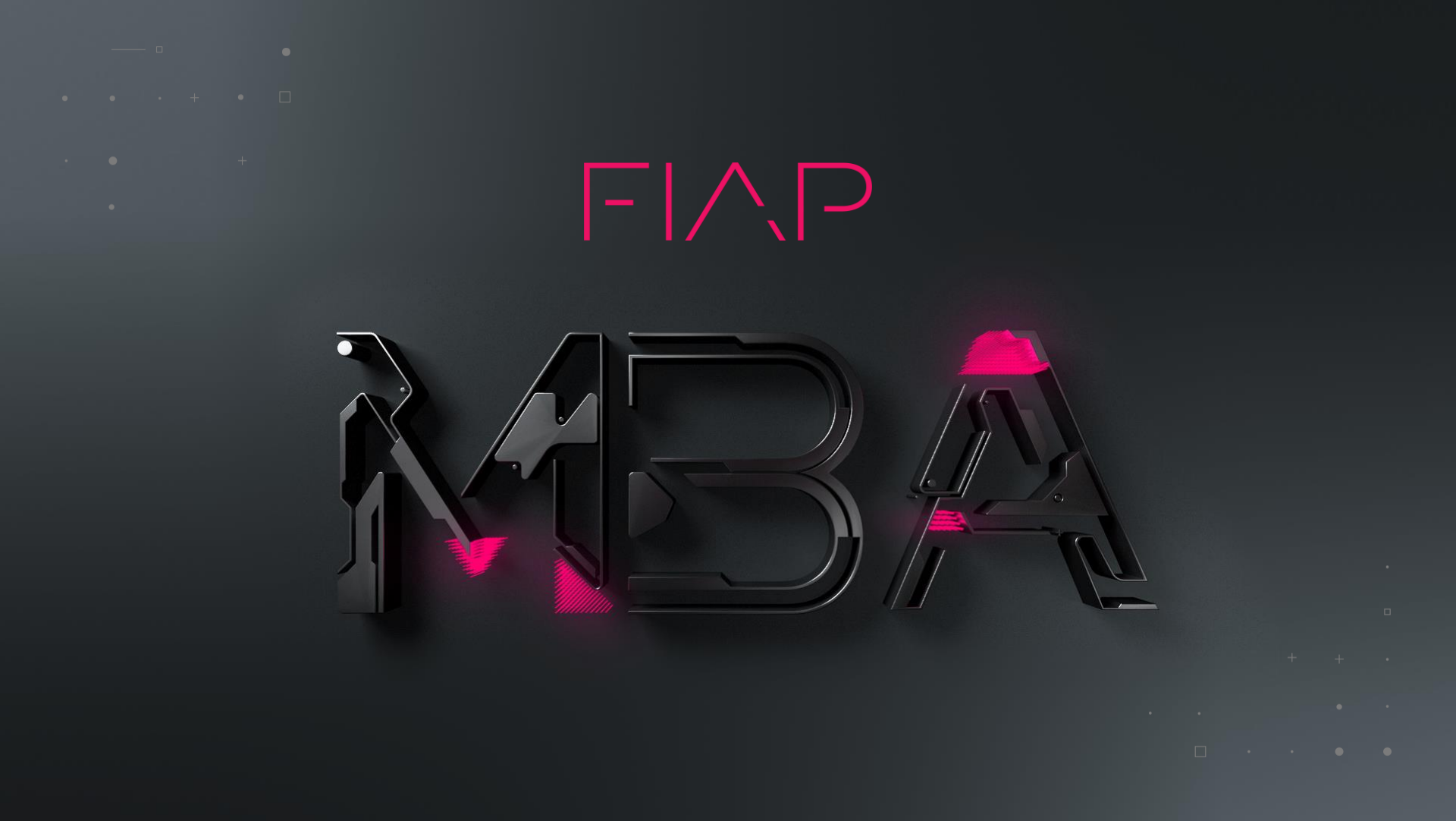

FIAP
MBA

BUSINESS INTELLIGENCE & ANALYTICS

Data Mining & Prescriptive Analytics

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

profadelaide.alves@fiap.com.br

ADELAIDE ALVES DE OLIVEIRA
PROFESSORA

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

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

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

PONTOS DA DISCIPLINA

1	**AULA 1**	INTRODUÇÃO – Teoria + Exemplos + Atividades
2	**AULA 2**	TÉCNICAS SUPERVISIONADAS – Predição e Estimação
3	**AULA 3**	TÉCNICAS SUPERVISIONADAS – Classificação
4	**AULA 4**	TÉCNICAS NÃO SUPERVISIONADAS – Ferramentas e Aplicabilidades + Cases e Exemplos de Sucesso + Atividades
5	**AULA 5**	CONTEÚDO – Ferramentas e Aplicabilidades + Cases e Exemplos de Sucesso + Atividades

