

FIAP

MBA

ADELAIDE ALVES DE OLIVEIRA

PROFESSORA



profadelaide.alves@fiap.com.br

Formação Acadêmica

- Bacharel em Estatística – UNICAMP
- Mestre em Ciências – FSP/USP

Atividades Profissionais

- Diretora Técnica Estatística da empresa **SD&W** - www.sdw.com.br
- Professora dos cursos de MBA, formatos: presencial, online e EAD das disciplinas de Fundamentos Estatísticos, DataMining, Análise Preditiva e Machine Learning dos cursos: Big Data, Data Science, Business Intelligence & Analytics, Digital Data Marketing, IA & ML e Engenharia de Dados e nos Shift: People Analytics e Python Journey

O QUE SABEMOS DE MACHINE LEARNING:



O QUE É ML?

MACHINE LEARNING

- O Machine Learning (Aprendizado de Máquinas), é uma área da Inteligência Artificial, tem por objetivo à busca de um conjunto de regras e procedimentos para permitir que as máquinas possam agir e tomar decisões baseadas em dados, ao invés de serem explicitamente programadas para realizar uma determinada tarefa.

Ao invés de criar um programa especificando os passos para executar sua tarefa, no aprendizado de máquinas utilizamos algoritmos que aprendem uma tarefa conforme seu treinamento.

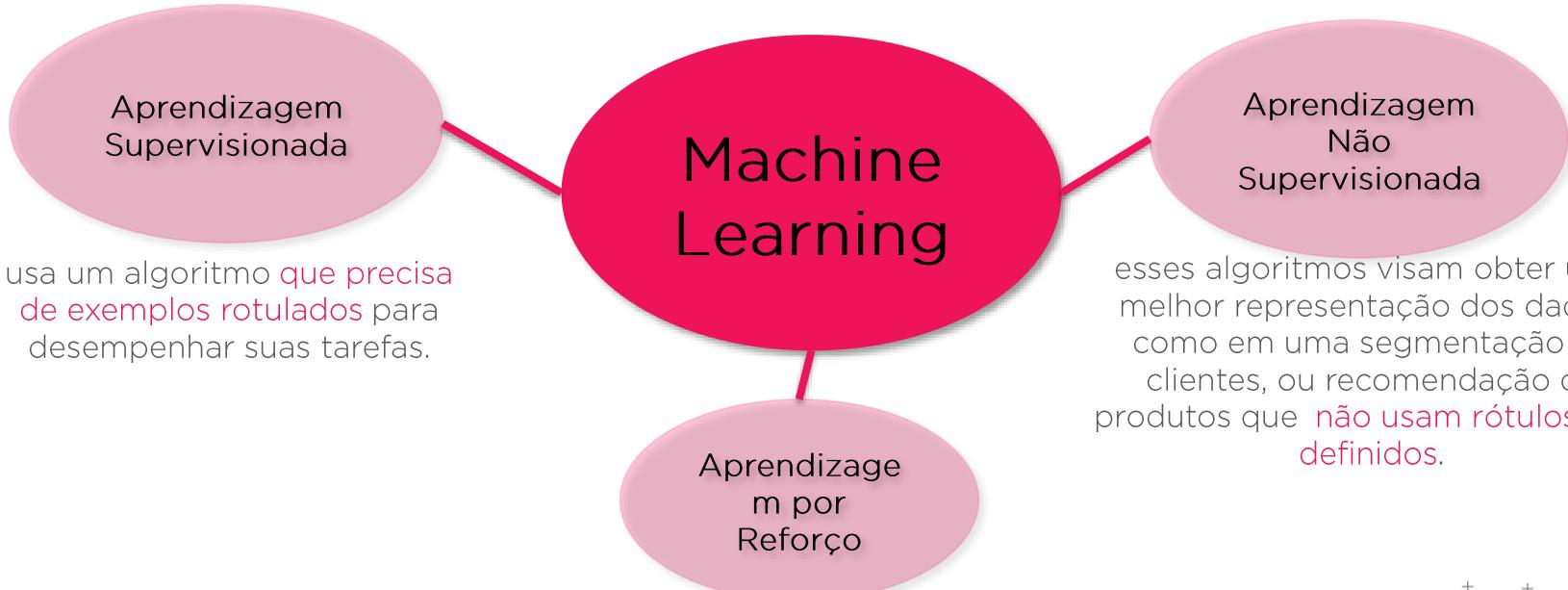
Dessa forma, ao analisarem um grande volume de informações, elas são capazes de identificar padrões e de tomar decisões com o auxílio de modelos. Isso torna as máquinas capazes de fazer predições por meio do processamento de dados.

O QUE SABEMOS DE MACHINE LEARNING:



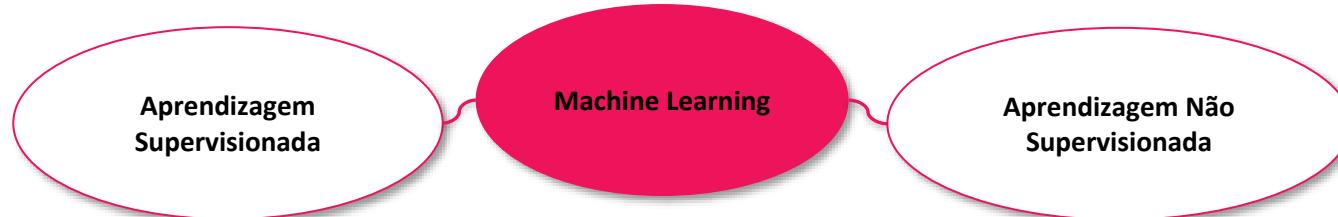
QUAIS AS
TÉCNICAS DE ML?

- TÉCNICAS de **MACHINE LEARNING**



conseguem desenvolver uma política de ações visando uma determinada recompensa. Aqui entram, por exemplo, máquinas que aprendem a jogar xadrez e tomam decisões embasadas em cada estado do jogo para maximizar a recompensa final (vencer a partida).

• ALGORITMOS de **MACHINE LEARNING**



- 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).
- 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.



• ALGORITMOS de **MACHINE LEARNING**

- e também, Aprendizado por Reforço

A Aprendizagem Por Reforço (ou *Reinforcement Learning*) é o treinamento de modelos de aprendizado de máquina para tomar uma sequência de decisões em um ambiente incerto e potencialmente complexo. No aprendizado por reforço, o sistema de inteligência artificial enfrenta uma situação. O computador utiliza tentativa de erro e acertos para encontrar uma solução para o problema. Para que a máquina faça o que o programador deseja, a inteligência artificial recebe recompensas ou penalidades pelas ações que executa. Seu objetivo é maximizar a recompensa total.

Muito usada em Games e Robótica.

Exemplo: AlphaGo.



O QUE SABEMOS DE MACHINE LEARNING:



QUAIS AS
DIFERENÇAS ENTRE
MODELOS
ESTATÍSTICOS E
MODELOS DE ML?



- ALGORITMOS de **MACHINE LEARNING**

- Machine Learning

- Baseado em algoritmos
- O objetivo é identificar o que funciona
- Interessa em prever os resultados de amostras futuras
- Foco na praticidade ➔ Desenvolve em uma amostra e aplica em outra
- Algoritmos para tomada de decisão
- Limitações/grandes desafios:
 - Tendência ao sobre ajuste
 - Dados influenciados por erros de medição e fatores aleatórios
 - Ajuste perfeito para um grupo de dados e pode não funcionar bem para outro
 - Algoritmo preconceituoso

Estatística Tradicional

- Inferência estatística
- Conjunto de técnicas estatísticas baseadas em I.C. e erro padrão

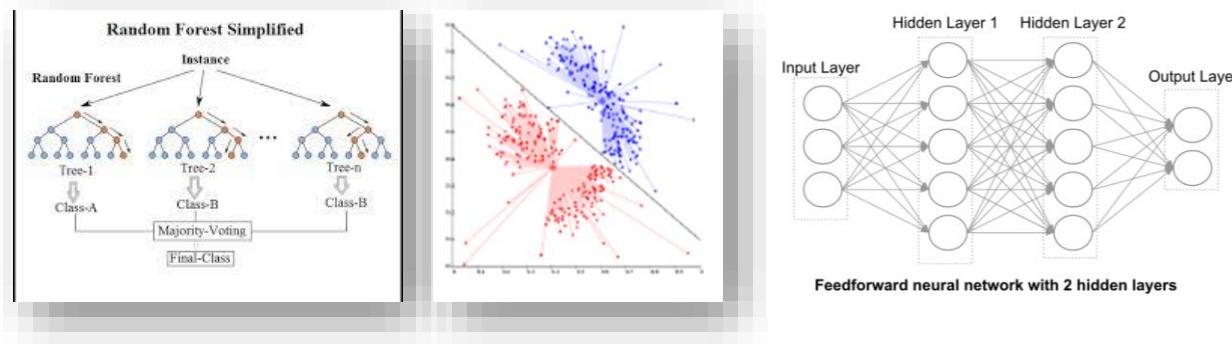
Interesse nas variáveis explicativas:

Exemplo:

- Quais os fatores de risco para o AVC?
- Quais as variáveis que impactam no churn ou na análise de crédito?



- ALGORITMOS de **MACHINE LEARNING**
- Problemas **práticos** de predição (para tomada de decisão)
 - Pouco interesse em interpretar os modelos
 - Liberdade para modelar a complexidade do mundo real



Se machine learning não se importa muito com interpretação, então se importa de fato com o quê?

➔ **Performance preditiva** (ou seja, acurácia das decisões)

O QUE SABEMOS DE MACHINE LEARNING:



QUAIS SÃO AS
ETAPAS DE ML?

• CICLO ANALÍTICO

Entender o problema de Negócio



Coletar DADOS



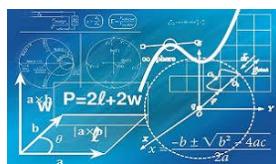
Explorar/
Visualizar



Feature
Engineering



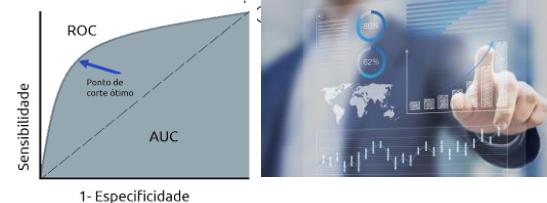
Feature Extraction / Selection



Machine Learning



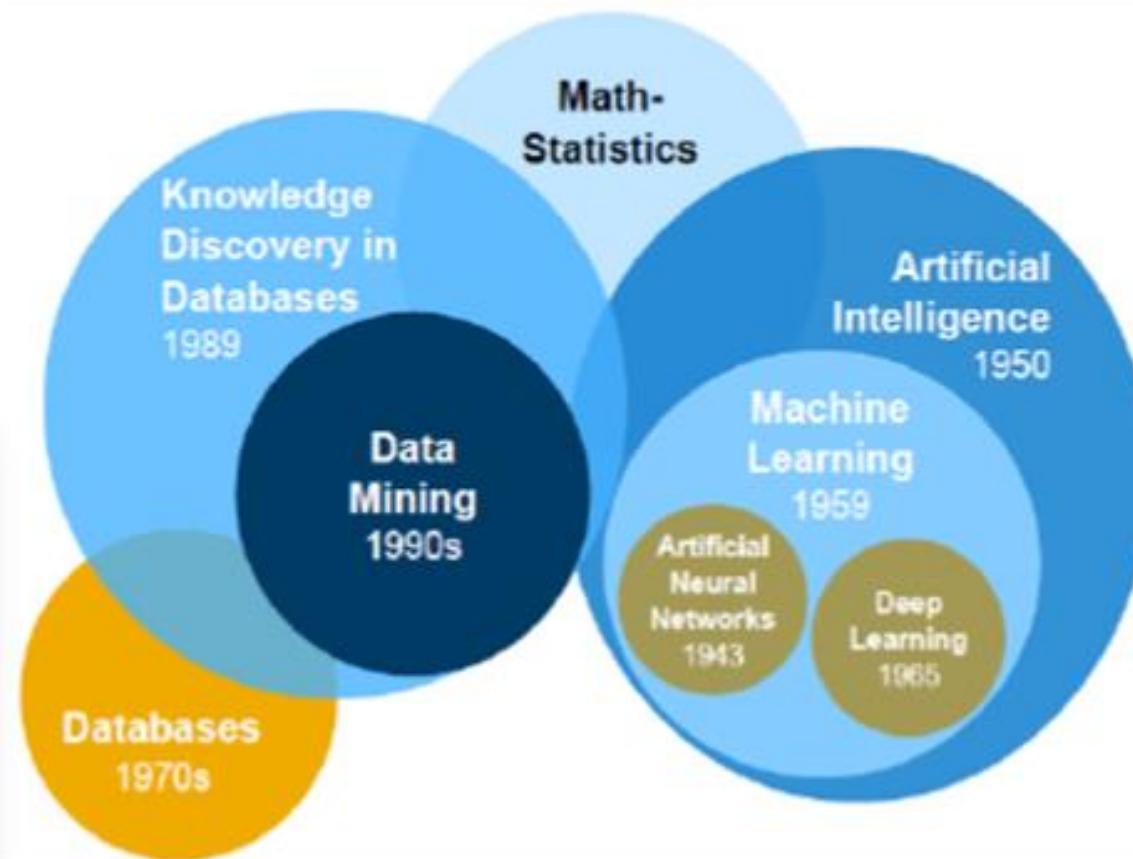
Validação / Monitoramen



Deploy / Implement

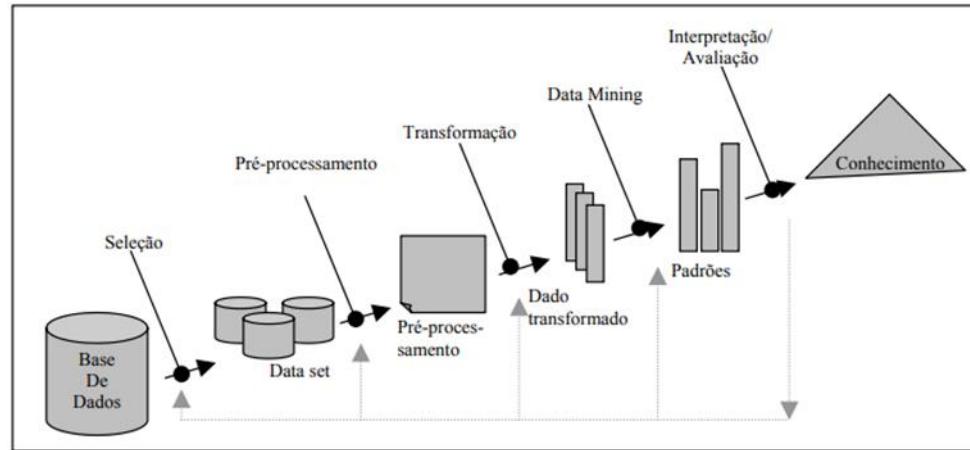


LEMBRANDO.....



PROCESSO KDD

KNOWLEDGE DISCOVERY IN DATABASES



Fonte: Processo de KDD. Adaptado de Fayyad et al. (1996a).

- CICLO
- **ANALÍTICO**

A seleção dos dados varia de acordo com os objetivos do estudo.

A escolha das variáveis de entrada que formam o conjunto de dados-alvo requer uma maior atenção, sempre com foco na busca dos objetivos do estudo.

Pré-selecionar variáveis que sejam preditoras plausíveis (bom senso do pesquisador).

