIBUIÇÃO DE CARGOS



# PERFIL DOS OPERADORES

## BASE COMPLETA – 60 MIL OPERADORES

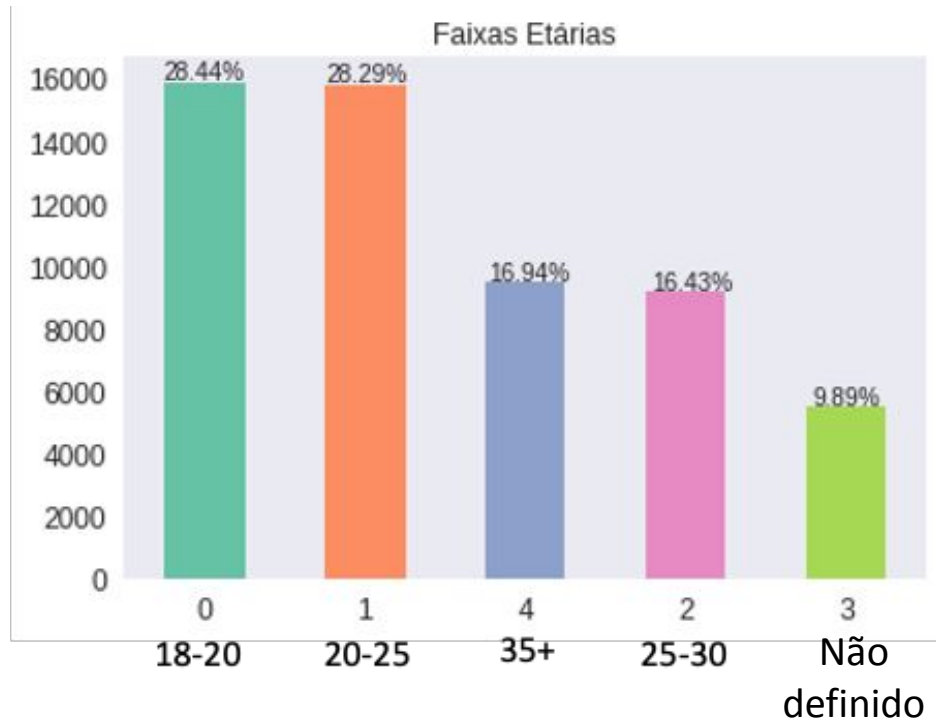
---



# PERFIL DOS OPERADORES

## BASE COMPLETA – 60 MIL OPERADORES

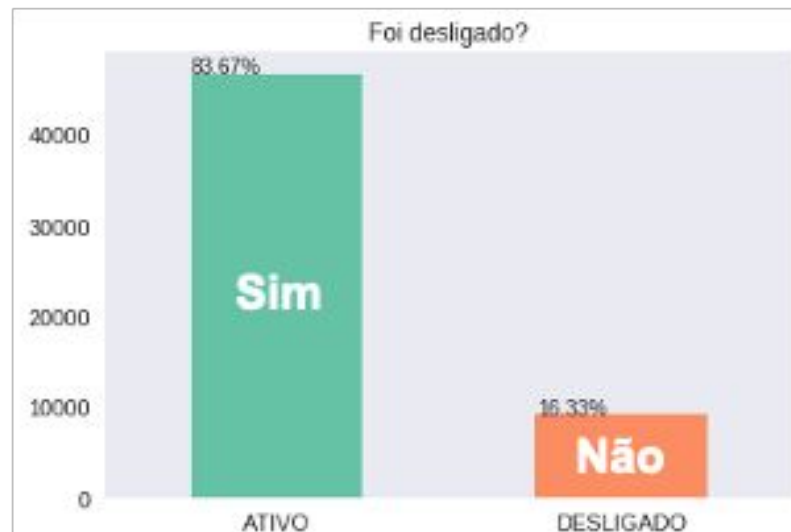
---



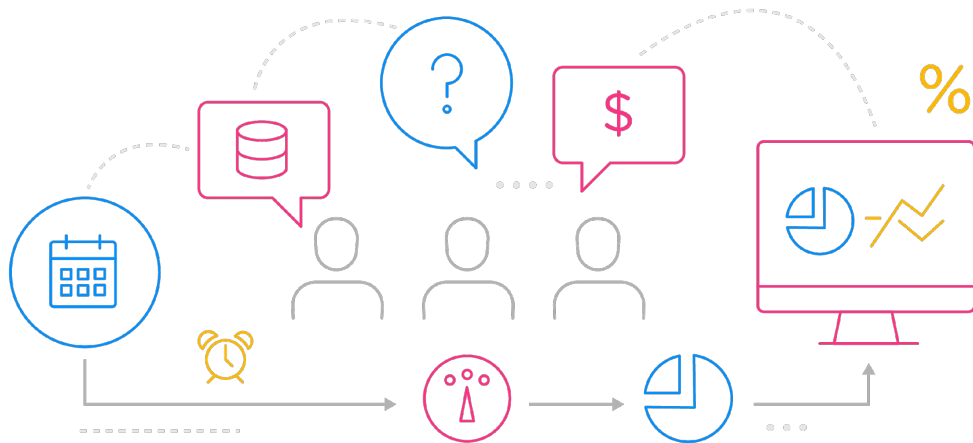
# PERFIL DOS OPERADORES

## BASE COMPLETA – 60 MIL OPERADORES

---



# MODELO PARA TURNOVER



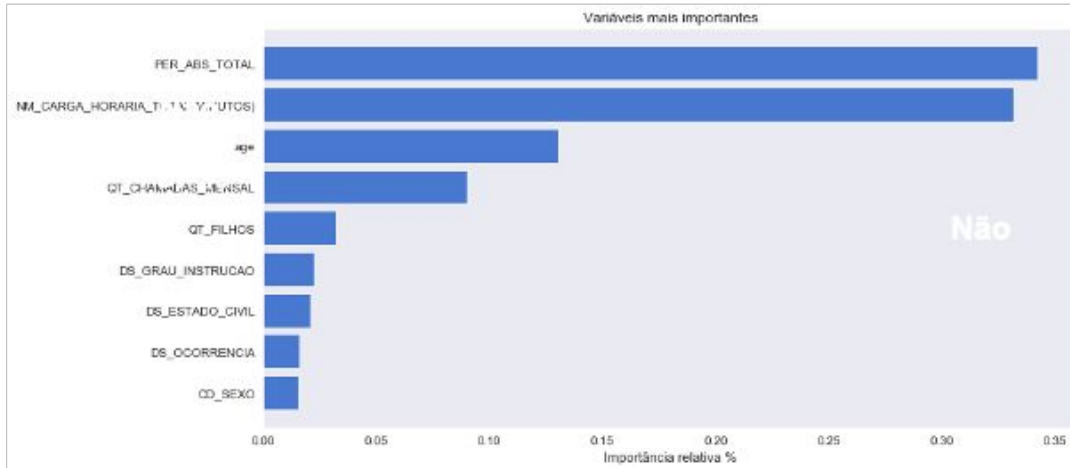
## Objetivos

- Entender, por meio de métodos matemáticos e de machine learning, quais variáveis têm maior influência na saída do operador.
- Mecanismo de previsão de Turnover.



