
FIAP
MBA

Machine Learning I

profregina.bernal@fiap.com.br
reginabernal@terra.com.br

Dra. Regina Tomie Ivata Bernal
Cientista de Dados na área da Saúde

Formação Acadêmica:

- Estatístico - UFSCar
- Mestre em Saúde Pública – FSP/USP
- Doutor em Ciências – Epidemiologia - FSP/USP

Atividades Profissionais:

- Professora de pós-graduação na FIAP
- Consultora externa da SVS/MS
- Cientista de Dados em Saúde

Objetivos da Disciplina

- Ênfase na iniciação em Machine Learning.
- Casos de uso de Machine Learning.
- Apresentar os conceitos básicos e metodologias para desenvolvimento de técnicas em "Analytics" aplicada á diversos setores de atividades.
- Fornecer instrumentos, através da disseminação das técnicas estatísticas de análise de dados, para um desenvolvimento de cultura de análise dos profissionais que interagem com o processo de informação.
- Proporcionar o conhecimento necessário para reconhecer as técnicas Supervisionadas como seguintes técnicas: Regressão Linear Múltipla, Regressão Logística, e técnicas Não Supervisionadas como: Análise de Cluster ou Conglomerado (Segmentação), e, ainda, entender de que forma esses modelos auxiliam no aumento de vendas, em redução de custos e adquirir diferencial competitivo.
- Métricas de validação dos modelos.

Bibliografia

BERRY,M.J.A., LINOFF,G. **Data Mining Techniques For Marketing, Sales and Customer Support**. 3a. ed. New York: John Wiley & Sons, Inc., 2011.

CARVALHO, L.A.V., Datamining – **A mineração de dados no marketing**, **medicina, economia, engenharia e administração**. Rio de Janeiro: Editora Ciência Moderna, 2005.

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.

HAIR, J.F. / ANDERSON, R.E. / TATHAN, R.L. / BLACK, W.C. Análise multivariada de dados, 2009

IZBICKI, R.; SANSTOS, T.M dos. **Aprendizado de máquina : uma abordagem estatística [livro eletrônico].** São Carlos, SP, 2020

KROESE, D.P, BOTEV, Z.I., TAIMRE T., VAISMA R. **Data Science and Machine Learning**. Mathematical and Statistical Methods, 2020

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.

MINGOTI, S.A.; **Análise de dados através de métodos de estatística multivariada**, UFMG, 2005

.

TORGO, L. **Data Mining with R: Learning with Case Studies**, 2.a ed. Chapman and Hall/CRC , 2007

Avaliação da disciplina

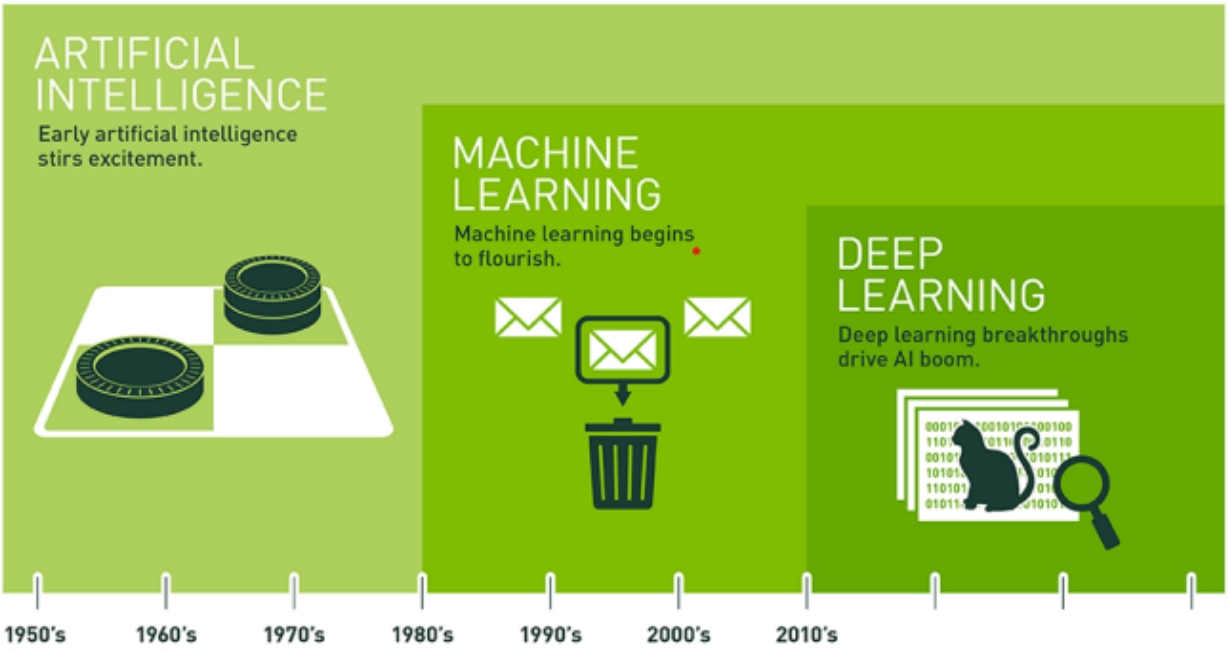
Avaliação	Peso
Listas de exercícios	0.5
Exercício prático em grupo	0.5

ESTATÍSTICA NA PRÁTICA

Fonte: Doris Fontes

De cada 10 vagas, duas são preenchidas por Bacharéis em Estatística.

INTRODUÇÃO: IA, ML & DL



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Machine Learning (Aprendizado de máquinas)

Aprendizado de máquina (AM) nasceu na década de 60 como um campo da inteligência artificial que tinha o objetivo de aprender padrões com base em dados. Originalmente, as aplicações de AM eram de cunho estritamente computacional. Contudo, desde o final dos anos 90, essa área expandiu seus horizontes e começou a se estabelecer como um campo por si mesma. Em particular, as aplicações de AM começaram a ter muitas intersecções com as de estatística.

A comunidade de AM é bastante interdisciplinar e utiliza ideias desenvolvidas em diversas áreas, sendo a estatística uma delas. Enquanto que até os anos 90 métodos criados pela comunidade estatística rapidamente começavam a ser incorporados em AM, mais recentemente o fortalecimento de AM fez com que a direção oposta começasse a ficar cada vez mais comum: métodos desenvolvidos por AM começaram a ser usados em estatística.

Fonte: Aprendizado de máquina : uma abordagem estatística [livro eletrônico] / Rafael Izbicki, Tiago Mendonça dos Santos. -- São Carlos, SP : Rafael Izbicki, 2020. PDF

Machine Learning (Aprendizado de máquinas)

In our present world of automation, cloud computing, algorithms, artificial intelligence, and big data, few topics are as relevant as *data science* and *machine learning*. Their recent popularity lies not only in their applicability to real-life questions, but also in their natural blending of many different disciplines, including mathematics, statistics, computer science, engineering, science, and finance.

To someone starting to learn these topics, the multitude of computational techniques and mathematical ideas may seem overwhelming. Some may be satisfied with only learning how to use off-the-shelf recipes to apply to practical situations. But what if the assumptions of the black-box recipe are violated? Can we still trust the results? How should the algorithm be adapted? To be able to truly understand data science and machine learning it is important to appreciate the underlying mathematics and statistics, as well as the resulting algorithms.

Fonte: Dirk P. Kroese, Zdravko I. Botev, Thomas Taimre, Radislav Vaisma. **Data Science and Machine Learning.** Mathematical and Statistical Methods, 2020

Algumas aplicações de Análises Estatísticas para tomada de decisão

- Financeiro
- Cartões de Crédito
- Seguros
- Indústria
- Varejo
- E-commerce
- Saúde
- Medicina
- Assistência Médica
- Telecom
- Aviação
- Ação Social
- Educação
- Utilies: Energia, Água
- Processos Civeis
- Fraudes
- .....

Aplicações de Machine Learning

DETECÇÃO DE
FRAUDES EM CARTÕES

SELF-DRIVING CARS

MERCADO FINANCEIRO

VENDAS /
RETENÇÃO DE CLIENTES

Aplicações de Machine Learning

RECOMENDAÇÕES

MEDICINA

TELECOMUNICAÇÕES

APLICAÇÃO DE MACHINE LEARNING

Ethical Implications Of Bias In Machine Learning

Adrienne Yapo
Bentley University
175 Forest St.
Waltham, MA 02452
adrienne.yapo@gmail.com

Joseph Weiss
Bentley University
175 Forest St.
Waltham, MA 02452
jweiss@bentley.edu

1.2. Machine learning algorithm bias

Although machine learning algorithms can produce numerous benefits to individuals, consumers, businesses, investors, the government, and society at large, recent research has uncovered many instances of bias in machine learning algorithms that have troubling implications and deleterious consequences.

1.3. Machine learning in the criminal justice system

Yet perhaps the most troubling incidents of bias in machine learning to date are unfolding in the criminal justice system. Consider the following statement from then U.S. Attorney General Eric Holder on the Sentencing Reform and Corrections Act of 2015:

Table 1: Disproportionate incarceration rates

Source: ***Propublica* analysis from Broward County, Florida**

Prediction Fails Differently for Black Defendants

	White	African American
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

"Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes."

Fonte: https://scholarspace.manoa.hawaii.edu/bitstream/10125/50557/paper0670.pdf

Publicidade Digital Moderna

Muitos Dados +

Métricas (+/- Fáceis de Medir) +

Velocidade

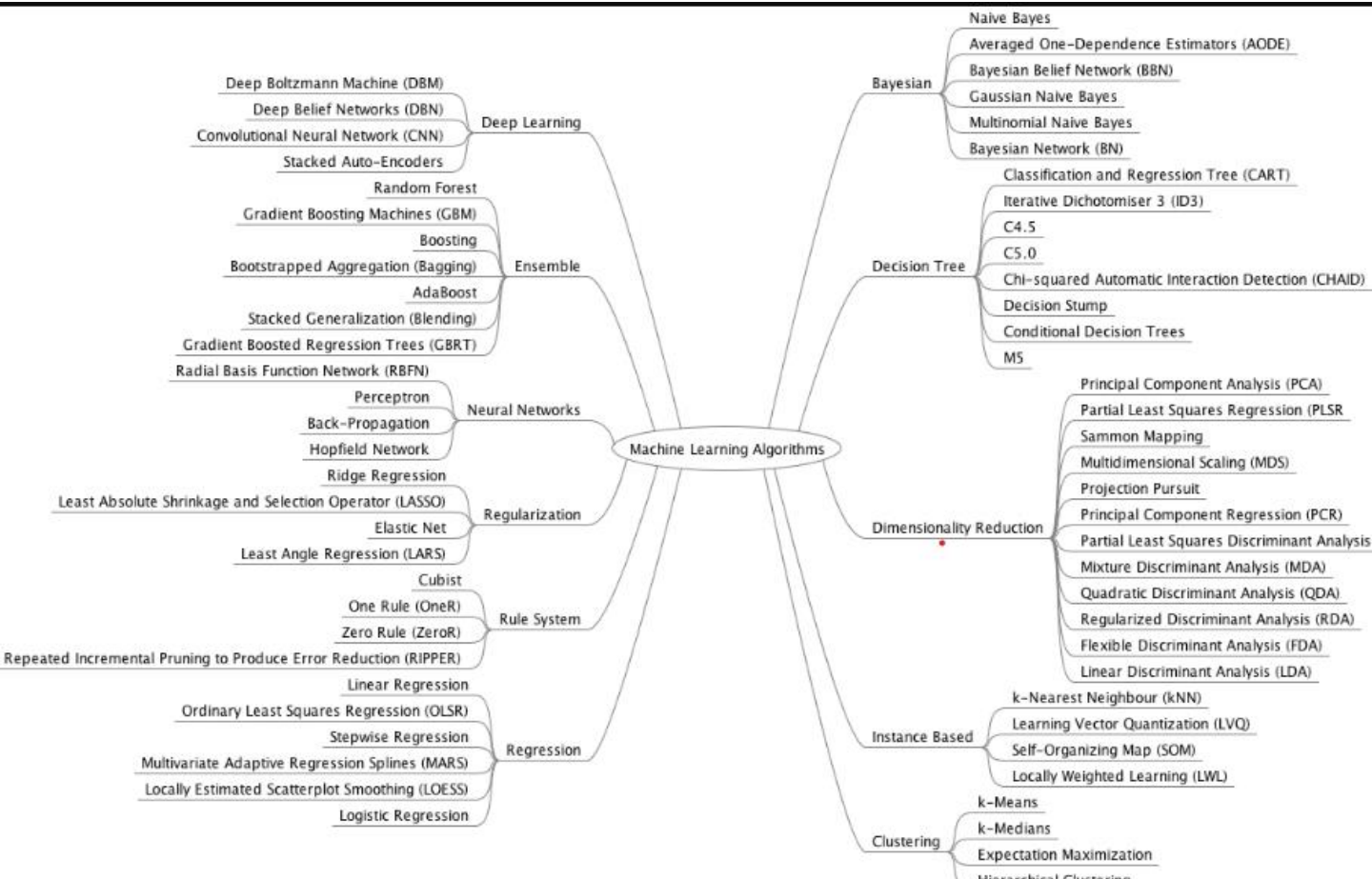
Resumo

Esta palestra mostra como funciona a publicidade digital moderna e como Data Science é uma ferramenta essencial para entender o perfil do consumidor e fazer campanhas mais assertivas.