Cuidado com vazamento de informação (“*data leakage*”).

Acontece quando os dados de treino apresentam informação escondida que faz com que o modelo aprenda padrões que não são do seu interesse.

Ter uma variável preditora escondida no conjunto de dados que “mede” a variável de interesse.

CICLO ANALÍTICO

- *Feature engineering* também conhecida como etapa de Pré-processamento de dados - uma das fases mais importantes do processo de construção de um modelo preditivo:

termo utilizado para definir um conjunto de técnicas utilizadas na criação e manipulação de features/variáveis/atributos, tendo como objetivo desenvolver um bom modelo de aprendizado de máquina. Etapa abrange a transformação matemática nas features existentes para extrair o máximo potencial dos dados e criação de novas features.

- *Feature selection* é o processo de seleção de um subconjunto de features relevantes para uso na construção do modelo.
 - *Feature extraction* cria novos recursos, projetando os dados de um espaço de alta dimensão para um espaço com menos dimensões.

O QUE SABEMOS DE MACHINE LEARNING:



QUAIS OS
CUIDADOS COM AS
FEATURES?

PRÓCESSO KDD

KNOWLEDGE DISCOVERY IN DATABASES



Preparação dos Dados : Transformação

Um dos objetivos principais da transformação de dados é converter o conjunto bruto de dados em uma forma padrão de uso.

Existem várias técnicas de transformações de dados. Essas técnicas usadas adequadamente sinalizam descobertas do estudo em análise.

TÉCNICAS DE TRANSFORMAÇÃO DOS DADOS

- ➔ Discretização - converter variáveis contínuas em categorizadas (variáveis discretas)

Exemplo: Faixas de Renda:
Até R\$ 1.500,00 = 1;
de R\$ 1.500,00 a R\$ 4.500,00 = 2
mais de R\$ 4.500,00 = 3



- ➔ Agregação - uma operação resumo é aplicada aos dados.

Exemplo: Vendas mensais calculadas a partir das vendas diárias.

- ➔ Construir novas variáveis - a partir de um conjunto de outras variáveis.

Exemplo de transformação: criar proporção, transformações logarítmicas, etc



- ➔ Alisamento - usada para remover valores discrepantes

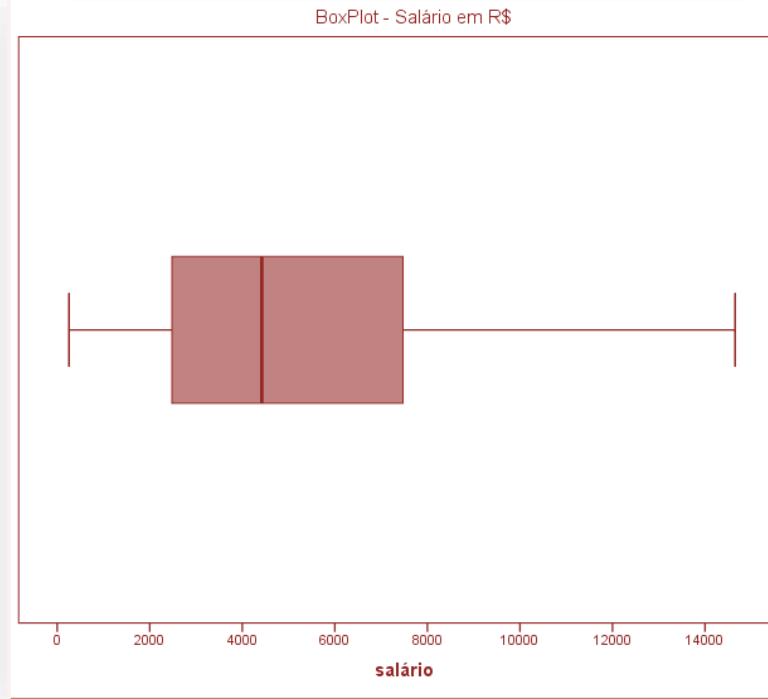
Exemplo: Médias móveis



TRANSFORMAÇÃO DOS DADOS: DISCRETIZAÇÃO

id	salário
1	4.763,75
2	7.391,72
3	729,33
4	2.376,28
5	1.887,72
6	1.207,36
7	4.745,39
8	3.635,80
9	8.119,15
10	2.356,41
11	13.502,54
12	2.655,92
13	3.920,45
14	853,32
15	12.819,59
16	10.088,13
17	4.414,62
18	7.293,00
19	11.445,93
20	8.339,63
21	4.858,72
22	1.616,16
23	1.339,24
24	7.108,82
25	2.054,73
26	1.441,01
27	8.981,38
28	8.753,71
29	3.426,82
38	1.049,08
39	9.072,00
40	3.273,02

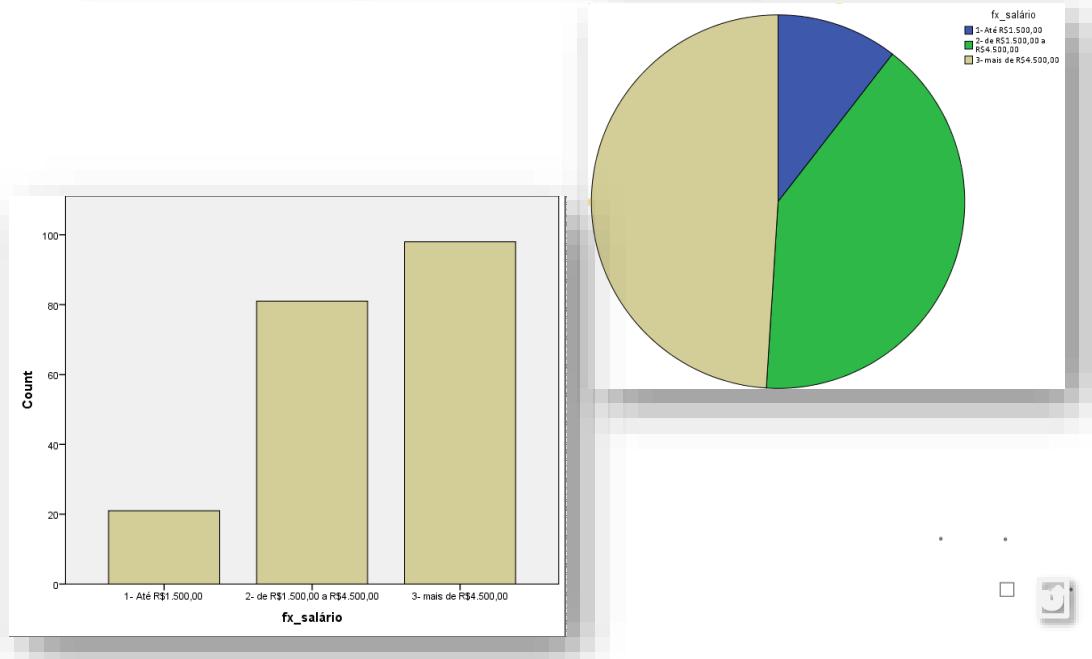
Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
salário	200	254,19	14.649,52	5.259,30	3.615,36
Valid N (listwise)	200				



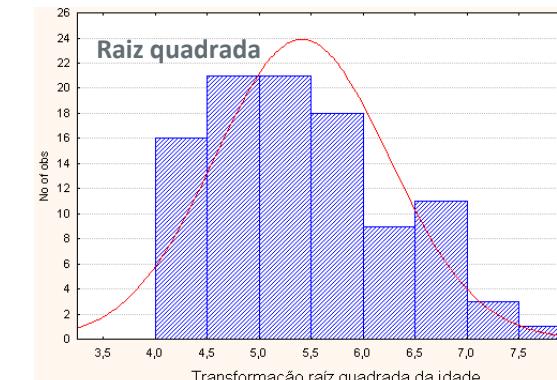
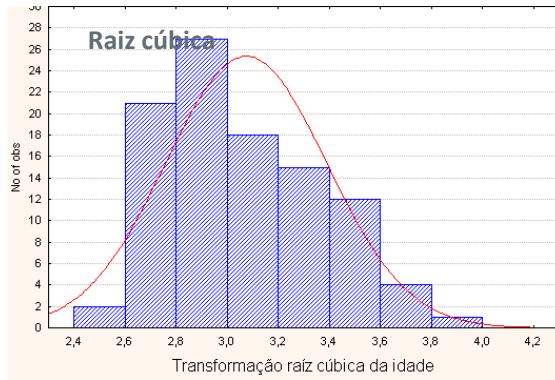
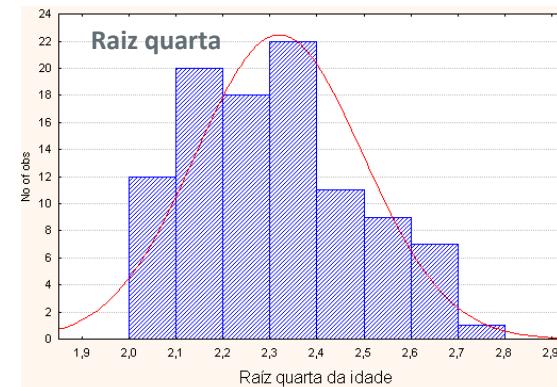
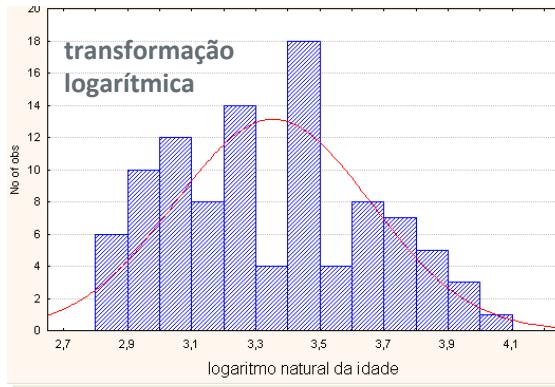
TRANSFORMAÇÃO DOS DADOS: DISCRETIZAÇÃO

id	salário	fx_sal
1	4.763,75	3
2	7.391,72	3
3	729,33	1
4	2.376,28	2
5	1.887,72	2
6	1.207,36	1
7	4.745,39	3
8	3.635,80	2
9	8.119,15	3
10	2.356,41	2
11	13.502,54	3
12	2.655,92	2
13	3.920,45	2
14	853,32	1
15	12.819,59	3
16	10.088,13	3
17	4.414,62	2
18	7.293,00	3
19	11.445,93	3
20	8.339,63	3
21	4.858,72	3
22	1.616,16	2
23	1.339,24	1
24	7.108,82	3
25	2.054,73	2
26	1.441,01	1
27	8.981,38	3
28	8.753,71	3
29	3.426,82	2
30	3.873,20	2
31	1.165,56	1
32	5.431,64	3
40	3.273,02	2

fx_salário	Frequency	Percent	Valid Percent	Cumulative Percent
1- Até R\$1.500,00	21	10,5	10,5	10,5
2- de R\$1.500,00 a R\$4.500,00	81	40,5	40,5	51,0
3- mais de R\$4.500,00	98	49,0	49,0	100,0
Total	200	100,0	100,0	

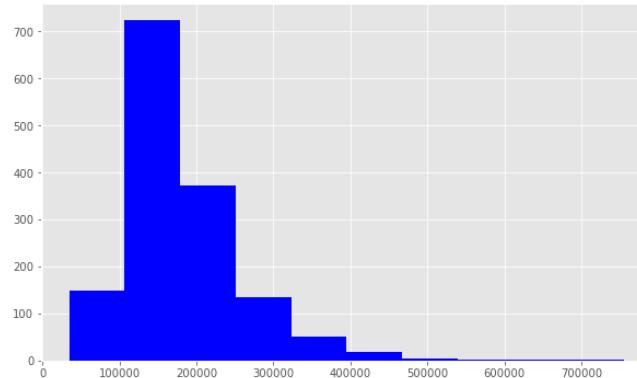


TRANSFORMAÇÃO DOS DADOS: CONSTRUÇÃO DE NOVAS VARIÁVEIS



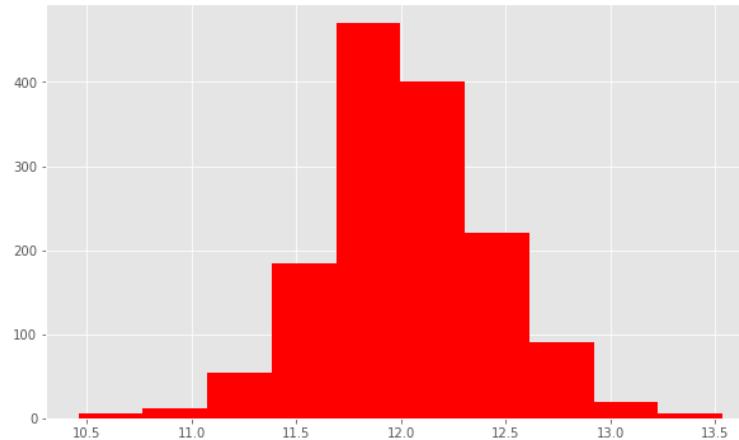
TRANSFORMAÇÃO DOS DADOS: CONSTRUÇÃO DE NOVAS VARIÁVEIS

Valor de imóveis de uma certa localidade



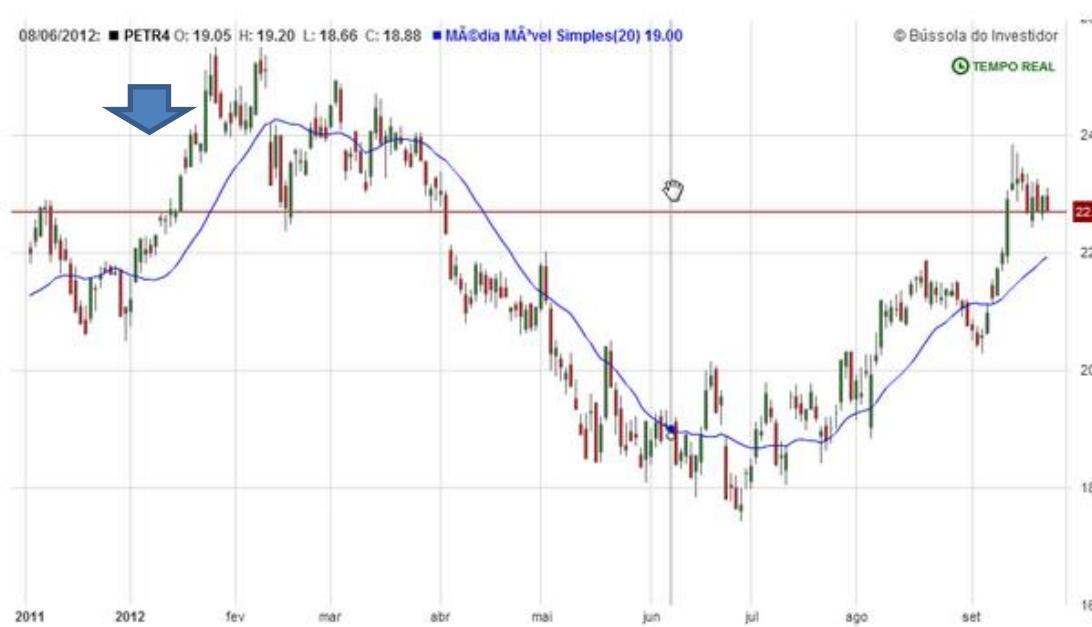
mean	180.921
std	79.442
min	34.900
25%	129.975
50%	163.000
75%	214.000
max	755.000