# MODELO PARA TURNOVER

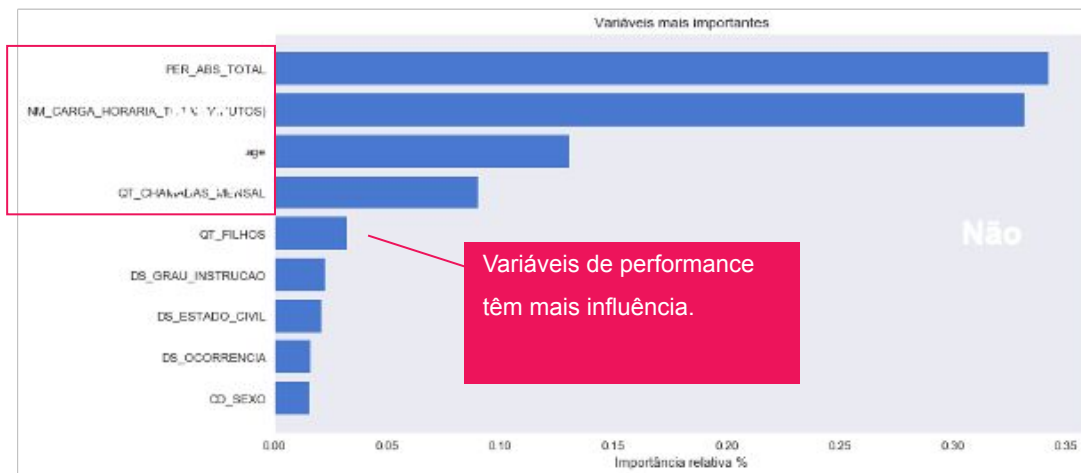
Variáveis de entrada									
Idade	Estado civil	Escolaridade	Número de filhos	Sexo	Carga horária	Absenteísmo total	Chamadas atendidas	Ocorrência disciplinar	Desligado
20, 25, ...	Solteiro Casado	Ens. Médio Ens. Superior	0, 1, 2, ...	M, F	(horas)	% do tempo trabalhado	Volume Mensal	Sim / Não	Sim / Não





# MODELO PARA TURNOVER

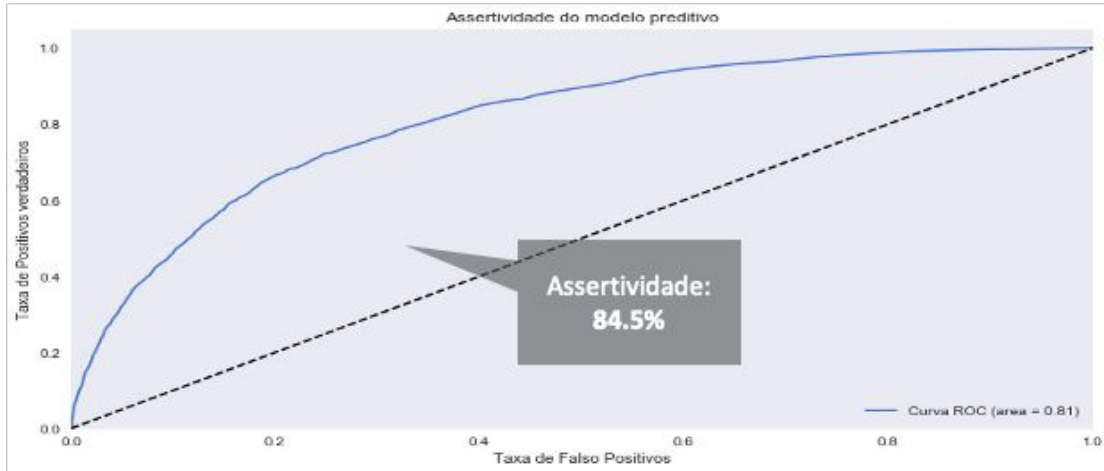
Variáveis de entrada									Desligado
Idade	Estado civil	Escolaridade	Número de filhos	Sexo	Carga horária	Absenteísmo total	Chamadas atendidas	Ocorrência disciplinar	
20, 25, ...	Solteiro Casado	Ens. Médio Ens. Superior	0, 1, 2, ...	M, F	(horas)	% do tempo trabalhado	Volume Mensal	Sim / Não	Sim / Não



# MODELO PARA TURNOVER

## Variáveis de entrada

Idade	Estado civil	Escolaridade	Número de filhos	Sexo	Carga horária	Absenteísmo total	Chamadas atendidas	Ocorrência disciplinar	Desligado
20, 25, ...	Solteiro Casado	Ens. Médio Ens. Superior	0, 1, 2, ...	M, F	(horas)	% do tempo trabalhado	Volume Mensal	Sim / Não	Sim / Não



# RISCO RELACIONADO A AÇÕES JUDICIAIS

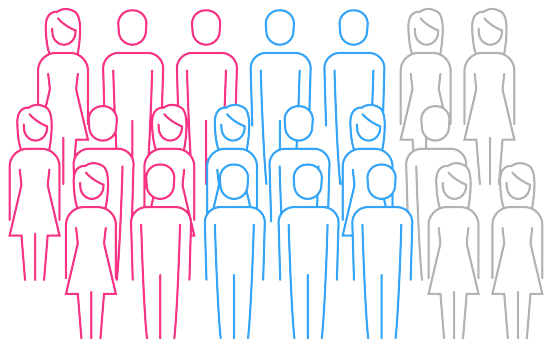


# PROJETO

---

1. Análise do perfil dos funcionários.
2. Estimar a probabilidade de entrada de processo trabalhista para cada funcionário.
3. Estimar a probabilidade de cada tipo de processo trabalhista para cada funcionário.

## Análise do perfil



## Análise de Cluster

Descrição dos grupos de funcionários:

- Idade
- Tempo de casa (meses)
- Site
- Cliente
- Operação: Trade Marketing, Call Center, Back Office e Administrativo



## DADOS DO PROJETO

---

Base geral:

- Dados fornecidos pelo RH.
- 817.363 funcionários.

Base do Jurídico:

- Dados fornecidos pelo Jurídico.
- 23.973 funcionários.

Risco:

- 59.459 ex-funcionários dentro do período legal de abrir processos.