Fonte: https://www.infoq.com/br/presentations/data-science-em-publicidade-digital/

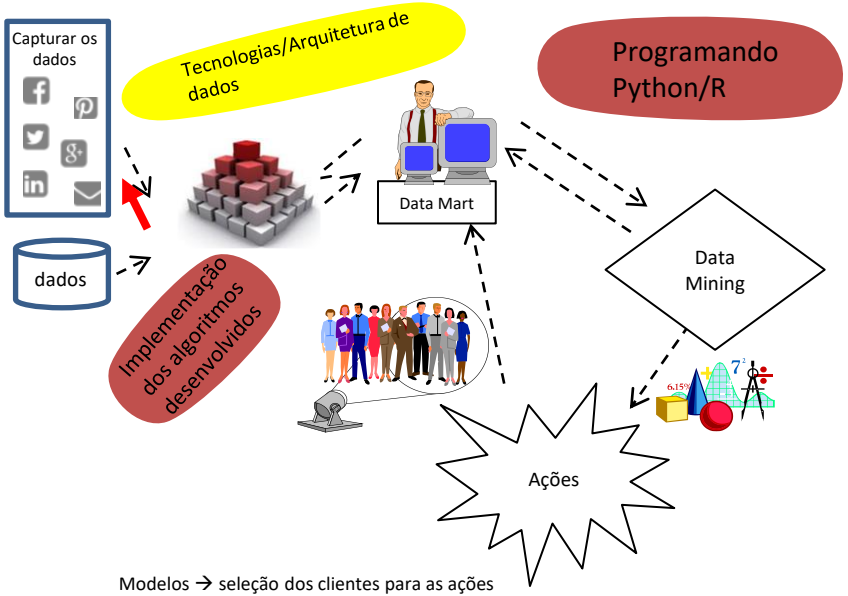
Algoritmos de Machine Learning

Machine Learning Algorithms Mindmap



Fonte: https://jixta.wordpress.com/2015/07/17/machine-learning-algorithms-mindmap/

Analytics

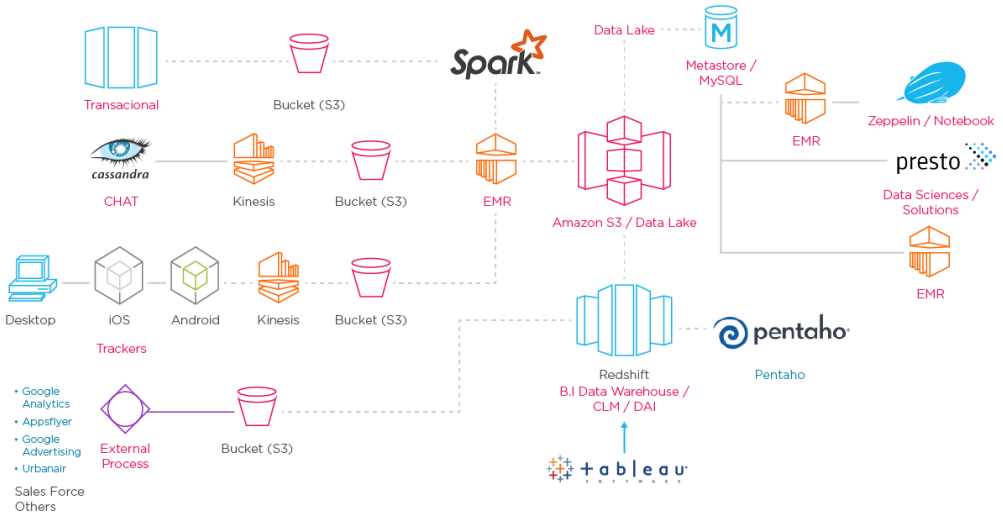


Modelos → seleção dos clientes para as ações
Segmentação dos clientes → comunicação customizada

- Conceitos Estatísticos
- Visualização dos dados
- Modelos IA & ML
- Deep Learning
- Sistema de recomendação

EXEMPLO

Plataforma de Dados



Fonte: https://www.infoq.com/br/presentations/data-science-na-olx

Data Mining

- A mineração de dados é um processo de análise detalhada de dados, para extrair e apresentar informações recentes, implícitas e que possam ser utilizadas para resolver um problema.

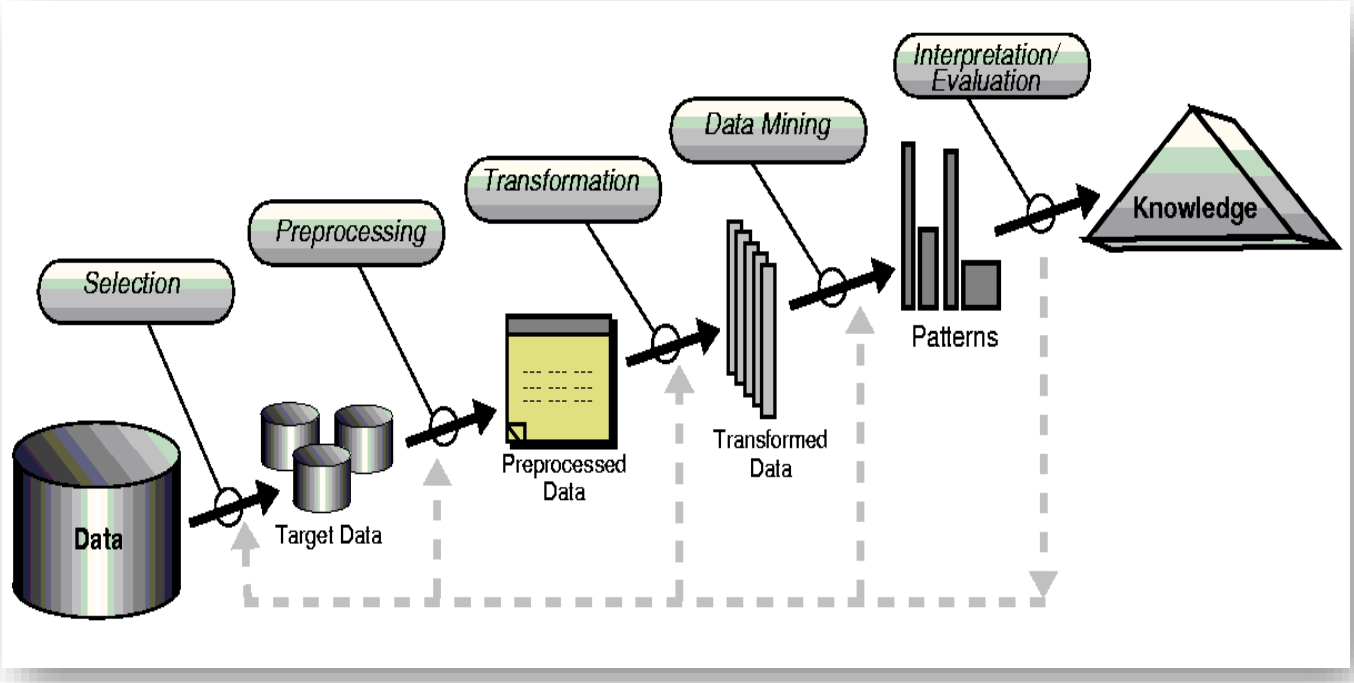
- Uso de técnicas, preferencialmente automáticas, de exploração de grandes quantidades de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertos a olho nú pelo ser humano *(Carvalho, 2001).*

Knowledge Discovery in Databases

Produzir conhecimento novo escondido em grandes bases de dados

Otimizar e Automatizar o processo de descrição de Tendências e de Padrões.

Utiliza-se um conjunto de técnicas estatísticas e de inteligência artificial.



Fonte: Berry & Linoff, 2004.

Modelos Preditivos

Estatística Tradicional

- Inferência estatística

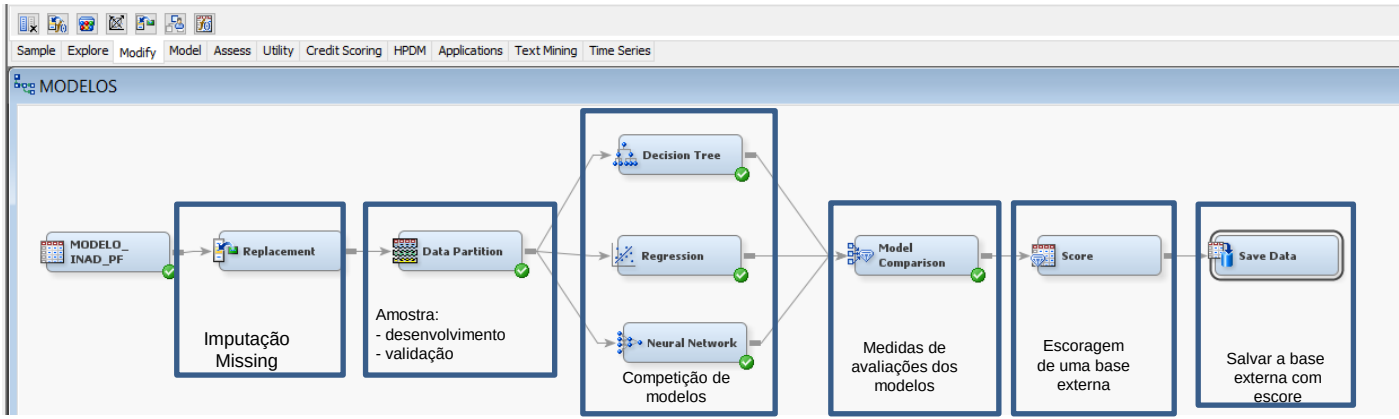
Conjunto de técnicas estatísticas baseadas em I.C. e erro padrão