transformação logarítmica



mean	12.01
std	0.40
min	10.52
25%	11.78
50%	12.00
75%	12.27
max	13.53

TRANSFORMAÇÃO DOS DADOS: MÉDIAS MÓVEIS

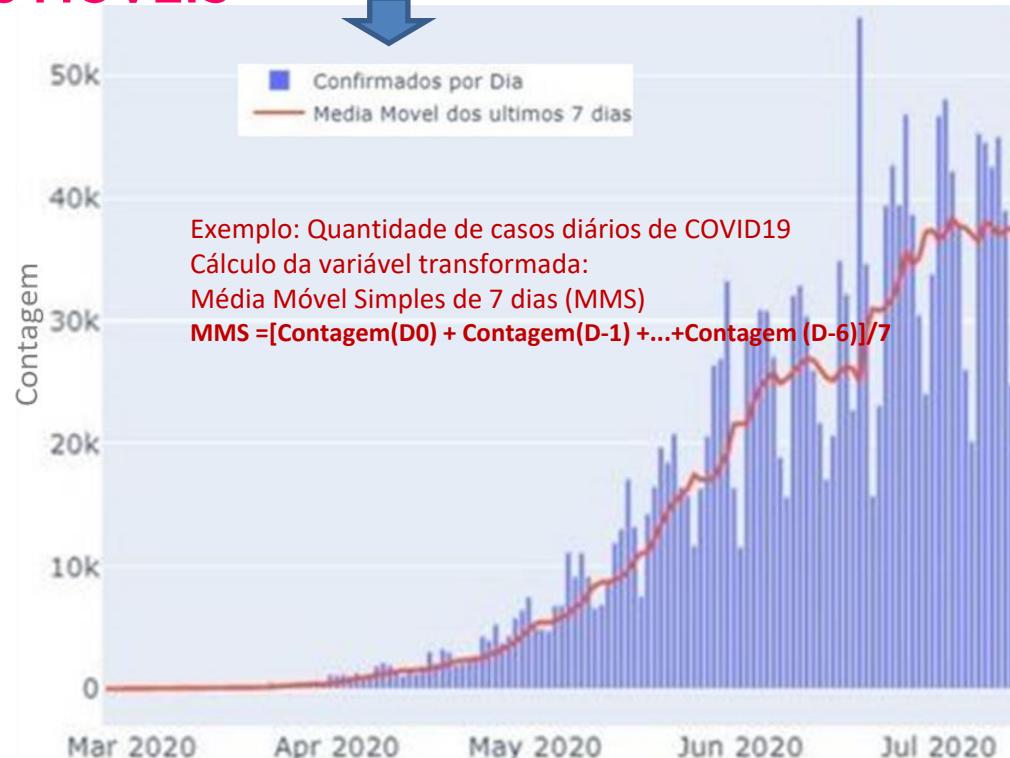


Média Móvel Simples
de 20 dias

$$\text{MMS} = [\text{Fecham.(D0)} + \text{Fecham.(D-1)} + \text{Fecham.(D-2)} + \dots + \text{Fecham.(D-19)}]/20$$

TRANSFORMAÇÃO DOS DADOS: MÉDIAS MÓVEIS

Exemplo



Série histórica de notificações diárias do COVID19

Fonte: COVID-19 BRASIL

TÉCNICAS DE TRANSFORMAÇÃO DOS DADOS

- ➔ Normalização Min-Max - transformação, onde os dados de um atributo são normalizados gerando valores entre 0,0 a 1,0.

$$\text{valor transformado} = \frac{\text{valor original} - \text{valor mínimo}}{\text{valor máximo} - \text{valor mínimo}}$$

Por exemplo: suponha que os valores mínimo e máximo da variável rendimento são R\$ 360,00 e R\$ 15.800,00, respectivamente. Transformar a variável rendimento na faixa [0,0; 1,0]. Um valor de rendimento igual a R\$ 5.300,00, transforma-se em:

$$w = \frac{5.300 - 360}{15.800 - 360} = 0,32$$

- ➔ Padronização - Transforma os valores em números de desvios padrões a partir da média. É dada por:

$$z = \frac{x - \bar{x}}{s}$$

Observação:

- A escala das variáveis pode afetar muito a qualidade das previsões.
- Alguns algoritmos darão preferência para utilizar variáveis com valores muito alto.
- Padronizar as variáveis contínuas para todas terem média de 0 e desvio-padrão de 1 ou normalizar gerando valores dentre 0,0 a 1,0.

TRANSFORMAÇÃO: Normatização Min-Max

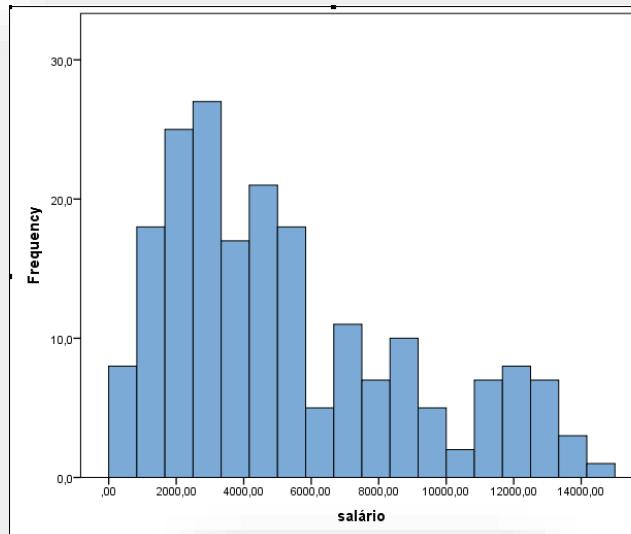
	id	salário
1	4.763,75	
2	7.391,72	
3	729,33	
4	2.376,28	
5	1.887,72	
6	1.207,36	
7	4.745,39	
8	3.635,80	
9	8.119,15	
10	2.356,41	
11	13.502,54	
12	2.655,92	
13	3.920,45	
14	853,32	
15	12.819,59	
16	10.088,13	
17	4.414,62	
18	7.293,00	
19	11.445,93	
20	8.339,63	
21	4.858,72	
22	1.616,16	
23	1.339,24	
24	7.108,82	
25	2.054,73	
26	1.441,01	
27	8.981,38	
28	8.753,71	
29	3.426,82	
30	3.873,20	
31	1.165,56	
32	5.431,64	
33	12.541,13	
39	9.072,00	
40	3.273,02	

$$sal_norm1 = \frac{salário - \text{mínimo}}{\text{máximo} - \text{mínimo}}$$

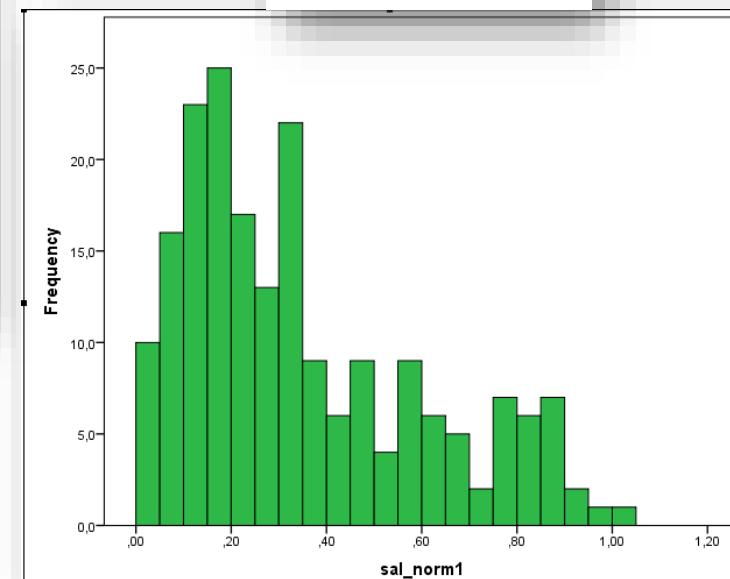
Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
salário	200	254,19	14649,52	5259,30	3615,36
sal_norm1	200	0,00	1,00	0,35	0,25
Valid N (listwise)	200				

id	salário	sal_norm1
1	4.763,75	0,31
2	7.391,72	0,5
3	729,33	0,03
4	2.376,28	0,15
5	1.887,72	0,11
6	1.207,36	0,07
7	4.745,39	0,31
8	3.635,80	0,23
9	8.119,15	0,55
10	2.356,41	0,15
11	13.502,54	0,92
12	2.655,92	0,17
13	3.920,45	0,25
14	853,32	0,04
15	12.819,59	0,87
16	10.088,13	0,68
17	4.414,62	0,29
18	7.293,00	0,49
19	11.445,93	0,78
20	8.339,63	0,56
21	4.858,72	0,32
22	1.616,16	0,09
23	1.339,24	0,08
24	7.108,82	0,48
25	2.054,73	0,13
26	1.441,01	0,08
27	8.981,38	0,61
28	8.753,71	0,59
29	3.426,82	0,22
30	3.873,20	0,25
31	1.165,56	0,06
32	5.431,64	0,36
33	12.541,13	0,85
39	9.072,00	0,61
40	3.273,02	0,21

TRANSFORMAÇÃO: Normatização Min-Max



Mean = 5259,3047
Std. Dev. = 3615,36167
N = 200



Mean = ,3477
Std. Dev. = ,25115
N = 200

TRANSFORMAÇÃO: Padronização

id	salário
1	4.763,75
2	7.391,72
3	729,33
4	2.376,28
5	1.887,72
6	1.207,36
7	4.745,39
8	3.635,80
9	8.119,15
10	2.356,41
11	13.502,54
12	2.655,92
13	3.920,45
14	853,32
15	12.819,59
16	10.088,13
17	4.414,62
18	7.293,00
19	11.445,93
20	8.339,63
21	4.858,72
22	1.616,16
23	1.339,24
24	7.108,82
25	2.054,73
26	1.441,01
33	12.541,13
34	5.889,54
35	2.585,15
36	5.146,24
37	718,91
38	1.049,08
39	9.072,00
40	3.273,02

$$z\text{salário} = \frac{\text{salário} - \text{média}}{\text{desvio}}$$

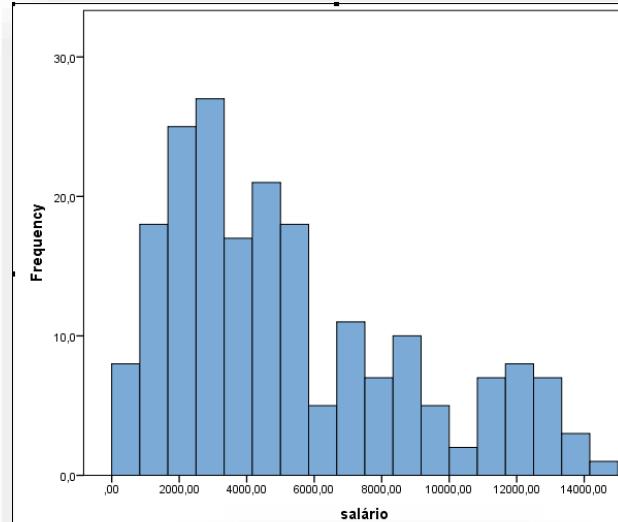
Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
salário	200	254,19	14.649,52	5.259,30	3.615,36
Zscore(salário)	200	-1,3844	2,5973	0,0000	1,0000
Valid N (listwise)	200				

Observação:

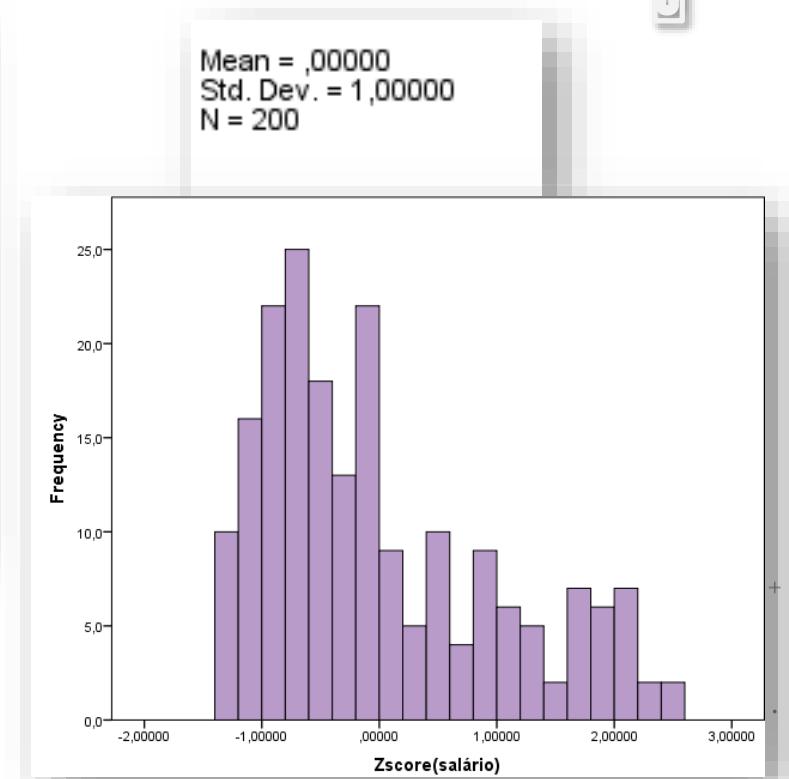
- A escala das variáveis pode afetar muito a qualidade das previsões.
- Alguns algoritmos darão preferência para utilizar variáveis com valores muito alto.
- Padronizar as variáveis contínuas para todas terem média de 0 e desvio-padrão de 1.

id	salário	Zsalário
1	4.763,75	-0,1371
2	7.391,72	0,5898
3	729,33	-1,2530
4	2.376,28	-0,7974
5	1.887,72	-0,9326
6	1.207,36	-1,1208
7	4.745,39	-0,1422
8	3.635,80	-0,4491
9	8.119,15	0,7910
10	2.356,41	-0,8029
11	13.502,54	2,2801
12	2.655,92	-0,7201
13	3.920,45	-0,3703
14	853,32	-1,2187
15	12.819,59	2,0912
16	10.088,13	1,3356
17	4.414,62	-0,2336
18	7.293,00	0,5625
19	11.445,93	1,7112
20	8.339,63	0,8520
21	4.858,72	-0,1108
22	1.616,16	-1,0077
23	1.339,24	-1,0843
24	7.108,82	0,5116
25	2.054,73	-0,8864
26	1.441,01	-1,0561
33	12.541,13	2,0141
34	5.889,54	0,1743
35	2.585,15	-0,7397
36	5.146,24	-0,0313
37	718,91	-1,2559
38	1.049,08	-1,1645
39	9.072,00	1,0546
40	3.273,02	-0,5494

TRANSFORMAÇÃO: Padronização



Mean = 5259,3047
Std. Dev. = 3615,36167
N = 200



Mean = ,00000
Std. Dev. = 1,00000
N = 200

TÉCNICAS DE **TRANSFORMAÇÃO DOS DADOS**

- A **normalização** pode ser aplicada quando a distribuição dos dados não é normal ou se o desvio padrão dos mesmos for muito pequeno;
- A **normalização** permite comparar a importância das features quando usamos modelos paramétricos como Regressão Linear e Regressão Logística;

- A **padronização** dos dados transforma os mesmos em uma distribuição normal padrão;
- A **padronização** é recomendada quando os dados estão em uma distribuição normal;

TÉCNICAS DE **TRANSFORMAÇÃO DOS DADOS**

➔ **Faixas com igual largura** - Transforma as variáveis em faixas de tamanhos fixos. A nova variável tem aproximadamente a mesma distribuição da variável original.