INTRODUÇÃO À MINERAÇÃO DE DADOS
E MODELOS PREDITIVOS

INTRODUÇÃO

“ O Índice de Churn da nossa empresa está em 12%”

“A taxa de conversão de leads é de 4%”

DADOS OU INFORMAÇÕES?

INTRODUÇÃO

...enfim, seus dados não servem para nada até que você saiba como tirar informações deles.

Conhecimento

Informação Estruturada e Contextualizada

Conjunto formado por experiências, valores e criatividade aplicado aos dados e informações

Informação

Dado com significado

Dados analisados com ajuda de ferramentas específicas, de modo a que sejam úteis para a tomada de decisão

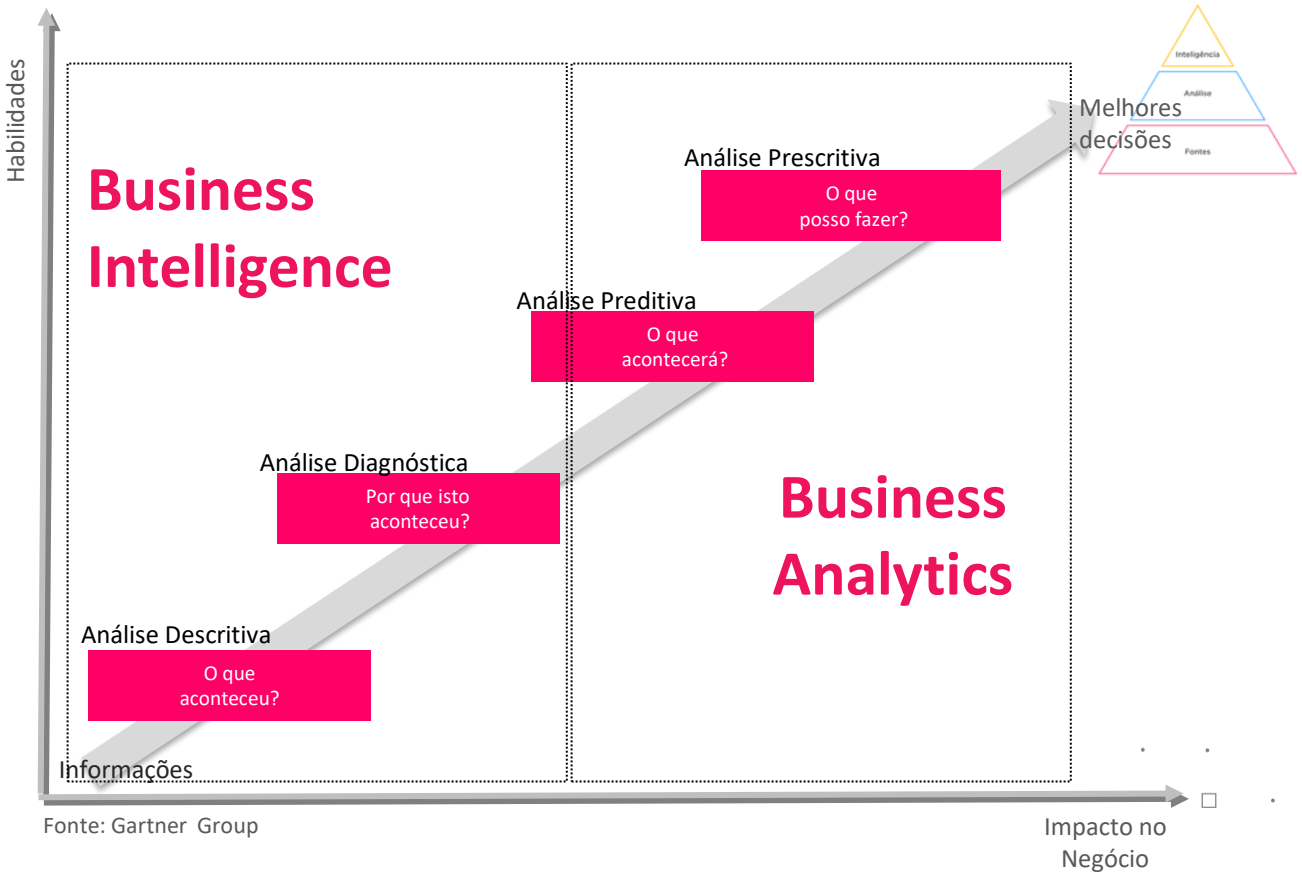
Dado

Valor sem significado

Fatos, números e estatísticas

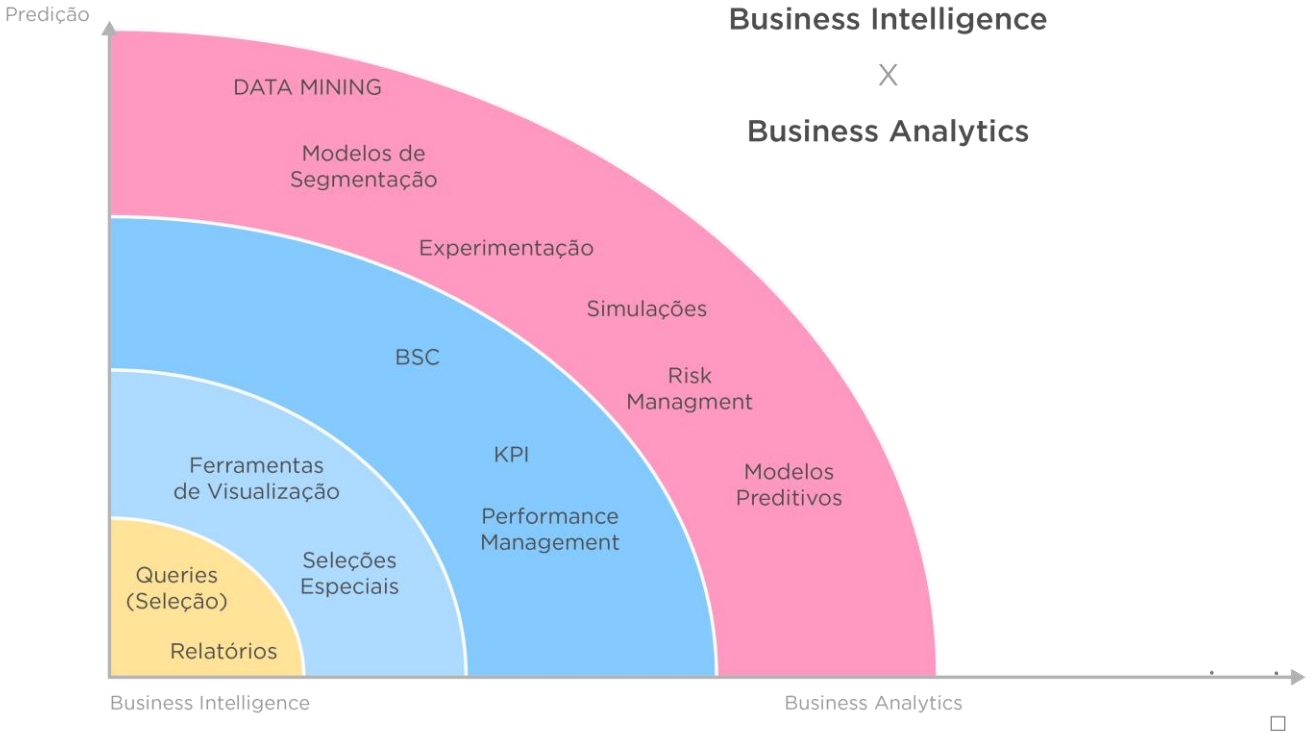
Ajudar o gestor a diminuir os riscos e aumentar as chances de sucesso!