Machine Learning

- Baseado em algorítmos
- 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

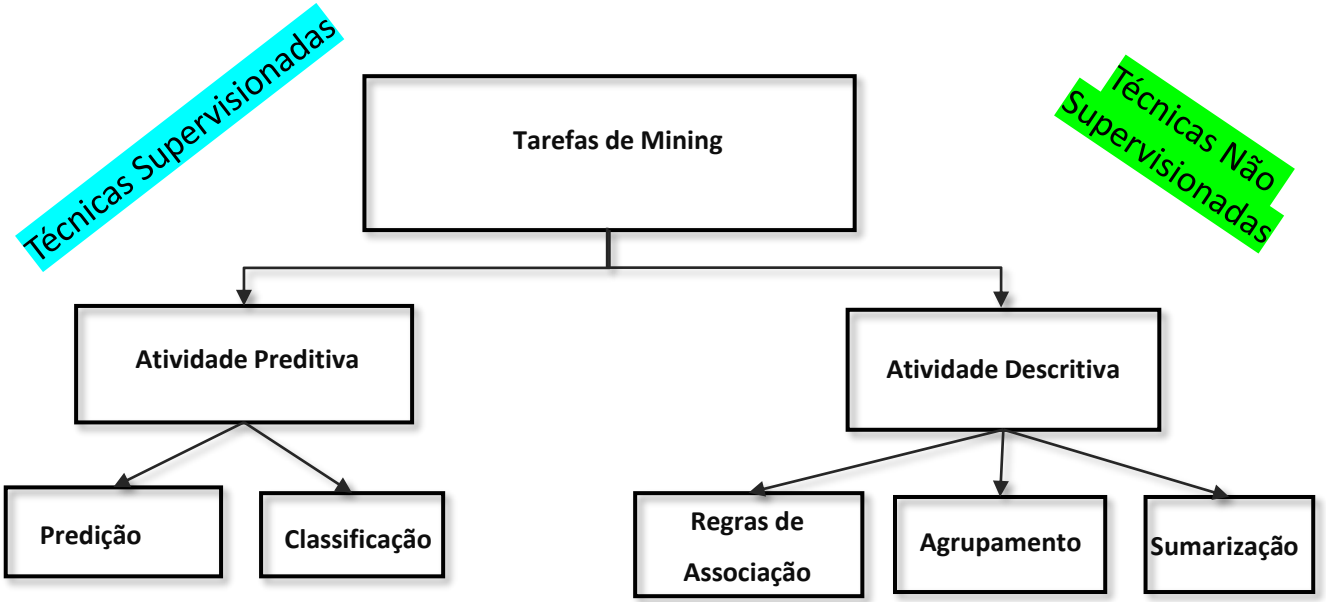
Enterprise Miner - SAS

Processo de modelagem e escoragem



Data Mining

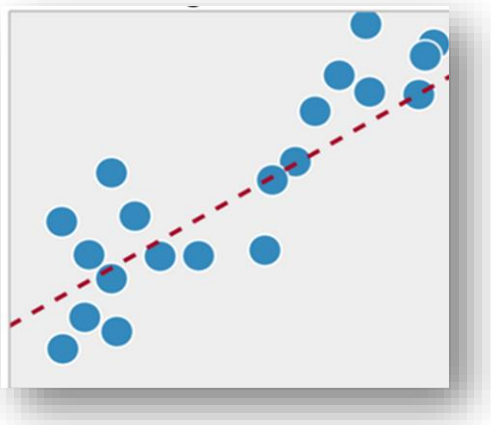
aplicações práticas de Data Mining se podem ser categorizadas de acordo com a tarefa que se pretende resolver



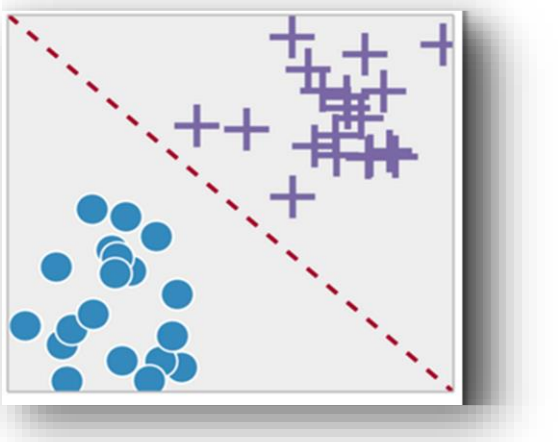
Data Mining: Mineração - Construção de Modelos

aplicações práticas de Data Mining se podem ser categorizadas de acordo com a tarefa que se pretende resolver

- **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.

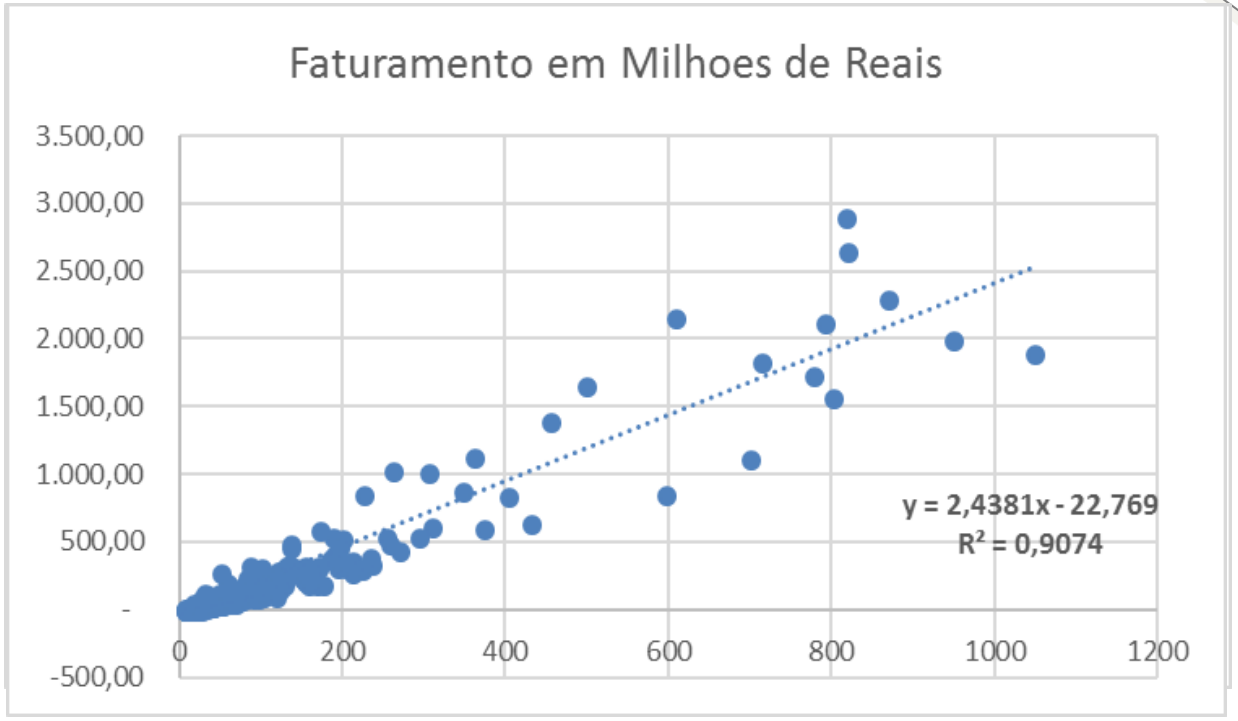


Fonte: Ronaldo Goldschmidt e Eduardo Bezerra - DataMining: conceitos, técnicas, algoritmos, orientações e aplicações. *Editora Elsevier.*

Técnica de Regressão: Regressão Linear Simples

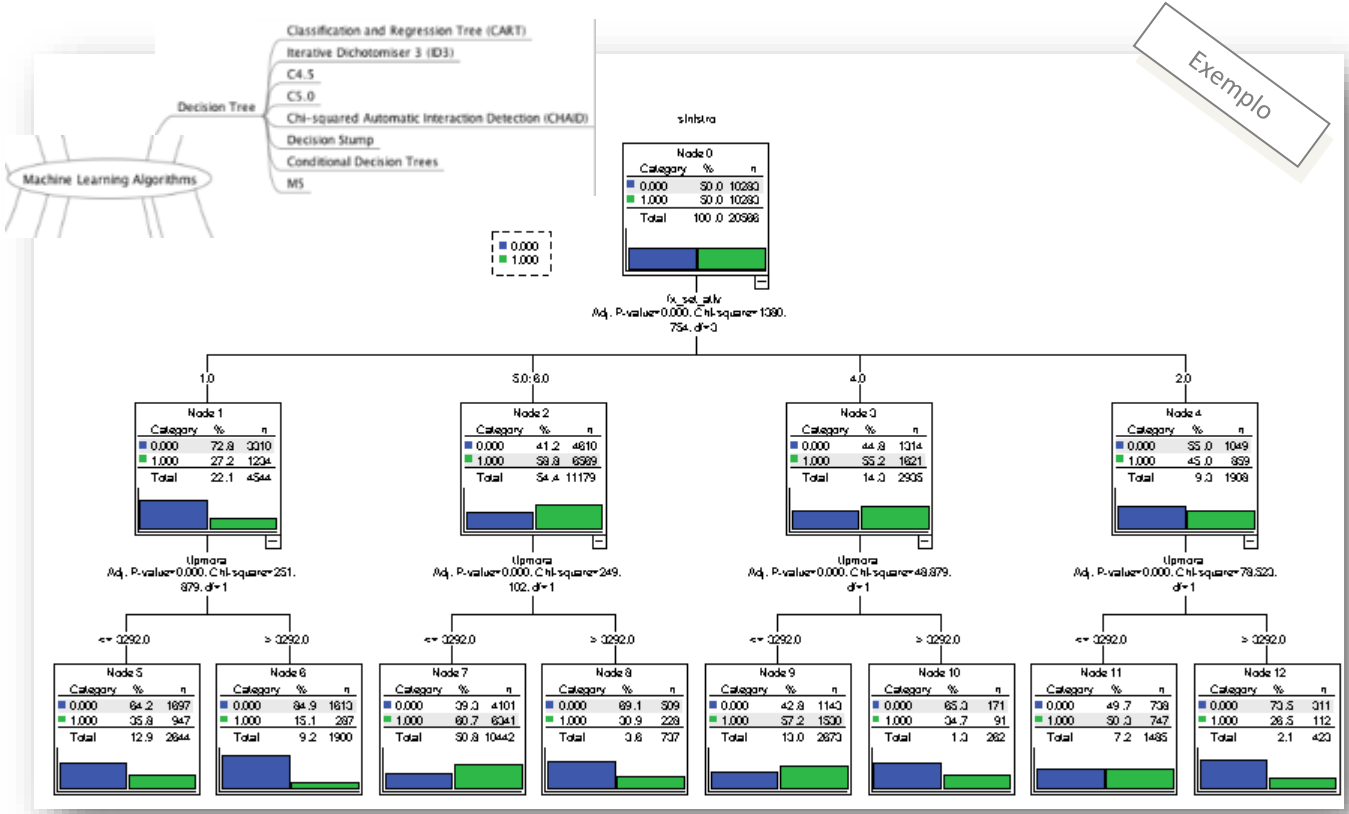
Exemplo

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



Fonte:ABRAS

Técnica de Classificação: Árvore de Decisão



Técnica de Classificação: Regressão Logística

FIAP MBA+

Exemplo

Modelo de Churn

Pesos definidos na modelagem

-0,24	Grupo 7	Origem	Grupo 1	0,29
-1,84	Grupo 1	Grupo de CEP	Grupo 6	1,34
-0,63	46 ou mais	Tempo de Base em meses	Menos de 12	0,73
	0	Atendimento Call Center	6 ou mais	
-0,59	0	Média de dias de Atraso	Mais de 24	0,88
	Mais de R$1000	Valor do Plano/Pacote	Menos de 160	
-0,11	Acima de 59 anos	Faixa Etária	18 a 23 anos	0,18
	2 ou mais	Dependentes	0	
0,23		Constante		0,23
4%	Propensão			98%

Técnica de Classificação: Regressão Logística

Clientes

Clientes com o evento(*)

