

FIAP

MBA

BUSINESS INTELLIGENCE & ANALYTICS

Data Mining & Prescriptive Analytics

Prof. Adelaide Alves
profadelaide.alves@fiap.com.br
2023



ADELAIDE ALVES DE OLIVEIRA PROFESSORA

Mestre em Ciências (FSP/USP), graduada em Estatística (Unicamp).

Diretora Técnica Estatística da empresa SD&W - www.sdw.com.br

Professora de Fundamentos Estatísticos, DataMining, Análise Preditiva e Machine Learning na FIAP dos cursos MBA Big Data (Data Science), MBA Business Intelligence & Analytics, MBA Digital Data Marketing, IA & ML e Shift em People Analytics e IA&ML



profadelaide.alves@fiap.com.br

TÉCNICAS SUPERVISIONADAS

PREVISÃO, ESTIMAÇÃO e CLASSIFICAÇÃO

ANÁLISE MULTIVARIADA

Análise Exploratória dos Dados

Análise de Discriminação de Estrutura

- Técnicas de dependência.
- Técnicas Multivariadas aplicáveis quando uma das variáveis pode ser identificada como dependente (variável *target*), e as restantes como variáveis independentes (ou preditoras).

Aprendizado Supervisionado

Análise Estrutural

- Técnicas de Interdependência.
- Técnicas Multivariadas que procuram agrupar dados com base em semelhança, permitindo assim a interpretação das estruturas dos dados. Não há distinção entre variáveis dependentes e independentes.

Aprendizado Não Supervisionado

MODELOS PREDITIVOS - AVALIAÇÃO

- Existem diversas métricas para determinar a qualidade de um modelo.
Dois exemplos muito utilizados:

- problema de estimação ou previsão:
(variável target quantitativa):
 - erro quadrático médio (MSE). Calculado por:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 = \frac{(\hat{y}_1 - y_1)^2 + \dots + (\hat{y}_n - y_n)^2}{n},$$

n é o número de observações,

y_i é o valor real e

\hat{y}_i é a predição do modelo.

Utilizamos a Raiz quadrada desse valor RMSE

Nesse caso, um modelo bom é aquele que possui o **menor erro quadrático médio**.

- problema de classificação:
(variável target categórica)
 - Percentagem de acertos do modelo

Acurácia: É a proporção de previsões corretas.

É dada por:

$$\text{Acurácia} = \frac{\text{Quantidade de Acertos}}{\text{Total}}$$

Nesse caso, um modelo bom é aquele que possui a **maior acurácia**.

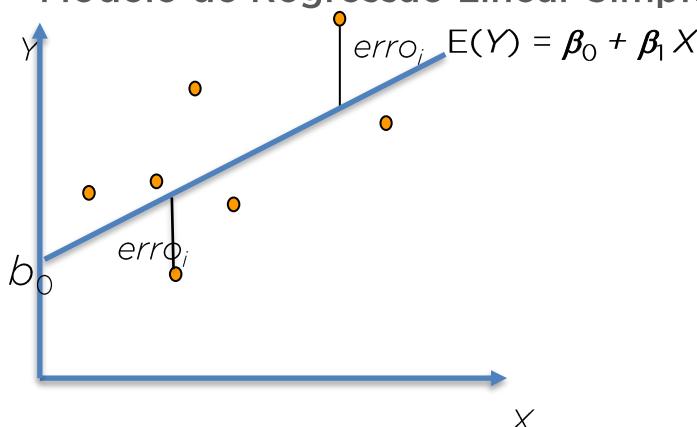
MODELOS PREDITIVOS - AVALIAÇÃO

- **Medidas de desempenho dos modelos**

→ problema de estimação ou previsão:

(variável target quantitativa):

Modelo de Regressão Linear Simples



Raiz do Erro quadrático médio(RMSE)

$$RMSE = \sqrt{\frac{5.743,3}{12}} = \sqrt{478,5} = 21,9$$

Id	Observado (A)	Estimado (B)	Erro (A-B)	Erro absoluto A-B	Erro^2
1	207	236	-28,7	28,7	822,8
2	289	265	24,0	24,0	576,1
3	285	272	13,5	13,5	181,9
4	292	278	14,0	14,0	195,2
5	269	285	-15,5	15,5	241,6
6	291	298	-6,6	6,6	43,2
7	331	304	26,9	26,9	724,4
8	283	307	-24,3	24,3	592,6
9	364	337	27,3	27,3	747,6
10	345	340	5,1	5,1	25,9
11	370	366	4,0	4,0	16,2
12	310	350	-39,7	39,7	1.575,1
			0,0	229,7	5.742,3

MODELOS PREDITIVOS - AVALIAÇÃO

- **Medidas de desempenho dos modelos**

➔ problema de classificação:
(variável target categórica/classes)

Id	Observado	Estimado
1	NÃO	SIM
2	SIM	SIM
3	NÃO	NÃO
4	SIM	SIM
5	SIM	NÃO
6	NÃO	NÃO
7	SIM	SIM
8	SIM	SIM
9	NÃO	NÃO
10	NÃO	SIM
11	SIM	SIM
12	NÃO	NÃO

		Estimado		
		NÃO	SIM	Total
Observado	NÃO	4	2	6
	SIM	1	5	6
	Total	5	7	12

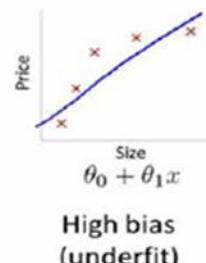
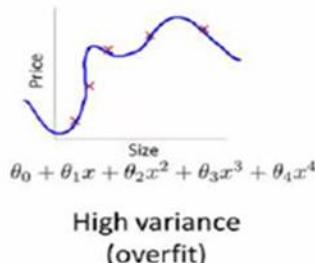
$$\text{Acurácia} = (4+5)/12 = 75,0\%$$

MODELOS PREDITIVOS

Alguns modelos são facilmente interpretáveis, outros muito complexos.

Problema: em geral alguns modelos têm sobreajuste (*overfitting*) quando há muitas variáveis preditoras, principalmente se forem colineares (baixo viés (*bias*) e alta variância (*variance*)).

- **Viés**, quando em alta, indica que o modelo **se ajusta pouco** aos dados de treino, causando o que é chamado de *underfitting*. O que significa que o MSE (raiz do erro quadrático médio) é alto, para a base de teste.
- **Variance**, em alta, diz que o modelo se ajusta demais aos dados (inclusive aos ruídos), causando por sua vez, *overfitting*, ou seja, se adaptam tão bem a amostra de treino que não conseguem



Um dos principais problemas a serem enfrentados na construção de modelos de predição é o de balancear a relação entre **bias** e **variance** (*bias-variance tradeoff*).

MODELOS PREDITIVOS

Para entender melhor a relação entre *bias* e *variance* é necessário notar que:

- algoritmos com alta *variance* tendem a ser mais complexos visto que conseguem se adaptar muito bem a qualquer conjunto de dados.
- algoritmos com alto *bias* são muito limitados por tudo aquilo que assumem sobre os dados, de forma que tem menor complexidade
- ou seja, ambos estão ligados ao nível de complexidade do modelo e são dependentes entre si. Em geral, modelo mais simples têm alto bias e baixa variance, enquanto modelos mais complexos têm baixo bias e alta variance.

Como, na prática, podemos avaliar essas situações:

➔ Solução: *Holdout* - Separar a sua base de dados em base de treino e base de teste.

- base de treino (*train data*) será utilizada para treinar seu modelo.
- base de teste (*test data*) refere-se à amostra de dados que será utilizada para avaliar o desempenho do seu modelo, medindo a capacidade do modelo de generalização (se ele funciona bem em outros dados)

MODELOS PREDITIVOS

→ Solução: *Validação Cruzada* - Uma outra maneira de verificar a performance e capacidade de generalização do seu modelo. Validação cruzada consiste na utilização de várias divisões de sua base de treino.

A validação cruzada nos fornece uma indicação melhor do quão bem o modelo se sairá com novos dados, já que, por meio das várias divisões, esta acaba testando o modelo na sua amostra de treino inteira, em contraste com o holdout que, por possuir apenas uma divisão, acaba dependendo de como os dados foram divididos entre as bases de treino e teste.

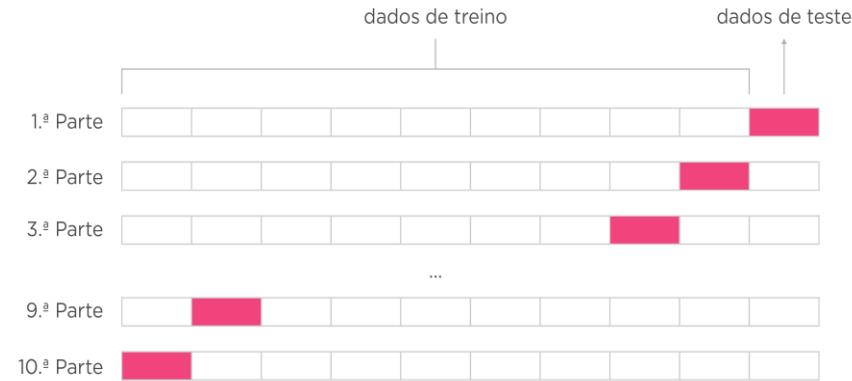
Além disso, validação cruzada é o método mais indicado quando possuímos poucos dados, já que a divisão em apenas duas bases pode acabar não fornecendo bases de treino e de teste boas o suficiente.

MODELOS PREDITIVOS

K-Fold- método de validação cruzada.

Dividir os dados em partes iguais e utilizar:

- Uma fração ($k-1$) delas para treinar o algoritmo com um hiperparâmetro;
- Outra parte testar (k) a sua predição.
- Depois dessa primeira iteração, um dos grupos que anteriormente era de treino torna-se o grupo de validação e o antigo grupo de validação passa a ser um grupo de teste. Esse processo se repete até que todos os k grupos tenham sido utilizados como grupo de validação. No final, a performance do modelo é calculada como a média de sua performance em cada iteração.



Seleção do hiperparâmetro com melhor performance >
definição do algoritmo com esse hiperparâmetro nos
dados de treino.

Fazer o mesmo para todos os algoritmos.

A única forma de saber qual o algoritmo de melhor performance é testando todos.

TÉCNICAS DE DISCRIMINAÇÃO PREVISÃO E ESTIMAÇÃO

Descobertas Supervisionadas de Relações

TÉCNICAS DE PREVISÃO: TÉCNICAS QUANTITATIVAS

Essas técnicas podem ser agrupadas em:

- Modelos de séries temporais: Enfoca os **padrões e suas mudanças**, desenvolvido por meio de sua **série**



- Modelos causais: Utiliza informações refinadas e específicas **sobre relações entre elementos do sistema**.

$$\text{Qualidade do Vinho} = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

- Variáveis preditoras como: tipo do vinho, acidez, ph, açúcar, ...

Utilização: As técnicas quantitativas são aplicadas nas condições:

- Informações do passado disponíveis;
- Informações quantificáveis em forma numérica;
- Assumir a hipótese de que algo dos padrões do passado irá se repetir no futuro (hipótese de continuidade).

PREVISÃO E ESTIMAÇÃO

REGRESSÃO
MULTIVARIADA

TÉCNICAS DE PREVISÃO: **TÉCNICAS QUANTITATIVAS**

O Modelo Causal permite:

- Expressar as relações de Causa-Efeito entre variáveis;
- Entender melhor os mecanismos geradores do fato em estudo;
- Simular situações de forma a se avaliar o seu impacto na previsão;
- Analisar situações independentes do tempo.

Modelo de Regressão:

Este modelo relaciona, funcionalmente, uma variável dependente às suas possíveis variáveis explicativas.

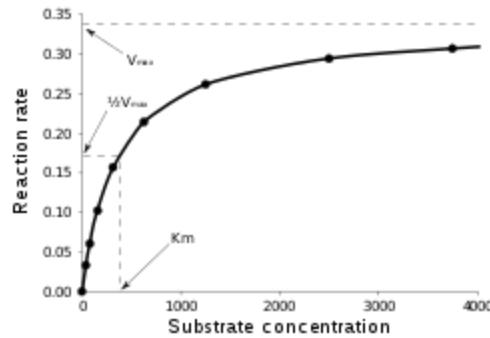
- Eficácia de propaganda sobre as vendas;
- Número de acidentes pela velocidade desenvolvida;
- Prever o tempo gasto no caixa de um supermercado em função do valor de compra;
- Satisfação do Cliente em função do tempo de relacionamento e intensidade de uso.

MODELOS DE REGRESSÃO

REGRESSÃO

Não-Linear

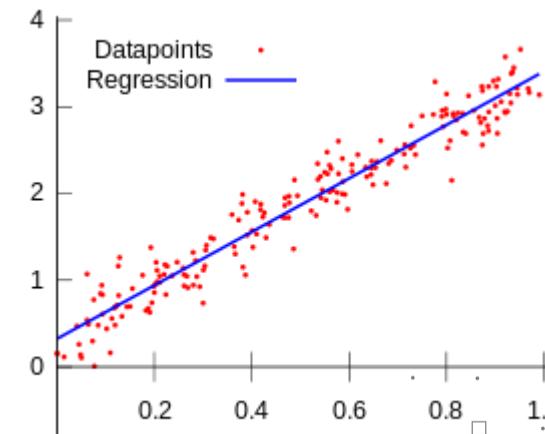
Linear



SIMPLES: uma variável explicativa

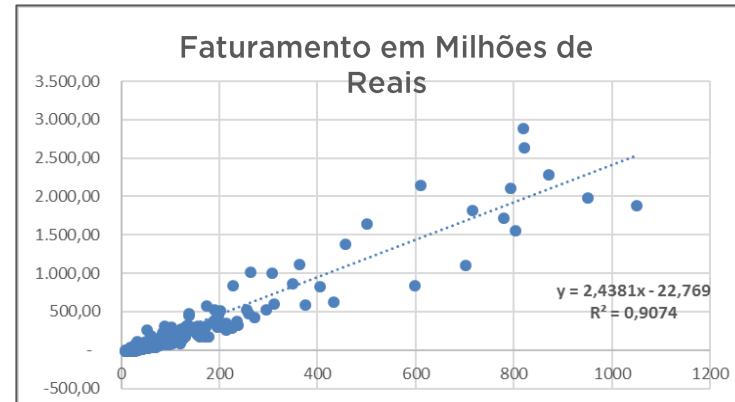
MÚLTIPLA: duas ou mais variáveis explicativas

Técnica estatística que relaciona, funcionalmente, uma variável dependente às suas possíveis variáveis explicativas



ANÁLISE DE REGRESSÃO

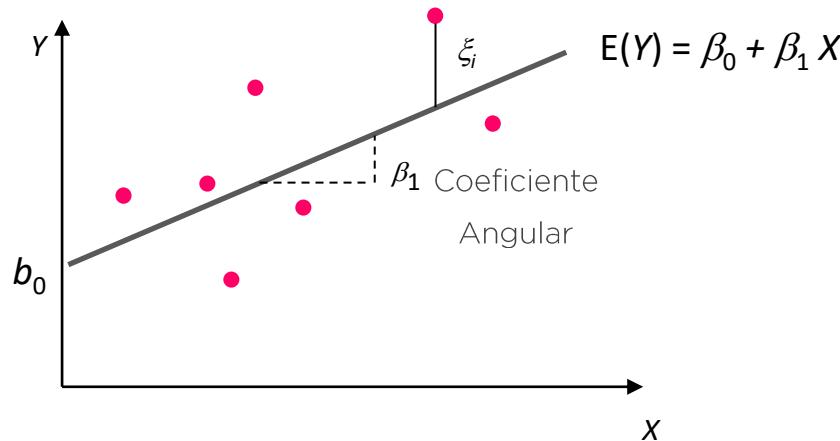
- Técnica Estatística que relaciona funcionalmente uma variável dependente às suas possíveis variáveis explicativas. Em outras palavras, consiste na obtenção de uma equação que tenta explicar variação da variável dependente pela variação do(s) nível(is) da(s) variável(is) independente(s).
 - Modelo Linear a Duas Variáveis.
 - Modelo Linear Múltiplo.