INTRODUÇÃO



Fonte: Gartner Group

INTRODUÇÃO


Predição
Business Intelligence
X
Business Analytics
DATA MINING
Modelos de Segmentação
Experimentação
Simulações
BSC
Risk Managment
KPI
Ferramentas de Visualização
Performance Management
Modelos Preditivos
Seleções Especiais
Queries (Seleção)
Relatórios
Business Intelligence
Business Analytics

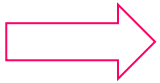
INTRODUÇÃO

... enfim, seus dados não servem para nada até que você saiba como tirar informações deles.

O que aconteceu?	DESCRITIVO	Quantos cancelamentos? Quantos clientes novos, quantos antigos? Qual região? Que tipo de cliente?
Por que isto aconteceu?	DIAGNÓSTICO	Qual a relação entre cancelamento voluntário por tempo de conta e tipo de cliente?
O que acontecerá?	PREDITIVO	Qual a probabilidade de um cliente cancelar o serviço para em uma certa região nos próximos 3 meses?
O que posso fazer?	PRESCRITIVO	Lista de ações para reter os clientes, segundo seu valor?

MODELOS ESTATÍSTICOS – PARA QUÊ?

Entendimento das estruturas dos dados



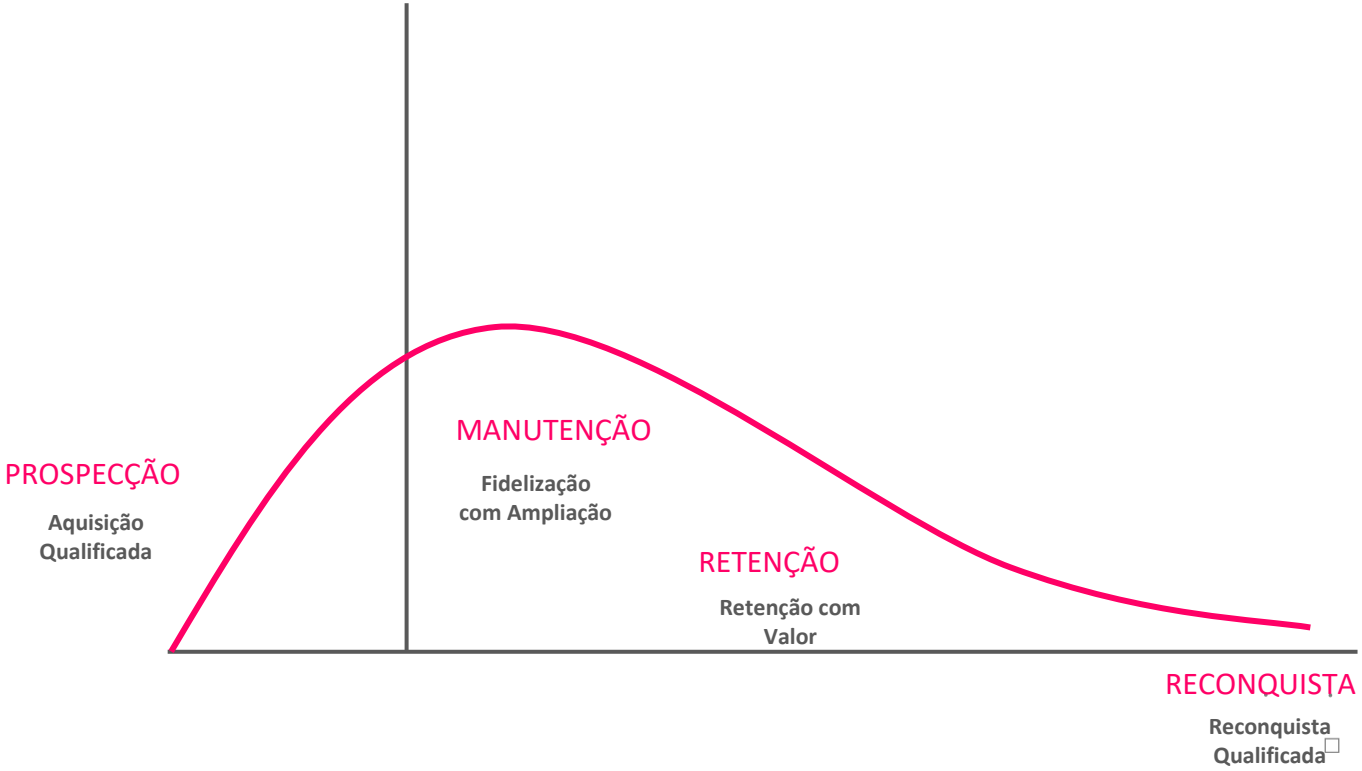
Reconhecimento do Cliente e Mercado

Definição das Estratégias

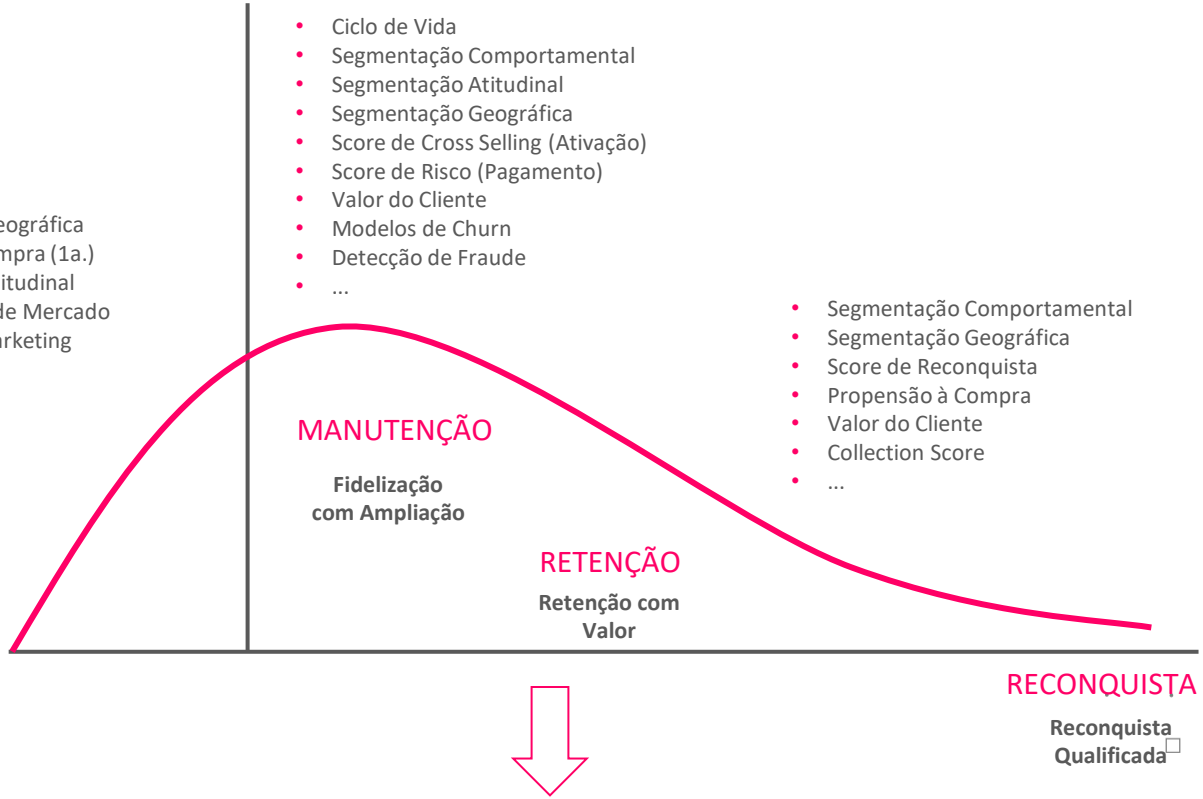
Predição da Resposta do Cliente/Observação