➔ **Faixas com igual altura** - Transforma as variáveis em decis, percentis, tal que o mesmo número de registros pertencem a uma mesma faixa. A nova variável tem distribuição uniforme.

➔ **Datas e Tempos** - Podendo ser transformada em variável tempo (o número de dias ou horas desde alguma data no passado). Neste caso os algoritmos tratam datas como números.

TÉCNICAS DE **TRANSFORMAÇÃO DOS DADOS**

→ Variáveis Categorizadas : Os algoritmos trabalham melhor com poucas categorias. Para reduzir o número de categorias pode-se usar atributos dos códigos, ou variáveis binárias para cada categoria.

“One-hot encoding”: Alguns algoritmos têm dificuldade em entender variáveis que têm mais de uma categoria. Acham que é uma variável contínua (0,1,2,3...)

➔ porém não tem significado contínuo. A solução é transformar todas as categorias em uma variável diferente de valores 0 e 1 (one-hot encoding)

Variável com n categorias ➔ criar n variáveis .

Pode trazer problemas em alguns modelos, como na regressão linear (solução criar *dummy*) (n-1 variáveis).

TÉCNICAS DE **TRANSFORMAÇÃO DOS DADOS**

→ **Variáveis Categorizadas** : Os algoritmos trabalham melhor com poucas categorias. Para reduzir o número de categorias pode-se usar atributos dos códigos, ou variáveis binárias para cada categoria.

Exemplo: Variável categórica SEXO {Masculino, Feminino}

- Variável Transformada: Masculino
- Masculino=1 se SEXO= “Masculino”; Masculino=0, caso contrário
- A variável Masculino é uma variável numérica, também chamada de variável DUMMY muito utilizada na construção de modelos.

TÉCNICAS DE TRANSFORMAÇÃO DOS DADOS

→ Variáveis Categorizadas : Os algoritmos trabalham melhor com poucas categorias. Para reduzir o número de categorias pode-se usar atributos dos códigos, ou variáveis binárias para cada categoria.

Quando conseguimos ordenar os dados usamos o Label Enconding

Quando não conseguimos ordenar os dados usamos o One Hot Enconding

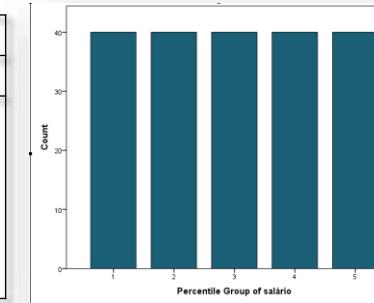
TRANSFORMAÇÃO: Categorização

id	salário
1	4.763,75
2	7.391,72
3	729,33
4	2.376,28
5	1.887,72
6	1.207,36
7	4.745,39
8	3.635,80
9	8.119,15
10	2.356,41
11	13.502,54
12	2.655,92
13	3.920,45
14	853,32
15	12.819,59
16	10.088,13
17	4.414,62
18	7.293,00
19	11.445,93
20	8.339,63
21	4.858,72
22	1.616,16
23	1.339,24
24	7.108,82
25	2.054,73
26	1.441,01
34	5.889,54
35	2.585,15
36	5.146,24
37	718,91
38	1.049,08
39	9.072,00
40	3.273,02

.Faixas com igual altura

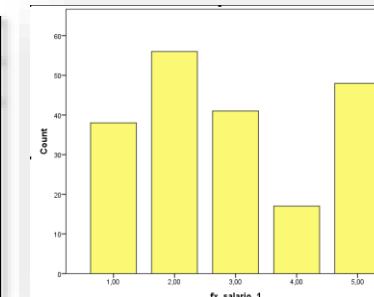
	salário			
	Minimum	Maximum	Count	%
Group of salário	1	254,19	2.180,63	40 20,0%
	2	2.197,28	3.426,82	40 20,0%
	3	3.557,88	5.083,60	40 20,0%
	4	5.129,45	8.474,94	40 20,0%
	5	8.543,09	14.649,52	40 20,0%

5 faixas com 20% das observações



.Faixas com igual largura 5 faixas variando : R\$ 2.000,00

	salário			
	Minimum	Maximum	Count	%
fx_sal_1	1	254,19	1.980,70	38 19,0%
	2	2.054,73	3.920,45	56 28,0%
	3	4.056,32	5.889,54	41 20,5%
	4	6.166,71	7.662,03	17 8,5%
	5	8.063,00	14.649,52	48 24,0%



TÉCNICAS DE TRANSFORMAÇÃO DOS DADOS

Redução de dimensionalidade - (combinar várias variáveis em uma única) - são comumente usadas (Análise de Componentes Principais - ACP).
Exemplo: IDH

Análise de Componentes Principais:

Técnica de aprendizado não supervisionado.

O objetivo é encontrar combinações lineares das variáveis que incluem a maior quantidade possível de variância original.

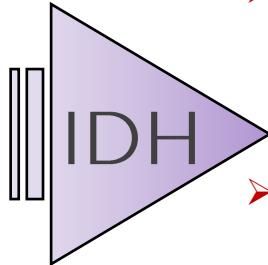
Esta transformação é definida de forma que o primeiro componente principal tem a maior variância possível, e cada componente seguinte, por sua vez, tem a máxima variância sob a restrição de ser ortogonal a (i.e., não correlacionado com) os componentes anteriores.

Quanto maior a dimensão dos dados (número de variáveis) maior o risco de sobre ajuste do modelo.

Uma das razões pela qual a ACP é tão utilizada, é o fato de que cria componentes principais não correlacionadas. (alguns algoritmos conseguem melhor performance preditiva com variáveis com baixa correlação).

Outra forma de diminuir a presença de variáveis com alta colinearidade é excluí-las. Variáveis colineares trazem informação redundante(tempo perdido). Aumentam a instabilidade dos modelos.

TÉCNICAS: Redução de dimensionalidade



- Acesso ao conhecimento: educação
 - Taxa de alfabetização da população acima de 15 anos
 - Proporção de pessoas com acesso aos níveis de ensino primário
- Direito a uma vida longa e saudável: longevidade
 - Expectativa de vida ao nascer
- Direito a um padrão de vida digno:
 - Renda PIB *per capita*

Exemplo

TÉCNICAS: Redução de dimensionalidade

MUNICÍPIO	UF	Esp_Vida	Tx_alfab	Tx_freq_esc	rendacapita	IDH_M	Class_UF	Class_BR
São Caetano do Sul	SP	78,18	97,01	98,57	834,00	0,919	1	1
Águas de São Pedro	SP	77,44	97,06	85,75	954,65	0,908	2	2
Santos	SP	72,27	96,44	92,62	729,62	0,871	3	6
Vinhedo	SP	74,87	94,08	79,73	627,47	0,857	4	15
Jundiaí	SP	73,94	94,99	88,46	549,96	0,857	5	17
Ribeirão Preto	SP	74,40	95,56	84,21	539,84	0,855	6	22
Santana de Parnaíba	SP	71,35	92,06	87,55	762,05	0,853	7	25
Campinas	SP	72,22	95,01	87,54	614,86	0,852	8	26
Saltinho	SP	77,35	95,78	80,34	406,27	0,851	9	28
Ilha Solteira	SP	75,80	94,77	90,74	390,05	0,850	10	33
São José dos Campos	SP	73,89	95,42	89,20	470,01	0,849	11	36
Araçatuba	SP	74,52	93,69	85,34	503,17	0,849	12	41
Paulínia	SP	73,30	93,93	89,37	503,34	0,847	13	44
Presidente Prudente	SP	73,58	93,81	89,58	482,62	0,846	14	47
São João da Boa Vista	SP	76,92	93,56	79,51	408,33	0,843	15	56
Valinhos	SP	71,91	94,42	84,54	569,31	0,842	16	63
São Carlos	SP	73,08	94,36	89,61	456,25	0,841	17	65
São Paulo	SP	70,66	95,11	85,48	610,04	0,841	18	68
Americana	SP	72,46	95,62	87,15	473,23	0,840	19	71
Pirassununga	SP	75,16	93,95	84,33	402,30	0,839	20	77
Taubaté	SP	72,73	95,18	85,12	460,86	0,837	21	87
Piracicaba	SP	72,95	94,95	84,05	455,87	0,836	22	93
Santo André	SP	70,61	95,55	88,59	512,87	0,836	23	94
Caçapava	SP	74,88	93,88	86,86	363,53	0,835	24	96
Cordeirópolis	SP	76,82	93,28	77,86	367,03	0,835	25	97
Tremembé	SP	74,47	94,43	84,83	383,76	0,834	26	99
São José do Rio Preto	SP	71,31	94,61	85,55	512,01	0,834	27	102
São Bernardo do Campo	SP	69,93	95,02	91,93	505,45	0,834	28	106
Sertãozinho	SP	74,40	91,62	87,98	397,11	0,833	29	109
Catanduva	SP	75,38	92,40	82,51	385,10	0,832	30	114

O QUE SABEMOS DE MACHINE LEARNING:



COMO
SELECCIONAR
FEATURES?

DATA MINING **SELEÇÃO DE VARIÁVEIS**

- Na fase de mineração de dados normalmente se trabalha com uma grande quantidade de variáveis.
- Para selecionar quais variáveis são “importantes” nesta fase pode-se usar os seguintes métodos:
- Métodos automáticos:
 - 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.
- As árvores de decisão também são utilizadas para seleção de variáveis. O conjunto de variáveis que aparecem na árvore formam o conjunto de variáveis selecionadas.
- Utilizar testes como Análise de Correlação, Chi-Quadrado, GiNI, Entropia entre outros



RELEMBRANDO... ALGUMAS MÉTRICAS



CORRELAÇÃO

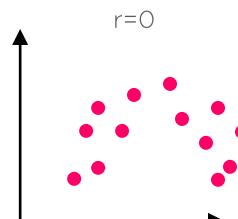
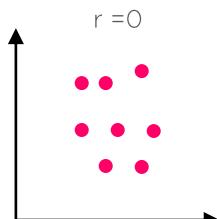
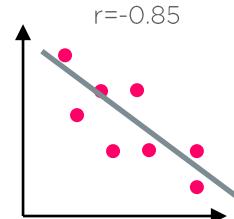
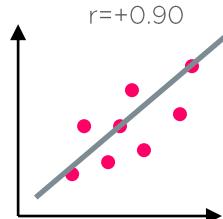


RELEMBRANDO...

ANÁLISE EXPLORATÓRIA DOS DADOS

Coeficiente de correlação (r) representa a relação linear entre duas variáveis.

Valores de r e suas implicações.



Correlação Linear Simples (r de Pearson)

$$\frac{\sum_{i=1} (X_i - \bar{X}) * (Y_i - \bar{Y})}{\sqrt{\sum_{i=1} (X_i - \bar{X})^2 * \sum_{i=1} (Y_i - \bar{Y})^2}}$$

- Para avaliar a correlação entre variáveis, é importante conhecer a magnitude ou força tanto quanto a significância da correlação.

CUIDADO

É importante lembrar que o conceito de correlação refere-se a uma associação numérica entre duas variáveis, não implicando necessariamente numa relação de *causa-efeito*. Portanto, mesmo que duas variáveis apresentem-se matematicamente relacionadas, não significa que deva existir uma relação lógica entre elas.



RELEMBRANDO... ALGUMAS MÉTRICAS



CHI-QUADRADO



• TESTE DE HIPÓTESE - NÃO PARAMÉTRICOS

- Exemplo:⁺ Após uma pesquisa de satisfação estamos interessados em verificar se a preferência pela operadora(OP) estava associada com o fator regional.

Estado e Operadora	OP1	OP2	OP3	OP4	Total	
SP	214 33%	237 37%	78 12%	119 18%	648	100%
Sul	51 17%	102 34%	126 42%	22 7%	301	100%
RJ	111 18%	304 51%	139 23%	48 8%	602	100%
Total	376 24%	643 42%	343 22%	189 12%	1551	100%

H_0 : São independentes

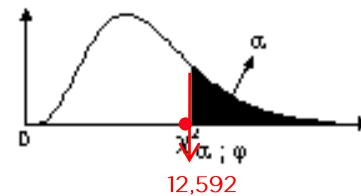
H_a : Não são independentes

$$\alpha = 5\%$$

$$\text{Graus de liberdade} = (I-1)(c-1) = 3 \cdot 2 = 6$$

➔ Se tivesse independência todos os estados teriam a mesma distribuição

➔ Número esperado em SP: $648 \cdot 0.24 = 157$; Sul: $301 \cdot 0.24 = 73$; RJ: $602 \cdot 0.24 = 146$;



• TESTE DE HIPÓTESE - NÃO PARAMÉTRICOS

- Exemplo: Valores Observados

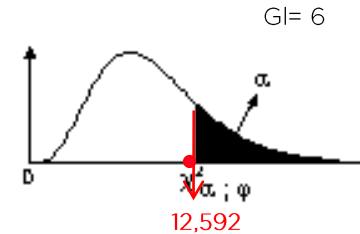
Estado e Operadora	OP1	OP2	OP3	OP4	Total
SP	214	237	78	119	648
Sul	51	102	126	22	301
RJ	111	304	139	48	602
Total	376	643	343	189	1.551

Valores Esperado

Estado e Operadora	OP1	OP2	OP3	OP4	Total
SP	157	269	143	79	648
Sul	73	125	67	37	301
RJ	146	250	133	73	602
Total	376	643	343	189	1.551

$$\chi^2_{obs} = (214-157)^2/157 \dots \dots (48-73)^2/73 = 173.24$$

Rejeitamos H_0 ao nível de 5%, isto é os dados trazem evidência de uma forte dependência entre os fatores: Operadora de Celular e Região



$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$



RELEMBRANDO... ALGUMAS MÉTRICAS



ENTROPIA

RELEMBRANDO... ALGUMAS MÉTRICAS

Entropia

Quantidade necessária de informação para identificar a classe de um caso

Critério de seleção de variável:

$$\text{Entropia}(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

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

v ∈ valores (A)

onde:

S é o conjunto de amostras (registros)

c é o número de valores possíveis da classe

p_i é a proporção de amostras da classe *i* em relação ao total de amostras

• ÁRVORE DE DECISÃO

- Segmento: Área Financeira
- A área de crédito deseja avaliar a propensão de um cliente tornar-se inadimplente.

Exemplo

Variável Resposta: 0 = Adimplente
1 = Inadimplente

Variáveis preditoras	Categorias
Tempo de relacionamento	1-ate 1 ano
	2-1 a 3 anos
	3-3 a 8 anos
	4-mais 8 anos
Valor da fatura	1-Ate R\$250
	2-R\$250 a R\$800
	3-R\$800 a R\$1499
	4-R\$1500 e mais
% gastos com alimentação	1-Ate 10%
	2-10% a 20%
	3-20% a 30%
	4-30% e mais

• ÁRVORE DE DECISÃO

- Segmento: Área Financeira
- A área de crédito deseja avaliar a propensão de um cliente tornar-se inadimplente.