as

MODELO DE REGRESSÃO

LINEAR SIMPLES



$$Y_i = \beta_0 + \beta_1 X_i + \xi_i$$

Inclinação
Populacional

↑

Intercepto
Populacional

• MODELO REGRESSÃO MÚLTIPLA

- O Modelo que relaciona Y com várias variáveis independentes

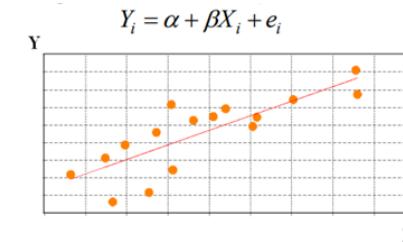
- Modelo Linear Simples: $Y = B_0 + B_1 X + e$

X = variáveis independentes

Y = variável dependente

B_0 = constante

B_1 = coeficientes de regressão



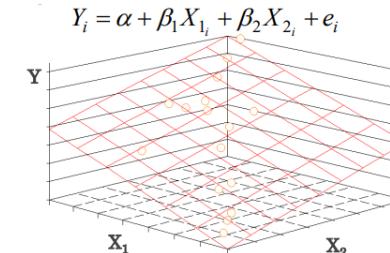
- Modelo Linear Múltiplo: $Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + \dots + B_n X_n + e$

$X_1, X_2, X_3, \dots, X_n$ = variáveis independentes

Y = variável dependente

B_0 = constante

$B_1, B_2, B_3, \dots, B_n$ = coeficientes de regressão associados às n variáveis



REGRESSÃO **LINEAR MÚLTIPLA**

- Com os dados de uma amostra, podemos calcular as estimativas dos parâmetros (B) não conhecidos. Usando para isso o ajuste pelo método dos mínimos quadrados.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

Que minimiza $SSE = \sum (\bar{y} - y)^2$

REGRESSÃO **LINEAR MÚLTIPLA**

- Exemplo:

Estimar o valor de venda de um imóvel em função das características do imóvel: área em m², quantidade de dormitórios, quantidade de suítes, andar, varanda gourmet, área verde nas proximidades, distância de metrô ou transporte público, etc

PREVISÃO E ESTIMAÇÃO

SÉRIES
TEMPORAIS

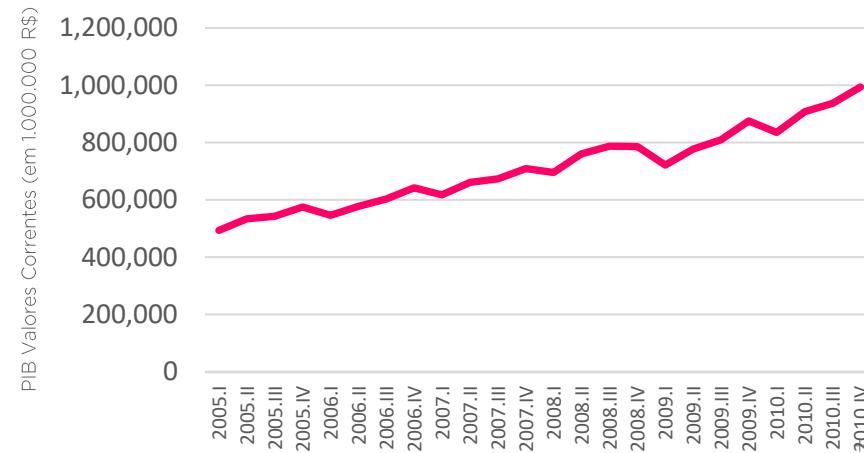
MODELOS DE SÉRIES TEMPORAIS

- Considerações gerais

Uma série temporal é qualquer conjunto de observações ordenadas no tempo.

Exemplos:

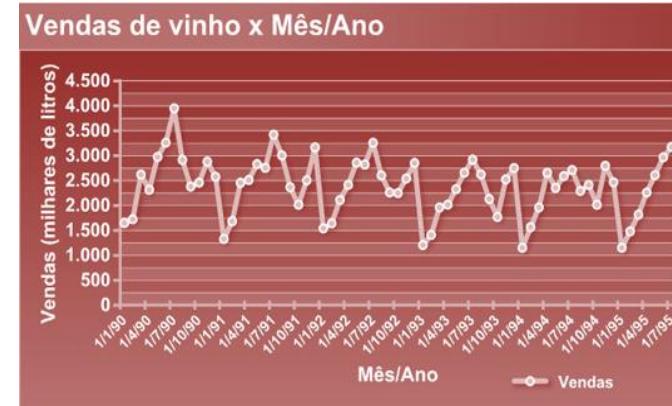
- faturamento da campanha
- número de pedidos
- produção mensal
- estoque mensal



TÉCNICAS DE PREVISÃO QUANTITATIVAS

MODELOS DE SÉRIES TEMPORAIS

O objetivo é identificar os padrões e suas mudanças, desenvolvido através de sua série histórica.

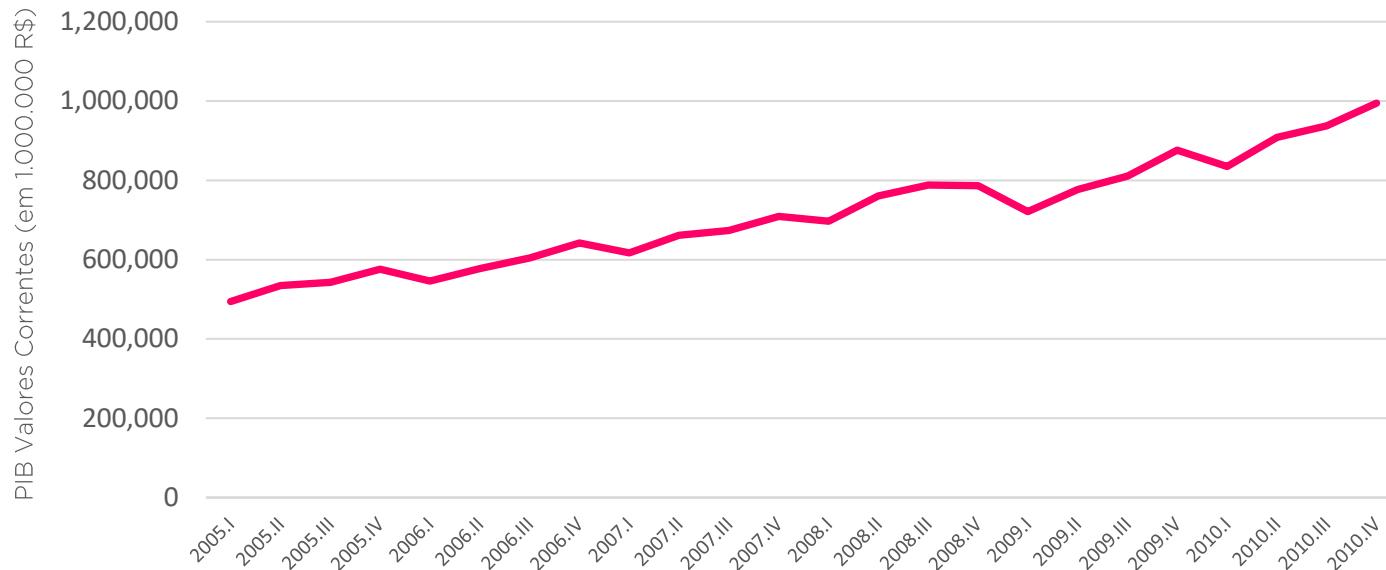


Utilização: As técnicas quantitativas são aplicadas nas condições:

- Informações históricas de pelo menos dois anos disponíveis;
- Informações quantificáveis em forma numérica;
- Assumir a hipótese de que algo dos padrões do passado irá se repetir no futuro (hipótese de continuidade).

SÉRIES TEMPORAIS

PIB, Valores Correntes Trimestrais



SÉRIES TEMPORAIS

- Análise de Séries Temporais visa identificar e explicar:

Série = T + S + C + A

T: Tendência

S: Sazonalidade

C: Ciclo

A: Aleatório

Tendência - evolução do fenômeno de interesse.

Sazonalidade - regularidade ou variação sistemática na série de dados.

Padrões Cíclicos - repetição de padrão num prazo superior a 2 anos.

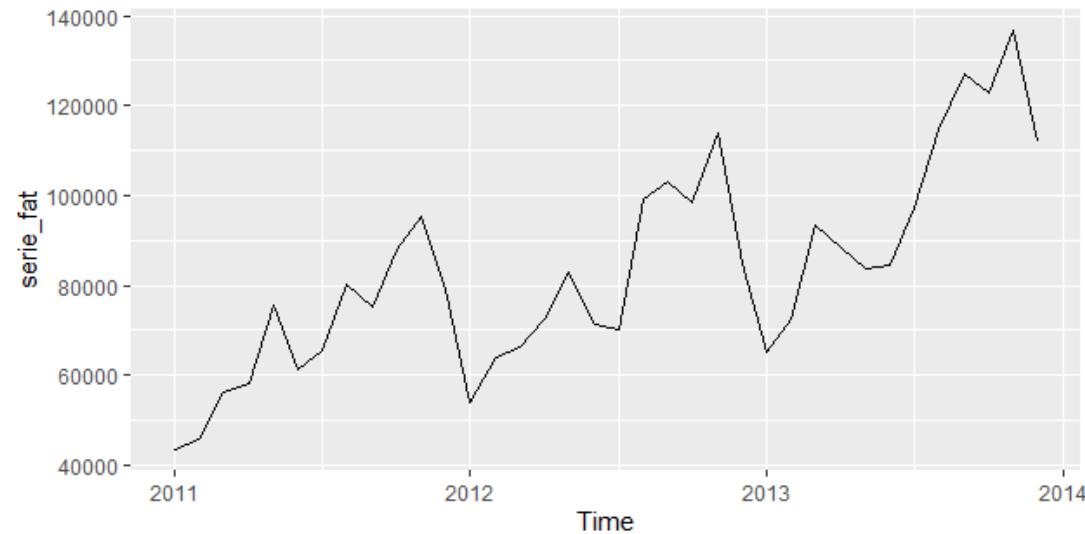
Aleatório - comportamento não explicável pelos três componentes anteriores (Erro Aleatório).

SÉRIES TEMPORAIS

- Principais objetivos ao analisar uma série temporal:
 - Investigar o mecanismo gerador da série temporal; por exemplo, analisando uma série de altura de ondas, queremos saber como estas ondas foram geradas;
 - Fazer previsões de valores futuros (curto ou longo prazo) da série;
 - Descrever apenas o comportamento da série;
 - Procurar periodicidade relevante nos dados.

SÉRIES TEMPORAIS

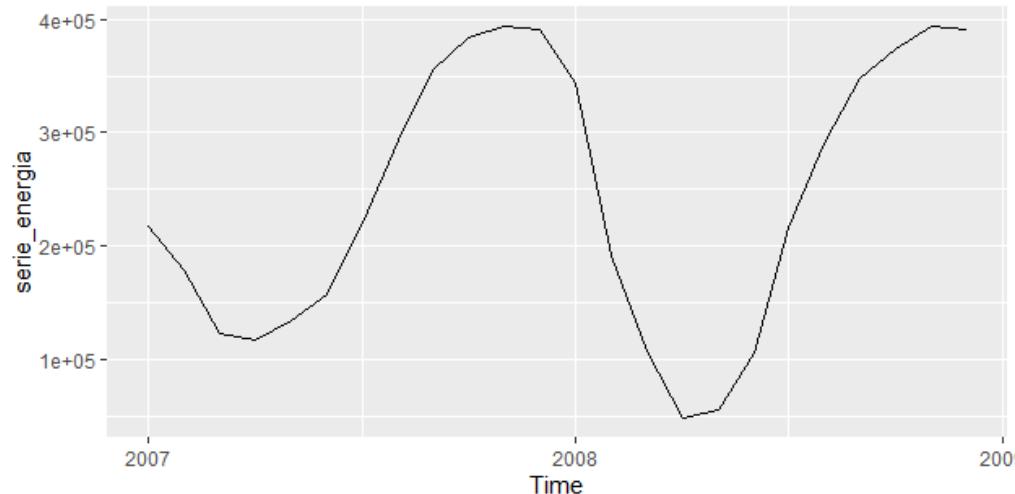
Série temporal do faturamento (R\$)



A série apresenta tendência? Sazonalidade?

SÉRIES TEMPORAIS

- Série temporal do consumo de energia (Kw/h) de empresas do setor Agricultura



A série apresenta tendência? Sazonalidade?

SÉRIES TEMPORAIS

FREQUÊNCIA DA SÉRIE

UNIDADE DE ANÁLISE	FREQUÊNCIA
Anual	1
Mensal	12
Diária	365
Trimestral	4
Semanal	52

SÉRIES TEMPORAIS

Exemplo 1:

Ano	Mes	Faturamento
2011	1	43484
2011	2	45859
2011	3	56254
2011	4	58224
2011	5	75403
2011	6	61255
2011	7	65601
2011	8	80099
2011	9	75017
2011	10	87932
2011	11	95266
2011	12	79175
2012	1	54085
2012	2	63808
2012	3	66330
2012	4	72442
2012	5	83072
2012	6	71321
2012	7	70095
2012	8	99071
2012	9	103100
2012	10	98380
2012	11	113751
2012	12	84933

Exemplo 2:

Período	Proporção de vendas
17/01 a 23/01	34.1
24/01 a 30/01	27.9
31/01 a 06/02	26.7
07/02 a 13/02	15.4
14/02 a 20/02	37.0
21/02 a 27/02	25.0
28/02 a 06/03	46.7

Exemplo 3:

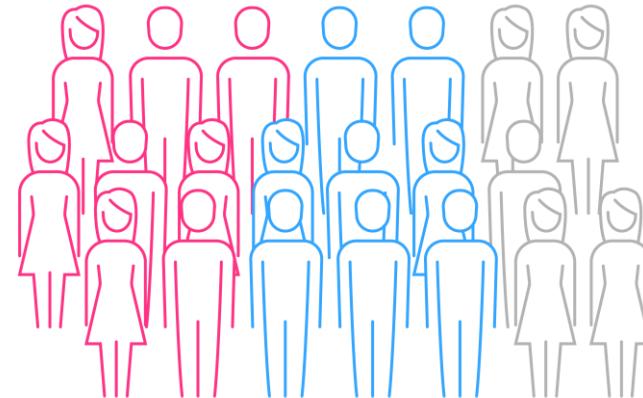
instant	dteday	Bikes alugadas
1	01/01/2011	985
2	02/01/2011	801
3	03/01/2011	1349
4	04/01/2011	1562
5	05/01/2011	1600
6	06/01/2011	1606
7	07/01/2011	1510
8	08/01/2011	959
9	09/01/2011	822
10	10/01/2011	1321
11	11/01/2011	1263
12	12/01/2011	1162
13	13/01/2011	1406
14	14/01/2011	1421
15	15/01/2011	1248
16	16/01/2011	1204
17	17/01/2011	1000

TÉCNICAS DE DISCRIMINAÇÃO CLASSIFICAÇÃO

Descobertas Supervisionadas de Relações

TÉCNICAS DE DISCRIMINAÇÃO MÉTODO DE CLASSIFICAÇÃO

- 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

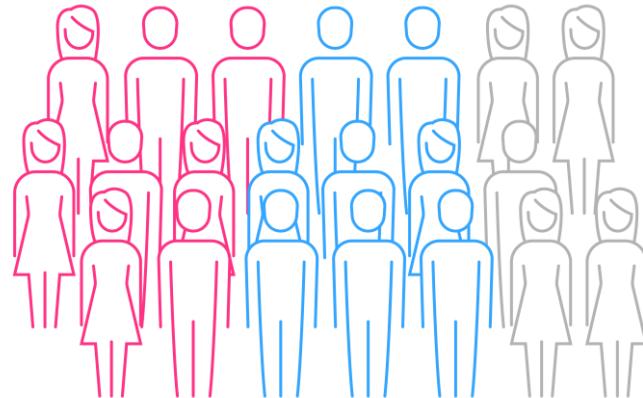


Como separar grupos previamente definidos? Como definir critérios, funções das variáveis que discriminem os grupos?