Estabelecimento das Ações

USO DOS MODELOS NO CICLO DO CLIENTE

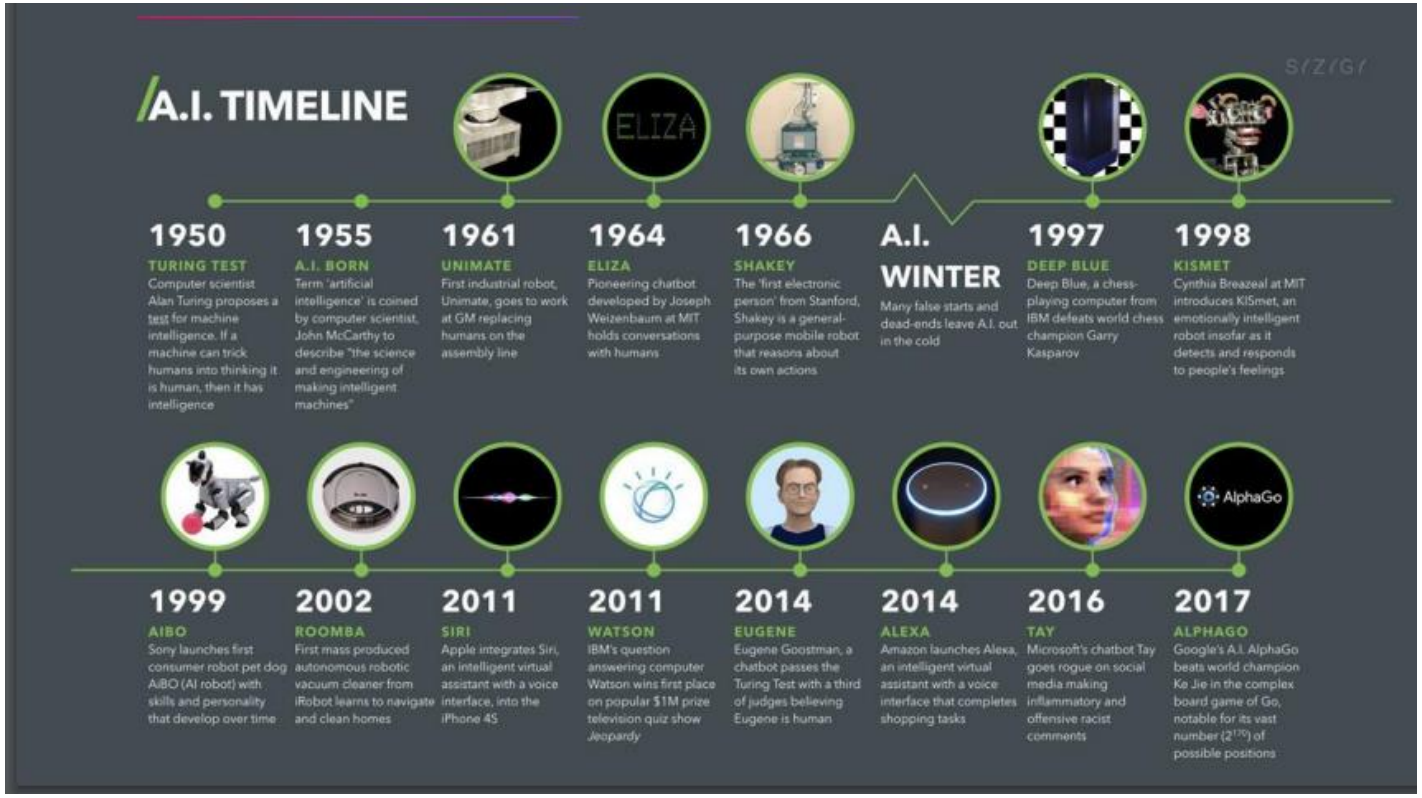