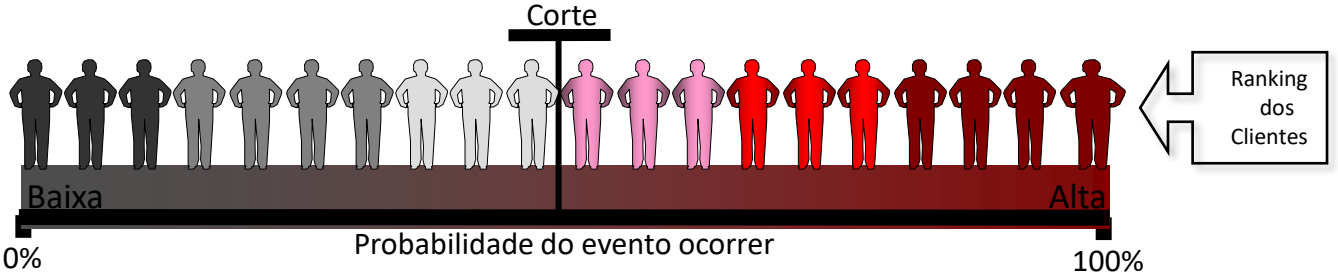
(*) Evento (exemplos)

- Aquisição
- Cancelamento
- Pagamento

Modelo

Pontuação dos Clientes com Probabilidade

Regra de Decisão

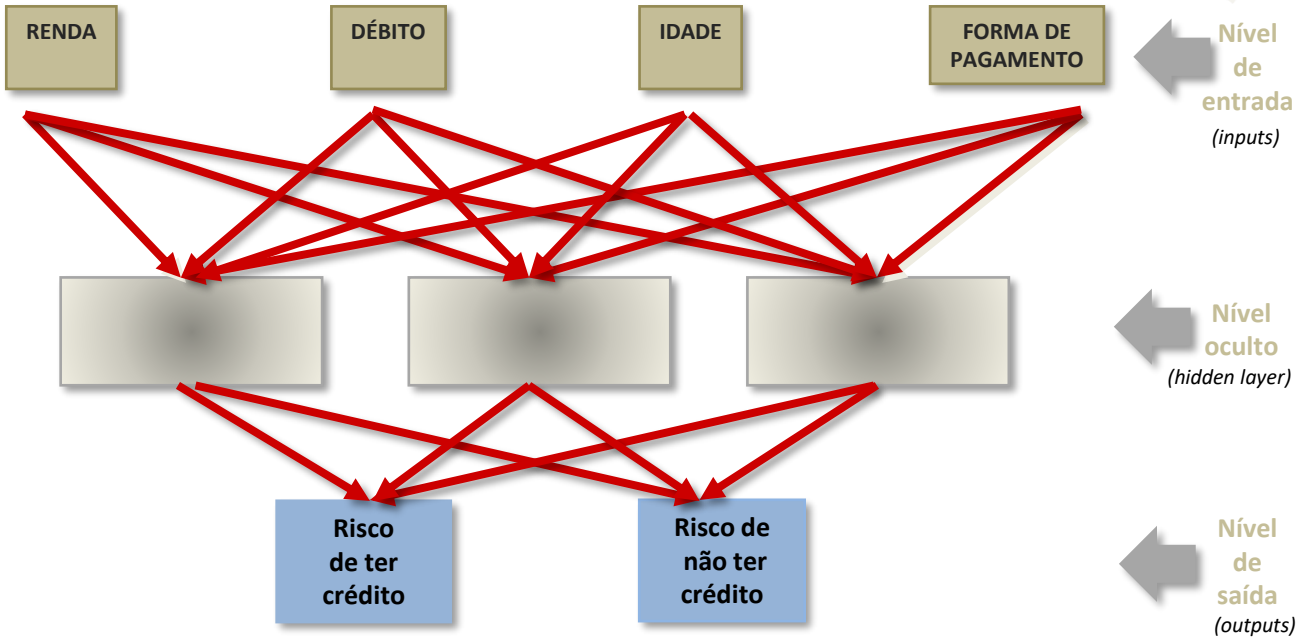


Técnica de Classificação: Redes Neurais

FIAP MBA+

Exemplo

Exemplo: risco de crédito



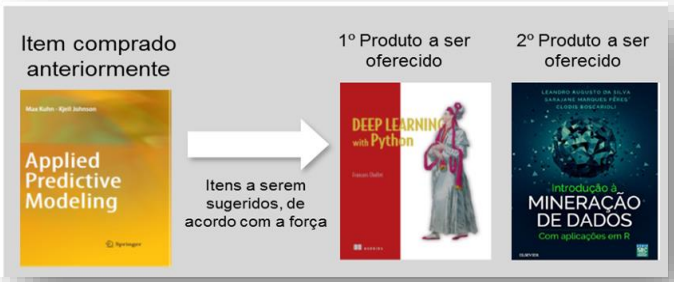
As redes neurais usam dados de entrada.
Atribui pesos nas conexões entre os atributos (neurônios).
E obtém um resultado (risco de ter ou não crédito) - nível de saída.

Data Mining

aplicações práticas de Data Mining se podem ser categorizadas de acordo com a tarefa que se pretende resolver

● **Descoberta de Associações:** Nessa 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.



Fonte: Ronaldo Goldschmidt e Eduardo Bezerra - DataMining: conceitos, técnicas, algoritmos, orientações e aplicações. *Editora Elsevier.*

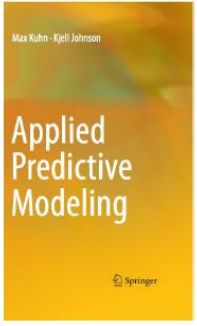
Técnica Descoberta de Sequências

Quais associações são significativas ?

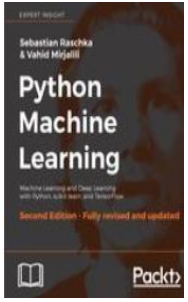
Item comprado anteriormente

Itens a serem sugeridos de acordo coma força

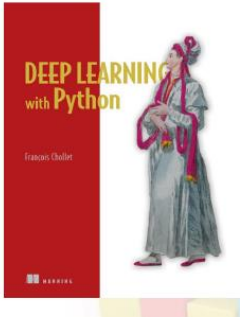
1.o produto



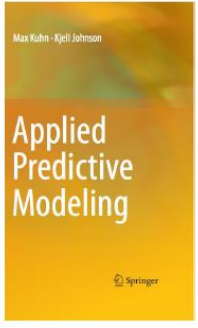
2.o produto



3.o produto

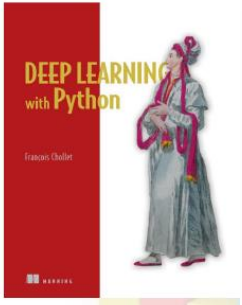


Item comprado anteriormente



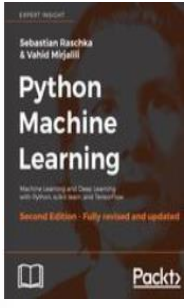
Itens a serem sugeridos de acordo coma força

1.o produto



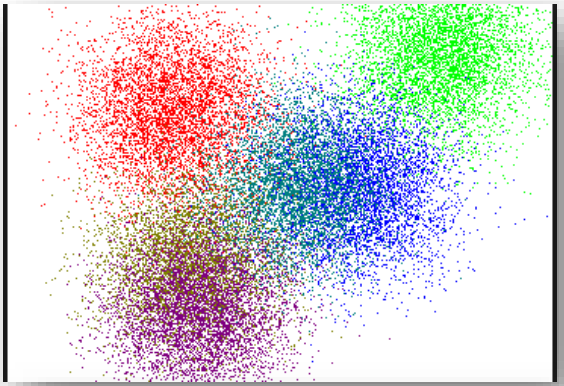
2.o produto

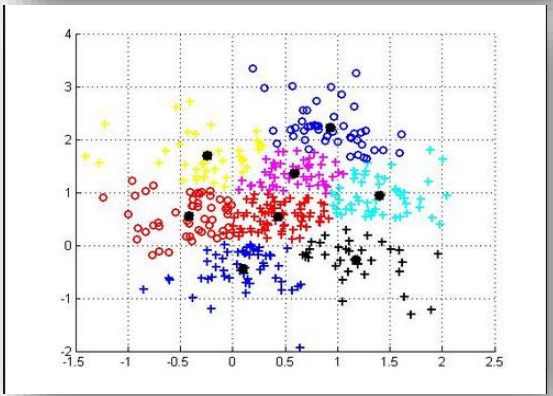
3.o produto



Data Mining

aplicações práticas de Data Mining se podem ser categorizadas de acordo com a tarefa que se pretende resolver





● **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.

Fonte: Ronaldo Goldschmidt e Eduardo Bezerra - DataMining: conceitos, técnicas, algoritmos, orientações e aplicações. *Editora Elsevier.*

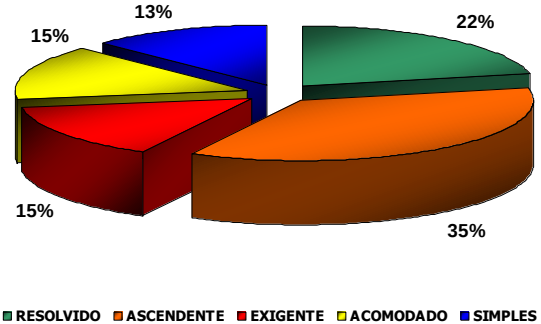
Técnica de Agrupamento: Análise de Clusters

FIAP MBA+

Exemplo

Segmentação Comportamental do Cliente

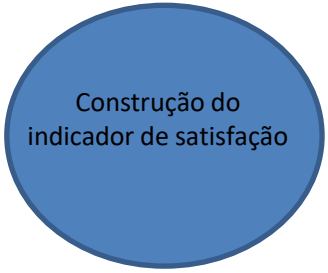
- Cadastro
- Gasto com Voz
- Gasto com Dados
- Pagamento
- Serviços Contratados
- Crédito
- Mercado (geográfico)



Data Mining

aplicações práticas de Data Mining se podem ser categorizadas de acordo com a tarefa que se pretende resolver

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



Exemplo

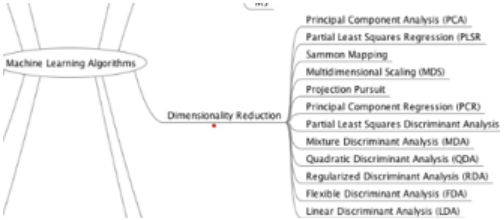
- Velocidade de acesso à internet
- Utilidade / Adequação da internet
- Estabilidade da conexão
- Disponibilidade de acesso à Internet
- Valores pelo acesso da internet
- Interesse dos(as) atendentes
- Solução dada pela empresa
- Conhecimento dos(as) atendentes
- Rapidez com que é dada a resposta
- Tempo para ser atendido
- Conhecimento dos tipos de serviços
- Informações apresentadas nos manuais
- Utilidade das informações na mídia
- Informações sobre os serviços e planos
- Informações sobre as áreas de cobertura
- Diversidade e facilidade de aquisição
- Utilidade / adequação dos serviços
- Tempo de espera para ser atendido
- Tempo do atendimento
- Conhecimento e preparo atendentes
- Solução dos problemas
- Valor dos descontos de horários
- Preço da ligação
- Modernidade da empresa
- Quantidade de vezes que não funciona
- Frequência que ocorre queda da ligação
- Cobertura no Estado
- Fazer e receber ligações na sua cidade
- Qualidade das ligações em áreas internas
- Qualidade das ligações entre celulares

Fonte: Ronaldo Goldschmidt e Eduardo Bezerra - DataMining: conceitos, técnicas, algoritmos, orientações e aplicações. *Editora Elsevier.*

Técnica de Sumarização : Análise Fatorial

Exemplo

- Entendimento das variáveis latentes
- Criação de Indicadores



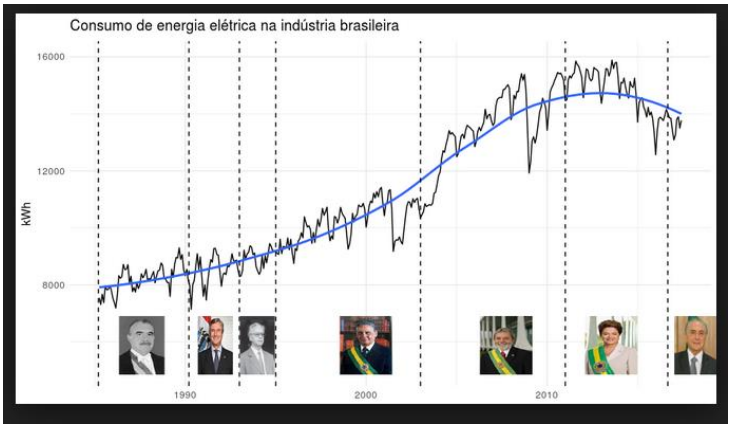
IDH

- 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*

Data Mining

aplicações práticas de Data Mining se podem ser categorizadas de acordo com a tarefa que se pretende resolver

● **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.