TÉCNICAS DE DISCRIMINAÇÃO

Métodos de Classificação

- Dado um conjunto de treinamento onde cada registro contém um conjunto de atributos, e um dos atributos é a nossa variável de interesse é tipo categórica/classes.
- Encontrar um modelo para determinar o valor do atributo/classe em função dos valores de outros atributos.
- Objetivo: definir a classe de novos registros, a classe deve ser atribuída o mais corretamente possível.



• MODELOS DE AQUISIÇÃO

- Adquirir **prospects** com os mesmos perfis dos bons clientes da empresa;
- Campanhas sobre os clientes da concorrência;
- Estimular os clientes à aquisição de novos produtos/serviços (*cross selling*).

TIPOS DE MODELOS

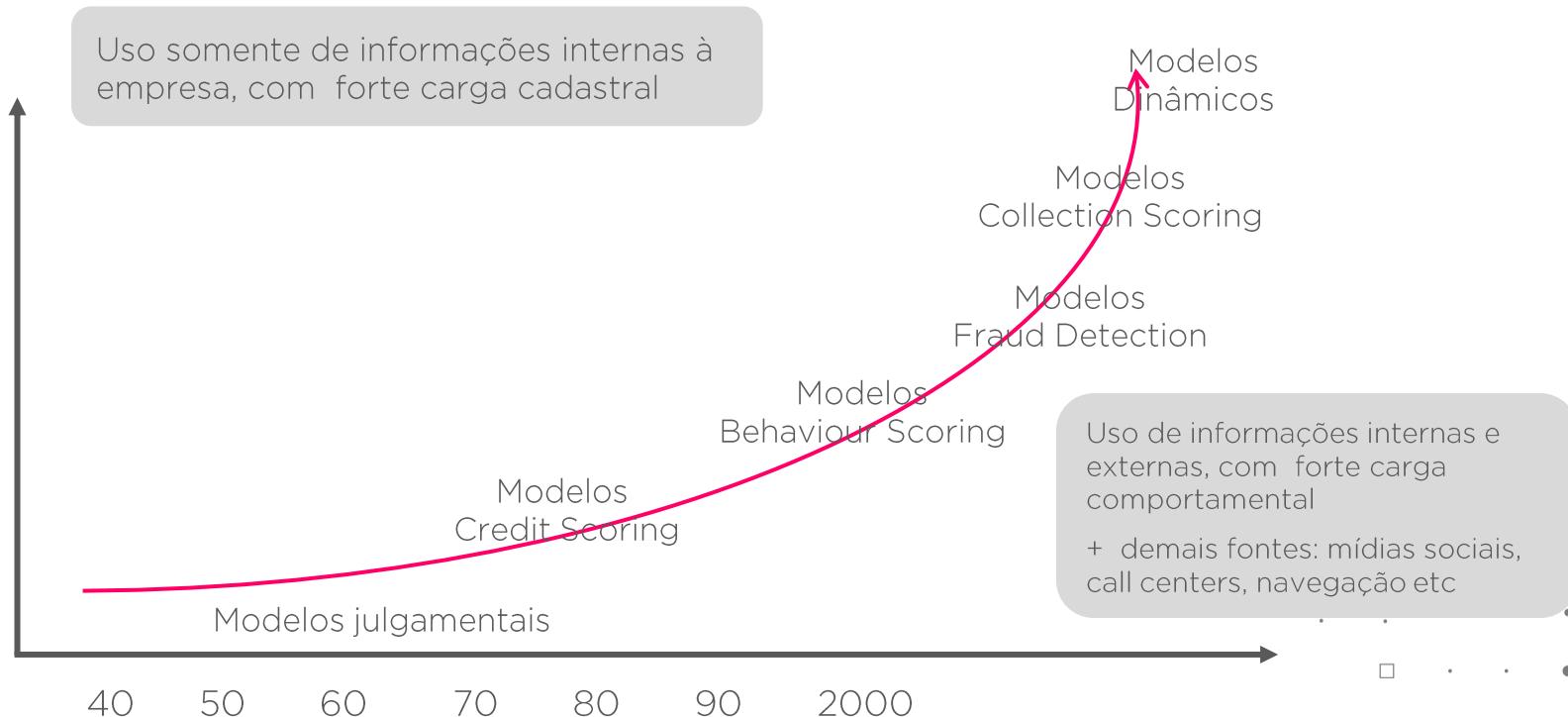
- MODELOS DE RETENÇÃO ou MODELOS DE CHURN

Objetivo:

- Identificar na base de dados de clientes prováveis a cancelar o relacionamento com a empresa.
- Oferecer suporte a área de relacionamento e permitir que campanhas de fidelização sejam direcionadas a clientes com risco real de interromper o relacionamento com a instituição.

- Melhores resultados nas campanhas realizadas;
- Redução de custos de abordagens indesejadas;
- Satisfação dos clientes;
- Maior credibilidade.

EVOLUÇÃO DAS FERRAMENTAS DE GESTÃO DE RISCO



TIPOS DE **MODELOS**

- **Modelo de *Credit Scoring***
 - Considera informações/dados do contrato (tempo de relacionamento recente);
 - Probabilidade de o novo cliente vir a ser inadimplente.
 - **Modelo de Inadimplência (*Behaviour Scoring*)**
 - Considera dados de utilização/comportamento dos clientes;
 - Probabilidade de o cliente vir a ser tornar um inadimplente.
 - **Modelo de Cobrança (*Collection Scoring*)**
 - Considera dados de utilização dos clientes e do mercado;
 - Probabilidade de um cliente pagar.
 - **Modelo de *Churn* e fraude/anomalias/ abusos**
 - Considera dados de utilização dos clientes e do mercado;
 - Probabilidade de o cliente cancelar a “conta/serviço/produto”.

• MODELOS PREDITIVOS

• Mercado Financeiro - Exemplos

Objetivo:

- Identificar na base de dados correntistas prováveis a cancelar/inativar o relacionamento (conta corrente) com o banco;

Dimensões:

- Utilização: diretamente relacionadas à geração de receita de cada correntista (dados transacionais).

Exemplos: produto adquirido, quantidade de cheques emitidos, saldo médio, tempo de relacionamento, conta conjunta, etc.

- Demográficas: informações descritivas do cliente.

Exemplos: sexo, idade, endereço, profissão, estado civil, renda, etc

- Definição da janela de tempo de análise
- Planejamento amostral (técnicas estatísticas aliadas às restrições do Banco)

Benefícios:

- Realizar ações fidelizadoras sobre os correntistas propensos a cancelar/inativar sua conta corrente

- TÉCNICAS DE DISCRIMINAÇÃO
- **MÉTODO DE CLASSIFICAÇÃO**

Classificadores eager (espertos)

A partir da amostragem inicial (conjunto de treinamento), constroem um modelo de classificação capaz de classificar novos registros.

Uma vez pronto o modelo, o conjunto de treinamento não é mais utilizado na classificação de novos objetos (registros)

Classificadores lazy (preguiçosos)

Cada novo registro é comparado com todo o conjunto de treinamento e é classificado segundo a classe do registro que é mais similar. Também conhecido como: Aprendizado baseado em exemplo (*Instance-based Learning*):

Outros Métodos

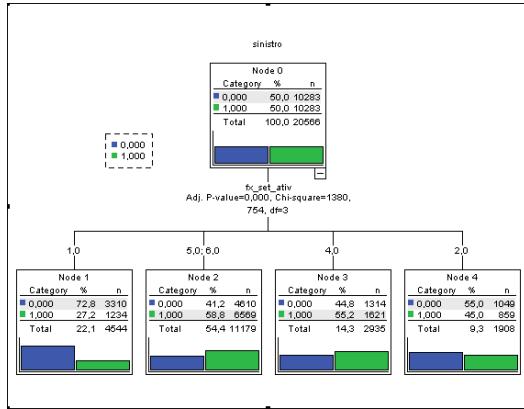
- Árvores e Regras de Decisão
- Redes Neurais
- Redes Bayesianas e Naïve Bayes
- SVM-Máquinas de Vetores de Suporte

- Método kNN (k-nearest-neighbor)

- Algoritmos Genéticos
- Conjuntos Fuzzy

Técnicas de Classificação:

Árvore de Decisão

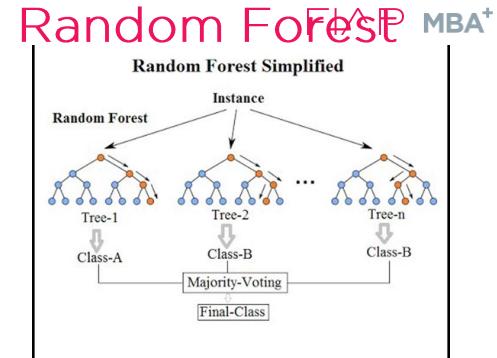
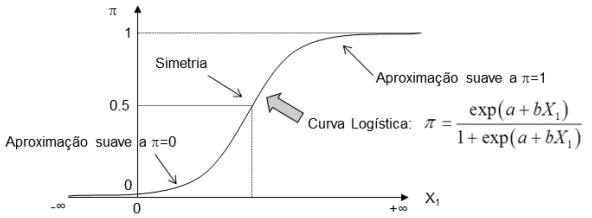
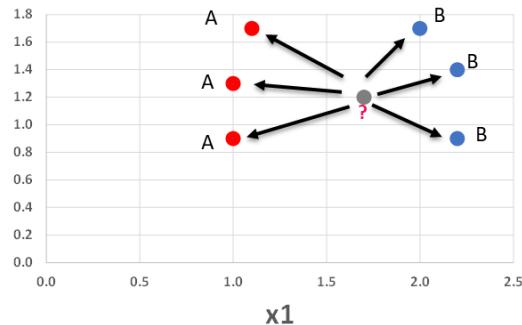


Regressão Logística

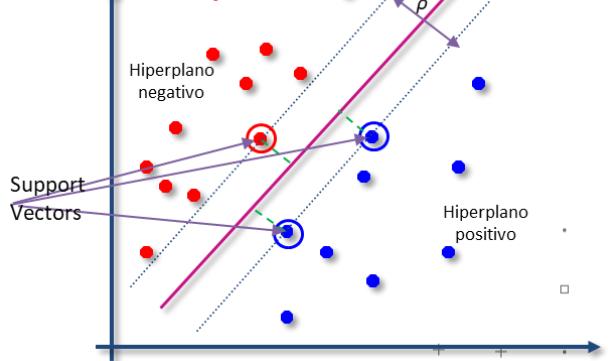
variável	categoria	Coeficientes
fatura em atraso	até 3 dias	-1,276
	3 a 15 dias	-0,611
	de 15 a 30 dias	0,580
	mais de 30 dias	1,308
Tempo de cliente	até 1 ano	0,580
	de 1 a 3 anos	0,401
	de 3 a 8 anos	-0,264
	mais de 8 anos	-0,718
valor da fatura	Até R\$250	0,262
	R\$ 250 a R\$ 800	0,103
	R\$ 800 a 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

K-NN K-Nearest Neighbors

Qual a distância euclidiana entre os pontos?



SVM-Suport Vector Machine



Naive Bayes

$$P(y | x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^n P(x_i | y)}{P(x_1, \dots, x_n)}$$

TÉCNICAS DE **CLASSIFICAÇÃO**

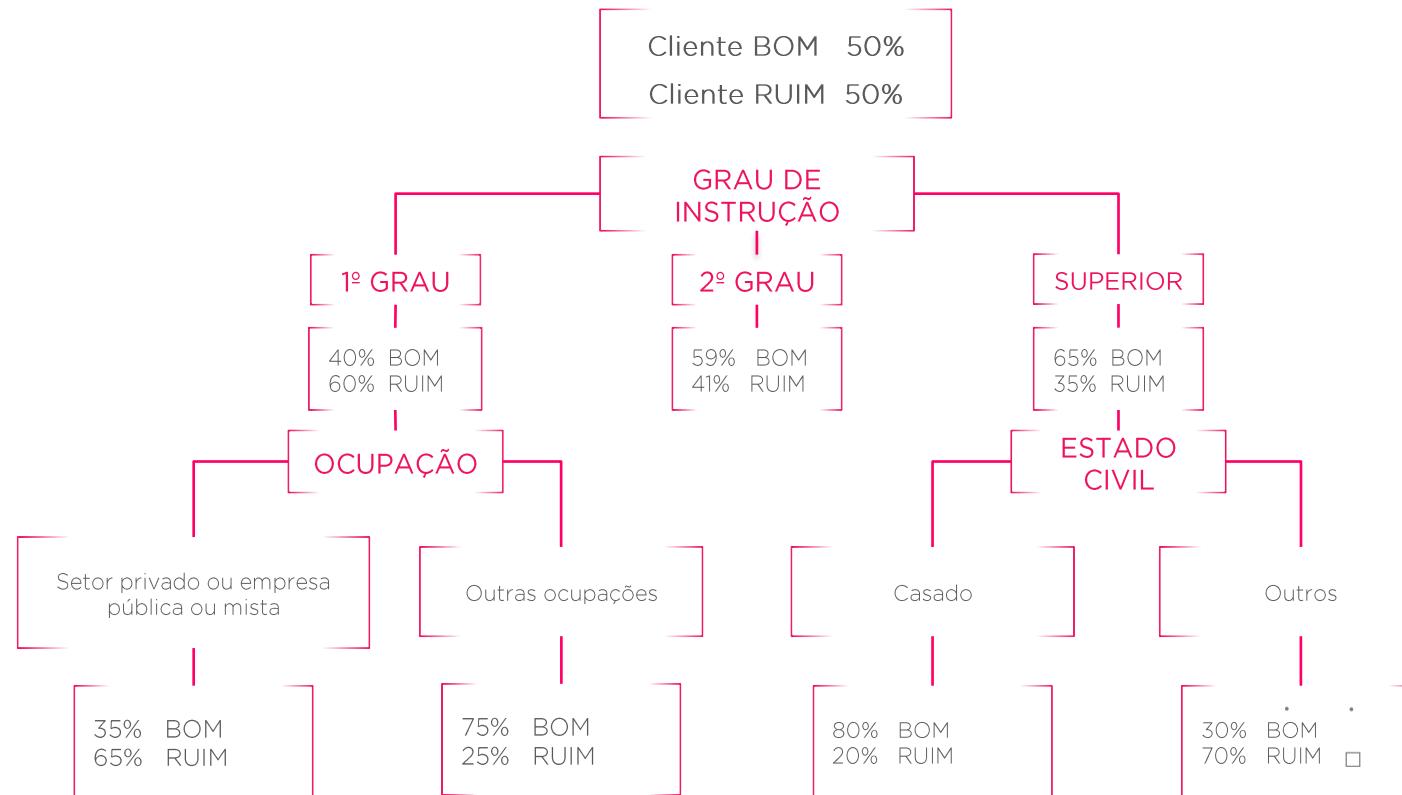
ÁRVORES DE
DECISÃO

TÉCNICAS DE **DISCRIMINAÇÃO**

ÁRVORES DE DECISÃO

- Metodologia estatística de fácil interpretação e utilização.
 - São estruturas de dados compostas de um nó raiz e vários nós filhos, que por sua vez têm seus filhos também e se interligam por ramos, cada um representando uma regra. Os nós que não possuem filhos são chamados de nós folhas e os que têm são chamados de nós pais, ou de decisão.
- Têm como objetivo encontrar regras que discriminem dois grupos previamente conhecidos.
- Exemplo: Encontrar uma regra que trace perfil de pessoas mais propensas a aderir a um certo produto.