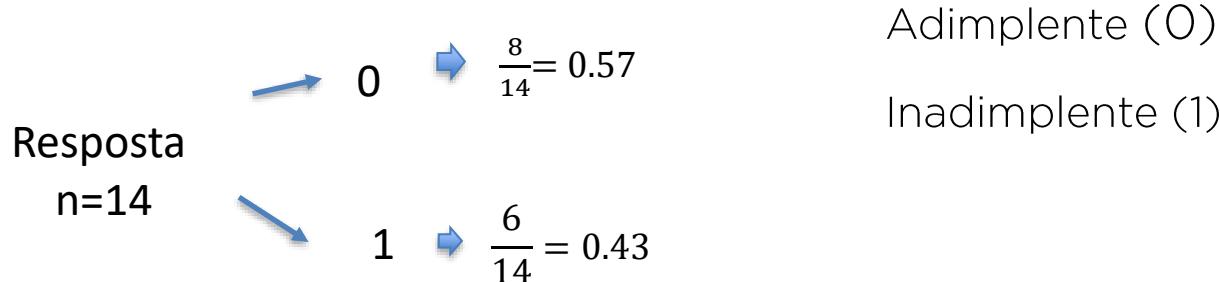
Exemplo

cliente	Resposta	TempoRel	valorFatura	GastosAlim
2642	0	4-mais 8 anos	3-R\$800 a R\$1499	2-10% a 20%
5620	0	2-1 a 3 anos	3-R\$800 a R\$1499	4-30% e mais
8238	0	4-mais 8 anos	2-R\$250 a R\$800	1-Ate 10%
8272	1	1-ate 1 ano	3-R\$800 a R\$1499	2-10% a 20%
9724	0	2-1 a 3 anos	2-R\$250 a R\$800	4-30% e mais
13021	0	3-3 a 8 anos	1-Ate R\$250	2-10% a 20%
15361	0	4-mais 8 anos	1-Ate R\$250	4-30% e mais
15606	1	2-1 a 3 anos	2-R\$250 a R\$800	1-Ate 10%
18010	0	4-mais 8 anos	4-R\$1500 e mais	4-30% e mais
20060	1	2-1 a 3 anos	1-Ate R\$250	3-20% a 30%
20342	1	1-ate 1 ano	1-Ate R\$250	2-10% a 20%
20569	0	1-ate 1 ano	2-R\$250 a R\$800	3-20% a 30%
21191	1	4-mais 8 anos	2-R\$250 a R\$800	1-Ate 10%
24385	1	4-mais 8 anos	1-Ate R\$250	4-30% e mais

• ÁRVORE DE DECISÃO

- Segmento: Área Financeira
- A área de crédito deseja avaliar a propensão de um cliente tornar-se inadimplente.

Exemplo



Entropia

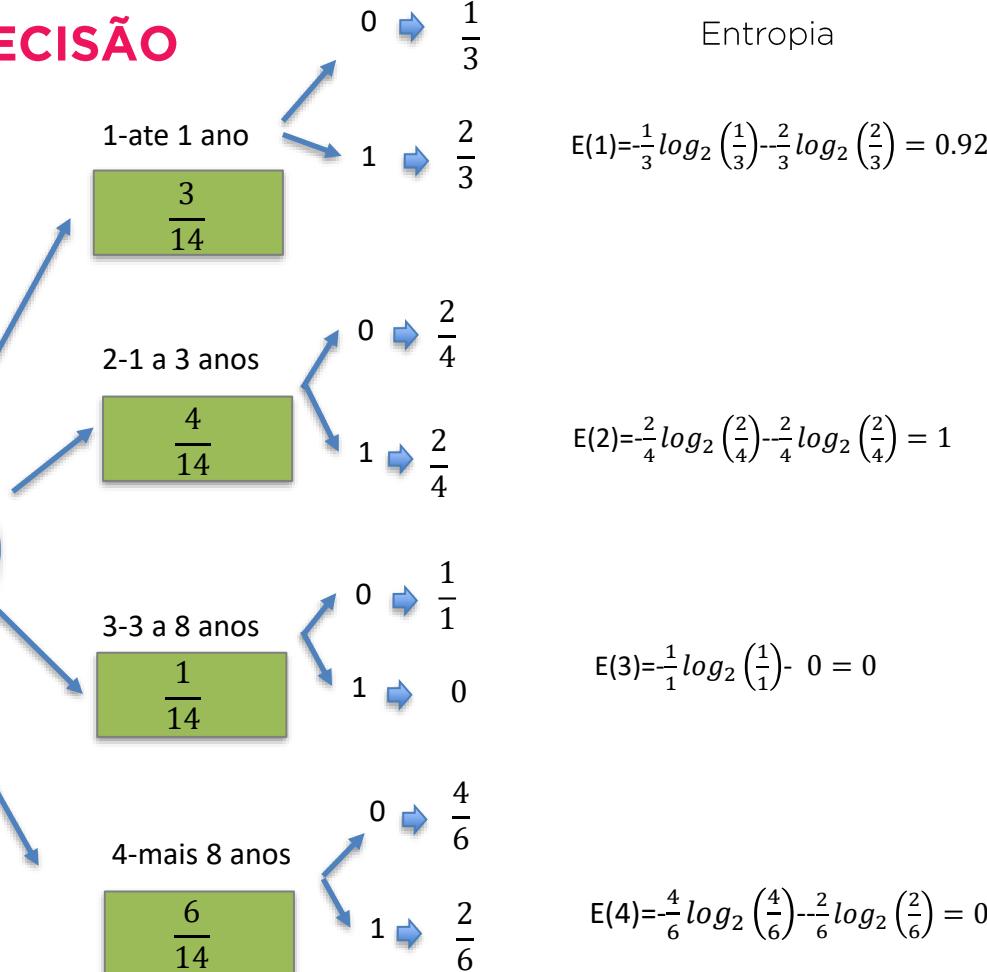
$$E(S) = -\frac{8}{14} \log_2 \left(\frac{8}{14} \right) - \frac{6}{14} \log_2 \left(\frac{6}{14} \right) = 0.98$$

Inadimplente

• ÁRVORE DE DECISÃO

Exemplo

Tempo de
relacionamento
 $n=14$



ÁRVORE DE DECISÃO

- Segmento: Área Financeira

Exemplo

$$\text{Ganho}(\text{tempo de relacionamento}) = 0.98 - \frac{3}{14} * 0.92 - \frac{4}{14} * 1.0 - \frac{1}{14} * 0 - \frac{6}{14} * 0.92 = -0.039$$

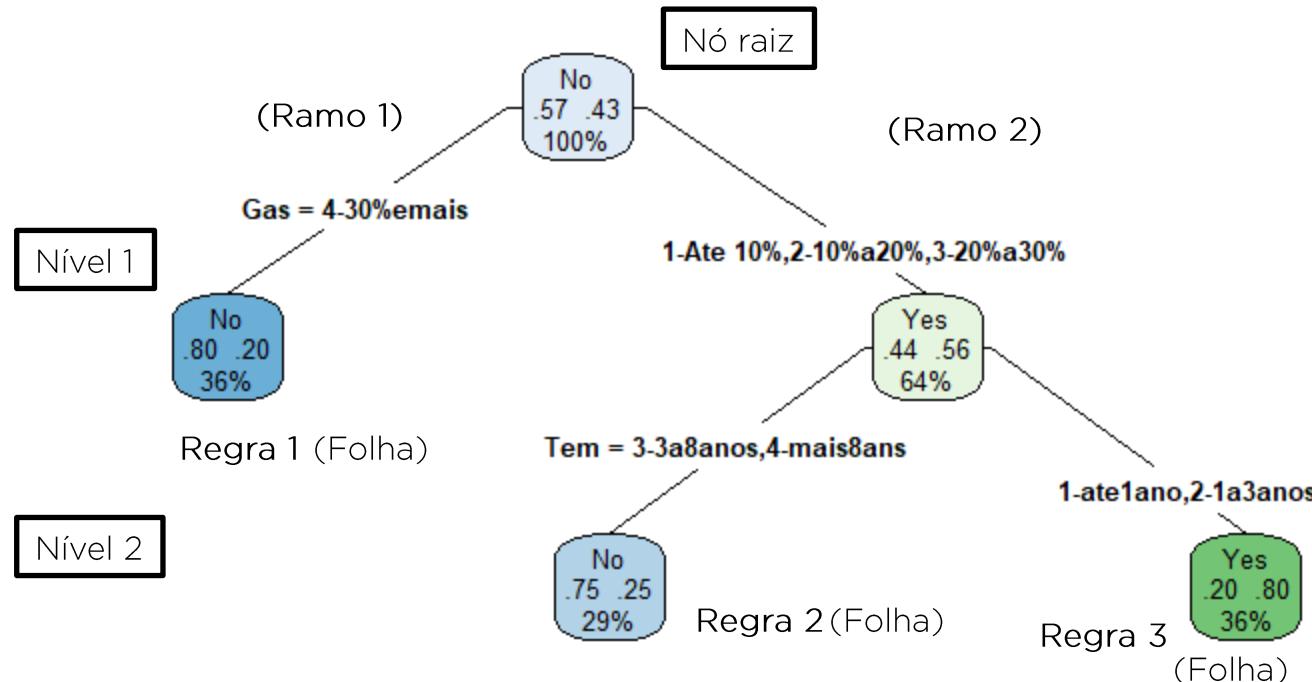
Variável	0
Tempo de relacionamento	-0.039
Valor da fatura	0.095
% gasto com alimentação	0.10



• ÁRVORE DE DECISÃO

- Segmento: Área Financeira
- A área de crédito deseja avaliar a propensão de um cliente tornar-se inadimplente.

Exemplo



RELEMBRANDO...

ALGUMAS MÉTRICAS

GINI

ÍNDICE DE GINI

- Ele é um índice de dispersão estatística que mede a heterogeneidade dos dados e é utilizado tanto para a seleção de atributos como também em análises econômicas e sociais para verificar a distribuição de renda em um certo país.
- O índice GINI para um conjunto de dados S, que contém n registros, cada um com uma classe Ci é dado pela equação:

$$gini(S) = 1 - \sum_{i=1}^k p(C_i|n)^2$$

Onde:

p_i : probabilidade relativa da classe C_i em S.

n : número de registros no conjunto S.

k : número de classes.

ÍNDICE DE GINI

- Se o conjunto S for particionado em dois ou mais subconjuntos Si, O índice GINI dos dados particionados será definido pela equação:

$$gini(S) = 1 - \sum_{i=1}^k p(C_i|n)^2$$

Onde: n_i : número de registros no subconjunto S_i .

n : número de registros no conjunto S.

Quando este índice é igual a zero, o conjunto de dados é puro, ou seja, todos os registros pertencem a uma mesma classe.

Por outro lado, quando ele se aproxima do valor um, o conjunto apresenta os registros distribuídos igualmente entre todas as classes.

- **ÍNDICE DE GINI**

Quando se utiliza o critério Gini na indução de árvores de decisão binárias, tende-se a isolar num ramo os registros que representam a classe mais frequente, assim, utilizando o atributo com menor valor do índice para a classificação, já, ao utilizar-se da entropia, balanceia-se o número de registros em cada ramo.



O QUE SABEMOS DE MACHINE LEARNING:



QUAIS OS TIPOS
DE MODELOS?

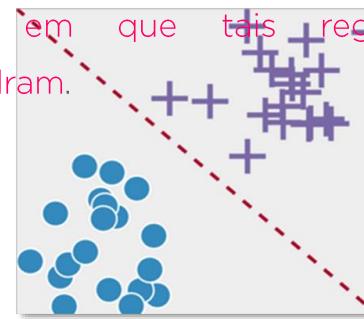
Tipos de problemas resolvidos com
MACHINE LEARNING

CONSTRUÇÃO DE MODELOS

- **Regressão:** Compreende a busca por uma função que mapeie os registros de um banco de dados em um intervalo de valores numéricos reais. Esta tarefa é similar à tarefa de Classificação, com a diferença de que o **atributo alvo assume valores numéricos.**

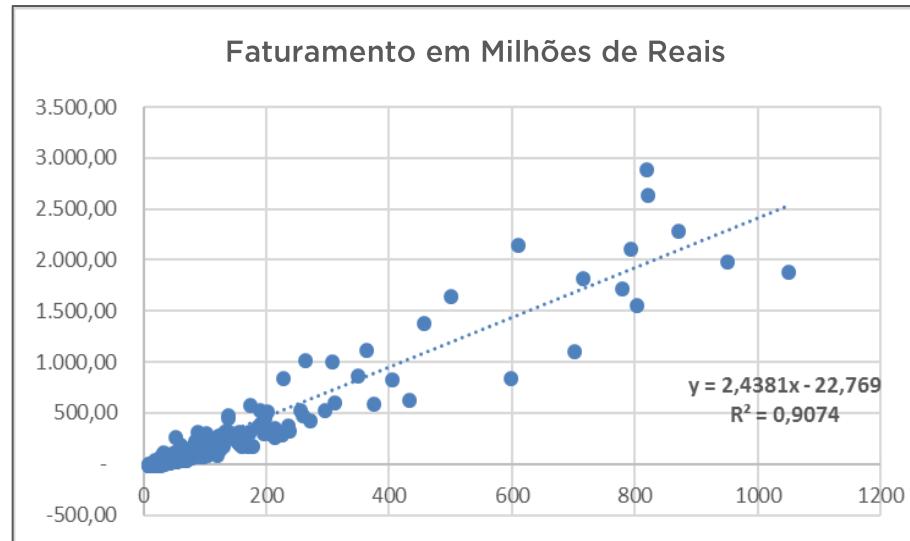


- **Classificação:** A tarefa de Classificação consiste em descobrir uma função que mapeie um conjunto de registros em um conjunto de classes. Uma vez descoberta, tal função pode ser aplicada a novos registros de forma a prever a classe em que tais registros se enquadram.



TÉCNICA DE REGRESSÃO: REGRESSÃO LINEAR SIMPLES

Exemplo: Faturamento anual (em milhões de Reais) por número de checkouts.