# COMPARAÇÃO: BASE GERAL E BASE DO JURÍDICO

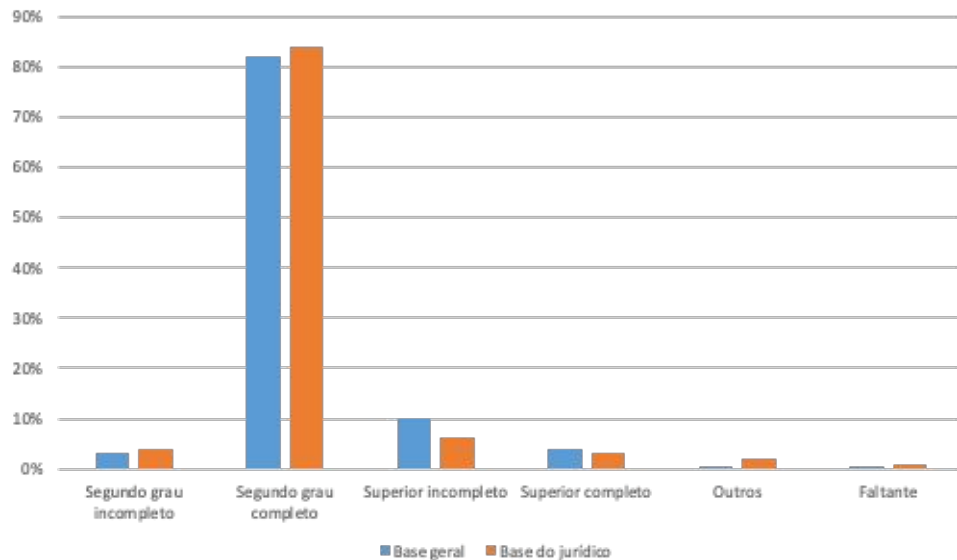
Base geral 817.363 funcionários
<b>Gênero</b> <ul style="list-style-type: none"><li>• 72% Feminino</li><li>• 28% Masculino</li></ul>
<b>Idade:</b> 26 anos
<b>Quantidade de dependentes:</b> 1
<b>Tempo de casa:</b> 1 ano
<b>Estado civil</b> <ul style="list-style-type: none"><li>• 82% solteiros</li><li>• 15% casados</li><li>• 3% outros</li></ul>

Base do Jurídico 23.973 funcionários admitidos
<b>Gênero</b> <ul style="list-style-type: none"><li>• 78% Feminino</li><li>• 22% Masculino</li></ul>
<b>Idade:</b> 29 anos
<b>Quantidade de dependentes:</b> 1
<b>Tempo de casa:</b> 2,5 anos
<b>Estado civil</b> <ul style="list-style-type: none"><li>• 78% solteiros</li><li>• 18% casados</li><li>• 2% divorciados</li><li>• 2% outros</li></ul>



# COMPARAÇÃO: BASE GERAL E BASE DO JURÍDICO

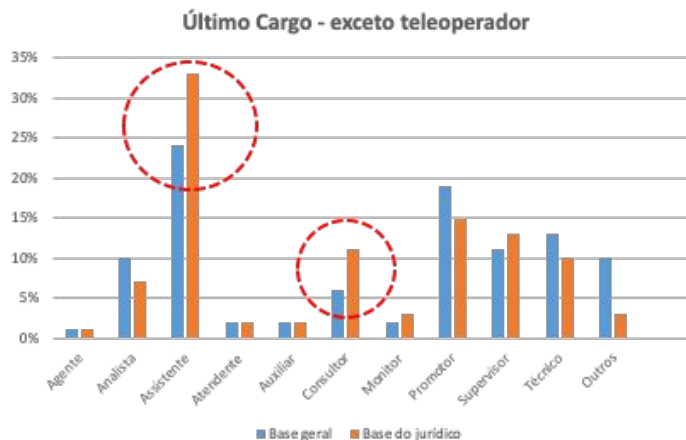
COMPARAÇÃO: BASE GERAL E BASE DO JURÍDICO  
- ESCOLARIDADE



Proporções próximas  
em todas as categorias.

# COMPARAÇÃO: BASE GERAL E BASE DO JURÍDICO – ÚLTIMO CARGO

Base geral	Base do Jurídico
Último cargo	Último cargo
79% Teleoperador	72% Teleoperador
21% Outros	28% Outros



Proporções maiores de  
assistentes e de consultores



## COMPARAÇÃO: BASE GERAL E BASE DO JURÍDICO – ESTADO

	Base geral	Base do Jurídico	Variação
Bahia	6%	6%	0%
Goiás	4%	5%	25%
Minas Gerais	3%	10%	233%
Rio de Janeiro	13%	9%	-31%
Rio Grande do Sul	3%	5%	67%
São Paulo	67%	62%	-7%
Outros	2%	3%	50%

Destaque para Minas Gerais,  
Rio Grande do Sul e Goiás.

	Total na base geral	Total na base do jurídico	Probabilidade estimada de abertura de processo	Variação em relação a SP
Bahia	50836	1497	3%	0%
Goiás	36550	1276	3,50%	17%
Minas Gerais	24534	2292	9%	200%
Rio de Janeiro	108097	2043	2%	-33%
Rio Grande do Sul	26907	1247	5%	67%
São Paulo	550599	14974	3%	-



## COMPARAÇÃO: BASE GERAL E BASE DO JURÍDICO – ESTADO

	Base geral	Base do Jurídico	Variação
Voluntária	45%	29%	-37%
Involuntária	40%	37%	-8%
Justa Causa	4%	19%	367%
Outras	11%	16%	49%

	Base geral	Base do Jurídico	Probabilidade estimada de abertura de processo	Risco	Total
Voluntária	369993	6844	2%	688	34434
Involuntária	325782	8755	3%	635	21184
Justa Causa	32549	4446	14%	462	3307
Outras	89039	3874	4%	21	534



# COMPARAÇÃO: BASE GERAL E BASE DO JURÍDICO – MEDIDAS DISCIPLINARES

	Base geral	Base do Jurídico
Nenhuma	80%	53%
Termo de notificação	2%	1%
Advertência	10%	18%
Suspensão de 1 dia	4%	10%
Suspensão de 2 dias	2%	6%
Suspensão de 3 dias	2%	12%

Tendência de aumento da probabilidade conforme aumenta a gravidade da medida disciplinar

	Total na base geral	Total na base do Jurídico	Probabilidade estimada de abertura de processo	Variação em relação a Nenhuma
Suspensão de 3 dias	18359	2794	15,0%	650%
Suspensão de 2 dias	13278	1366	10,0%	400%
Suspensão de 1 dia	29828	2362	8,0%	300%
Advertência	82804	4392	5,0%	150%
Termo de notificação	13391	304	2,0%	0%
Nenhuma	659703	12755	2,0%	-

# PERFIS GERAIS

## DENTRO DOS GRUPOS

---

### Trade Marketing (Geral e Jurídico: 4%)

- 85% Mulheres
- 87% Solteiros, 12% Casados e 1% Outros
- 25 anos
- 1 dependente
- 1 ano e meio de casa