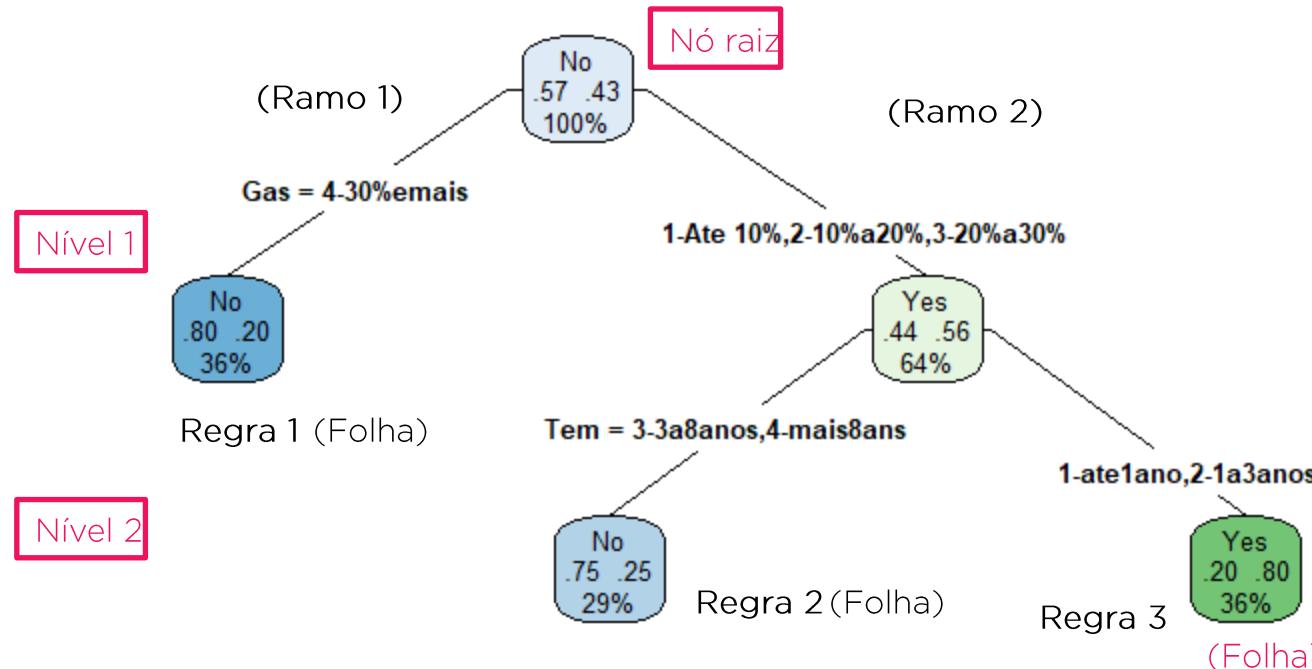
ÁRVORES DE DECISÃO - EXEMPLO



ÁRVORES DE DECISÃO - EXEMPLO

- Segmento: Área Financeira

A área de crédito deseja avaliar a propensão de um cliente tornar-se inadimplente.



ÁRVORES DE DECISÃO - PARÂMETROS

- Qual preditor e qual valor dividir os dados
- Profundidade e complexidade da árvore
- Resultado de cada folha

Tipos de Variáveis

- Todos os tipos de variáveis: Quantitativas e Categóricas (nominais ou ordinais)

ÁRVORES DE DECISÃO - ALGORITMOS

- Algoritmos utilizados para implementar uma árvore de decisão:
 - CHAID: CHi-square Automatic Interaction Detector
 - Algoritmo de Hunt-Szymanski (1976): diff (file comparison)
 - 1R: regras usam só um atributo (Holte, 1993);
 - ID3: binária categórica (Ross Quinlan, 1986);
 - C4.5 similar a ID3 porém permite atributos numéricos;
 - CART: Classification And Regression Trees (similar a C4.5 porém permite regressão);
 - CART: C5.0: última versão que usa menos memória;

ÁRVORES DE DECISÃO - CRITÉRIOS

- Como selecionar os valores limites dos atributos para realizar a melhor partição do dados?
- A seleção dos “nós” a serem utilizados na árvore é baseada na Teoria da Informação de Shannon, mais especificamente nos conceitos de entropia e ganho de informação
- Existem varias métricas (chamadas de “medidas de impureza”):
 - Entropia;
 - Índice de Gini;
 - Erro de classificação;

Obs: CART usa Gini ID3 e C4.5 usa Entropia

No ScikitLearn existe um hiperparâmetro chamado “criterion” onde se pode definir o uso de Gini ou Entropia (default=gini)

<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>

ÁRVORES DE DECISÃO - CRITÉRIOS

$$\text{Entropy}(t) = - \sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t),$$

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2,$$

$$\text{Classification error}(t) = 1 - \max_i [p(i|t)],$$

Onde:

i : classe alvo entre “ c ” classes possíveis e,

t : nó da árvore,

$p(i|t)$: frequência que a classe i aparece dentro do nó.

ÁRVORES DE DECISÃO - CRITÉRIOS

- Independente da métrica de impureza, em todas as árvores o que se busca é o **Ganho de Informação**.

Ganho de informação: É a redução esperada da entropia ao utilizarmos um atributo na árvore

O ganho de informação é dado por:

$$\text{Ganho}(S, A) = \text{Entropia}(S) - \sum \left(\frac{|S_v|}{|S|} * \text{Entropia}(S_v) \right)$$

Onde:

$\text{Ganho}(S, A)$ é o ganho do atributo A sobre o conjunto S

S_v = subconjunto de S para um valor do atributo A

$|S_v|$ = número de elementos de S_v

$|S|$ = número de elementos de S

EXEMPLO - **MODELO DE PROPENSÃO**

Segmento: Seguro Residencial

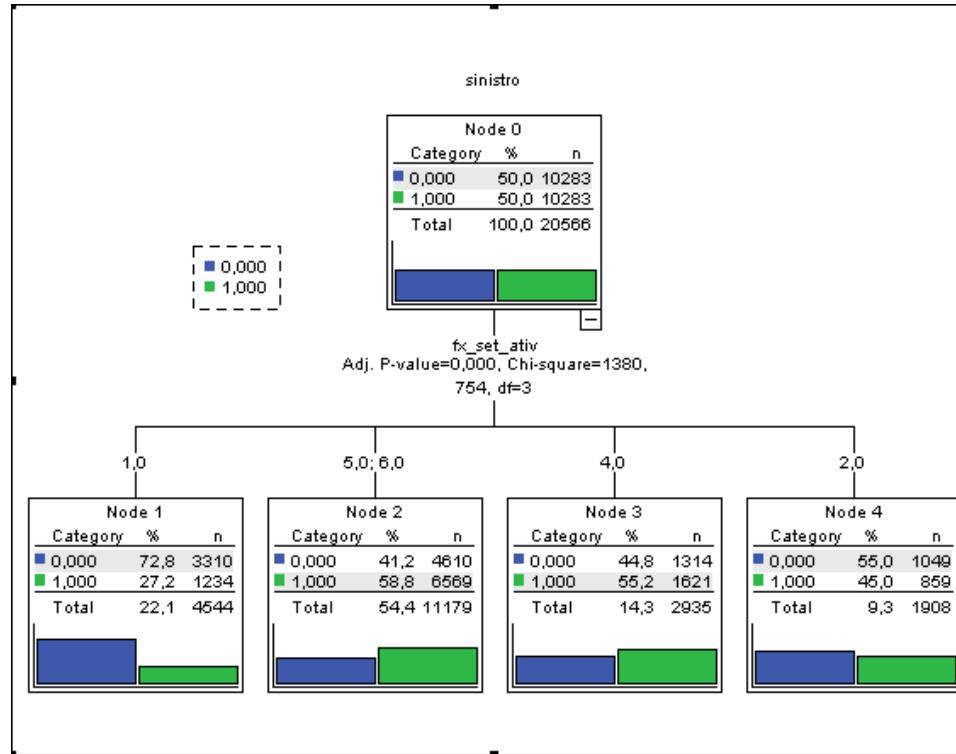
A área de Seguros deseja avaliar a propensão de um novo cliente sinistrar na apresentação de uma proposta.

EXEMPLO - MODELO DE PROPENSÃO

apolice	parcelas	qtde_cob	tpconstr	tipmora	clasmora	corretor	current	uf	set_ativ	Impseg(R \$)	sinistro
925578	6	9	6	casa	moradia	2	N	MS	90	100000	1
395699	1	9	6	apto	moradia	1	S	ES	26	30000	0
863771	11	9	6	casa	moradia	1	S	SP	24	200000	0
892165	11	9	6	casa	moradia	1	S	MG	27	30000	0
923092	1	9	6	casa	veraneio	2	N	SP	90	70000	0
1003098	4	9	6	casa	veraneio	1	S	SP	7	150000	1
955644	11	9	6	casa	moradia	1	S	MG	11	30000	1
987421	1	9	6	casa	moradia	2	N	SP	90	65000	1
744959	4	9	6	casa	veraneio	1	S	RS	18	70000	1
920814	11	9	6	casa	moradia	2	S	SP	90	100000	0
395550	2	9	6	casa	moradia	1	S	ES	26	20000	0
972615	6	9	6	casa	veraneio	2	N	SP	90	87500	1
958900	11	9	6	casa	moradia	1	S	MG	23	85000	1
911272	4	9	6	casa	veraneio	2	N	SP	90	150000	0
895508	11	9	6	casa	moradia	1	S	MG	33	50000	0
374234	1	9	6	apto	moradia	1	N	DF	6	30000	0
883254	11	9	6	casa	moradia	1	S	SP	24	100000	0
727885	3	9	6	casa	moradia	2	S	RS	90	180000	1
327315	11	9	6	casa	moradia	1	S	BA	21	20000	0
910241	11	9	6	apto	moradia	1	S	SP	49	50000	0
956554	10	9	6	casa	moradia	1	S	MG	27	70000	1
1000162	3	9	6	casa	moradia	2	S	MS	90	80000	1
920421	1	9	6	casa	veraneio	1	S	SP	1	40000	1

EXEMPLO - MODELO DE PROPENSÃO

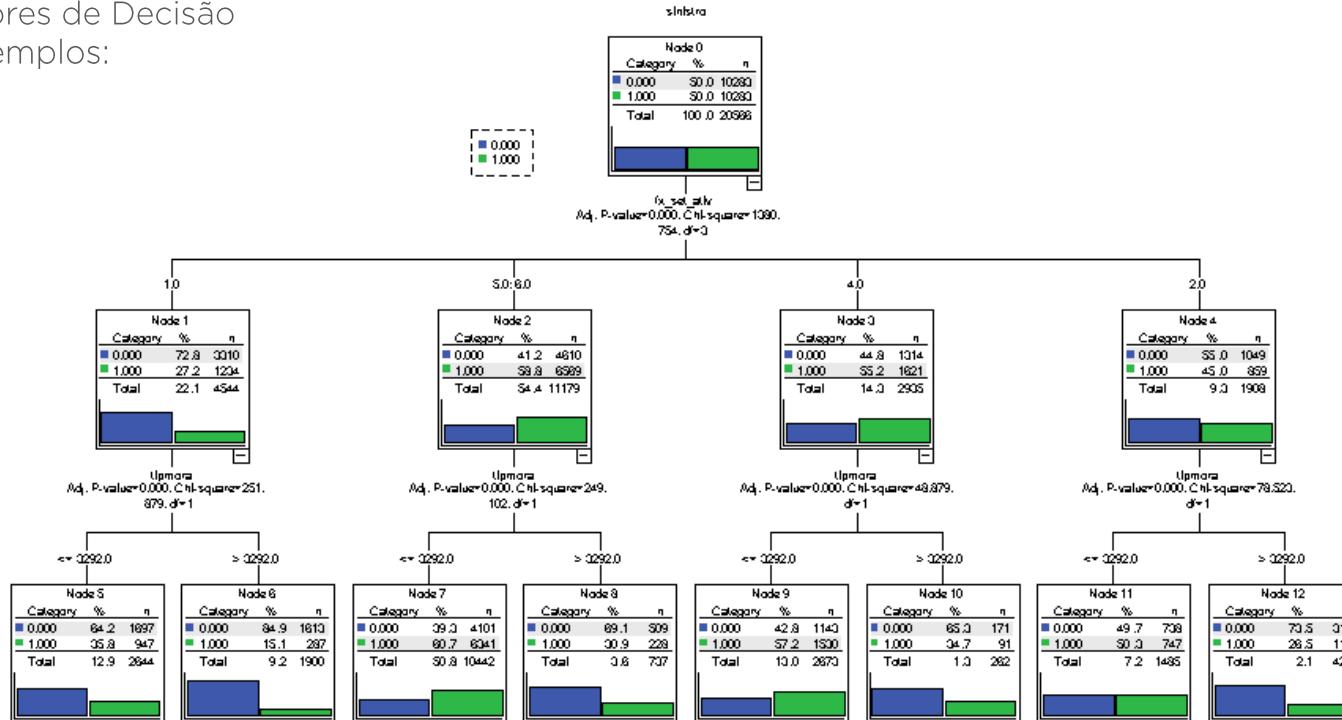
Árvores de Decisão
- exemplos:



EXEMPLO - MODELO DE PROPENSÃO

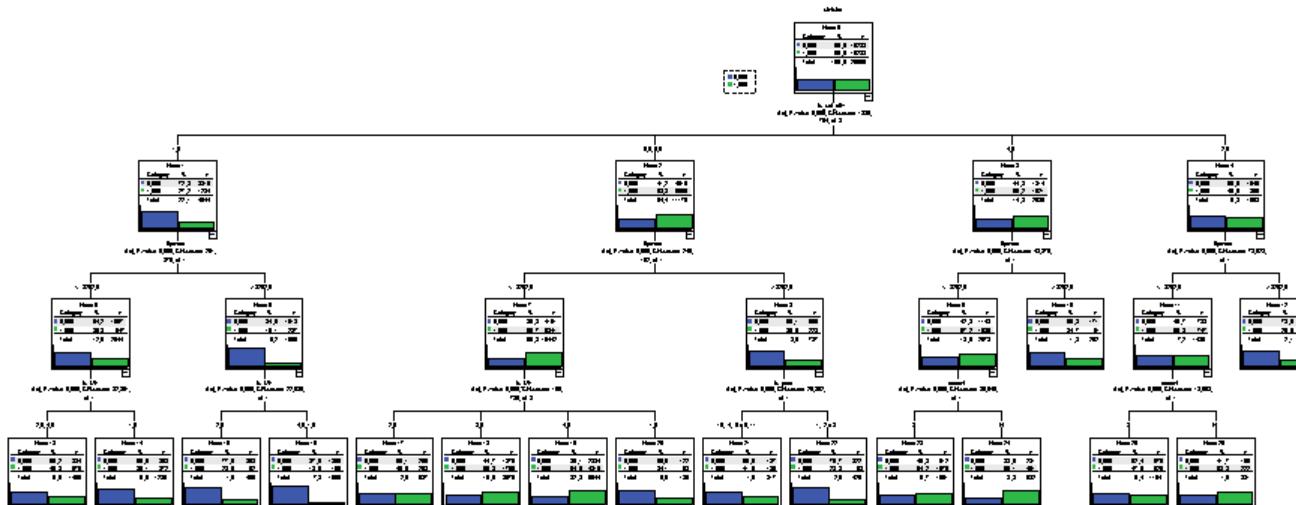
Árvores de Decisão

- exemplos:



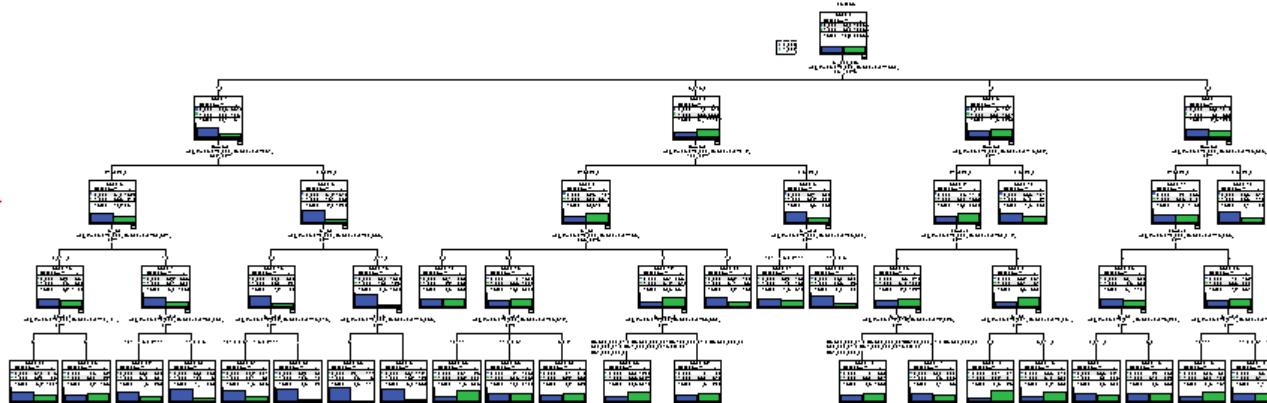
EXEMPLO - MODELO DE PROPENSÃO

Árvores de Decisão
- exemplos:



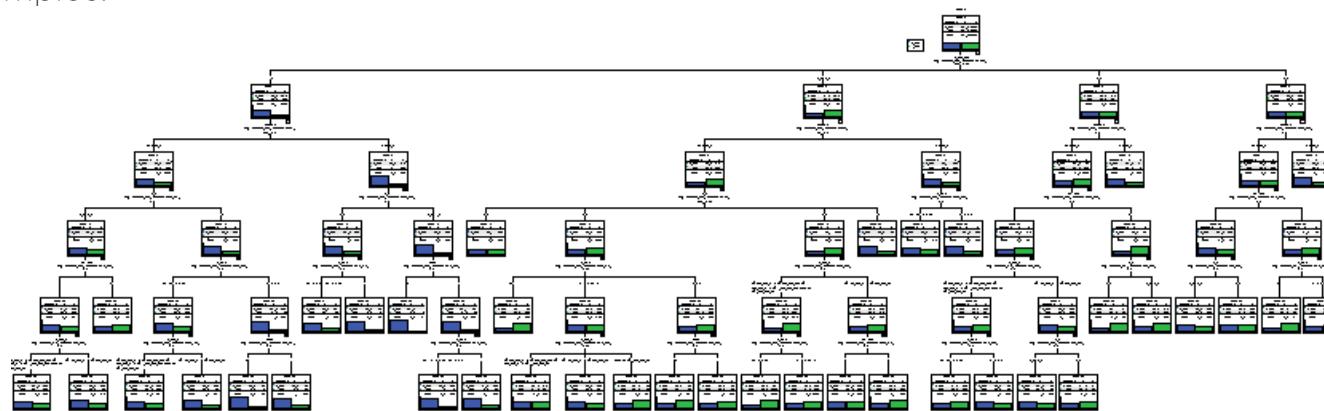
EXEMPLO - MODELO DE PROPENSÃO

Árvores de Decisão
- exemplos:



EXEMPLO - MODELO DE PROPENSÃO

Árvores de Decisão
- exemplos:



AVALIAÇÃO DO **MODELO**

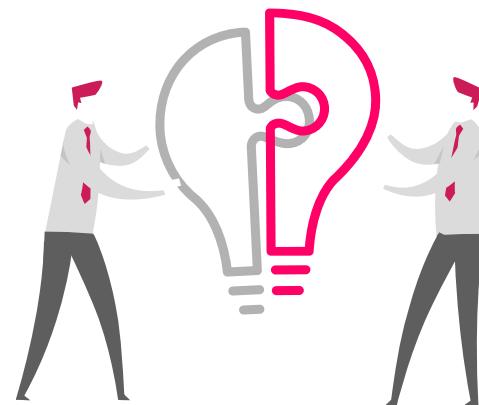
Exemplo

Classification				
Observed		Predicted		
		0	1	Percent Correct
0		5.137	999	83,7%
1		1.208	4.412	78,5%
Overall Percentage		81,0%	81,5%	81,2%

Growing Method: EXHAUSTIVE CHAID
Dependent Variable: Resposta

EXERCITANDO

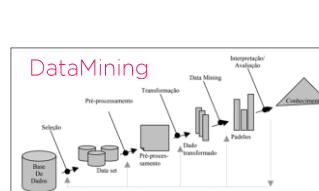
ÁRVORE DE
DECISÃO



Inadimplência

- MÉTODOS DE ENSEMBLE LEARNING
- INTRODUÇÃO

Para
resolver
um
problema



(*) No caso de problema supervisionado com variável target categórica.



Modelo 1 Árvore de Decisão 75%

Modelo 2 KNN 73%

Modelo 3 Regressão Logística 72%

Indicador de Performance
Acurácia(*), por exemplo:

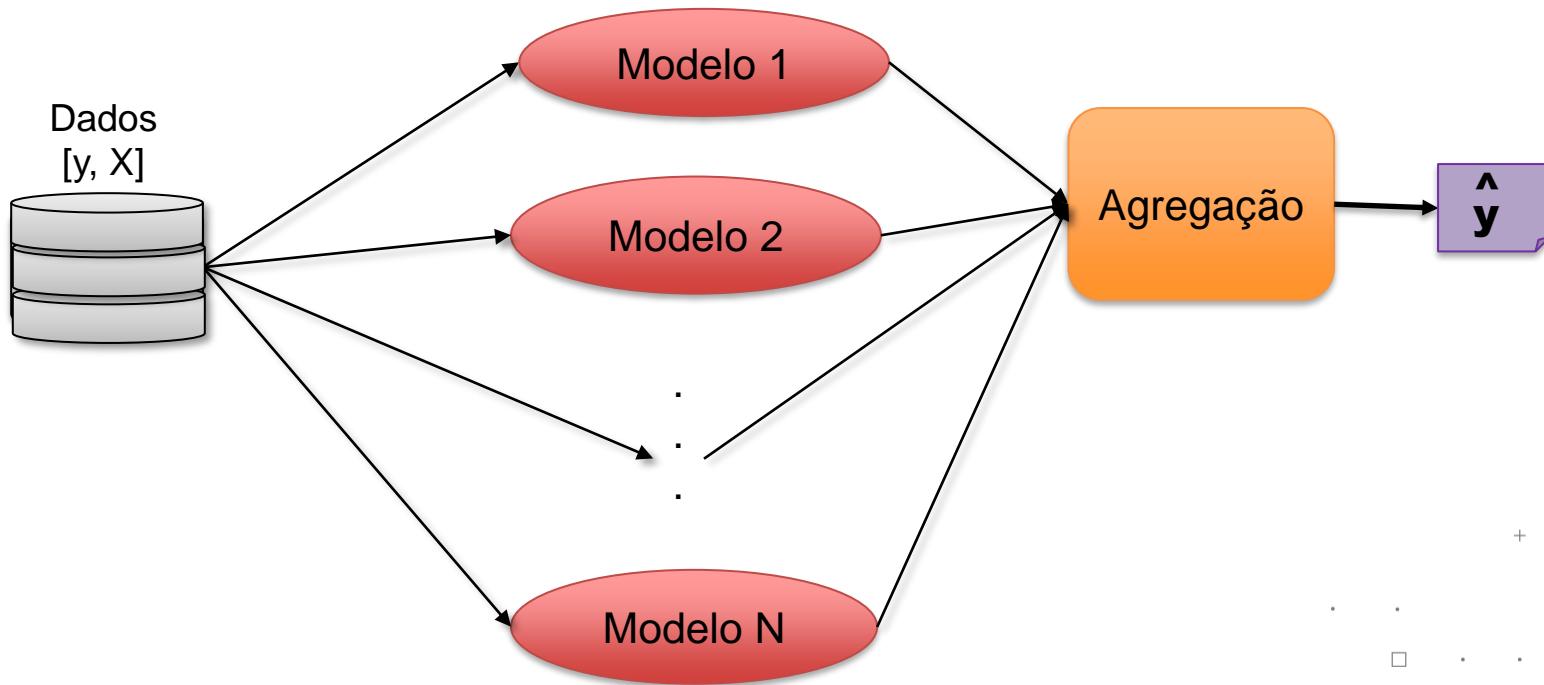


Ensemble Methods

Modelo combinado
Com acurácia de 78%, por exemplo



- MÉTODOS DE ENSEMBLE LEARNING
- INTRODUÇÃO



- MÉTODOS DE ENSEMBLE LEARNING
- INTRODUÇÃO

Junta múltiplos modelos mais “fracos”, com o objetivo de buscar a diminuir a suscetibilidade geral deles quanto o viés e a variância, tornando-os mais robustos.

Os métodos de Ensemble devem levar em conta a maneira com a qual eles agrupam os modelos, associando os algoritmos de forma a minimizar suas desvantagens individuais no modelo final.

Existem variados métodos de ensemble como: **Bagging, Boosting e Stacking,**

Stacking é de 1992, Bagging e Boosting são de 1996.

- MÉTODOS DE ENSEMBLE LEARNING
- **CARACTERÍSTICAS PRINCIPAIS**

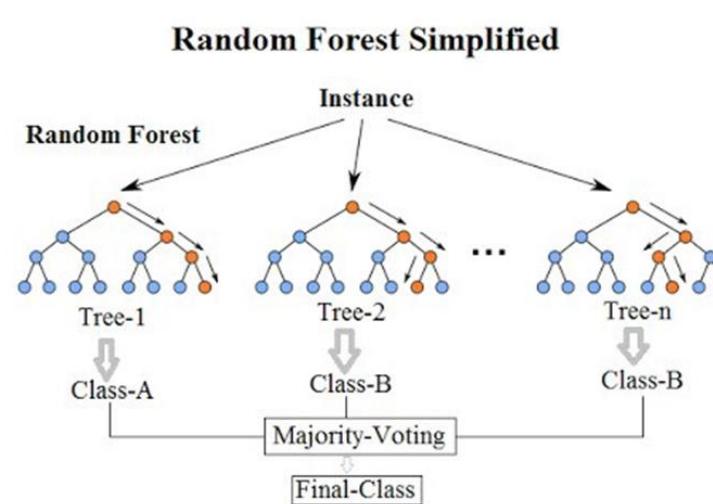
- **Bagging** → geralmente usa mesmo tipo de modelos individuais, cada um de forma independente em relação ao outro, de forma paralela. O algoritmo final é então feito a partir de algum tipo de resultado médio do que foi obtido a partir dos modelos bases.
- **Boosting** → geralmente usa mesmo tipo de modelos individuais, que são aplicados de forma sequencial (o posterior depende do antecessor) e depois combinados no modelo final.
- **Stacking** → geralmente usa tipos diferentes de modelos individuais, treinando-os em paralelo. É então aplicado um modelo no output dos weak learners (podendo incluir ou não as features utilizadas para treiná-los).

- MÉTODOS DE ENSEMBLE LEARNING
- **BAGGING – RANDOM FOREST**

- Random Forest é uma técnica de bagging.

Usa **diversas árvores de decisão como modelos individuais**, além de fazer uma seleção aleatória de casos e de variáveis. As árvores são extremamente interpretáveis, entretanto costumam ter um poder preditivo muito baixo quando comparados aos demais estimadores. Uma forma de contornar isso é através **da combinação da predição fornecida por diversas árvores**.

Cada árvore tenta estimar uma classificação e isso é chamado como “voto”. Idealmente, consideramos cada voto de cada árvore e escolhemos a classificação mais votada (estatística: Moda). No caso de problemas de regressão funciona similarmente, cada árvore tenta estimar a variável target e depois é considerada a média dos valores estimados em cada árvore.



MÉTODOS DE ENSEMBLE LEARNING

BAGGING

- Como funciona:
 - Treina **modelos individuais** usando uma amostra aleatória para cada;
 - Agrega os **modelos individuais depois de treinados** com suas respectivas amostras;
 - No caso de problemas de regressão usa a média e no caso de classificação a moda;
- Vantagem:
 - ajuda reduzir a variância (amostragem aleatória);
 - pode reduzir o viés(pois estamos usando média e moda para combinar os modelos);
 - fornece estabilidade e robustez (alto número de estimadores usados).
- Desvantagem:
 - tem um custo computacional alto (usa muito espaço e tempo - cada nova iteração é criada uma amostra diferente)
 - A técnica só funciona se o modelo base já tem uma boa performance. Usar o bagging em um modelo base ruim pode fazer com que o modelo final fique ainda pior. Como os modelos individuais usam o mesmo algoritmo, o bagging pode não reconhecer alguns padrões.

TÉCNICAS DE **CLASSIFICAÇÃO**

REGRESSÃO
LOGÍSTICA

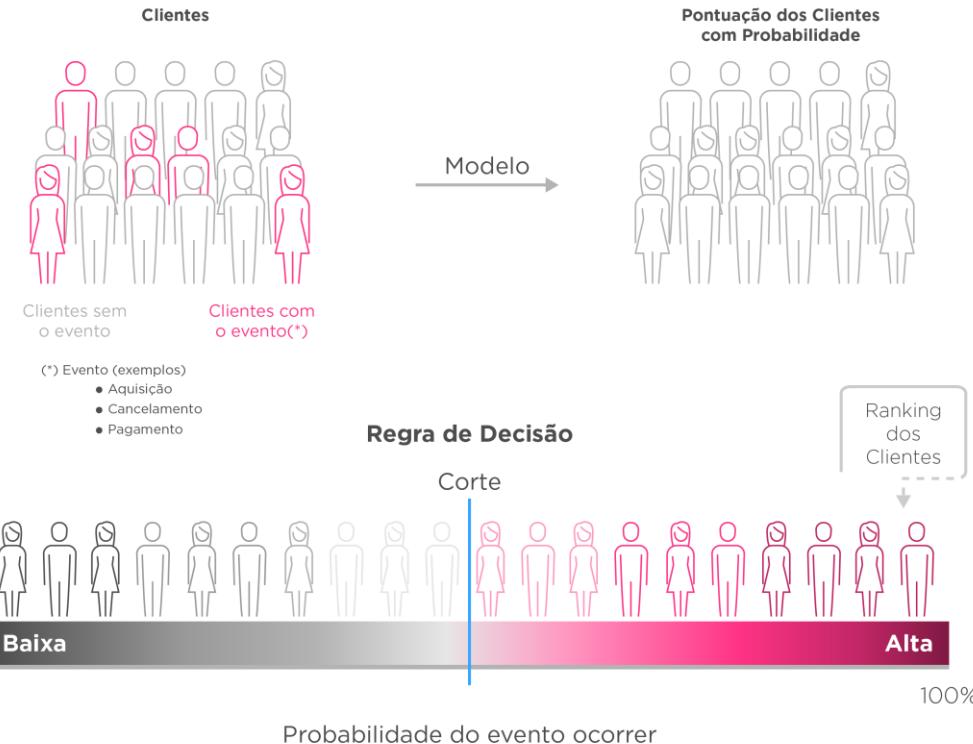
ANÁLISE DE DISCRIMINAÇÃO **DE ESTRUTURA**

REGRESSÃO LOGÍSTICA

Encontrar uma **função logística**, formada por meio de ponderações das variáveis (atributos), cuja resposta permita estabelecer a **probabilidade de ocorrência** de determinado evento e a **importância das variáveis** (peso) para essa ocorrência.

ANÁLISE DE DISCRIMINAÇÃO DE ESTRUTURA

REGRESSÃO LOGÍSTICA



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:

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

ANÁLISE DE REGRESSÃO LOGÍSTICA

Exemplo: Preferência por canal de futebol

Sexo	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

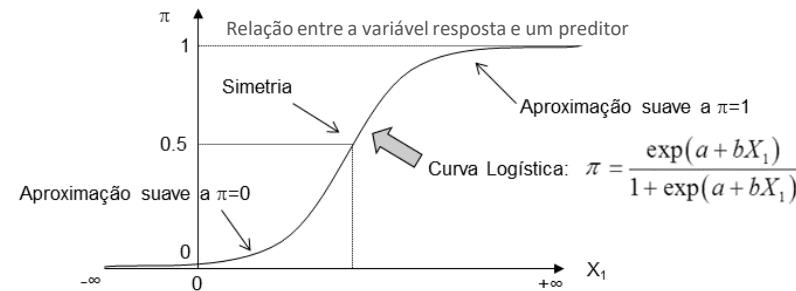
$$G = a + B_1 X_1 + B_2 X_2 + \dots + B_n X_n$$

G: logit da resposta de preferência (sim) a :

Intersecção B_1, B_2, \dots, B_n : 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

- Método de Estimação dos Coeficientes

- Regressão Linear: Método dos Mínimos Quadrados

- É o método que determina a linha reta mais adequada, minimizando a soma dos quadrados das diferenças entre os valores estimados de Y por meio da reta de regressão e os valores observados de Y.