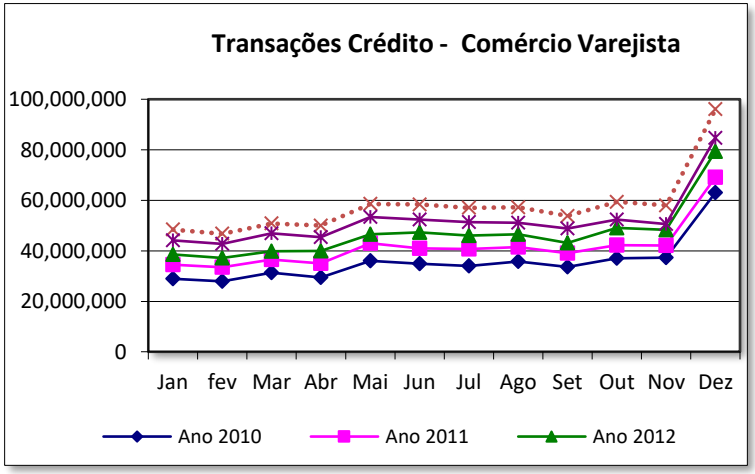
Fonte: Ronaldo Goldschmidt e Eduardo Bezerra - DataMining: conceitos, técnicas, algoritmos, orientações e aplicações. *Editora Elsevier.*

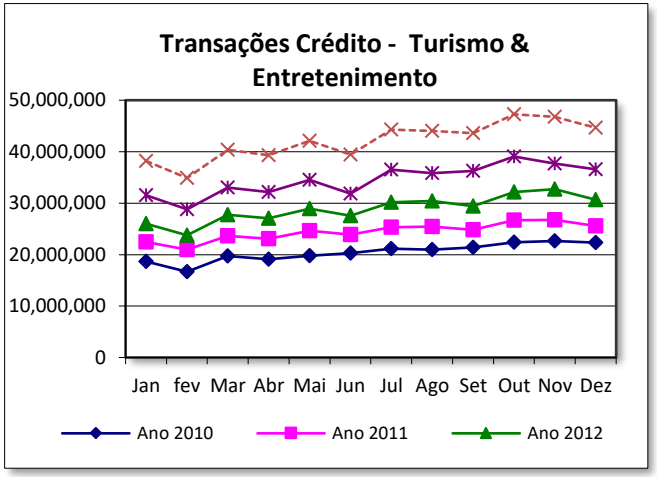
Modelos de Séries Temporais

Previsão

Exemplo

Quantidade de transações mensais com cartões de crédito



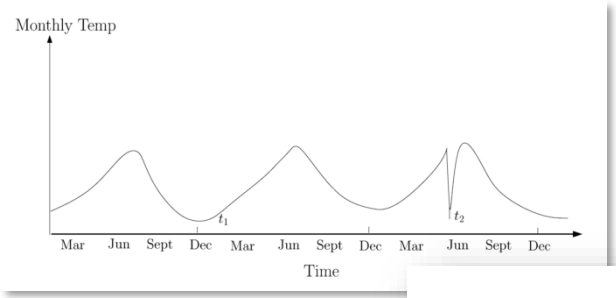


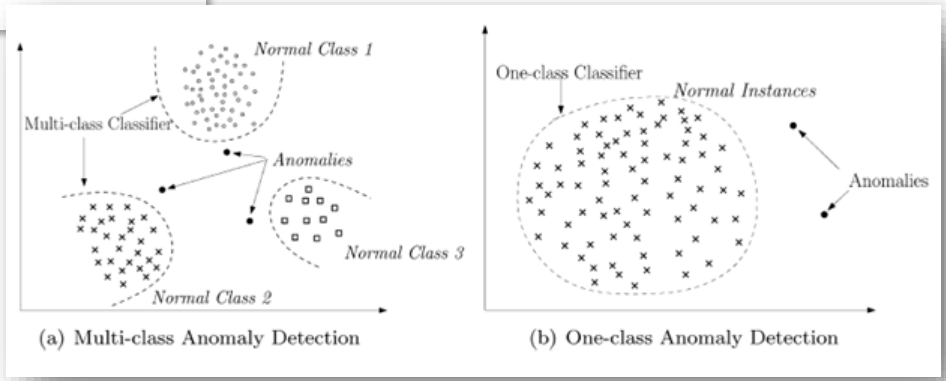
Fonte:ABECS

Data Mining

aplicações práticas de Data Mining se podem ser categorizadas de acordo com a tarefa que se pretende resolver

● **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).





Fonte: Ronaldo Goldschmidt e Eduardo Bezerra - DataMining: conceitos, técnicas, algoritmos, orientações e aplicações. *Editora Elsevier.*

Análise Multivariada

FIAP MBA+

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

Análise Estrutural

Técnicas de Interdependência

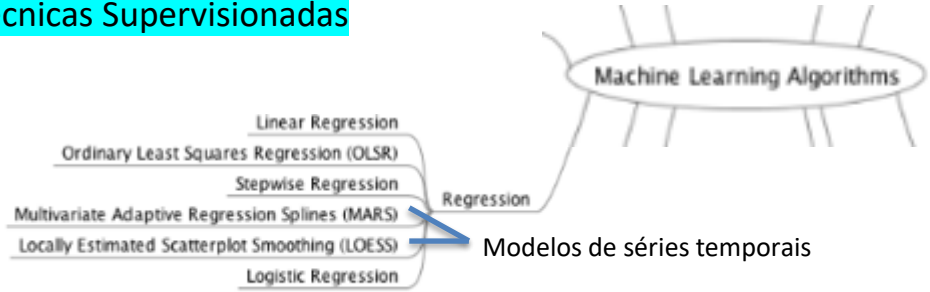
Técnicas Multivariadas que procuram agrupar dados com base em semelhança, permitindo assim a interpretação das estruturas dos dados. Não há distinção entre variáveis dependentes e independentes.

Análise Supervisionada

Análise Não Supervisionada

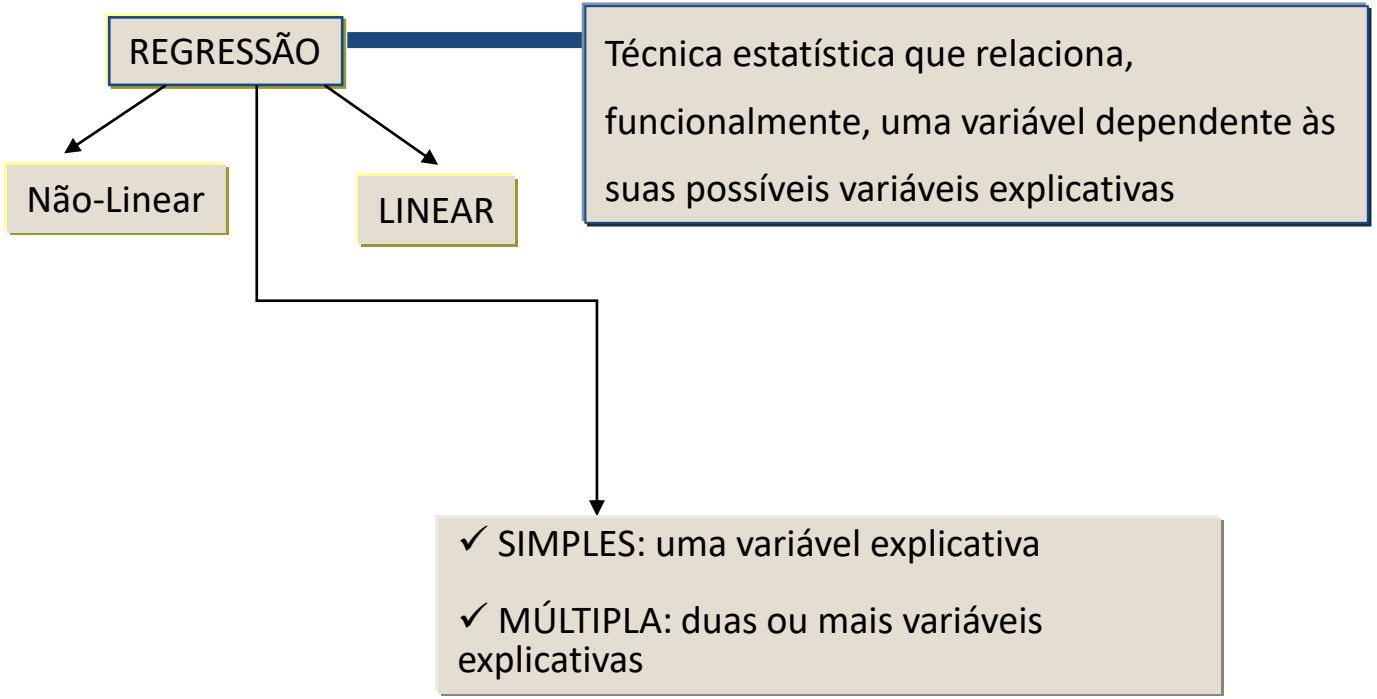
Machine Learning Algorithms Mindmap

Técnicas Supervisionadas



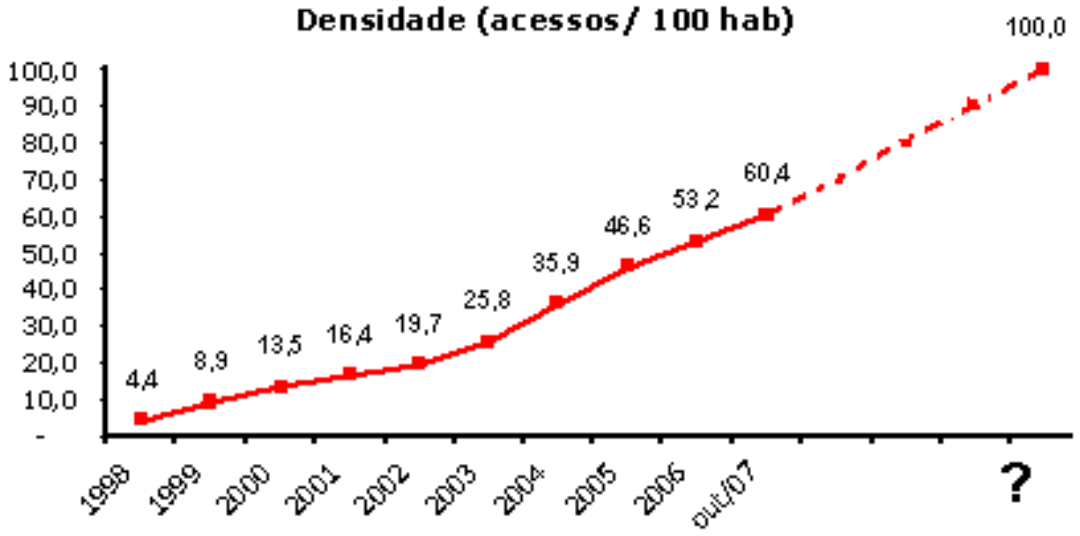
Predição (Resposta (y) é quantitativa)	Classificação (Resposta (y) é qualitativa)
Linear Regression	Logistic Regression
Ordinary Least Square Regression (OLSR)	
Stepwise Regression	
Multivariate Adaptive Regression Splines (MARS)	
Locally Estimated Scatterplot Smoothing (LOESS)	

Modelos de Regressão



Regressão Linear

✓ Exemplo: Quando o Brasil vai ter 100 celulares para cada 100 habitantes?