USO DOS MODELOS NO CICLO DO CLIENTE

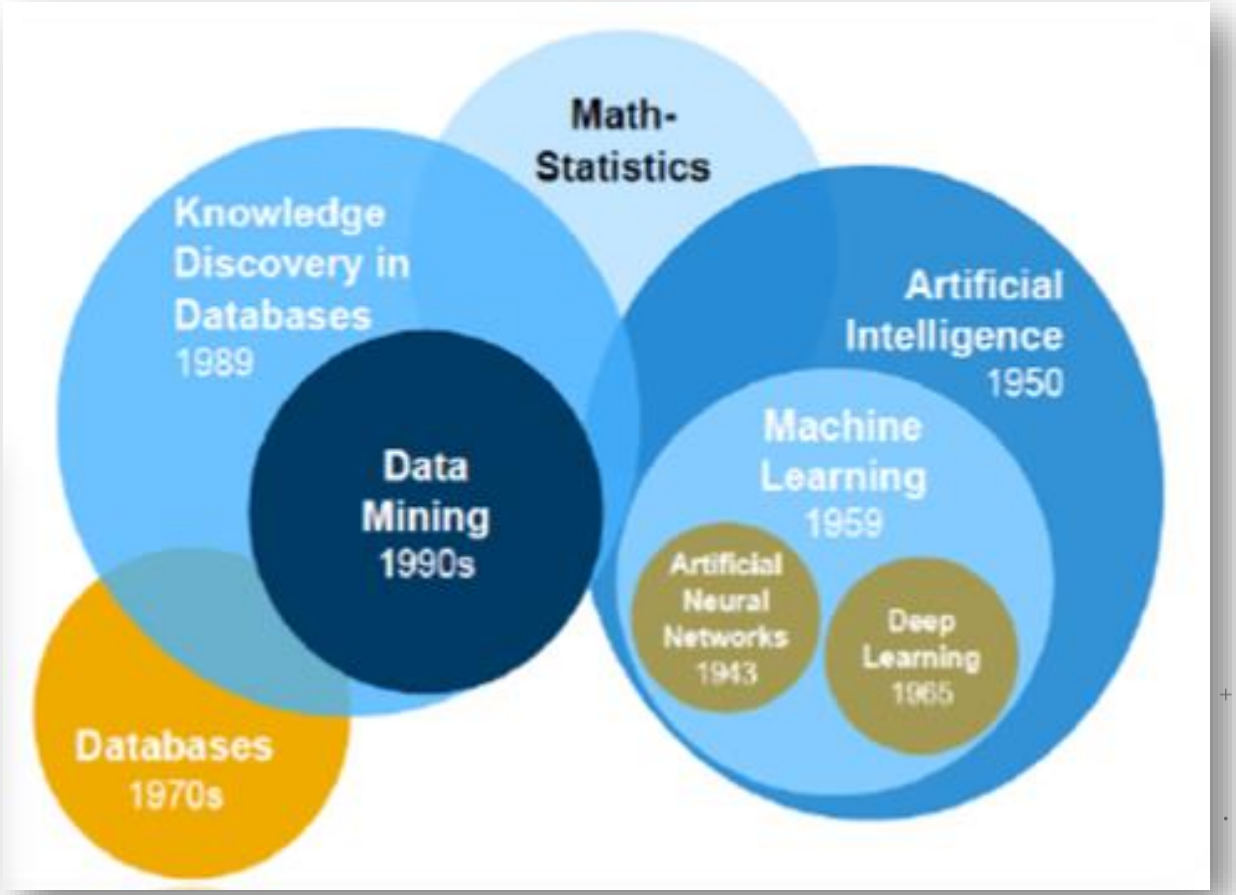


INTRODUÇÃO À MACHINE LEARNING

INTRODUÇÃO: HISTÓRICO