- L

Consiste em determinar uma função, denominada função de verossimilhança $[L(y, \theta)]$, que é a função de probabilidade de ocorrência de um específico conjunto de dados e estimar os parâmetros que a maximizam.

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.

ANÁLISE DE REGRESSÃO LOGÍSTICA

Qualificação do Ajuste do Modelo

- Matriz de Classificação
- Estatística de Ajuste
- Verossimilhança : $-2 \log$ Verossimilhança
- Significância do Modelo : Qui-quadrado (similar ao F regressão)
- Ganho no Modelo (significância)

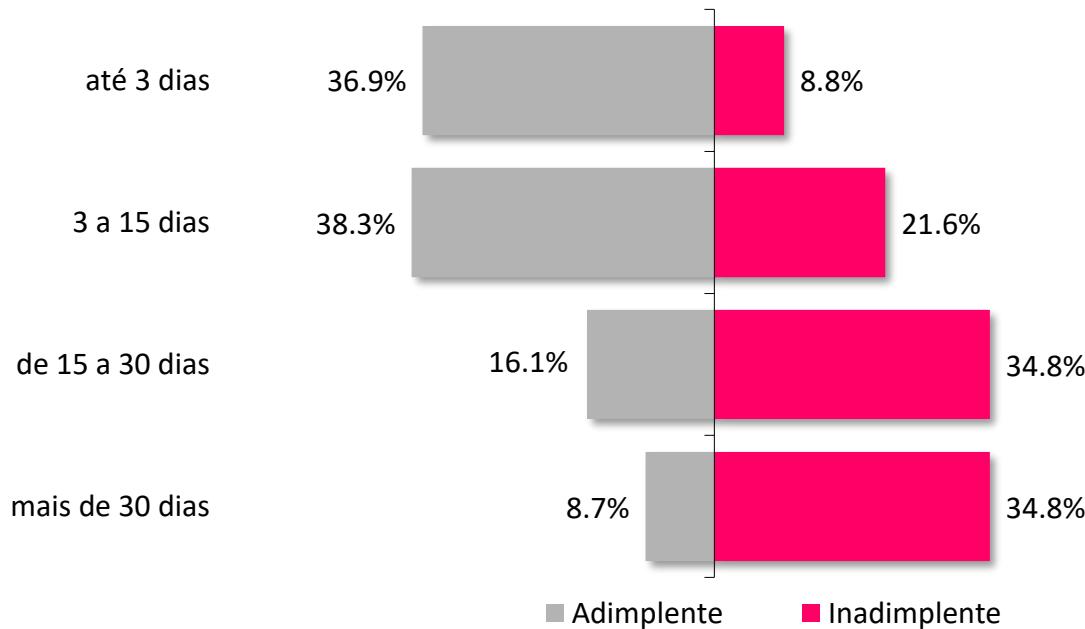
EXEMPLO - **MODELO DE INADIMPLÊNCIA**

- Segmento: Cartões de Crédito

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

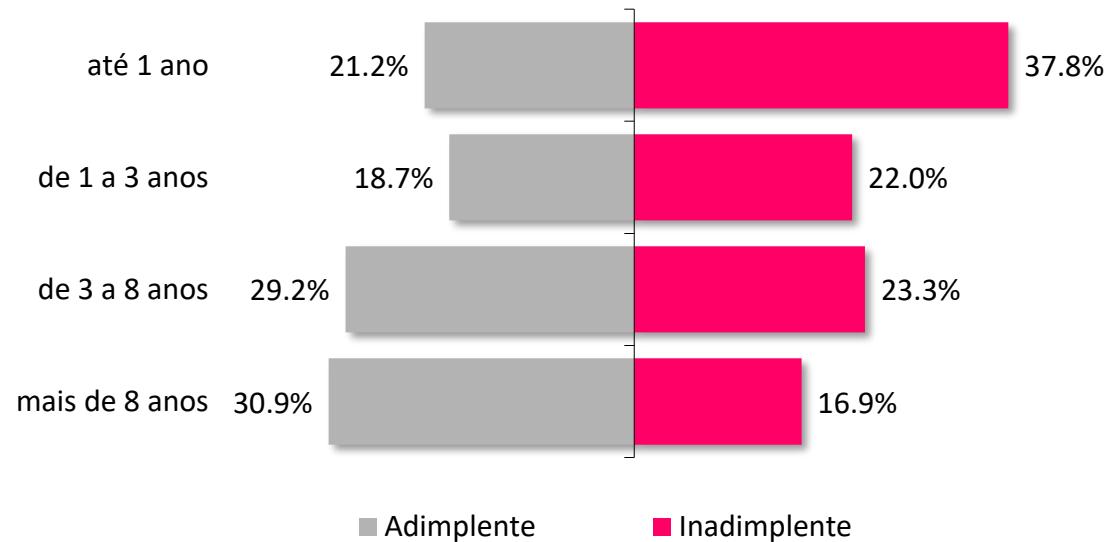
EXEMPLO - MODELO DE INADIMPLÊNCIA

Média de dias com pagamentos em atraso nos últimos 6 meses



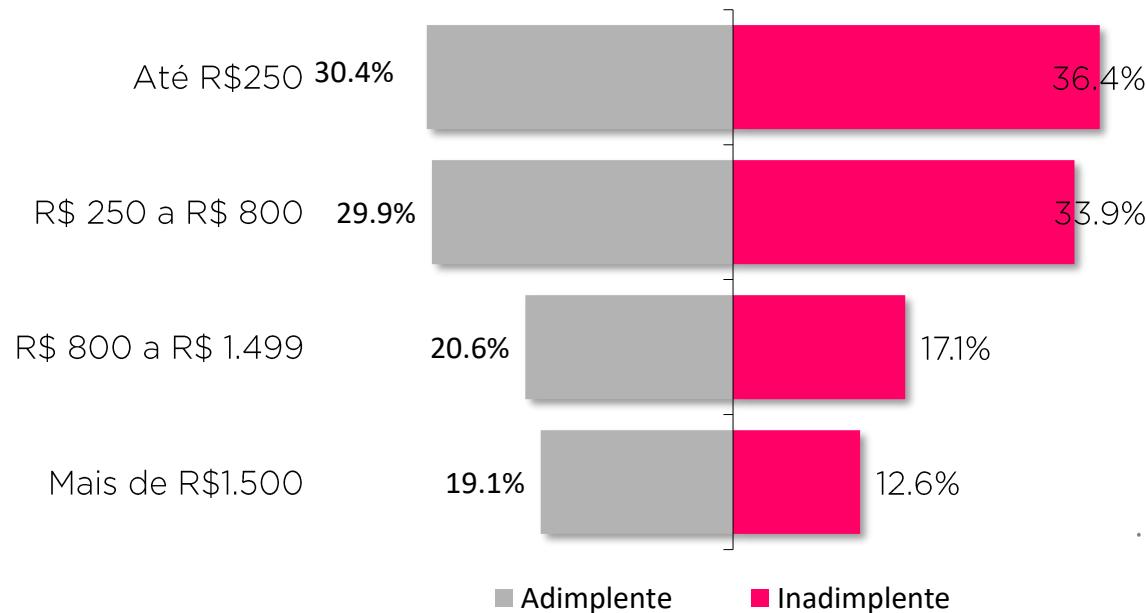
EXEMPLO - MODELO DE INADIMPLÊNCIA

Tempo de relacionamento em anos



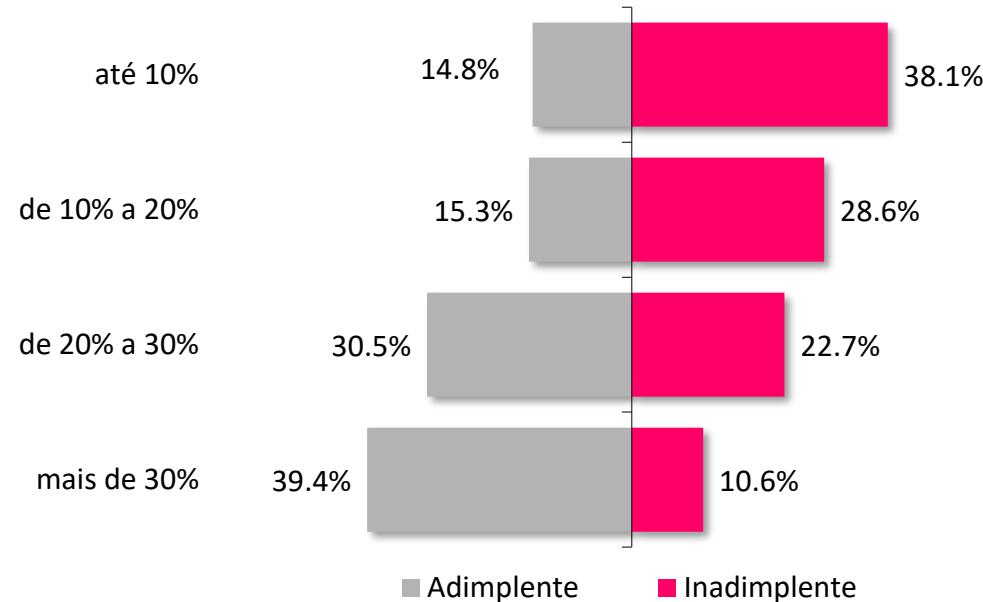
EXEMPLO - MODELO DE INADIMPLÊNCIA

Valor Médio da Fatura Mensal



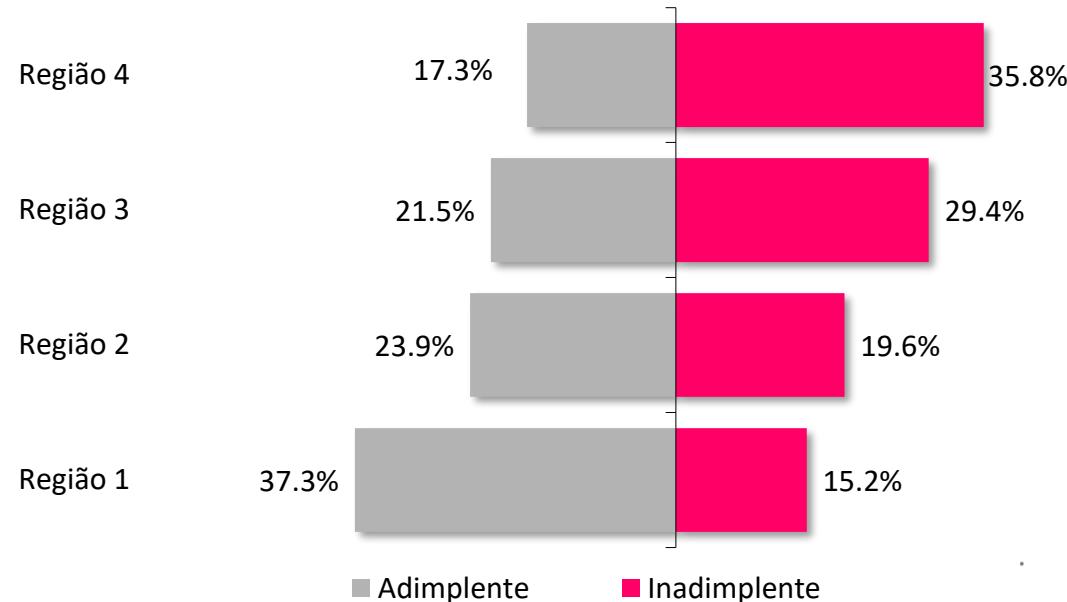
EXEMPLO - MODELO DE INADIMPLÊNCIA

Percentual dos gastos em alimentação



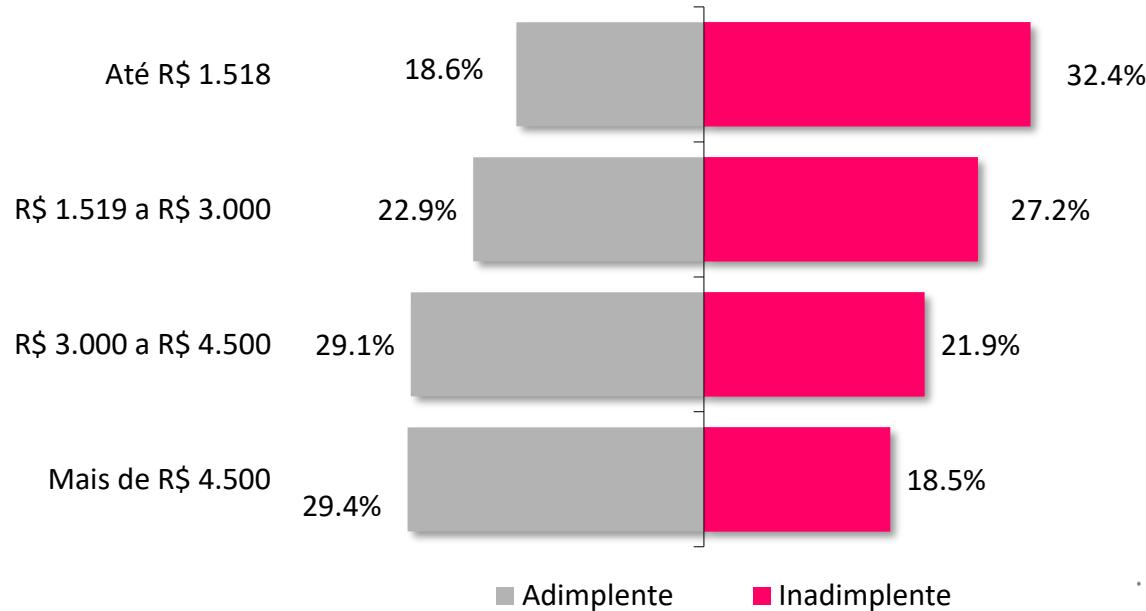
EXEMPLO - MODELO DE INADIMPLÊNCIA

Regiões de Risco



EXEMPLO - MODELO DE INADIMPLÊNCIA

Renda média mensal



EXEMPLO - MODELO DE INADIMPLÊNCIA

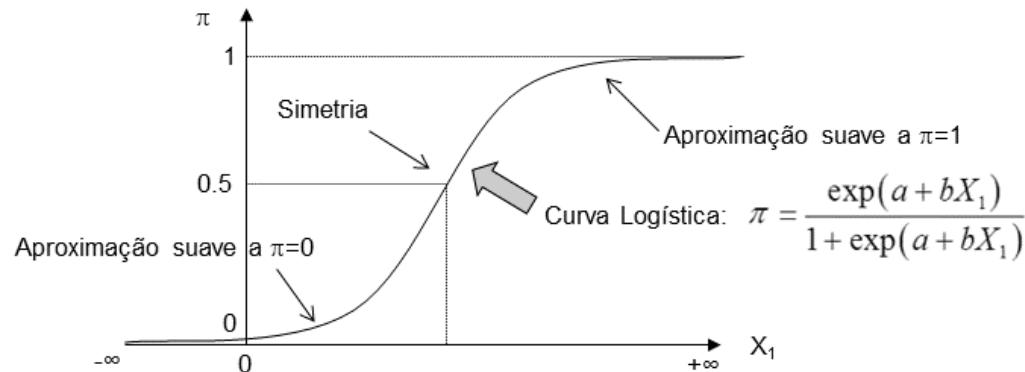
- Tabela de Coeficientes do Modelo

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

EXEMPLO - MODELO DE INADIMPLÊNCIA

→ Soma dos respectivos pesos da tabela dos pesos (x)