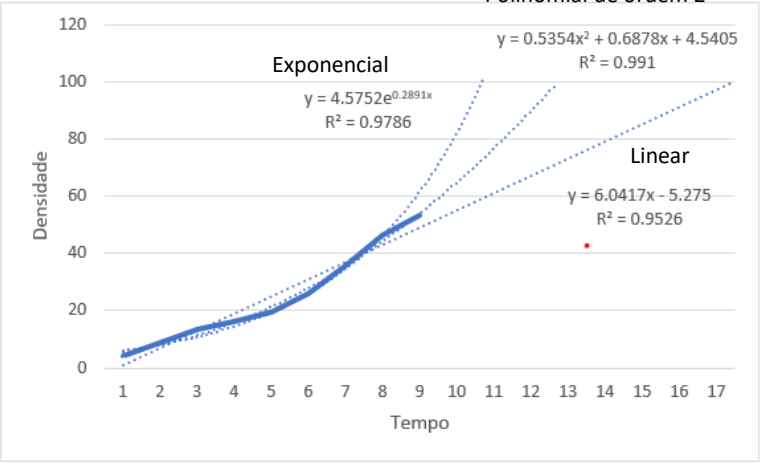
Fonte: http://www.teleco.com.br/comentario/com237.asp

Modelos de Regressão

✓ Exemplo: Quando o Brasil vai ter 100 celulares para cada 100 habitantes?

Ano	Tempo (X)	Densidade (Y)	linear	exponencial	polinomial2
1998	1	4.4	0.77	6.11	5.76
1999	2	8.9	6.81	8.16	8.06
2000	3	13.5	12.85	10.89	11.42
2001	4	16.4	18.89	14.54	15.86
2002	5	19.7	24.93	19.42	21.36
2003	6	25.8	30.98	25.93	27.94
2004	7	35.9	37.02	34.62	35.59
2005	8	46.6	43.06	46.22	44.31
2006	9	53.2	49.10	61.72	54.10
out/2007	10	60.4	55.14	82.41	64.96
2008	11		61.18	110.03	76.89
2009	12		67.23		89.89
2010	13		73.27		103.96
2011	14		79.31		
2012	15		85.35		
2013	16		91.39		
2014	17		97.43		
2015	18		103.48		

Polinomial de ordem 2



Exponencial

$$y = \alpha e^{\beta x}$$

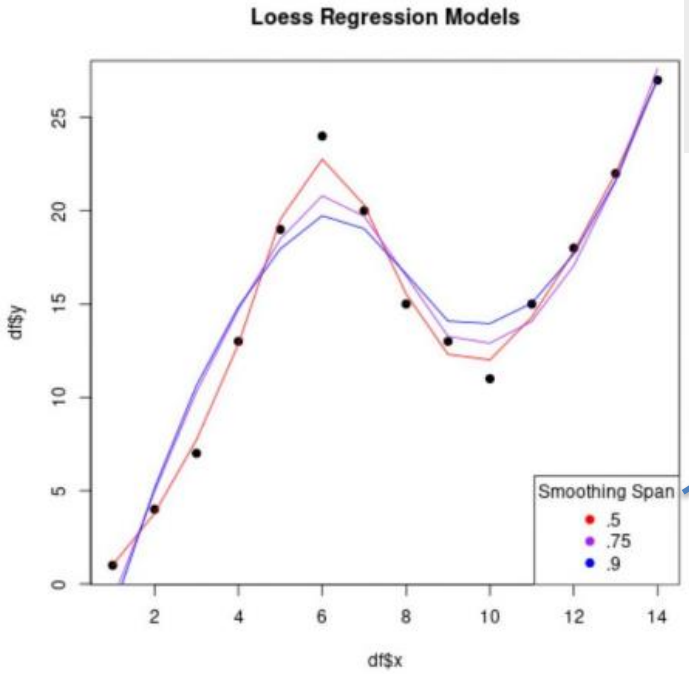
Linear

$$y = \beta_0 + \beta_1 x$$

Polinomial de ordem 2

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

Modelos de Regressão




```r
library(caret)

#define k-fold cross validation method
ctrl <- trainControl(method = "cv", number = 5)
grid <- expand.grid(span = seq(0.5, 0.9, len = 5), degree = 1)

#perform cross-validation using smoothing spans ranginf from 0.5 to 0.9
model <- train(y ~ x, data = df, method = "gamLoess", tuneGrid=grid, trControl = ctrl)
```

Hiperparâmetro do modelo

Método K-fold para encontrar o melhor parâmetro do modelo

Fonte: https://www.statology.org/loess-regression-in-r/

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.

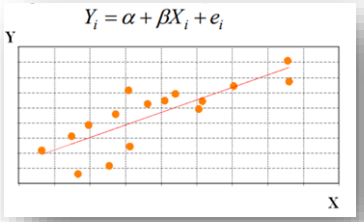
MODELO DE REGRESSÃO:

Esse 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

Modelo Regressão Linear Simples e Múltipla

O Modelo que relaciona Y com várias variáveis independentes



- Modelo Linear Simples: $Y=B_o + B_1X + e$

  X = variáveis independentes

  Y = variável dependente

  B_o = constante

  B_1 = coeficientes de regressão

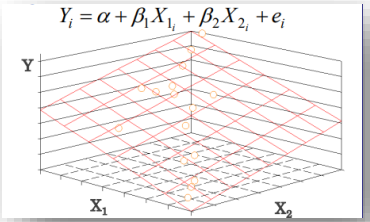
- Modelo Linear Múltiplo: $Y=B_o+B_1X_1+B_2X_2+B_3X_3+...+B_nX_n+ e$

  $X_1,X_2,X_3,....,X_n$ = variáveis independentes

  Y = variável dependente

  B_o = constante

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



Análise de Variância

A variabilidade total observada na variável dependente está dividido em componentes:

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

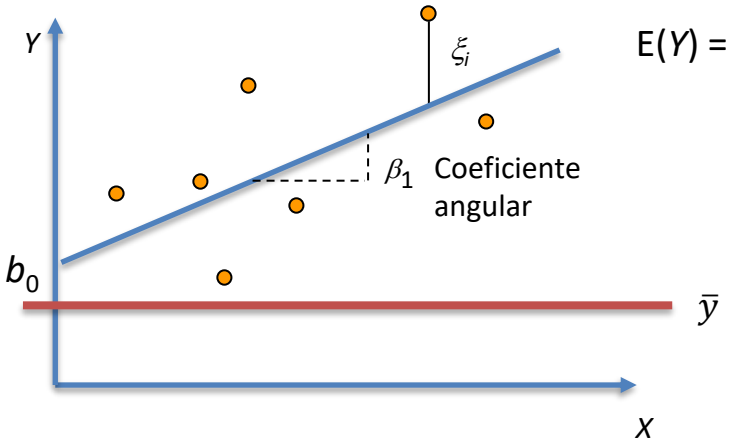
Soma de quadrados total	Soma de quadrados residual	Soma de quadrados regressão
SQTot	SQRes	SQReg

Análise de Variância

Podemos resumir todas essas informações numa única tabela anova:

Fonte	gl	SQ	QM	F
Regressão	p - 1	SQReg	$QMReg = \frac{SQReg}{p-1}$	$\frac{QMReg}{Se^2}$
Resíduo	n - p	SQRes	$Se^2 = \frac{SQRes}{n-p}$	
Total	n - 1	SQTOT	$S^2 = \frac{SQTot}{n-1}$	

Análise de Variância (ANOVA)



$E(Y) = \beta_0 + \beta_1 X$

Teste de hipóteses:

H_0: regressão = $\bar{y}$

H_1: regressão ≠ $\bar{y}$

Critério de decisão:

Se p-valor < 0.05 então rejeito H0
Se p-valor >= 0.05 então não rejeito H0

```
(Intercept) 1214.6 161.2 7.537 0.000000000000143 *** temp 6640.7 305.2 21.759 <
0.0000000000000002 ***
```

Análise de Variância

Acurácia do modelo

Coeficiente de determinação (R^2): Multiple R-squared

$$R^2 = \frac{SQReg}{SQTot}$$

Coeficiente de determinação ajustado (R^2): Adjusted R-squared

$$R_a{}^2 = 1 - \frac{n-1}{n-(p+1)}(1-R^2)$$

Onde: n = número de observações
p = número de variáveis preditoras

ANÁLISE DE ASSOCIAÇÃO

Analisa o comportamento conjunto de duas variáveis qualitativas apresentada em tabela bivariada.

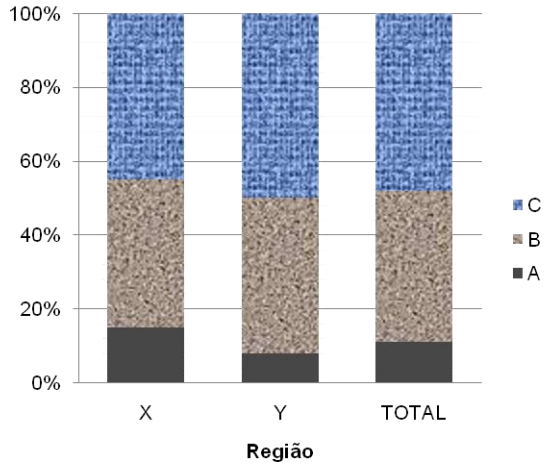
- Teste Qui-Quadrado (Variáveis Qualitativas)
- Correlação de Pearson (Variáveis Quantitativas)

Existe associação entre as vendas de produto e região?

Exemplo:

Tabela 1 – Distribuição de vendas segundo produto e região. 2018

Produto	Região				Total	
	X		Y			
	N	%	N	%	N	%
A	300	15	200	8	500	11
B	800	40	1000	42	1800	41
C	900	45	1200	50	2100	48
Total	2000	100	2400	100	4400	100



Existe associação entre as vendas de produto e região?

		REGIAO		
		X	Y	TOTAL
Produto	A	300	200	500
	B	800	1000	1800
	C	900	1200	2100
		2000	2400	4400

Chi Square for R by C Table

Chi Square=	49.12
Degrees of Freedom=	2
p-value=	<0.0000001

Cochran recommends accepting the chi square if:
1. No more than 20% of cells have expected < 5.
2. No cell has an expected value < 1.

In this table:
None of 6 cells have expected values < 5.
No cells have expected values < 1.

Using these criteria, this chi square can be accepted.

Expected value = row total*column total/grand total

Rosner, B. Fundamentals of Biostatistics. 5th ed. Duxbury Thompson Learning. 2000; p. 395

Teste de independência qui-quadrado

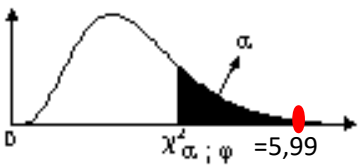
H_0: independentes

H_1: dependentes

α= 5%

Conclusão:
Rejeito H0, portanto há associação.