### Back Office (Geral: 2% - Jurídico: 3%)

- 77% Mulheres
- 79% Solteiros, 19% Casados e 2% Outros
- 30 anos
- 1 dependente
- 4 anos de casa

### Administrativo (Geral e Jurídico: 3%)

- 65% Mulheres
- 66% Solteiros, 31% Casados
- 33 anos
- 2 dependente
- 4 anos de casa

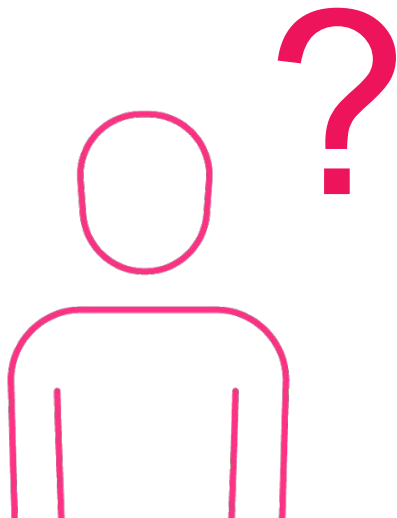
### Call Center (Geral: 91% - Jurídico: 90%)

- 80% Mulheres
- 79% Solteiros, 18% Casados e 3% Outros
- 29 anos
- 1 dependente
- 2 anos e meio de casa



## QUICK WIN

---



### Grupo de análise

Perfil de risco:

- 1.239 ex-funcionários
- 652 funcionários
- Tipo de demissão:
- 1.806 ex-funcionários

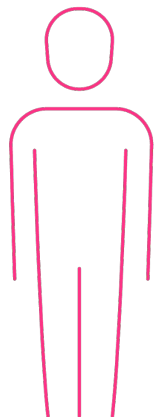
### Possíveis alavancas para abertura de processo trabalhista

- Gênero
- Tempo de casa
- Último cargo
- Estado
- Medidas disciplinares
- Tipo de demissão



# ESTIMAÇÃO

---



- Idade: 34
- Tempo de casa: 2 meses
- Site: Santos
- Cliente: Vivo
- Operação: Call Center
- Cargo: Teleoperador

Probabilidade de entrada de processo trabalhista:  
80%

- Hora extra: 50%
- FGTS: 30%
- Assédio moral: 20%



# MODELO DE INADIMPLÊNCIA



## ESTIMAÇÃO

---

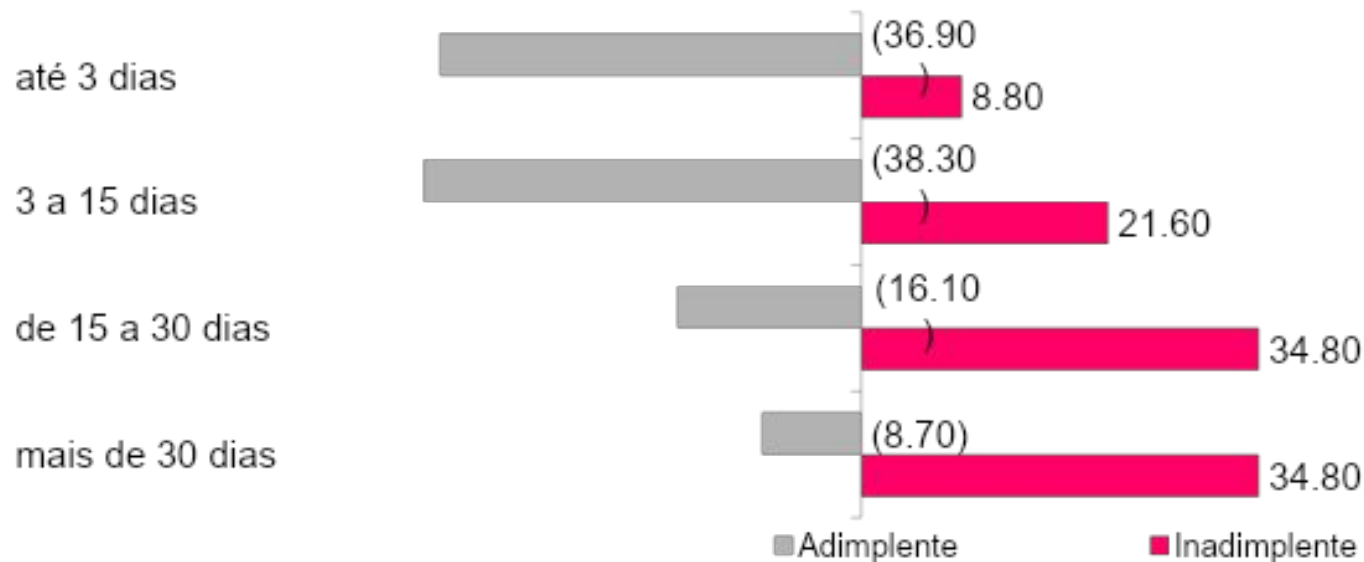
A área de crédito deseja avaliar a propensão ao risco de seus clientes e implementar políticas de redução da inadimplência.