Fonte:ABRAS

TÉCNICA DE REGRESSÃO: REGRESSÃO LINEAR MÚLTIPLA

Estimar o valor de imóveis a partir de suas características
Variáveis:

Valor do Imóvel [Valor]: Valor do imóvel

Área [Area]: Utilizou-se a área total do apartamento em metros quadrados;

Idade Aparente [IA]: Idade aparente em anos

Andar [Andar]: É o número do andar do apartamento;

Suítés [Suites]: Número de suítés;

Vista Panorâmica [Vista]: A variável ambiental vista panorâmica é uma variável dicotômica: se o apartamento tiver vista panorâmica a variável vista assume valor igual a 1, se não tiver vista seu valor será 0;

Sem Ruído na rua [Sem Ruído]: A variável ambiental Sem Ruído é uma variável dicotômica: se o apartamento está localizado em rua onde o nível de ruído está abaixo do que é considerado não prejudicial terá valor 1, se tiver nível de ruído acima terá valor 0;

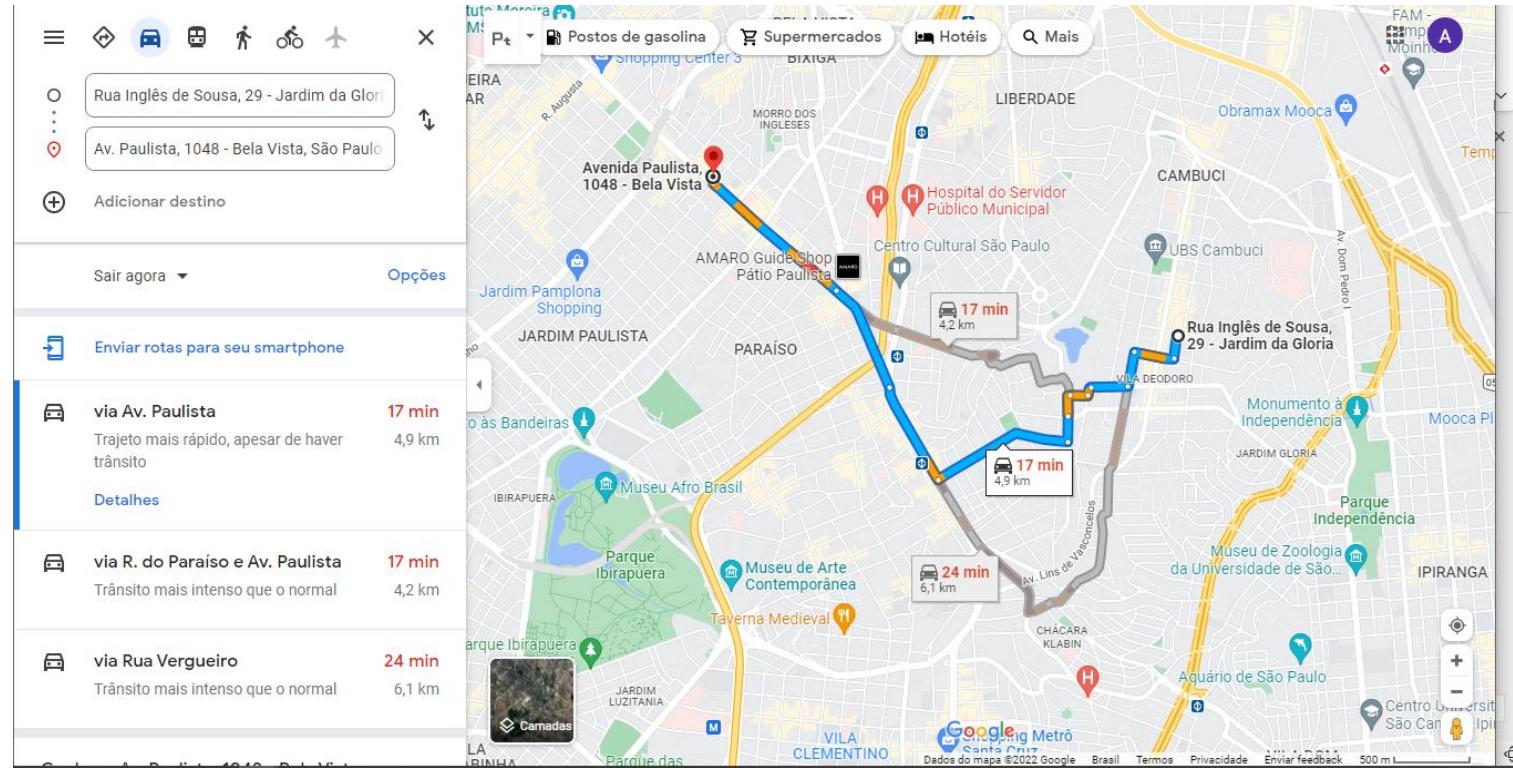
Distância a Avenida Beira Mar [Dist. BM]: A distância é medida em metros, pelo eixo da rua do prédio onde os apartamentos estão localizados até a Avenida Beira Mar;

Área Verde a uma distância de 200 metros [AV 200m]: Área

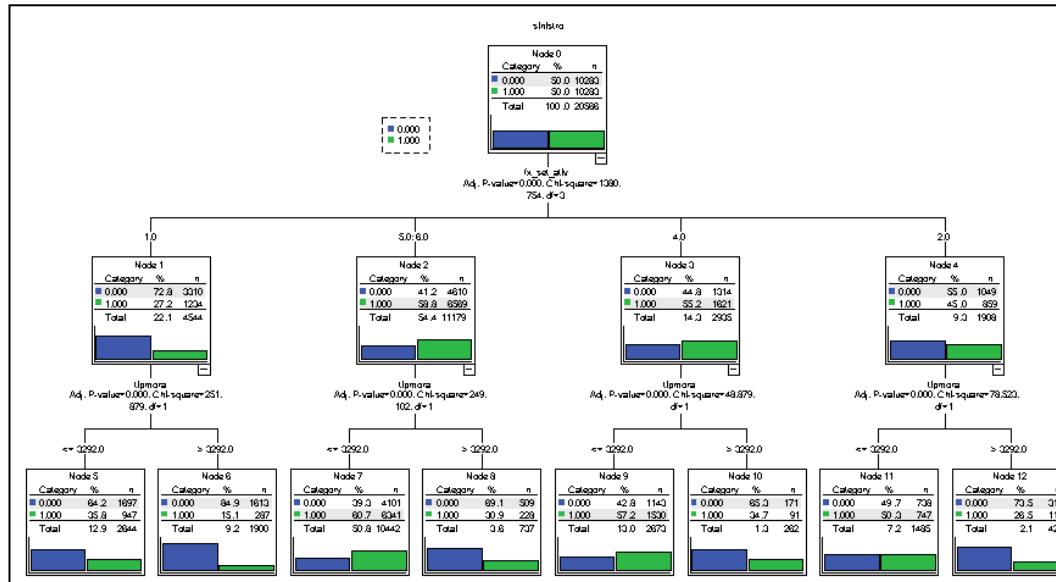
Trecho do arquivo de

Ordem	Valor	Área	m2	IA	Andar	Suítés	Vista	Dist.BM	SemRuído	AV200m
1	160,000	167.81		1	5	1	1	294	1	0
2	67,000	128.80		1	6	0	0	1.505	1	0
3	190,000	217.37		1	8	1	0	251	0	1
4	110,000	180.00		12	4	1	0	245	0	0
5	70,000	120.00		15	3	1	0	956	1	0
6	75,000	160.00		18	2	0	1	85	0	1
7	95,000	155.00		5	3	1	0	1.401	1	0
8	135,000	165.00		1	2	1	1	148	0	1
9	110,000	150.00		10	4	1	0	143	0	0
10	115,000	185.00		15	5	1	0	831	0	0
11	325.669	392.40		1	4	2	0	421	1	1
12	362.400	392.40		1	8	2	0	421	1	1
13	163.798	225.60		1	2	1	0	397	1	0
14	261.250	312.82		1	3	2	0	319	1	0
15	276.870	304.35		1	5	4	0	461	1	1
16	284.626	304.35		1	7	4	0	461	1	1
17	95,000	161.00		6	3	1	0	143	0	0

TÉCNICA DE REGRESSÃO:



- TÉCNICA DE CLASSIFICAÇÃO:
- ÁRVORE DE DECISÃO
- Estimar se a pessoa vai sinistrar ou não
Exemplo de Modelo



- TÉCNICA DE CLASSIFICAÇÃO:
REGRESSÃO LOGÍSTICA
- Estimar se o cliente vai cancelar ou não

Exemplo de Modelo de Churn: Pesos definidos na modelagem



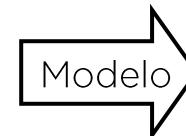
- TÉCNICA DE CLASSIFICAÇÃO:
REGRESSÃO LOGÍSTICA



Clientes
Clientes com o evento(*)

(*) Evento (exemplos)

- Aquisição
- Cancelamento
- Pagamento



Pontuação dos Clientes
com Probabilidade



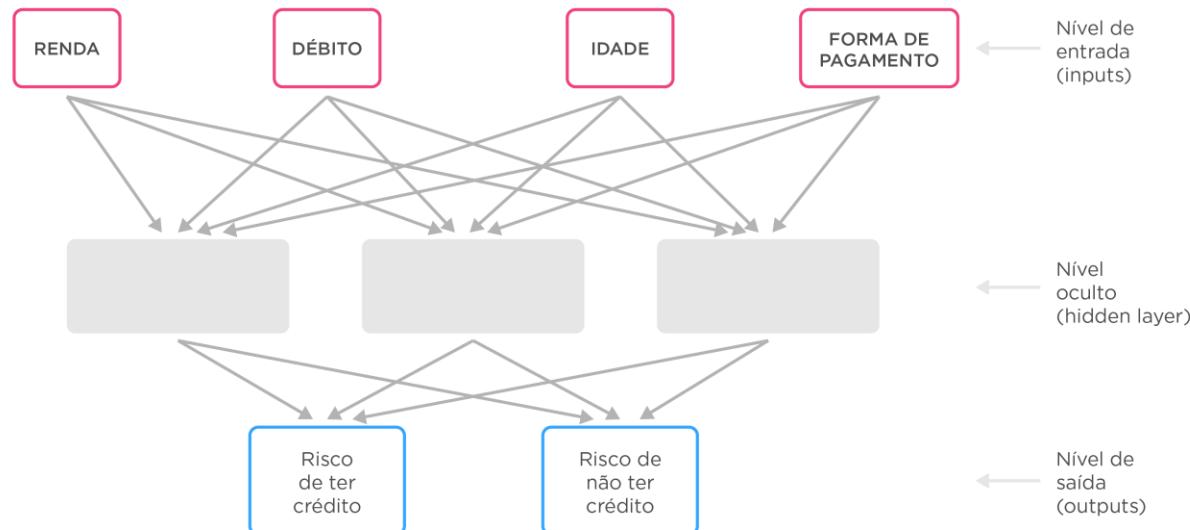
Regra de Decisão

Corte



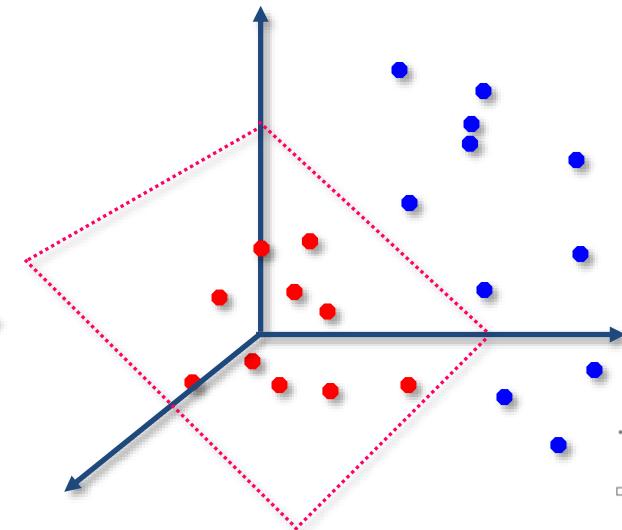
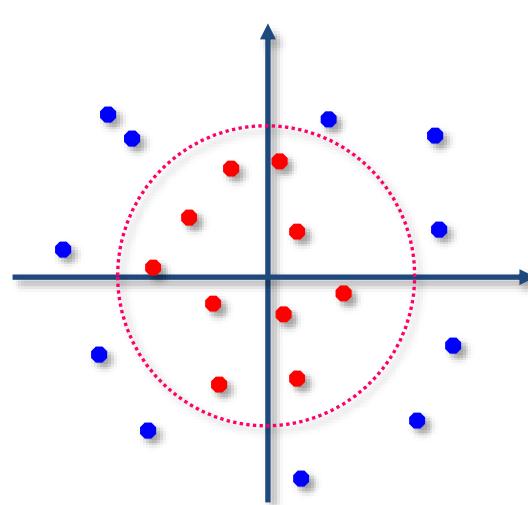
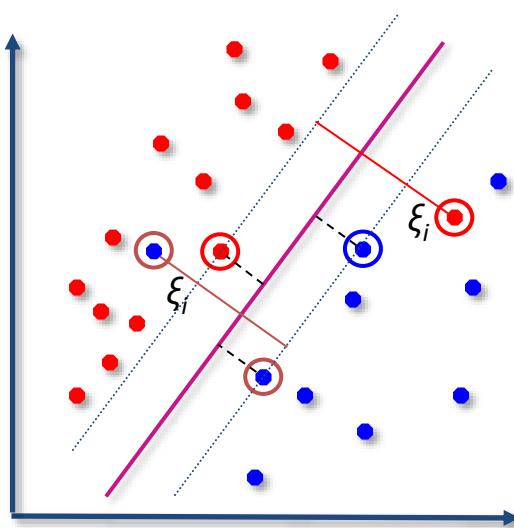
TÉCNICA DE CLASSIFICAÇÃO: REDES NEURAIS

Exemplo: risco de crédito



As redes neurais usam dados de entrada.
Atribui pesos nas conexões entre os atributos (neurônios).
E obtém um resultado (risco de ter ou não crédito) - nível de saída.

- TÉCNICA DE CLASSIFICAÇÃO:
- **SVM – Support Vector Machine**



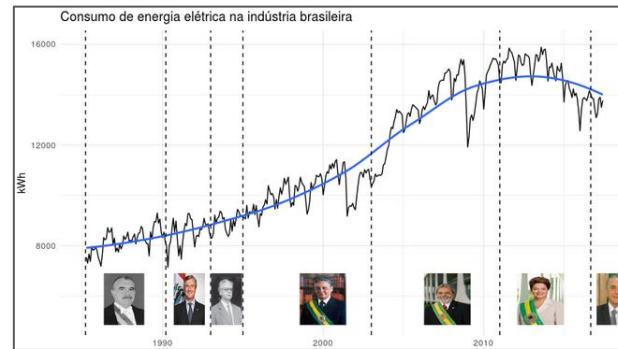
• TÉCNICA DE CLASSIFICAÇÃO:

Principal Social Promoções

Detalhes da Mensagem	Conteúdo da Mensagem	Data
Anahp	O Observatório 2022 está chegando! Saiba quando será o lançamento - Visualizar como página web...	10:04
JOSI ^{1%}	PAUTA REUNIÃO 26 MARÇO 2022 - EQUIPE 7 C - Boa noite queridos casais ! Boa noite Frei Ademir... Pauta Reunião ...	20 de mar.
Recibos da Uber	Sua viagem de sexta-feira à tarde com a Uber - Total R\$ 11,9618 de março de 2022 Obrigado por vi...	18 de mar.
Doméstica Legal	Jornada parcial de domésticos: quanto pagar de salário? - Conheça nossa calculadora de jornada ...	18 de mar.
Anahp	Saiba como foi o Café da Manhã com o Medportal sobre conhecimento de impacto para lideranças -	18 de mar.
Uber 2	Último dia para aproveitar! 🚗 - Corre que ainda tem 15%OFF em Cervejas selecionadas. Vem ver! ...	18 de mar.
Anahp	A Anahp quer te fazer um convite: vamos fazer de 2022 o ano de ouvir a saúde? - Visualizar como ...	17 de mar.
Marineide	PAUTA p/ 26 MARÇO- PRÉVIA - Boa tarde. Td bem? Segue prévia da Pauta para observações, alter...	17 de mar.
Anahp	Pauta Reunião ... Daqui a pouco: participe do debate sobre saúde, política e Eleições 2022 - Visualizar como página ...	17 de mar.
Uber 2	Semana do Consumidor tá acabando 🕒 - Mas liga o turbo e não perde tempo 🔥 Uber Aproveitar a...	17 de mar.
Prevent Senior	+Rapidez para realizar exames na rede credenciada - Se você não estiver visualizando a mensage...	16 de mar.

MODELOS DE SÉRIES **TEMPORAIS**

- **Previsão de Séries Temporais:** Uma série temporal é um conjunto de observações de um fenômeno (variável numérica) ordenadas no tempo. A previsão de uma série temporal tem como objetivo inferir valores que a variável da série deverá assumir no futuro considerando como base valores passados dessa série.