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



EXEMPLO - 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%

AVALIAÇÃO DO **MODELO**

- Exemplo

Classification				
Observed		Predicted		
		0	1	Percent Correct
0		5.137	999	83,7%
1		1.208	4.412	78,5%
Overall Percentage		81,0%	81,5%	81,2%

Growing Method: EXHAUSTIVE CHAID
Dependent Variable: Resposta

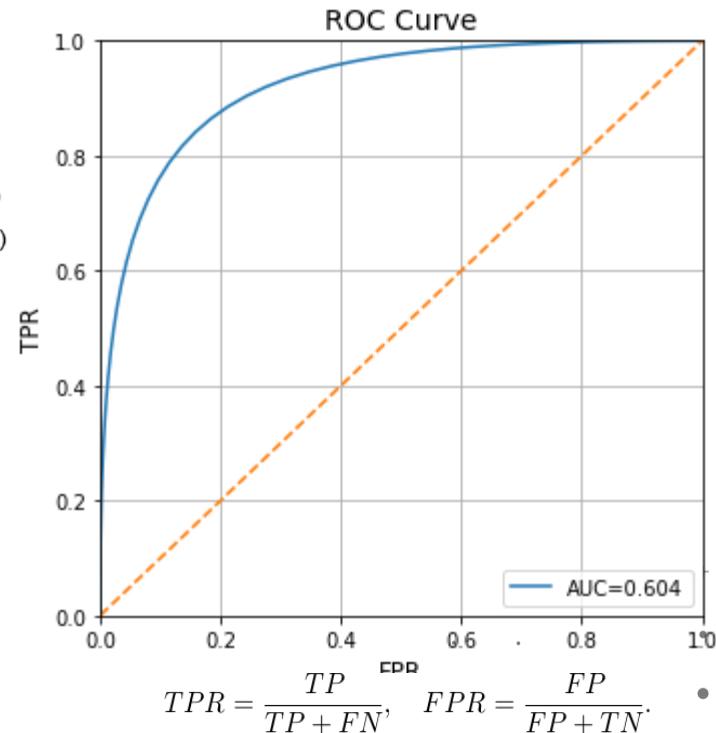
- TÉCNICAS DE CLASSIFICAÇÃO
- Qualificação do Ajuste do Modelo
- Qualificação do Ajuste do Modelo

		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 = $n4 / (n3+n4)$

- Acurácia: É a proporção de previsões corretas: $(n1+n4) / (n1+n2+n3+n4)$
- A curva ROC plota (chamado de sensibilidade) versus (chamado de 1-especificidade) para todos os possíveis pontos de corte entre 0 e 1.
- Uma forma bastante utilizada para determinar o ponto de corte .



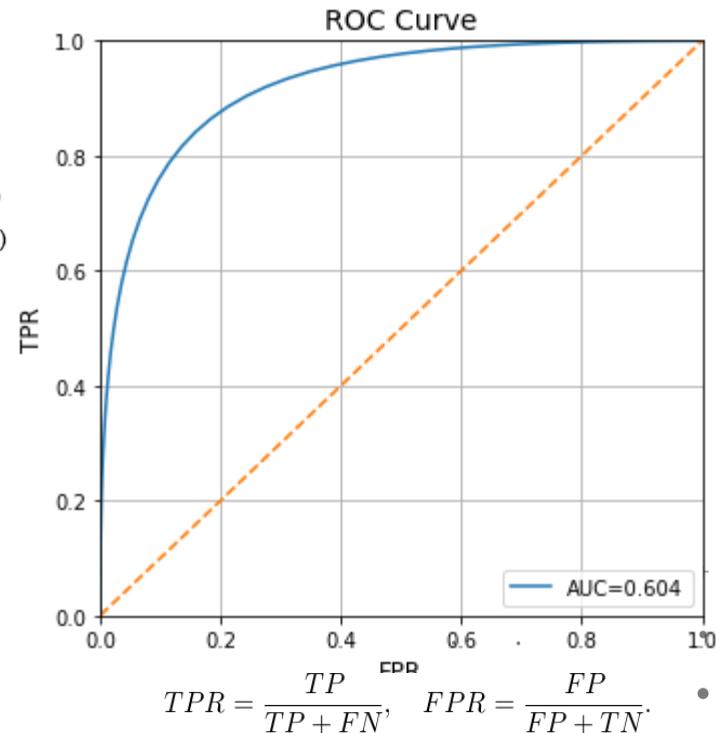
- TÉCNICAS DE CLASSIFICAÇÃO
- Qualificação do Ajuste do Modelo
- Qualificação do Ajuste do Modelo

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

Sensibilidade = $n1 / (n1+n2)$

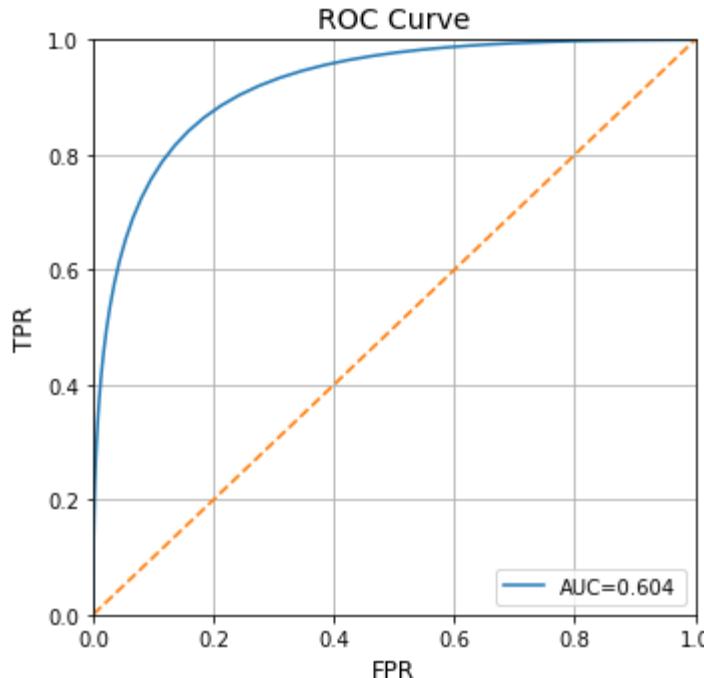
Especificidade = $n4 / (n3+n4)$

- Acurácia: É a proporção de previsões corretas: $(n1+n4) / (n1+n2+n3+n4)$
- A curva ROC plota (chamado de sensibilidade) versus (chamado de 1-especificidade) para todos os possíveis pontos de corte entre 0 e 1.
- Uma forma bastante utilizada para determinar o ponto de corte .



TÉCNICAS DE CLASSIFICAÇÃO

Qualificação do Ajuste do Modelo



- A curva ROC plota (chamado de sensibilidade) versus (chamado de 1-especificidade) para todos os possíveis pontos de corte entre 0 e 1.
- Uma forma bastante utilizada para determinar o ponto de corte .

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}.$$

TÉCNICAS DE CLASSIFICAÇÃO

Qualificação do Ajuste do Modelo

Matriz de Confusão

		Classe Preditada	
		positivo	negativo
Classe Esperada	positivo	Verdadeiros Positivos (VP)	Falsos Negativos (FN)
	negativo	Falsos Positivos (FP)	Verdadeiros Negativos (VN)

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 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 previsões corretas, sem considerar o que é positivo e o que negativo e sim o acerto total. É dada por: $(VP+VN)/(VP+FN+FP+VN)$

ANÁLISE DE REGRESSÃO LOGÍSTICA

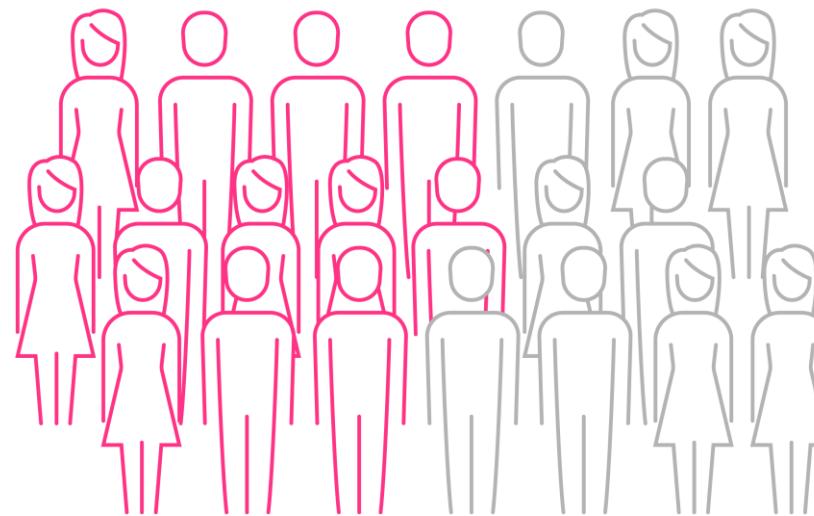
- Exemplo: Modelo Cross-Selling
- Propensão à Compra de um Produto

Objetivo:

Estabelecer público-alvo para a venda qualificada de um determinado Produto X, com uso dos mailing's internos do cliente, por meio do desenvolvimento de modelos preditivos.

MODELOS CROSS SELLING

- Propensão de compra do Produto X



SEM PRODUTO X

COM PRODUTO X

MODELOS **CROSS SELLING**

- Implementação
 - Propensão de compra do Produto X

Algoritmo Matemático

Para associar uma probabilidade de compra de um produto X a cada cliente, os seguintes passos devem ser tomados:

1. Identificar as variáveis, associando os respectivos coeficientes;
2. Somar os coeficientes encontrados no item 1, juntamente com a constante do modelo determinando o valor de Y;
3. Efetuar a operação matemática que se segue, para determinação final do score.

$$\text{Probabilidade} = 100 \times e^{-y} / (1 + e^{-y})$$

MODELOS **CROSS SELLING**

Implementação

- Propensão de compra do Produto X

Regra de Decisão Estatística

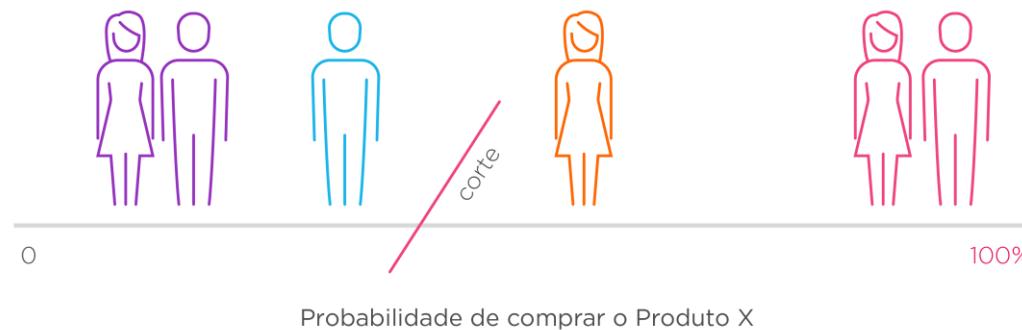
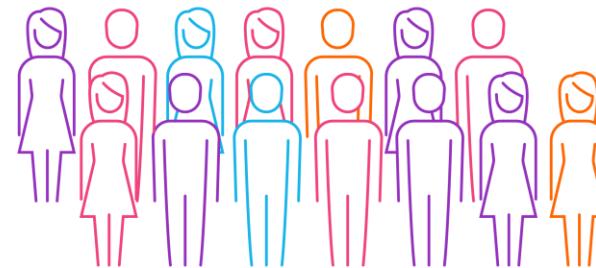
Após associar a cada indivíduo sua probabilidade de compra do produto, deve-se submetê-la à Regra de Decisão, ou seja, se a probabilidade obtida for menor ou igual ao valor de corte* o assinante pertencerá ao grupo que não irá adquirir o produto, caso contrário, se essa probabilidade for maior que o valor de corte, ele pertencerá ao grupo que irá adquirir.

* valor de corte é o valor de probabilidade que define os grupos, segundo análise de acertos do modelo.

MODELOS CROSS SELLING

- Propensão de compra do Produto X

Regra de Decisão

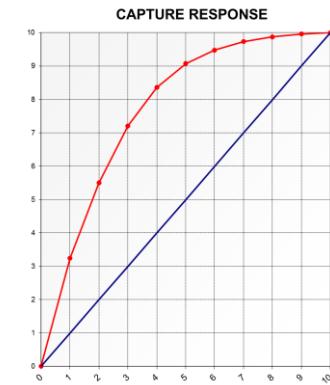


AVALIAÇÃO DO MODELO PREDITIVO

Decil	Clientes	Base sem utilização de modelos						
		Penetração			Lift	Lift Ac.	Capture	
		Qtde	%	% Ac.			%	% Ac.
1	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	10,0%
2	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	20,0%
3	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	30,0%
4	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	40,0%
5	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	50,0%
6	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	60,0%
7	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	70,0%
8	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	80,0%
9	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	90,0%
10	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	100,0%
Total	53.000	6.120	11,5%					

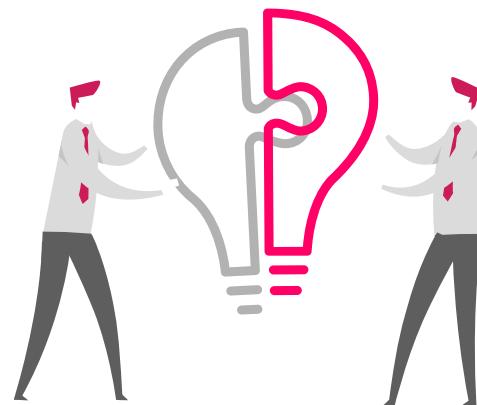
AVALIAÇÃO DO MODELO PREDITIVO

Decil	Clientes	Aplicando Modelo							
		Qtde	Qtde	Penetração		Lift	Lift Ac.	Capture	
				%	% Ac.			%	% Ac.
1	5.300	1.986	37,5%	37,5%	37,5%	3,24	3,24	32,4%	32,4%
2	5.300	1.379	26,0%	31,7%	26,0%	2,25	2,75	22,5%	55,0%
3	5.300	1.046	19,7%	27,7%	19,7%	1,71	2,40	17,1%	72,1%
4	5.300	706	13,3%	24,1%	13,3%	1,15	2,09	11,5%	83,6%
5	5.300	440	8,3%	21,0%	8,3%	0,72	1,82	7,2%	90,8%
6	5.300	244	4,6%	18,2%	4,6%	0,40	1,58	4,0%	94,8%
7	5.300	157	3,0%	16,1%	3,0%	0,26	1,39	2,6%	97,3%
8	5.300	87	1,6%	14,3%	1,6%	0,14	1,23	1,4%	98,8%
9	5.300	52	1,0%	12,8%	1,0%	0,08	1,11	0,8%	99,6%
10	5.300	24	0,5%	11,5%	0,5%	0,04	1,00	0,4%	100,0%
Total	53.000	6.120	11,5%						



EXERCITANDO

REGRESSÃO
LOGÍSTICA



Inadimplência

TÉCNICAS DE **CLASSIFICAÇÃO**

KNN-
K-Nearest Neighbors

KNN K-Nearest Neighbors

O que é?

É um algoritmo utilizado em Machine Learning para **classificar uma nova observação** utilizando medida de similaridade no espaço.

KNN K-Nearest Neighbors

Como funciona?

- Algoritmo de *Machine Learning* utilizado para classificação, mas pode ser utilizado para predição;
- Pressuposto: Elementos do mesmo grupo estão localizados próximos no espaço;
- Utilizado para imputação de dados faltantes;
- Medida de distância é utilizada para o cálculo da proximidade (sensibilidade a escala).

KNN K-Nearest Neighbors

Como funciona?