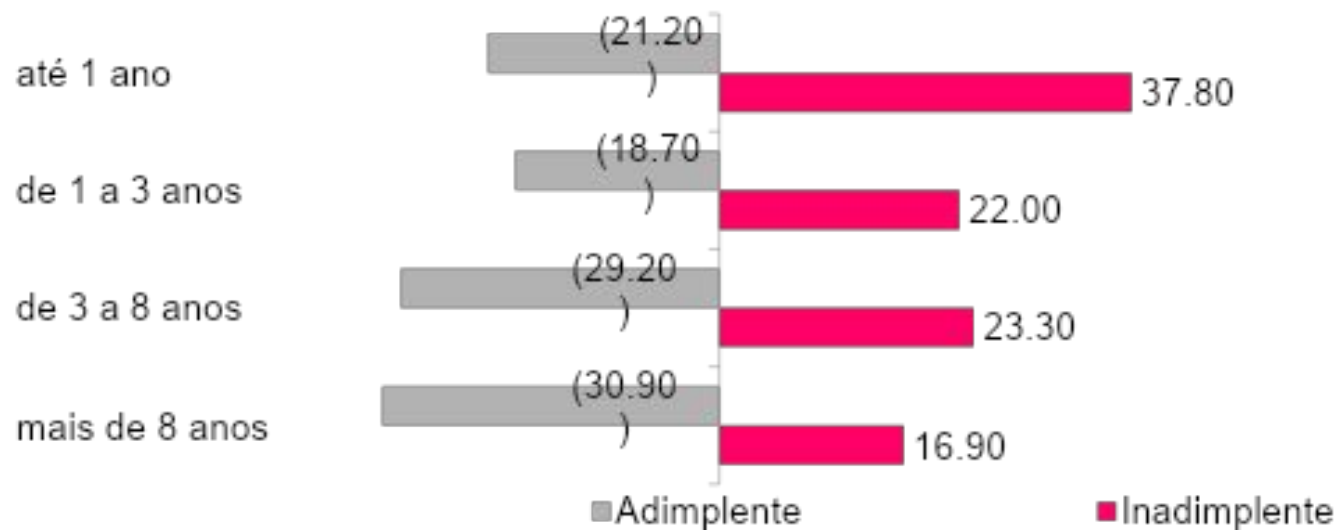
## MODELO DE INADIMPLÊNCIA

Média de dias com pagamentos em atraso nos últimos 6 meses



## MODELO DE INADIMPLÊNCIA

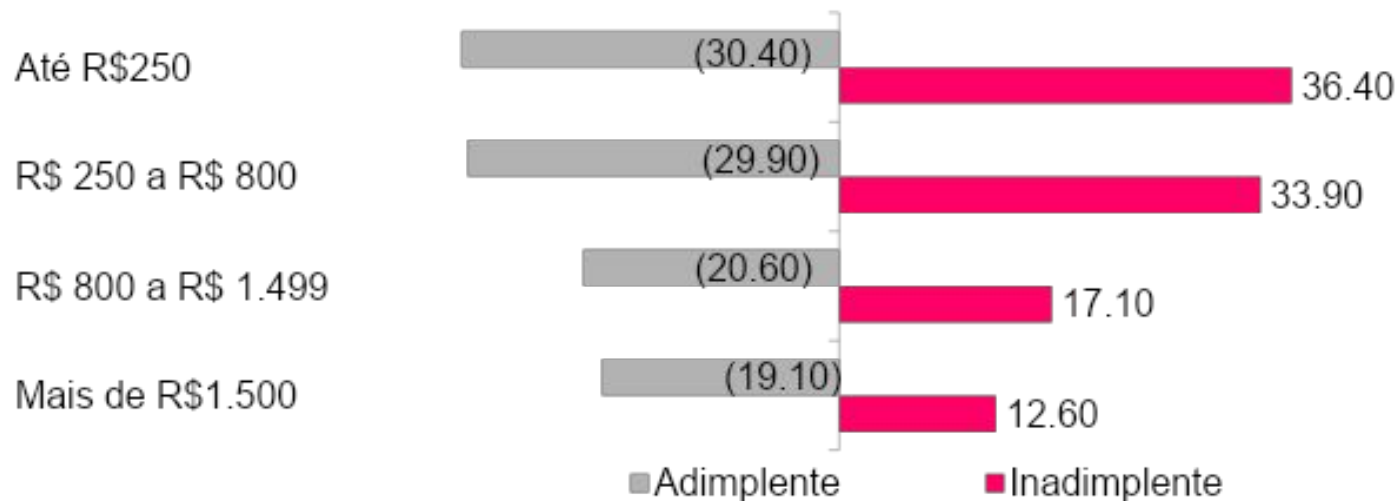
Tempo de relacionamento em anos



## MODELO DE INADIMPLÊNCIA

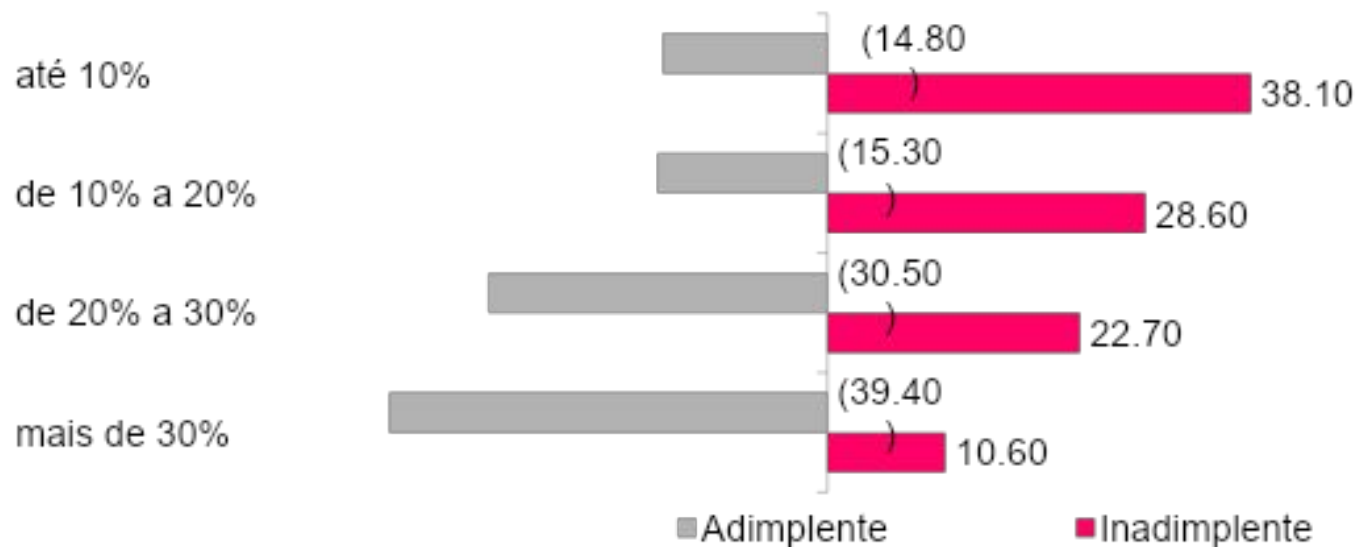
---

Valor médio da fatura mensal



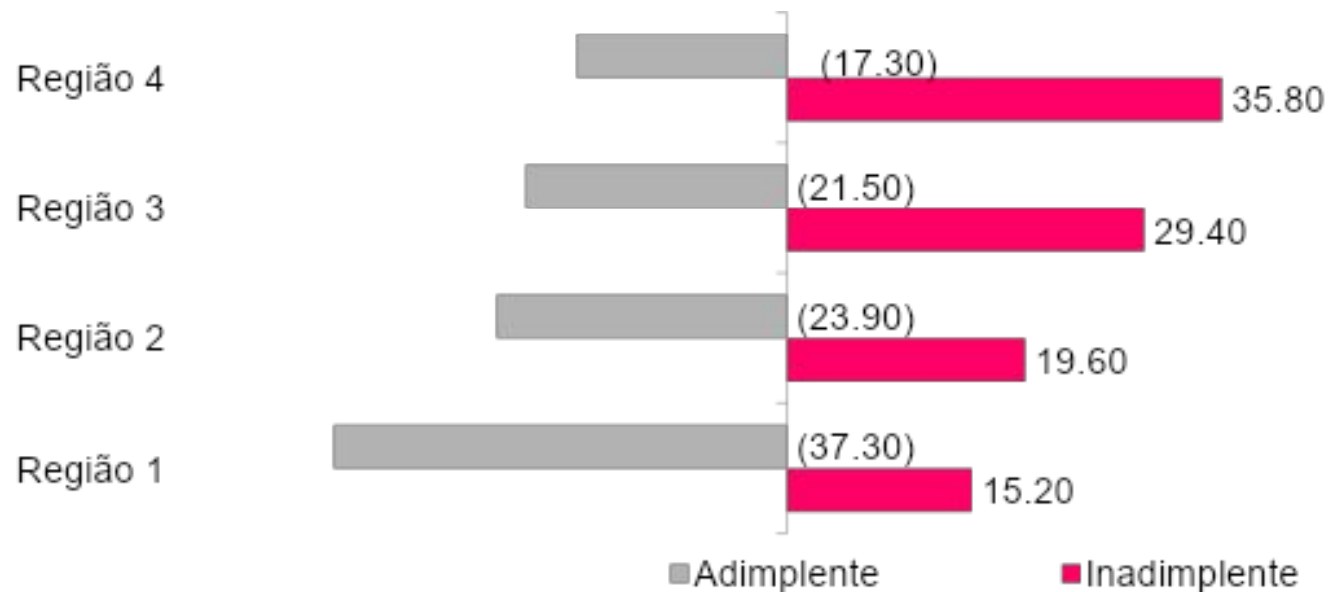
## MODELO DE INADIMPLÊNCIA

Percentual dos gastos em alimentação



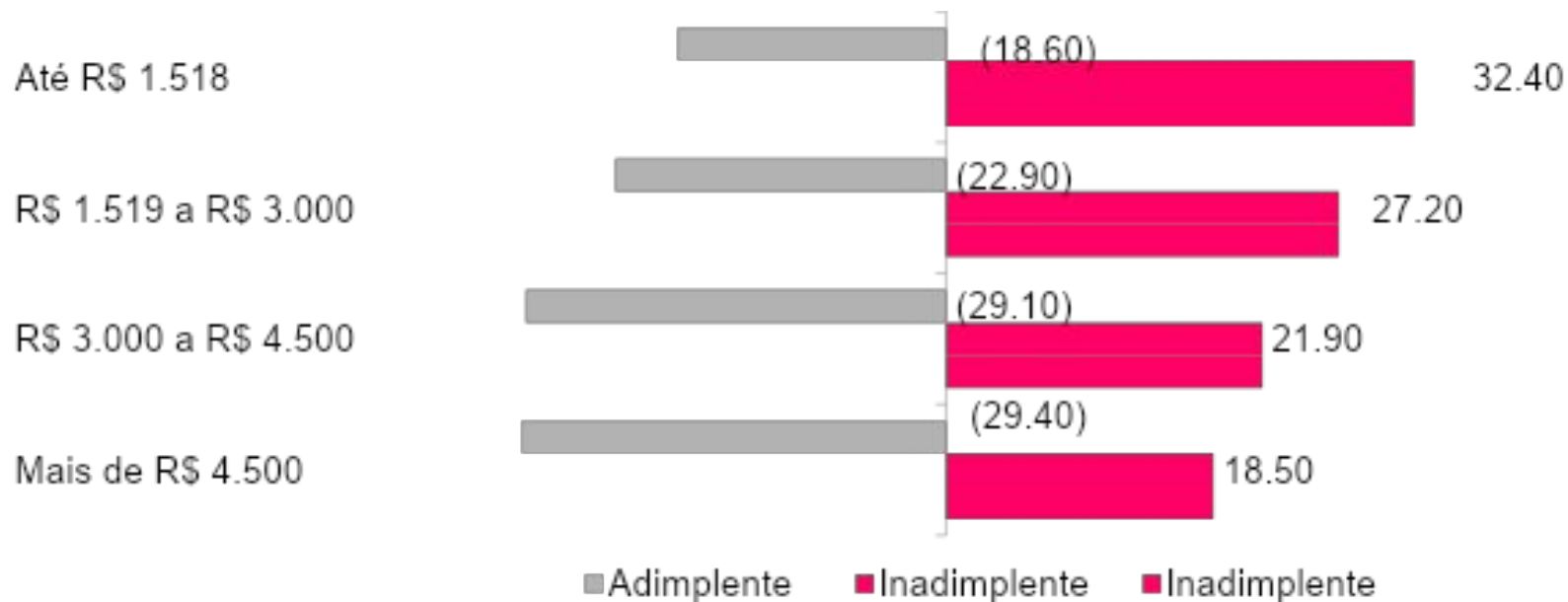
## MODELO DE INADIMPLÊNCIA

Regiões de risco



## MODELO DE INADIMPLÊNCIA

Renda média mensal



# MODELO DE INADIMPLÊNCIA

Tabela de  
coeficientes do  
modelo

Variável	Categoria	Coeficientes
Fatura em atraso	até 3 dias	-1,276
	3 a 15 dias	-0,611
	de 15 a 30 dias	0,580
	mais de 30 dias	1,308
Tempo de cliente	até 1 ano	0,580
	de 1 a 3 anos	0,401
	de 3 a 8 anos	-0,264
	mais de 8 anos	-0,718
Valor da fatura	Até R\$250	0,262
	R\$ 250 a R\$ 800	0,103
	R\$ 800 a R\$ 1.499	-0,105
	Mais de R\$1.500	-0,261
% de gasto com alimentação	até 10%	0,581
	de 10% a 20%	0,401
	de 20% a 30%	-0,264
	mais de 30%	-0,718
Região de risco	Região 4	1,067
	Região 3	0,371
	Região 2	-0,368
	Região 1	-1,069