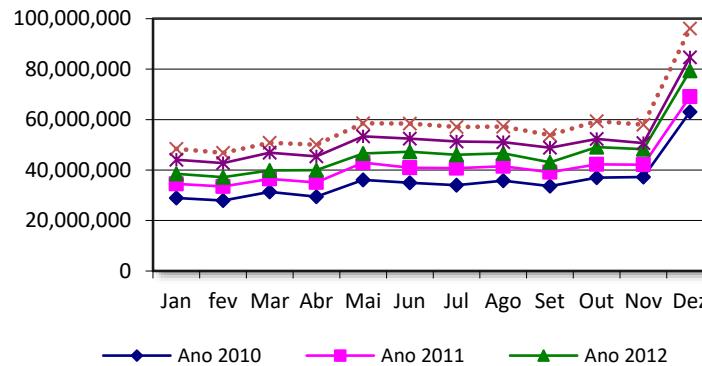


MODELOS DE SÉRIES TEMPORAIS

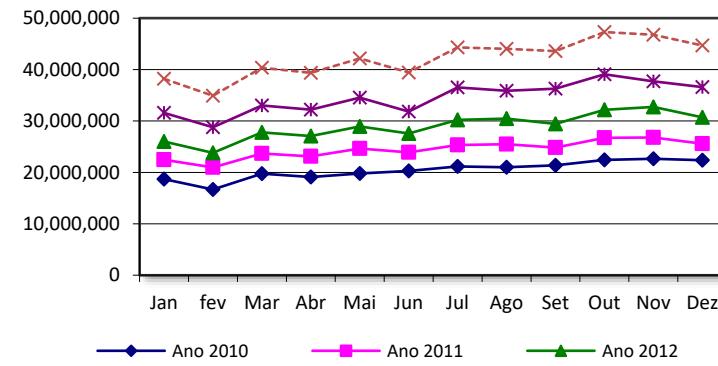
- Previsão

Quantidade de transações mensais com cartões de crédito

Transações Crédito - Comércio Varejista



Transações Crédito - Turismo & Entretenimento

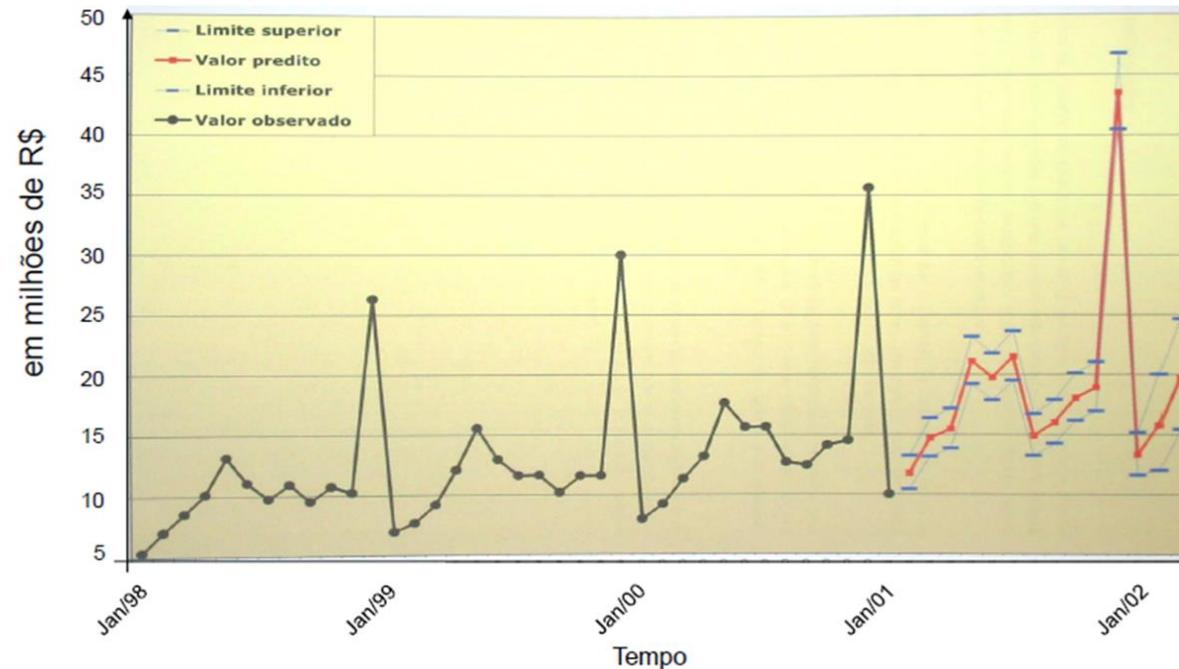


Fonte:ABECS

MODELOS DE SÉRIES TEMPORAIS

Exemplo: Varejo
(Vestuário)

→ Previsão de 12 meses para o faturamento mensal



- 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

• SÉRIES **TEMPORAIS**

Alguns exemplos de Modelos de séries temporais

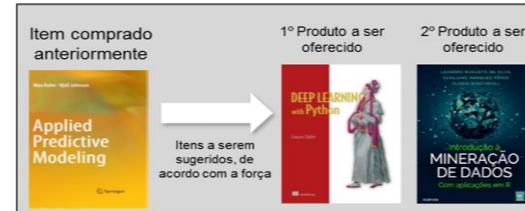
- Média Móvel
- Alisamento Exponencial
- Método de Decomposição
- Modelos ARIMA (Box Jenkins)

• TÉCNICA DESCOPERTA DE SEQUÊNCIAS

- Descoberta de Associações: Nesta tarefa, cada registro do conjunto de dados é normalmente chamado de transação. Cada transação é composta por um conjunto de itens. A tarefa de descoberta de associações compreende a busca por itens que frequentemente ocorrem de forma simultânea em uma quantidade mínima de transações do conjunto de dados.



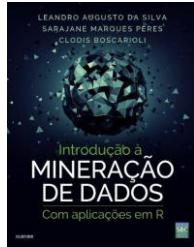
- Descoberta de Sequências: É uma extensão da tarefa de Descoberta de Associações cujo propósito é identificar itens frequentes considerando um determinado período de tempo. Consideremos o exemplo das compras no supermercado. Se o banco de dados possui a identificação do cliente responsável por cada compra, a descoberta de associações pode ser ampliada de forma a considerar a ordem em que os produtos são comprados ao longo do tempo.



• TÉCNICA DESCOBERTA DE SEQUÊNCIAS

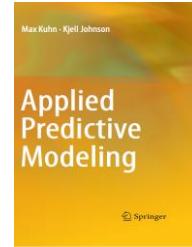
Exemplo Associações

Item comprado anteriormente

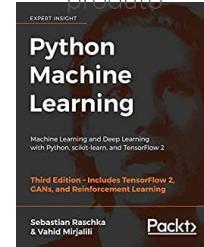


Itens a serem sugeridos de acordo com a força

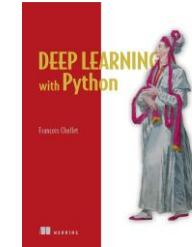
1º produto



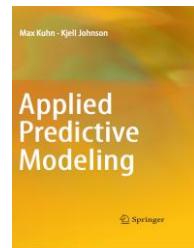
2º produto



3º produto

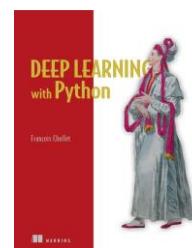


Item comprado anteriormente



Itens a serem sugeridos de acordo com a força

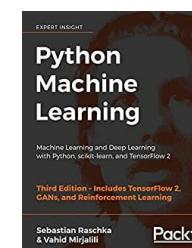
1º produto



2º produto

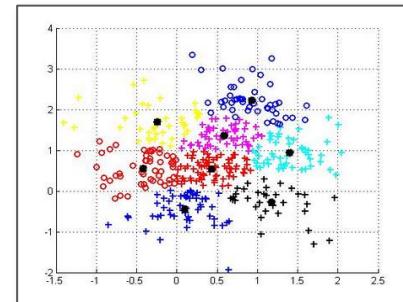
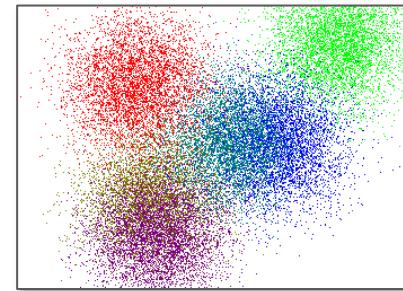


3º produto



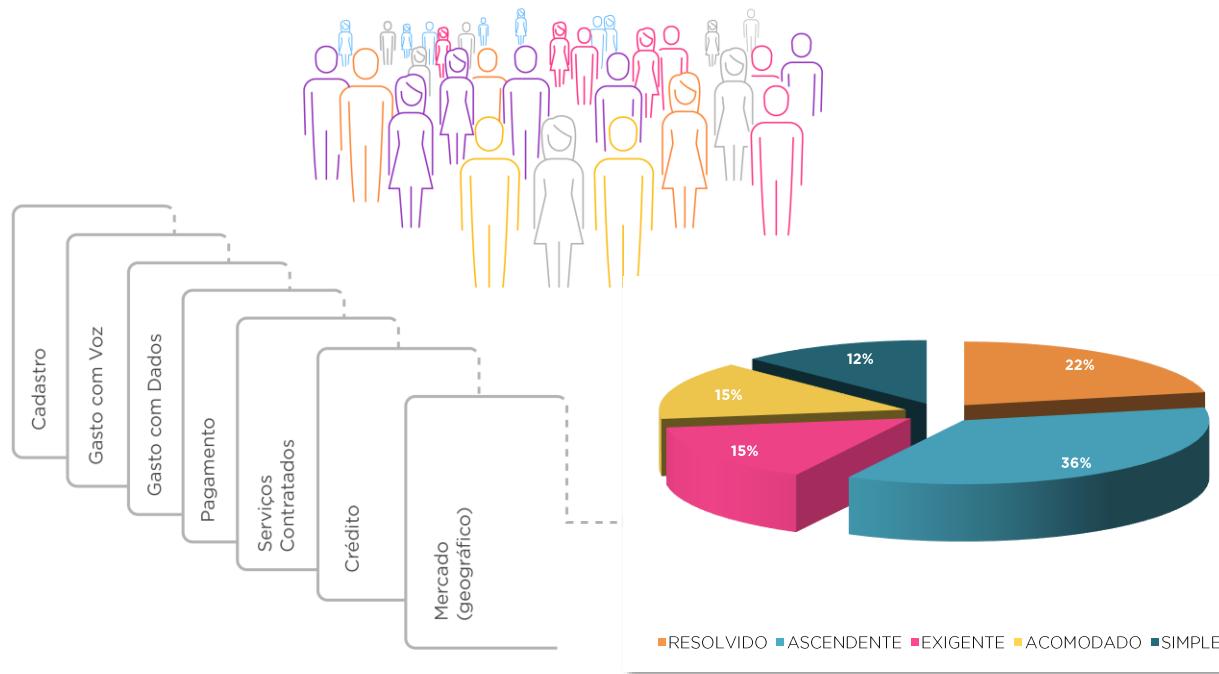
- TÉCNICA DE
- AGRUPAMENTO:
ANÁLISE DE CLUSTERS

- Agrupamento (Clusterização): Consiste em segmentar os registros do conjunto de dados em subconjuntos ou clusters, de tal forma que os elementos de um cluster compartilhem propriedades comuns que os distingam de elementos nos demais clusters. O objetivo nesta tarefa é maximizar a similaridade intracluster e minimizar a similaridade intercluster.



TÉCNICA DE AGRUPAMENTO: ANÁLISE DE CLUSTERS

Segmentação Comportamental do Cliente





• TÉCNICA REDUÇÃO DA DIMENSIONALIDADE:

•

- **Sumarização:** Consiste em identificar e indicar similaridades entre registros do conjunto de dados.

Exemplo: Construção do indicador de satisfação

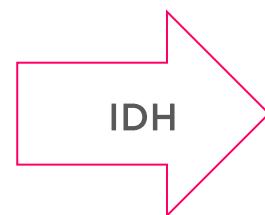


- TÉCNICA REDUÇÃO DA DIMENSIONALIDADE:
- **ANÁLISE FATORIAL**

- Entendimento das variáveis latentes.
- Criação de Indicadores.

Acesso ao conhecimento: educação.

- Taxa de alfabetização da população acima de 15 anos.
- Proporção de pessoas com acesso aos níveis de ensino primário.



Direito a uma vida longa e saudável: longevidade.

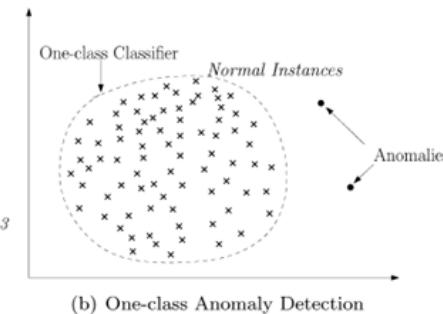
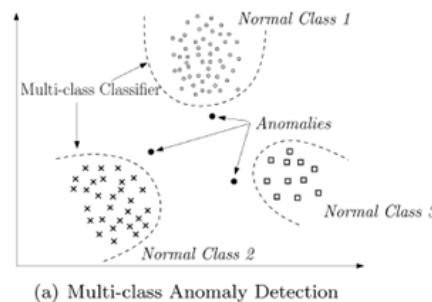
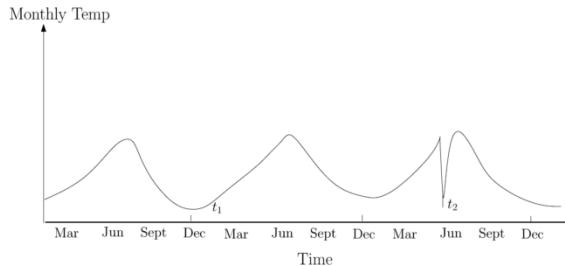
- Expectativa de vida ao nascer.

Direito a um padrão de vida digno:

- Renda PIB per capita.

DETECÇÃO de DESVIOS:

- Detecção de Desvios: Tal tarefa consiste em identificar registros do conjunto de dados cujas características destoem dos que se considera a norma no contexto em análise. Tais registros são denominados valores atípicos (outliers).



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).