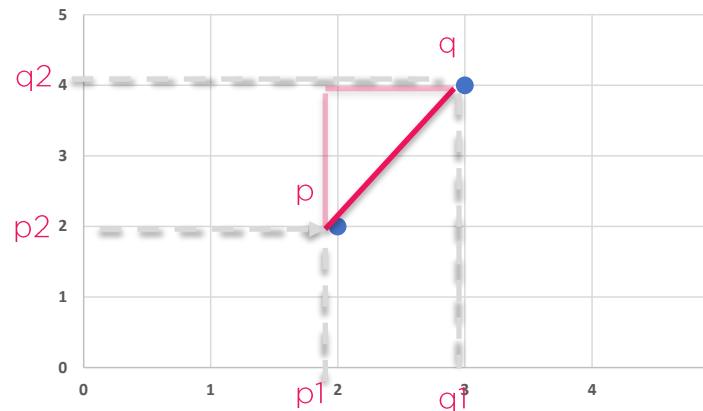
- Aprendizado baseado em exemplo (*Instance-based Learning*):
- Não constroem descrições gerais e explícitas (função alvo) a partir dos exemplos de treinamento;
- Armazena-se uma base de exemplos (*instances*) que é usada para realizar a classificação e de uma nova base;
- Em geral apresenta alto custo computacional.

KNN K-Nearest Neighbors

- Variáveis numéricas (variável qualitativa deve ser transformada em número)
- Medida de distância utilizada para o cálculo da proximidade

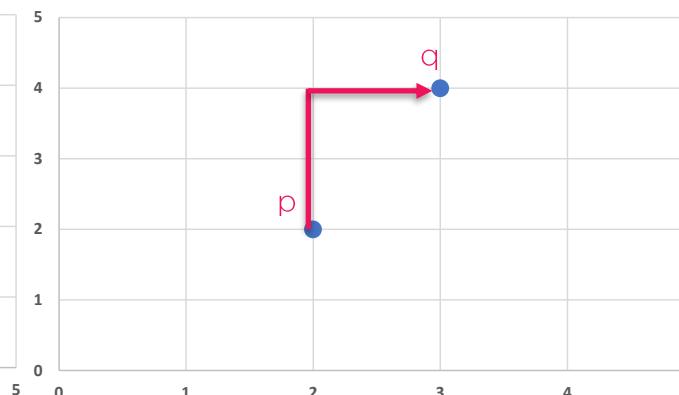
Distância Euclidiana

$$DE(\vec{p}, \vec{q}) = \sqrt{\sum (p_i - q_i)^2}$$



Distância Manhattan

$$DM(\vec{p}, \vec{q}) = \sum |p_i - q_i|$$

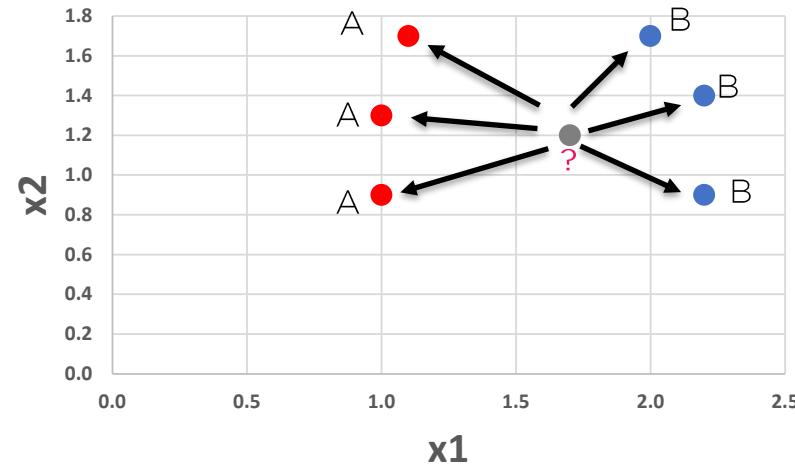


KNN K-Nearest Neighbors

- Exemplo:

Objeto	X1	X2
A	1.0	0.9
A	1.3	1.7
A	1.0	1.3
B	2.0	1.7
B	2.1	1.4
B	2.2	0.9
?	1.7	1.2

Qual a distância euclidiana entre os pontos?



O novo objeto:

pertence ao grupo A ou B ?

KNN K-Nearest Neighbors

Exemplo:

Objeto ? X1=1.7 X2=1.2

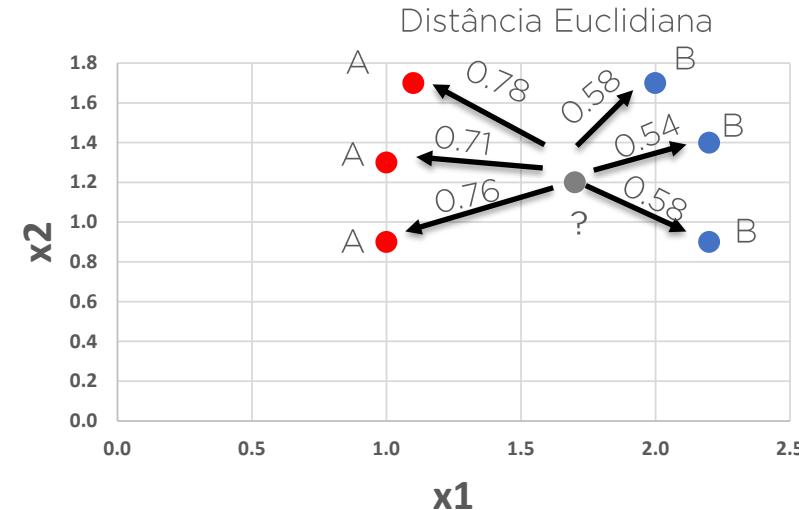
Objeto	x1	x2	Distância euclidiana (DE)	DE(x1,x 2)
A	1.0	0.9	$\sqrt{(1.0 - 1.7)^2 + (0.9 - 1.2)^2}$	0.76
A	1.1	1.7	$\sqrt{(1.1 - 1.7)^2 + (1.7 - 1.2)^2}$	0.78
A	1.0	1.3	$\sqrt{(1.0 - 1.7)^2 + (1.3 - 1.2)^2}$	0.71
B	2.0	1.7	$\sqrt{(2.0 - 1.7)^2 + (1.7 - 1.2)^2}$	0.58
B	2.2	1.4	$\sqrt{(2.2 - 1.7)^2 + (1.4 - 1.2)^2}$	0.54
B	2.2	0.9	$\sqrt{(2.2 - 1.7)^2 + (0.9 - 1.2)^2}$	0.58

K-Nearest Neighbors

Exemplo:

Objeto ? $x_1=1.7$ $x_2=1.2$

Objeto	x_1	x_2	Distância euclidiana (DE)	$DE(x_1, x_2)$
A	1.0	0.9	$\sqrt{(1.0 - 1.7)^2 + (0.9 - 1.2)^2}$	0.76
A	1.1	1.7	$\sqrt{(1.1 - 1.7)^2 + (1.7 - 1.2)^2}$	0.78
A	1.0	1.3	$\sqrt{(1.0 - 1.7)^2 + (1.3 - 1.2)^2}$	0.71
B	2.0	1.7	$\sqrt{(2.0 - 1.7)^2 + (1.7 - 1.2)^2}$	0.58
B	2.2	1.4	$\sqrt{(2.2 - 1.7)^2 + (1.4 - 1.2)^2}$	0.54
B	2.2	0.9	$\sqrt{(2.2 - 1.7)^2 + (0.9 - 1.2)^2}$	0.58



Hiperparâmetro
 $K = 1$

Menor distância
0.54
O novo objeto é
classificado como B

K-Nearest Neighbors

Exemplo:

Objeto ? $x_1 = 1.7$ $x_2 = 1.2$

Objeto	x_1	x_2	Distância euclidiana (DE)	$DE(x_1, x_2)$
A	1.0	0.9	$\sqrt{(1.0 - 1.7)^2 + (0.9 - 1.2)^2}$	0.76
A	1.1	1.7	$\sqrt{(1.1 - 1.7)^2 + (1.7 - 1.2)^2}$	0.78
A	1.0	1.3	$\sqrt{(1.0 - 1.7)^2 + (1.3 - 1.2)^2}$	0.71
B	2.0	1.7	$\sqrt{(2.0 - 1.7)^2 + (1.7 - 1.2)^2}$	0.58
B	2.2	1.4	$\sqrt{(2.2 - 1.7)^2 + (1.4 - 1.2)^2}$	0.54
B	2.2	0.9	$\sqrt{(2.2 - 1.7)^2 + (0.9 - 1.2)^2}$	0.58

Hiperparâmetro
 $K = 2$

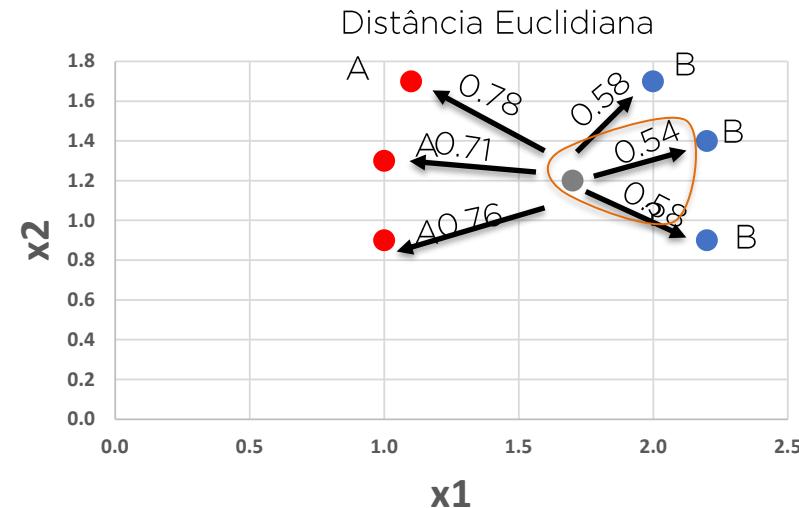


2



0

Para $k=2$, temos 2-NN e a classe do novo objeto seria B, pois temos 2 vizinhos na classe B e zero vizinhos na classe A



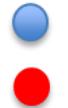
K-Nearest Neighbors

Exemplo:

Objeto ? $x_1 = 1.7$ $x_2 = 1.2$

Objeto	x_1	x_2	Distância euclidiana (DE)	$DE(x_1, x_2)$
A	1.0	0.9	$\sqrt{(1.0 - 1.7)^2 + (0.9 - 1.2)^2}$	0.76
A	1.1	1.7	$\sqrt{(1.1 - 1.7)^2 + (1.7 - 1.2)^2}$	0.78
A	1.0	1.3	$\sqrt{(1.0 - 1.7)^2 + (1.3 - 1.2)^2}$	0.71
B	2.0	1.7	$\sqrt{(2.0 - 1.7)^2 + (1.7 - 1.2)^2}$	0.58
B	2.2	1.4	$\sqrt{(2.2 - 1.7)^2 + (1.4 - 1.2)^2}$	0.54
B	2.2	0.9	$\sqrt{(2.2 - 1.7)^2 + (0.9 - 1.2)^2}$	0.58

Hiperparâmetro
 $K = 3$

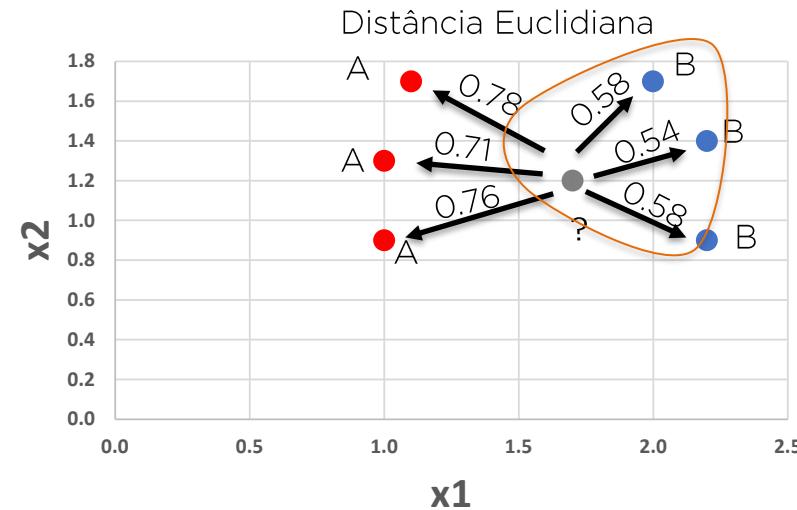


3



0

Para $k=3$, temos 3-NN e a classe do novo objeto seria B, pois temos 3 vizinhos na classe B e zero vizinhos na classe A



KNN K-Nearest Neighbors

Algoritmo KNN (k vizinhos mais próximos)

- Determinar o valor de K;
- Calcular a distância entre os elementos;
- Dentre os K vizinhos mais próximos descobrir o grupo mais frequente;
- Retorna o grupo mais frequente entre os vizinhos mais próximos, isto é, calcula a Moda dos K vizinhos.

KNN K-Nearest Neighbors

Hiperparâmetro K vizinhos:

- Determinar o valor para k vizinhos antes do treino;
- Testar diferentes valores para k vizinhos;
- Utilizar a técnica de validação cruzada no processo de encontrar o “melhor” valor de k vizinhos;
- O teste do hiperparâmetro é realizado na amostra de validação.

KNN K-Nearest Neighbors

Resumo do processo realizado pelo algoritmo KNN:

- Recebe um dado não classificado;
- Padronização das variáveis;
- Mede a distância do novo dado em relação a cada um dos outros dados que já estão classificados;
- Seleciona as K menores distâncias;
- Verifica a(s) classe(s) dos dados que tiveram as K menores distâncias e contabiliza a quantidade de vezes que cada classe que apareceu;
- Classifica esse novo dado como pertencente à classe que mais apareceu.

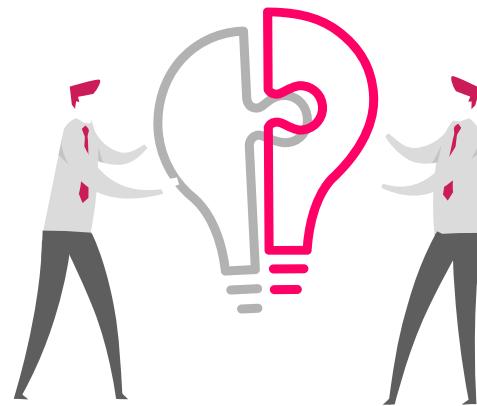
KNN K-Nearest Neighbors

Quando der empate entre as classes:

- Optar por k- de quantidade ímpar;
- Calcular o k-NN ponderado pela distância, possibilitando uma menor chance de empates.

EXERCITANDO

KNN



Inadimplência

BIBLIOGRAFIA

- KUHN, M. / JOHNSON K. *Applied Predictive Modeling*, 1st ed. 2013, Corr. 2nd printing 2018 Edition
- LESKOVEC, RAJAMARAM, ULLMAN. *Mining of Massive Datasets*, 2014. <http://mmds.org>.
- HAIR, J.F. / ANDERSON, R.E. / TATHAN, R.L. / BLACK, W.C. *Análise multivariada de dados*, 2009
- TORGÓ, L. *Data Mining with R: Learning with Case Studies*, 2.a ed. Chapman and Hall/CRC , 2007
- MINGOTI, S.A.; *Análise de dados através de métodos de estatística multivariada*, UFMG, 2005
- CARVALHO, L.A.V., *Datamining - A mineração de dados no marketing, medicina, economia, engenharia e administração*. Rio de Janeiro: Editora Ciência Moderna, 2005.
- BERRY,M.J.A., LINOFF,G. *Data Mining Techniques For Marketing, Sales and Customer Support*. 3a. ed. New York: John Wiley & Sons, Inc., 2011.
- DUNHAM, M.H. *Data Mining - Introductory and Advanced Topics*. Prentice Hall, 2002.
- DINIZ,C.A.R. , NETO F.L. *Data Mining: Uma Introdução*. São Paulo: XIV Simpósio Nacional de Probabilidade e Estatística. IME-USP, 2000.

OBRIGADA!



/adelaide-alves



profadelaide.alves@fiap.com.br

FIAP

Copyright © 2023 | Professor (a) Adelaide Alves de Oliveira

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