Renda mensal	Até R\$ 1.518	0,455
	R\$ 1.519 a R\$ 3.000	0,080
	R\$ 3.000 a R\$ 4.500	-0,122
	Mais de R\$ 4.500	-0,413
Constante		0,099

# MODELO DE INADIMPLÊNCIA

## Modelo Logístico

### Pesos definidos na modelagem

-1,276	Até 3 dias	Fatura em atraso	Mais de 30 dias	1,308
-0,718	Mais de 8 anos	Tempo de relacionamento	Até 1 ano	0,580
-0,261	Mais de R\$ 1.500	Valor da fatura	Até R\$ 250	0,262
-0,718	Mais de 30%	% de gasto com alimentação	Até 10%	0,580
-1,069	Região 1	Região de risco	Região 4	1,067
-0,413	Mais de R\$ 4.500	Renda mensal	Até R\$ 1.518	0,455
0,099		Constante		0,099
4%	Propensão			98%





# MODELOS CLASSIFICATÓRIOS E PREDITIVOS CONCEITOS



# REGRESSÃO LOGÍSTICA

---

- Desenvolvida na década de 1960.
- Modelos de regressão não linear são usados, em geral, em duas situações: casos em que as variáveis respostas são qualitativas e os erros não são normalmente distribuídos.
- Verifica a probabilidade de ocorrência do evento de interesse.
- O modelo de regressão não linear logístico binário é utilizado quando a variável resposta é qualitativa com dois resultados possíveis, por exemplo, sobrepeso de crianças (têm sobrepeso ou não têm sobrepeso).