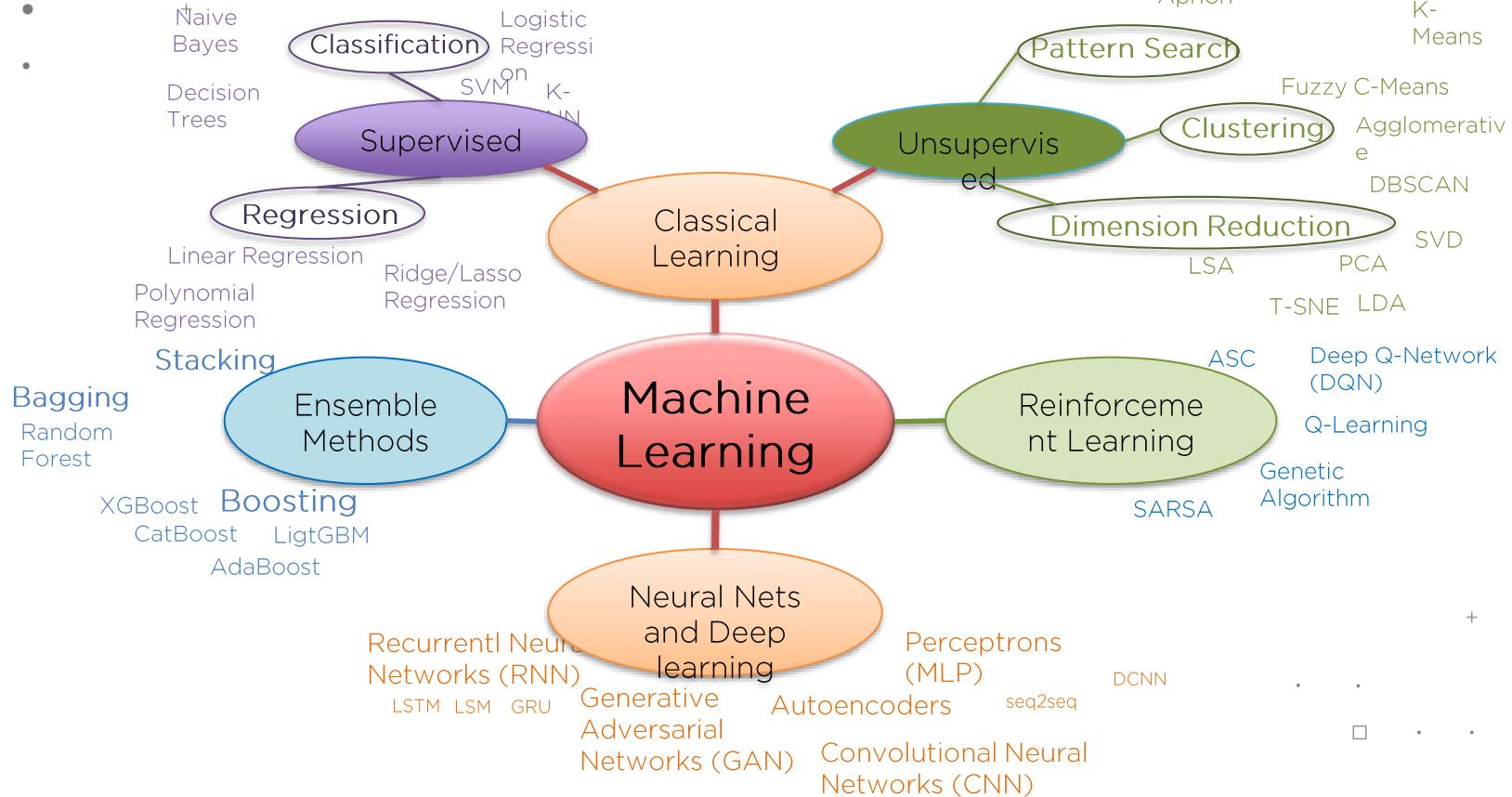
Análise Supervisionada

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.**

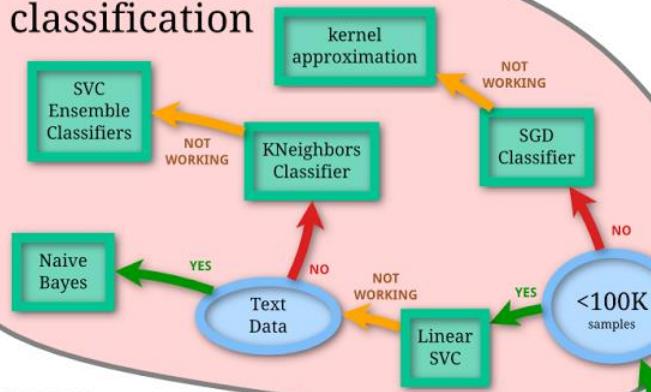
Análise Não Supervisionada

ALGORITMOS de MACHINE LEARNING

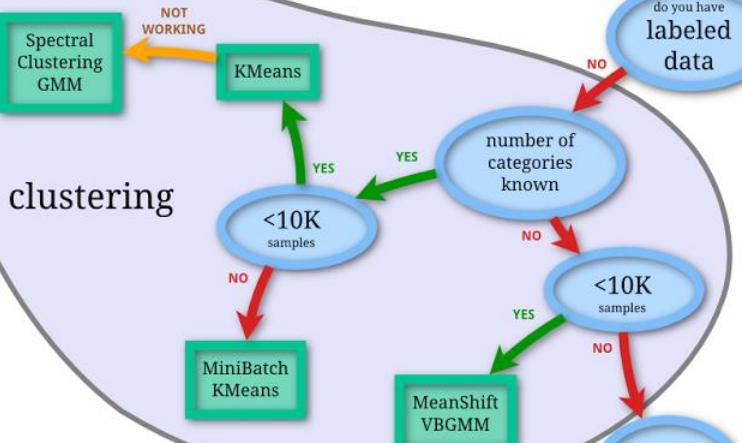


scikit-learn algorithm cheat-sheet

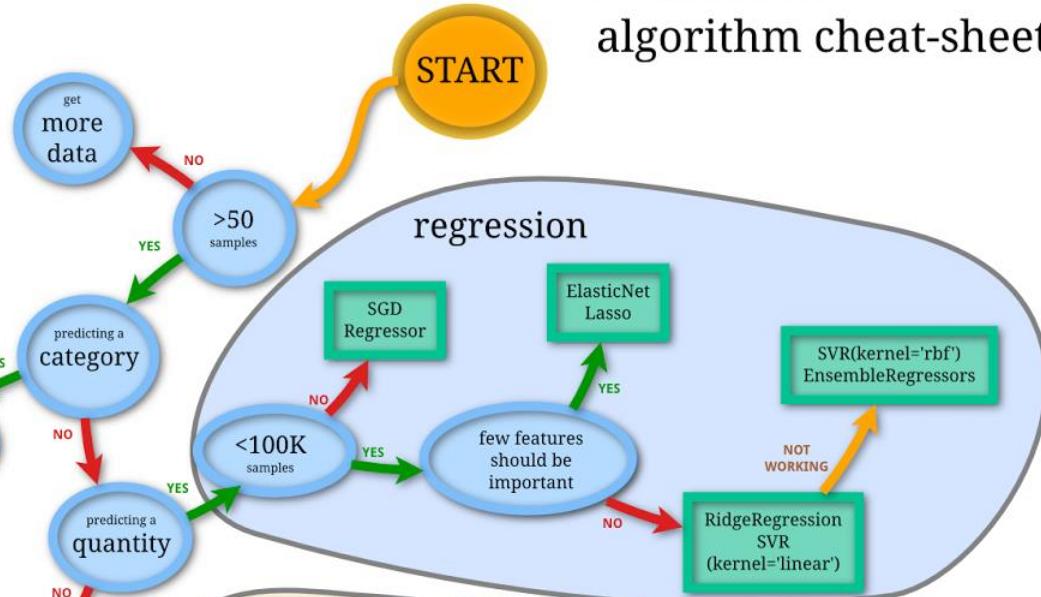
classification



clustering



regression



dimensionality reduction

O QUE SABEMOS DE MACHINE LEARNING:



QUAIS AS ETAPAS
NO TREINAMENTO
DOS MODELOS

• TREINAR MODELOS

- Escolher um ou mais modelos diferentes para treinar
- Separar uma parte dos dados para o treino
- Verificar se as características escolhidas estão sendo úteis para o modelo

• DESAFIOS DURANTE O TREINO



- Poucas amostras de dados
- Dados que não são representativos
- Baixa qualidade dos dados
- Features irrelevantes
- Overfitting ou Underfitting

- TESTAR O MODELO

-

- Separar uma parte dos dados para teste
- Avaliar o desempenho dos modelos
- Se necessário voltar para o passo de preparar e treinar novamente o modelo

O QUE SABEMOS DE MACHINE LEARNING:

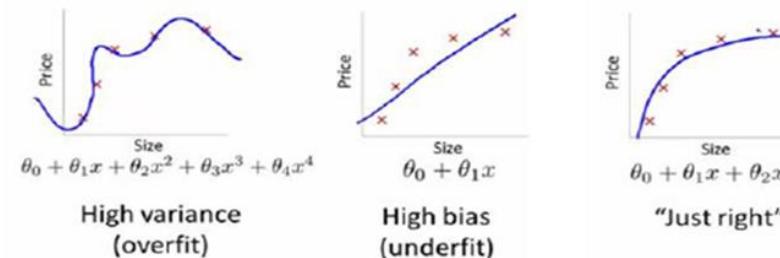


O QUE É
*bias-variance
tradeoff*

MODELOS PREDITIVOS

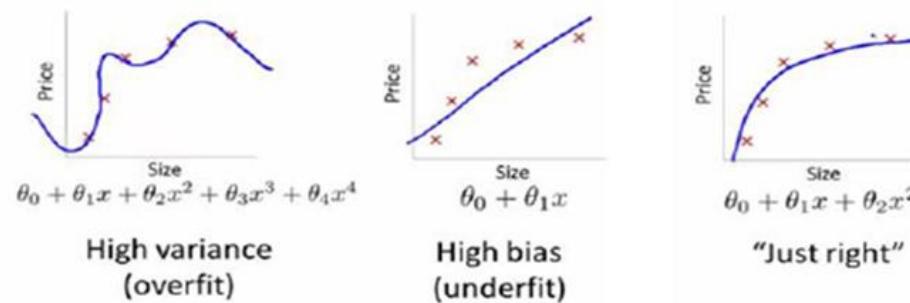
- Alguns modelos são facilmente interpretáveis, outros muito complexos.

Problema: em geral alguns modelos têm sobreajuste (*overfitting*) quando há muitas variáveis preditoras. Um modelo com overfitting tem mais coeficientes do que o necessário. É um modelo com pouca capacidade de generalização: ele terá alta acurácia para os dados de treinamento e acurácia ext



MODELOS PREDITIVOS

- **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 generalizar



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.

MODELOS PREDITIVOS

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 quanto 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.

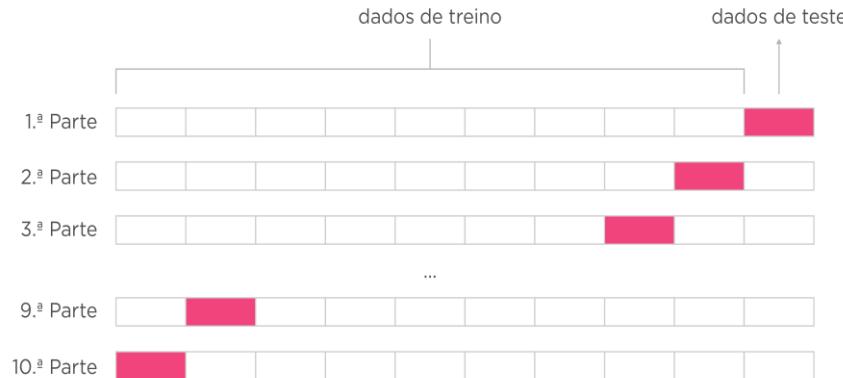
Além disso, validação cruzada é o método mais indicado quando possuímos poucos dados, já que a divisão em apenas duas bases pode acabar não fornecendo bases de treino e de teste boas o suficiente.

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



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.

O QUE SABEMOS DE MACHINE LEARNING:



COMO AVALIAR OS
MODELOS?

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:

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

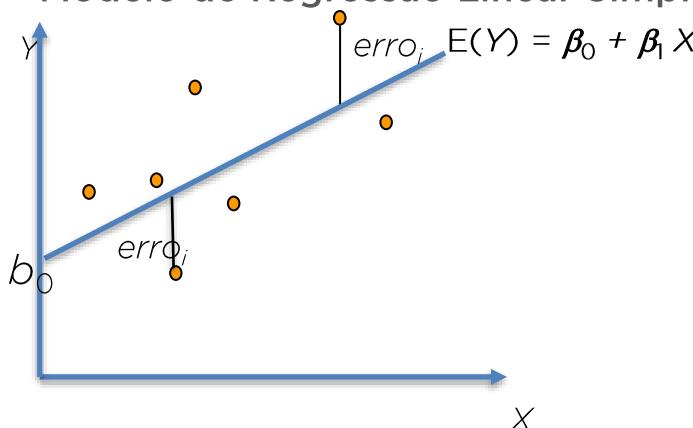
Nesse caso, um modelo bom é aquele
que possui **a maior acurácia.**

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

Mean Error (ME):

$$ME = \frac{\sum_{i=1}^n y_i - \hat{y}_i}{n}$$

Mean Absolute Error (MAE):

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$$

Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$$

Mean Percentage Error (MPE):

$$MPE = \frac{\sum_{i=1}^n \frac{y_i - \hat{y}_i}{y_i} * 100}{n}$$

Mean Absolute Percentage Error (MASE):

$$MAPE = \frac{\sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} * 100}{n}$$

- 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

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

ANÁLISE DE REGRESSÃO LOGÍSTICA

- Qualificação do Ajuste do Modelo

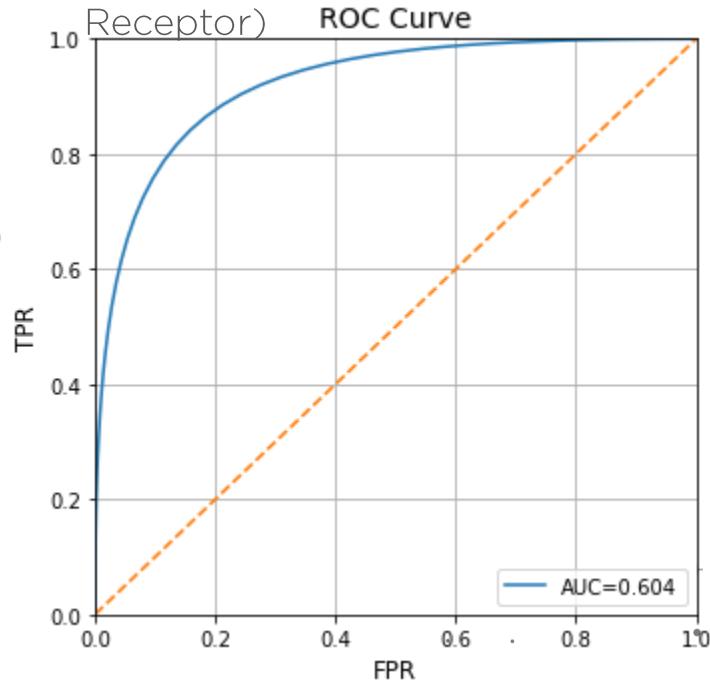
		Previsão do modelo		Total
		y=1	y=0	
Obs.	y=1	n1	n2	n1+n2
	y=0	n3	n4	n3+n4

$$\text{Sensibilidade} = n1 / (n1+n2)$$

$$\text{Especificidade} = n4 / (n3+n4)$$

- 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 .

Receiver operating characteristic
(Característica de Operação do Receptor)



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 ou revocação ou recall: $(VP / (VP + FN))$
- 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: $(VP+VN)/(VP+FN+FP+VN)$

TÉCNICAS DE CLASSIFICAÇÃO

Qualificação do Ajuste do Modelo

- Se a opção for escolher um algoritmo que como medida uma alta precisão (VP / (VP + FP))
 - teremos poucos falsos positivos
- Se a opção for escolher um algoritmo que como medida uma alta sensibilidade (VP / (VP + FN))
 - teremos poucos falsos negativos

Alternativa:

O F1 Score é a média harmônica da Precisão e da Revocação(recall);

$$F_1 = \left(\frac{2}{\frac{1}{recall} + \frac{1}{precisão}} \right) = 2 \times \frac{precision \times recall}{precision + recall}$$

Exercitando!!!



Identificar técnicas e modelos para solução de alguns
problemas

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!



/AdelaideAlves



profadelaide.alves@fiap.com.br

FIAP

Copyright © 2022 | 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.