Graus de liberdade=



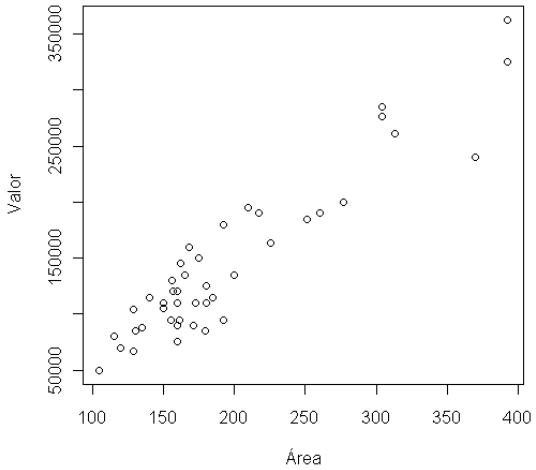
φ = graus de liberdade

ANÁLISE DE ASSOCIAÇÃO

Analisa o comportamento conjunto de duas variáveis qualitativas apresentada em tabela bivariada.

- Teste Qui-Quadrado (Variáveis Qualitativas)
- Correlação de Pearson (Variáveis Quantitativas)

Existe correlação entre o valor do imóvel e a área?



Teste Correlação de Pearson

H_0: r = 0 (ausência de correlação)

H_1: r≠ 0 (presença de correlação)

Erro de decisão: 0,05 ou 5%

R Console

```
> corr_t<-cor.test(Valor,Área,method="pearson",alternative="two.sided")
> corr_t

        Pearson's product-moment correlation

data:  Valor and Área
t = 17.0563, df = 41, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8845899 0.9651600
sample estimates:
      cor
0.9362024
```

Conclusão:

Rejeito H0, portanto há associação.

CORRELAÇÃO DE PEARSON

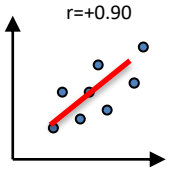
Correlação indica a força e a direção do relacionamento linear entre duas variáveis aleatórias. No uso estatístico geral, correlação se refere à medida da relação entre duas variáveis, embora correlação não implique causalidade. Nesse sentido geral, existem vários coeficientes medindo o grau de correlação, adaptados à natureza dos dados.

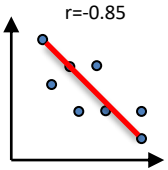
Vários coeficientes são utilizados para situações diferentes. O mais conhecido é o coeficiente de correlação de Pearson, o qual é obtido dividindo a covariância de duas variáveis pelo produto de seus desvios padrão. Apesar do nome, ela foi apresentada inicialmente por Francis Galton, em meados do século XVII.

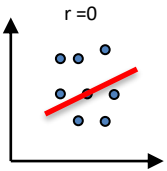
Coeficiente de correlação de Pearson, em geral é expresso por (R ou ρ).

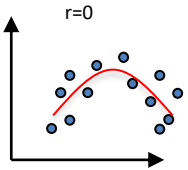
CORRELAÇÃO DE PEARSON

Análise de correlação









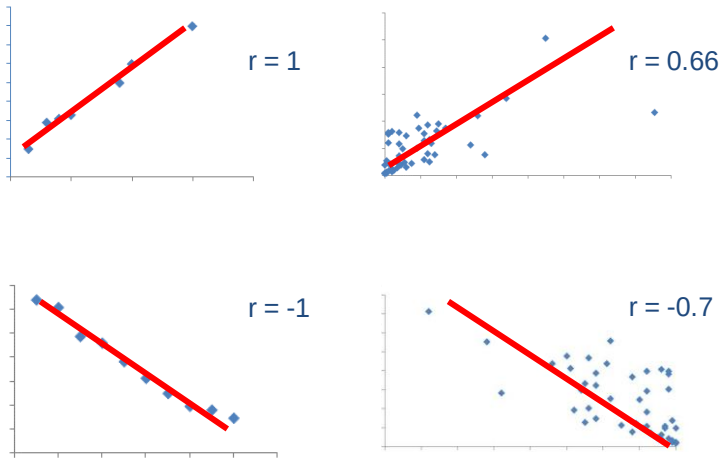
Correlação Linear Simples
(r de Pearson)

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

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

CORRELAÇÃO DE PEARSON

Análise de correlação



Interpretação dos resultados

Vemos na tabela abaixo que apenas os 3 primeiros resultados não Incluem dúvidas. Entretanto, nas demais situações o quadrado do valor de R informa a quantidade de variação conjunta para as duas variáveis.

R		**R^2**	**Variação conjunta %**
- 1	Correlação negativa perfeita	1	100
0	Independência	0	nenhuma
1	Correlação perfeita positiva	1	100
0,9	????	0,81	81
0,8		0,64	64
0,7		0,49	49
0,5	????	0,25	25
0,6		0,36	36
0,3	????	0,09	9

Medidas de desempenho dos modelos

Erro Médio (Mean error-ME): $ME=\frac{\sum_{i=1}^{n} y_i-\hat{y}_i}{n}$

Erro Médio Absoluto (Mean Absolut Error-MAE): $MAE=\frac{\sum_{i=1}^{n} |y_i-\hat{y}_i|}{n}$

Raiz do Erro Quadrático Médio (Root Mean Squared Error-RMSE): $RMSE=\sqrt{\frac{\sum_{i=}^{n}(y_i-\hat{y}_i)^2}{n}}$

Erro Percentual Médio (Mean Percent Error-MPE): $MPE = \frac{\sum_{i=1}^{n} \frac{y_i - \hat{y}_i}{y_i * 100}}{n}$

Erro Percentual Absoluto Médio (Mean Absolut Percent Error-MAPE): $MAPE = \frac{\sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i * 100}}{n}$

Sendo:

y_i = variável resposta

$\hat{y}_i$ = previsão do modelo

Data Mining

aplicações práticas de Data Mining se podem ser categorizadas de acordo com a tarefa que se pretende resolver

Exercitando!!!!!

Base
Vendas

Exemplo

Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa.

Projeção para 2019

		Vendas	Budget
jan/19	2019	**1.512.274**	91000
fev/19	2019		154240
mar/19	2019		169702
abr/19	2019		185081
mai/19	2019		199683
jun/19	2019		229192
jul/19	2019		238403
ago/19	2019		247253
set/19	2019		311114
out/19	2019		320442
nov/19	2019		373507
dez/19	2019		336157

Modelo de regressão linear simples

```
> summary(modelo)

Call:
lm(formula = Vendas ~ Budget_Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-655330 -256271  -30444  234875  743028

Coefficients:
                    Estimate  Std. Error t value       Pr(>|t|)
(Intercept)       1060550.396  151771.308    6.99 0.000000046312 ***
Budget_Advertising        4.964       0.524    9.47 0.000000000046 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 350000 on 34 degrees of freedom
Multiple R-squared:  0.725,     Adjusted R-squared:  0.717
F-statistic: 89.7 on 1 and 34 DF,  p-value: 0.0000000000458
```

Vendas janeiro/19 = 1060550 + 4.964 * 91000 = 1.512.274

```
> modelo <- lm(df$Vendas ~ df$Budget_Advertising)
> summary(modelo)

Call:
lm(formula = df$Vendas ~ df$Budget_Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-655330 -256271  -30444  234875  743028

Coefficients:
                          Estimate  Std. Error t value        Pr(>|t|)
(Intercept)           1060550.396  151771.308   6.988 0.0000000463123 ***
df$Budget_Advertising       4.964       0.524   9.473 0.0000000000458 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 349600 on 34 degrees of freedom
Multiple R-squared:  0.7252,    Adjusted R-squared:  0.7172
F-statistic: 89.75 on 1 and 34 DF,  p-value: 0.00000000004575
```

Resultado da ANOVA

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

Soma de quadrados total	Soma de quadrados residual	Soma de quadrados regressão
SQTot	SQRes	SQReg

Exemplo: y=Vendas e x =budget

```
# Análise de variância (ANOVA)
df$predito = 1060550.396 + 4.964*df$Budget_Advertising
df$residuo = df$Vendas - df$predito
df$residuo2 = df$residuo*df$residuo

df$reg = (df$predito - ybarra)*(df$predito - ybarra)

# Soma dos quadrados
sqreg = sum(df$reg) ; sqreg
sqres = sum(df$residuo2) ; sqres
sqtotal = sqreg+sqres

rquadrado = sqreg/sqtotal ; rquadrado
```

```
> # Soma dos quadrados
> sqreg = sum(df$reg) ; sqreg
[1] 10966753022839
> sqres = sum(df$residuo2) ; sqres
[1] 4154940166057
> sqtotal = sqreg+sqres
>
> rquadrado = sqreg/sqtotal ; rquadrado
[1] 0.7252331
>
```

Teste de Hipóteses

TESTANDO OS PARÂMETROS B'S

H_o: $B_i = 0$

H_1: $B_i \neq 0$

$$t = \frac{B_i}{erro\_padrao(B_i)}$$ com gl = n - p

Quando $t > t_{\alpha/2} \Rightarrow$ região de rejeição

$$IC : \overline{b}_i + t_{\alpha/2} Sb_i$$

Machine Learning

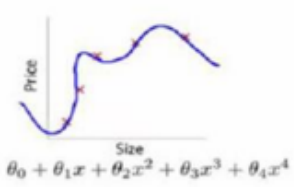
- Baseado em algorítmos
- O objetivo é identificar o que funciona
- Interessa em prever os resultados de amostras futuras
- Foco na praticidade➔ Desenvolve em uma amostra e aplica em outra
- Algoritmos para tomada de decisão
- Limitações/grandes desafios:
  - Tendência ao sobre ajuste
  - Dados influenciados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

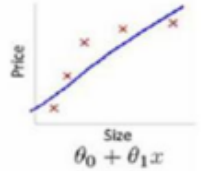
Machine Learning

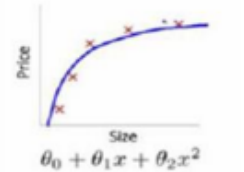
➢ O que é sobreajuste (overfitting) ?

Problema: em geral esses modelos têm sobreajuste 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, causando por sua vez, overfitting. Nesse caso o MSE é zero para os dados de teste, mas podemos dizer que não.







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

Machine Learning

➢ Como evitar o sobreajuste (overfitting) ?

Validação Cruzada

Dividir os dados em partes iguais e utilizar:

- Uma fração delas para treinar o algoritmo com um hiperparâmetro;
- Outra parte testar a sua predição.

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.