# MODELOS DE REGRESSÃO COM VARIÁVEIS RESPOSTAS BINÁRIAS

---

Em muitos estudos, a variável resposta tem duas possibilidades e, assim, pode ser representada pela variável indicadora, recebendo os valores 0 (zero) e 1 (um).

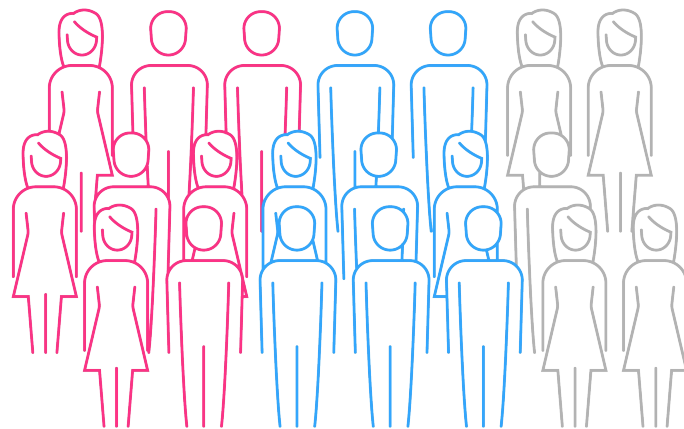
Exemplo:

Num estudo sobre a participação das esposas no mercado de trabalho, como função da idade da esposa, número de filhos e rendimento do marido, a variável resposta  $Y$  foi definida do seguinte modo: a mulher participa do mercado de trabalho ou não. Novamente, essas respostas podem ser codificadas como 1 e 0, respectivamente.



# TÉCNICAS DE DISCRIMINAÇÃO

- Como os *heavy users* **se diferem** em seu perfil demográfico dos *light users*?
- Quais são os clientes ativos que **se assemelham** aos clientes cancelados?
- Que **fatores ou atitudes** fazem com que os meus clientes **prefiram** o meu produto?
- Quais são as **características** que apresentam os clientes que compraram o produto de maior rentabilidade?



GRUPO A

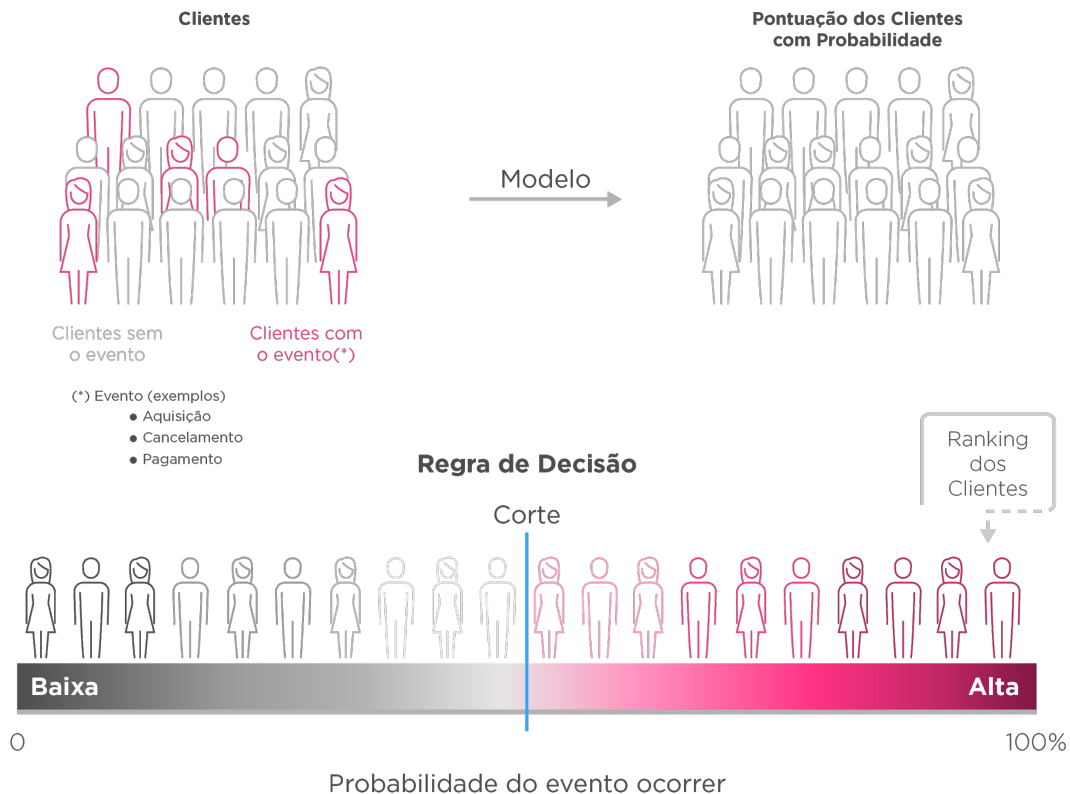
GRUPO B

GRUPO C

Como separar grupos **previamente definidos**? Como definir critérios, funções das variáveis que discriminem os grupos?



# TÉCNICAS DE DISCRIMINAÇÃO



# ANÁLISE DE REGRESSÃO LOGÍSTICA

---

## Probabilidade (lembrando...)

Sendo Y: a resposta à preferência por um evento (sim ou não),

- a probabilidade de:
  - Preferência (*ou sucesso*) será  $p$
  - Não preferência (*de fracasso*) será  $(1 - p)$

## “Chance de Ocorrência de um Evento”

- Chance = (probabilidade de sucesso) / (probabilidade de fracasso)

Exemplo, se a probabilidade de sucesso é 0,65:

a chance é igual a:  $p / (1 - p) = p / q = 0,65 / 0,35 = 1,86$



# ANÁLISE DE REGRESSÃO LOGÍSTICA

**Exemplo:** Preferência por canal de futebol

Gênero	Prefere	Não prefere	Total
Masculino	146	120	266
Feminino	110	124	234
Total	256	244	500

- **Chance** de preferir o canal de futebol entre **homens**:

$$- p1 / (1 - p1) = (146/266) / (120/266) = 0,55 / 0,45 = 1,22$$

- **Chance** de preferir o canal de futebol entre **mulheres**:

$$- p2 / (1 - p2) = (110/234) / (124/234) = 0,47 / 0,53 = 0,89$$

- **Razão de chances** de preferir canal de futebol **entre homens, em relação às mulheres**:

$$- [p1 / (1 - p1)] / [p2 / (1 - p2)] = 1,22 / 0,89 = 1,37$$



# ANÁLISE DE REGRESSÃO LOGÍSTICA

## Modelo de Regressão Logística

$$G = a + B_1 X_1 + B_2 X_2 + \dots + B_n X_n$$

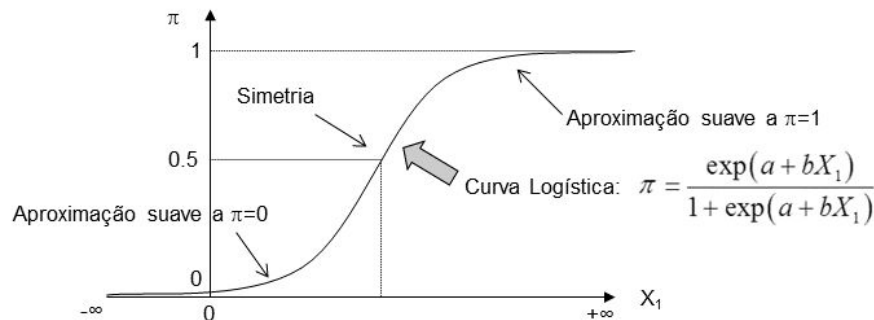
G: logit da resposta de preferência (sim)

a: Intersecção

B<sub>1</sub>, B<sub>2</sub>, ..., B<sub>n</sub>: coeficientes logísticos

- A função logística é dada pelo logito-inverso (anti-logit) que nos permite transformar o logito em probabilidade:

$$p = \frac{\exp(x)}{1 + \exp(x)}$$





# ANÁLISE DE REGRESSÃO LOGÍSTICA

---

## Seleção Conjuntos de Atributos (Variáveis)

- Variáveis Discriminantes
- Variáveis Não Discriminantes

## Instrumento para selecionar variáveis (atributos) significativos

### BACKWARD FORWARD STEPWISE

- **Backward Selection**: Procedimento constrói **adicionando todas as variáveis** e vai eliminando iterativamente uma a uma até que não haja mais variáveis.
- **Forward Selection**: Procedimento constrói iterativamente **adicionando variáveis uma a uma** até que não haja mais variáveis preditoras.
- **Stepwise**: Combinação de Forward Selection e Backward Elimination. Procedimento constrói iterativamente uma sequência de modelos pela adição ou remoção de variáveis em cada etapa.



# QUALIFICAÇÃO DO AJUSTE DO MODELO

## MEDIDAS DE AVALIAÇÃO

---

		Classe Predita	
		Positivo	Negativo
Classe Esperada	Positivo	Verdadeiros Positivos (VP)	Falsos Negativos (FN)
	Negativo	Falsos Positivos (FP)	Verdadeiros Negativos (VN)

# QUALIFICAÇÃO DO AJUSTE DO MODELO

## MEDIDAS DE AVALIAÇÃO

---

- Sensibilidade ou taxa de verdadeiros positivos:  $(VP / (VP + FN))$
- Especificidade ou taxa de verdadeiros negativos:  $(VN / (FP + VN))$
- Taxa de falsos positivos: % de falsos positivos dentre todos em que a classe esperada é a classe negativa  $(FP / (VN + FP))$
- Taxa de falsas descobertas: % de falsos positivos dentre a classe esperada é a classe positiva  $(FP / (VP + FP))$
- Preditividade positiva ou precisão: % de acertos ou verdadeiros positivos  $(VP / (VP + FP))$
- Preditividade negativa: % de verdadeiros negativos dentre todos classificados como negativos  $(VN / (VN + FN))$
- Acurácia: É a proporção de predições corretas, sem considerar o que é positivo e o que é negativo e, sim, o acerto total. É dada por:  $(VP + VN) / (VP + FN + FP + VN)$

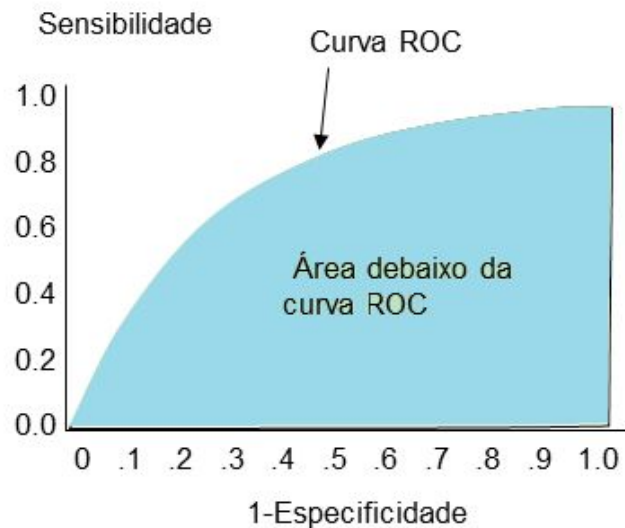
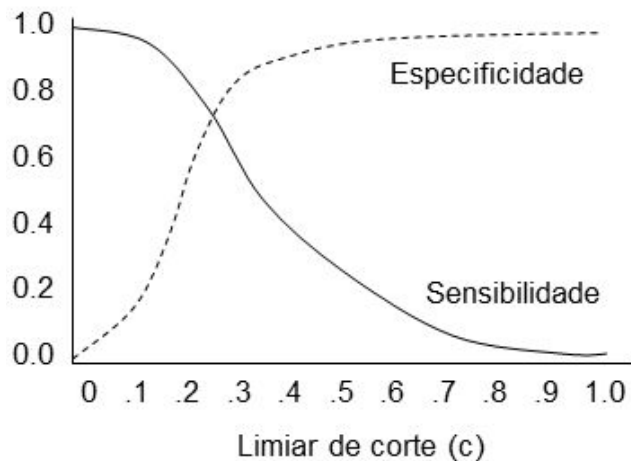


# ANÁLISE DE REGRESSÃO LOGÍSTICA

		Previsão do modelo		Total
		y = 1	y = 0	
Obs.	y = 1	n1	n2	n1 + n2
	y = 0	n3	n4	n3 + n4

Sensibilidade =  $n1 / (n1 + n2)$

Especificidade =  $n4 / (n3 + n4)$



## MEDIDAS DE AJUSTE

---

Área abaixo da curva ROC	Interpretação
Menor ou igual a 0,5	Não há discriminação
Entre 0,7 e 0,8	Discriminação aceitável
Maior que 0,8	Discriminação excelente

Quanto maior a área abaixo da Curva ROC, maior é a capacidade do modelo em discriminar os grupos de evento de interesse e de não interesse.



# OBRIGADO

 /andresilvadecarvalho



[lattes.cnpq.br/6876528572507972](https://lattes.cnpq.br/6876528572507972)

FIAP

Copyright © 2021 | Professor André Silva de Carvalho

Todos os direitos reservados. Reprodução ou divulgação total ou parcial deste documento, é expressamente proibido sem consentimento formal, por escrito, do professor/autor



# SHIFT

 FIAP