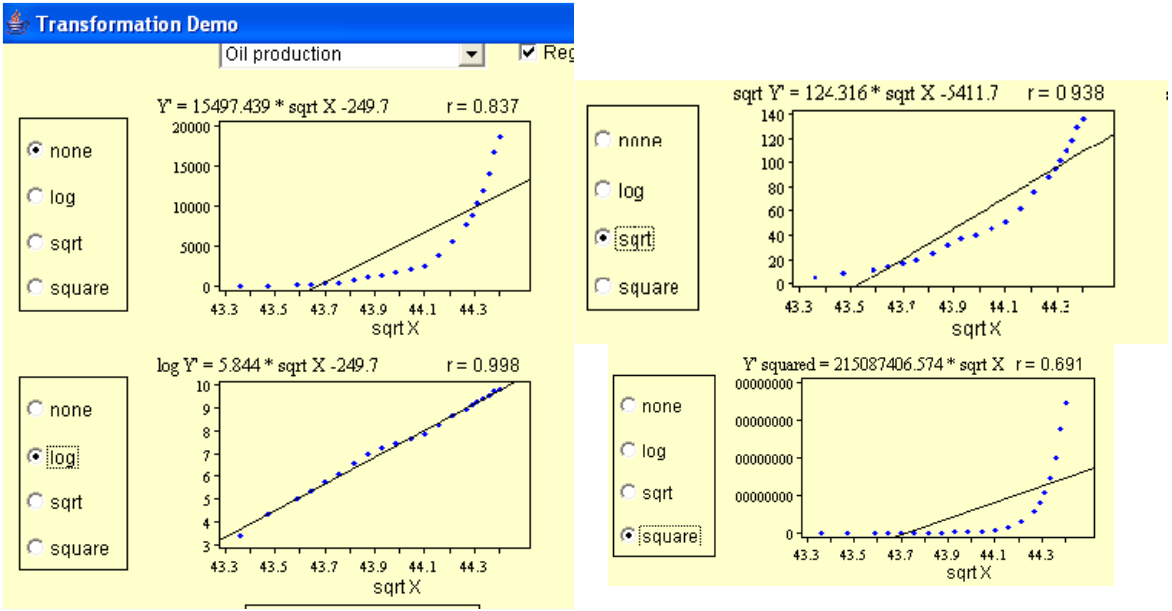
Regressão Linear

Pontos de atenção	Análise	O que fazer
Colinearidade: Correlação entre duas variáveis preditoras do modelo	Correlação de Pearson	- Escolher uma das variáveis ou - Criar a variável de iteração
Multicolinearidade: Qualquer variável preditora é correlacionada com um conjunto de outras variáveis preditoras	Fator de inflação da variância (VIF)	- Escolher uma das variáveis ou - Criar fatores usando a análise de componentes principais
Relação não linear entre a variável resposta e a preditora	Gráfico de dispersão	- Transformar a variável
Outliers	Resíduos padronizados	- Excluir os outliers da base de dados a cada nova rodada

Transformação de Variáveis

Quando o modelo não é conhecido, pode-se escolher a transformação examinando o gráfico x e y.

Exemplos:



Fonte: http://onlinestatbook.com/stat_sim/transformations/index.html

Normalização dos Dados

- Distribuição Normal Padronizada

$$X \sim N(\mu, \sigma^2) \Rightarrow \boxed{Z = \frac{X - \mu}{\sigma}} \Rightarrow Z \sim N(0,1)$$

- Máximo e Mínimo

$$X_p = \frac{X - X_{mínimo}}{X_{máximo} - X_{mínimo}}$$

Exemplos de aplicações da normalização dos dados:

Convolutional Neural Networks (CNNs) e Algoritmos de Machine Learning (Regressão, SVM, Cluster e outros)

Data Mining

aplicações práticas de Data Mining se podem ser categorizadas de acordo com a tarefa que se pretende resolver

Exemplo

Predição de tráfego

FIAP MBA+

Exemplo

Predição de Tráfego – Por que?

Pode demorar de 10 a 20 min para ter uma máquina no ar. Dá pra esperar tudo isso?

Evite falsas quedas de tráfego

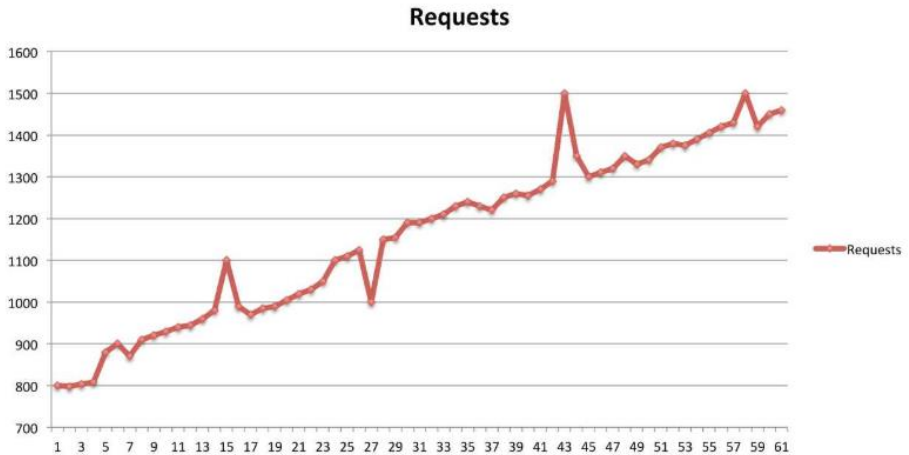
Fonte:https://www.infoq.com/br/presentations/data-science-em-publicidade-digital

Técnica de Regressão: Regressão Linear Simples

FIAP MBA+

Exemplo

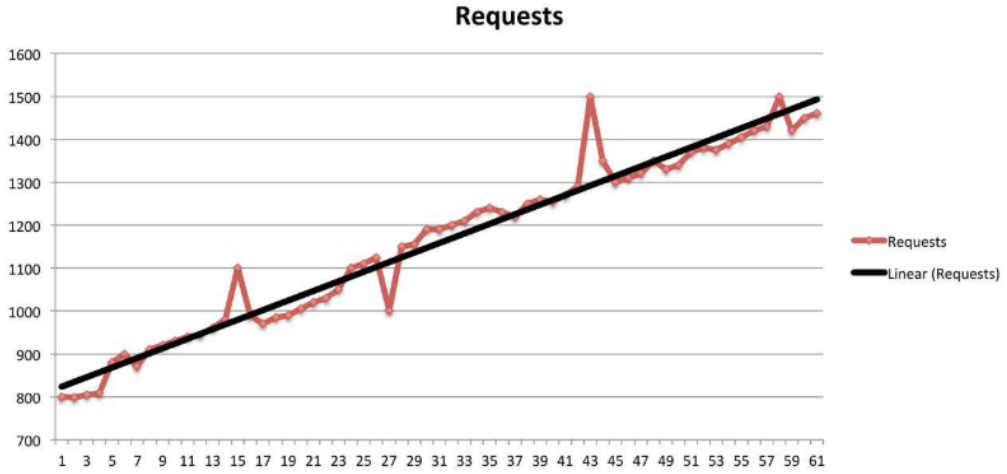
Predição de Tráfego



Fonte:https://www.infoq.com/br/presentations/data-science-em-publicidade-digital

Técnica de Regressão: Regressão Linear Simples

Predição de Tráfego



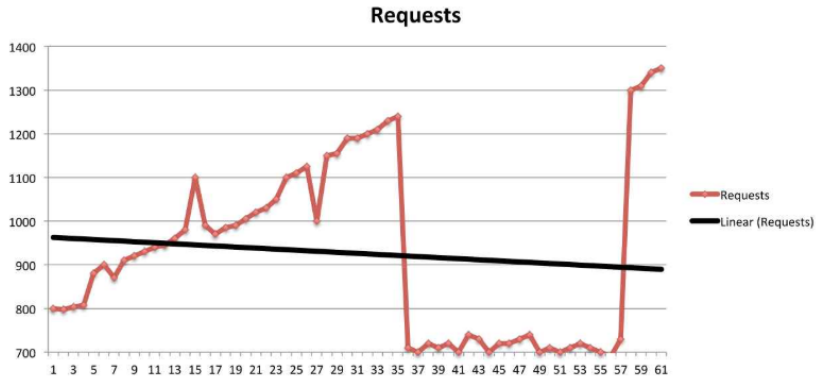
Fonte:https://www.infoq.com/br/presentations/data-science-em-publicidade-digital

Técnica de Regressão: Regressão Linear Simples

FIAP MBA+

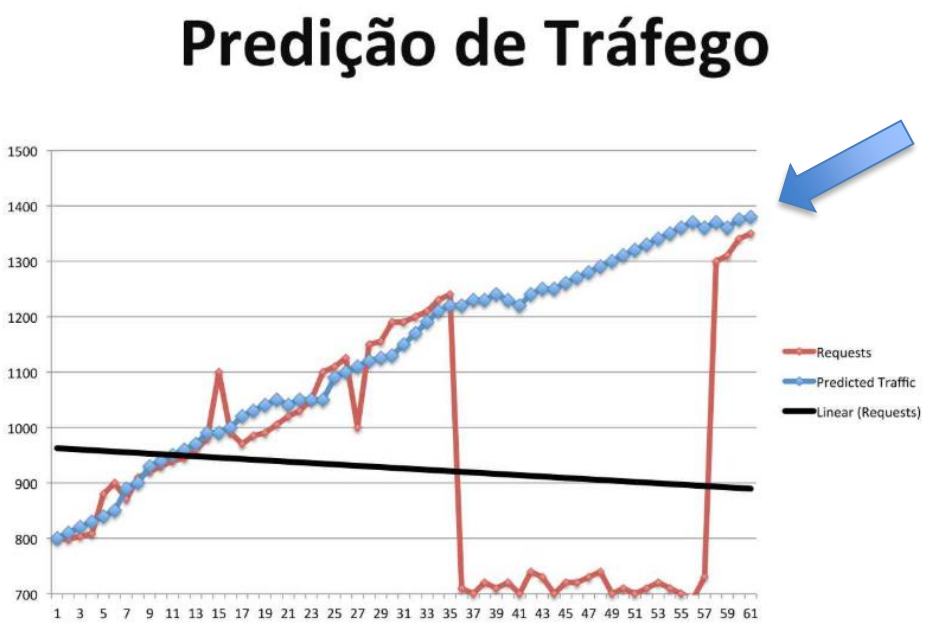
Exemplo

Predição de Tráfego



[Fonte:https://www.infoq.com/br/presentations/data-science-em-publicidade-digital](https://www.infoq.com/br/presentations/data-science-em-publicidade-digital)

Técnica de Regressão: Regressão Linear Simples



Aplicação da Regressão linear para imputação de dados

[Fonte:https://www.infoq.com/br/presentations/data-science-em-publicidade-digital](https://www.infoq.com/br/presentations/data-science-em-publicidade-digital)

Data Mining

aplicações práticas de Data Mining se podem ser categorizadas de acordo com a tarefa que se pretende resolver

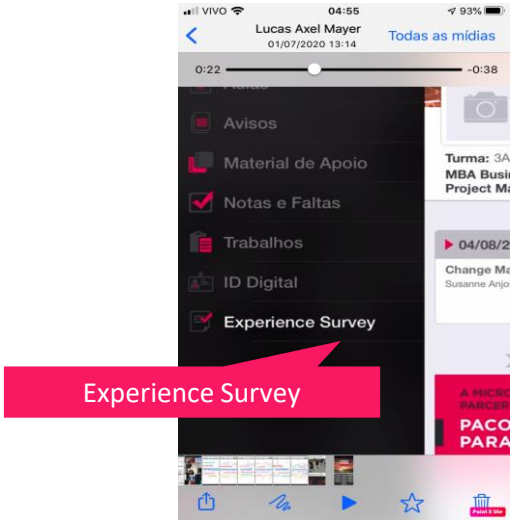
Exercitando!!!!!

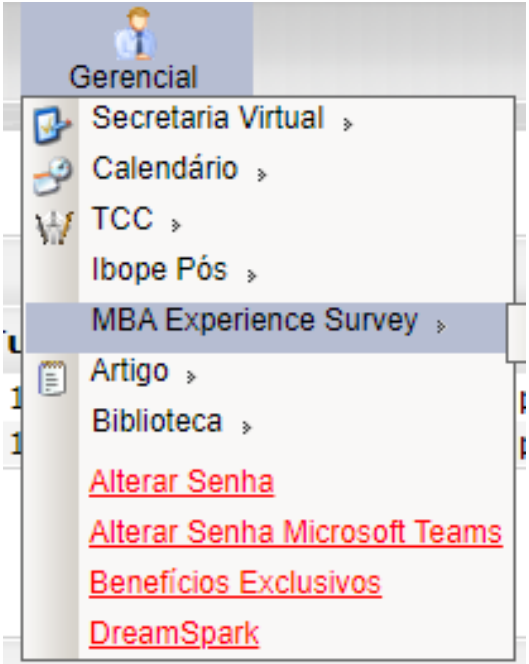
Base
Bike Sharing

O que você achou da aula de hoje?

Pelo aplicativo da FIAP

(Entrar no FIAPP, e no menu clicar em Experience Survey)





/ Regina T. I. Bernal

FIAP


FIAP