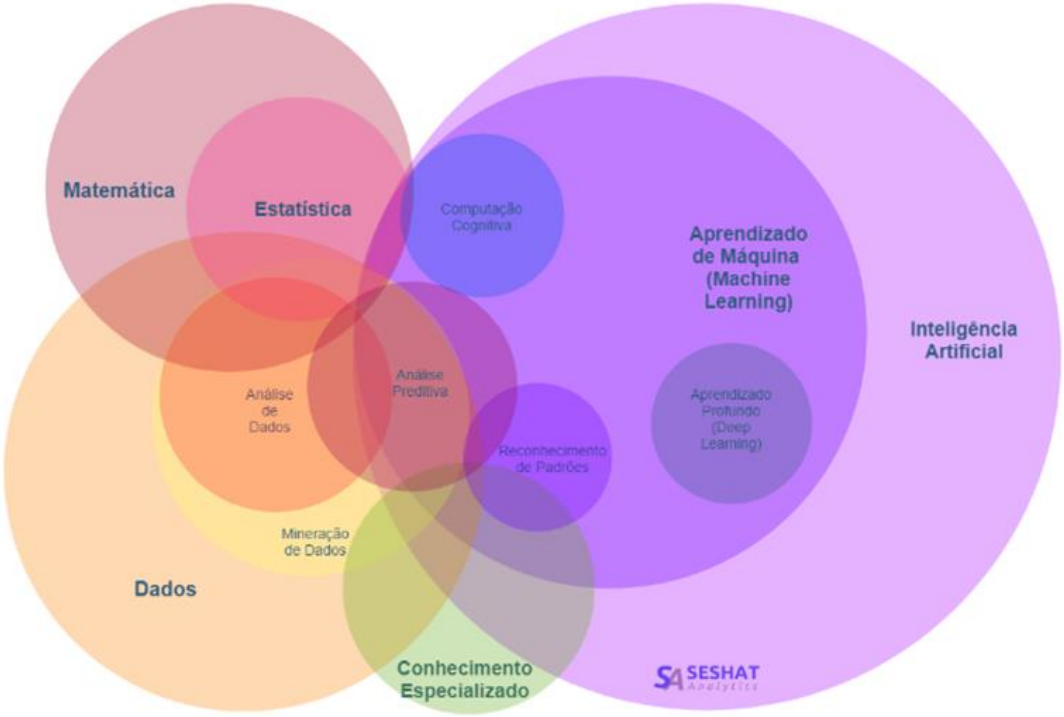


INTRODUÇÃO

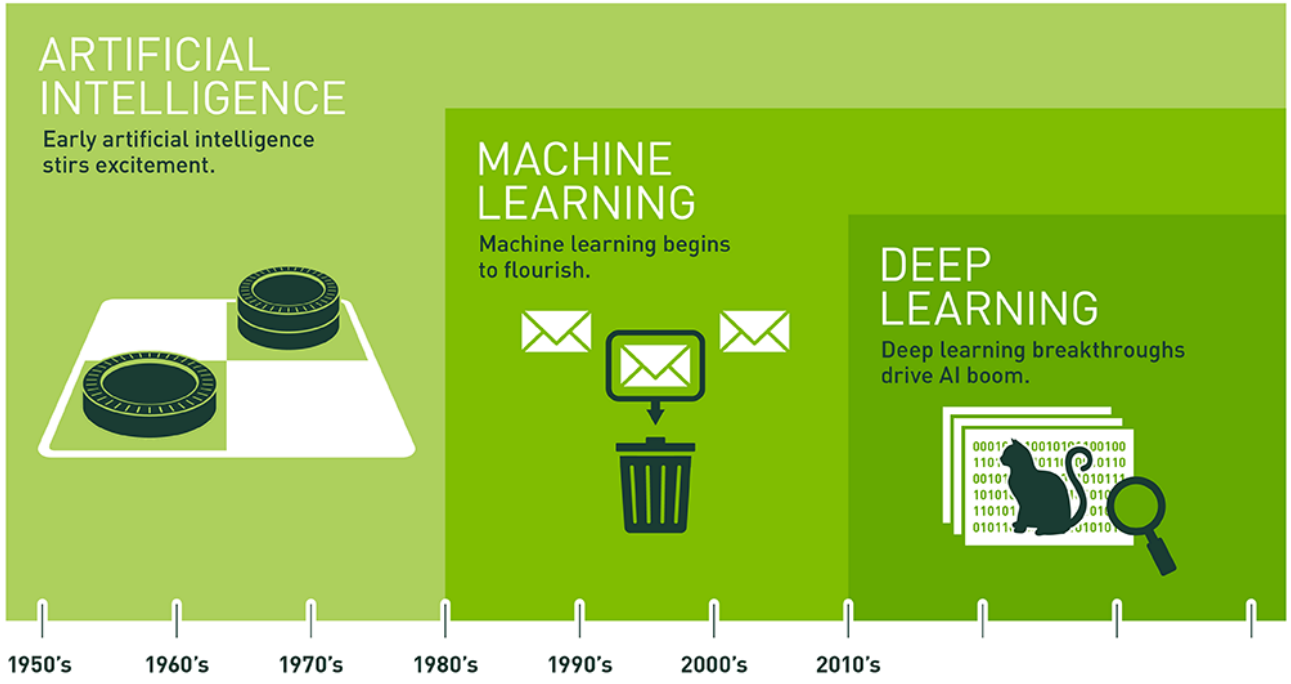


- SPSS: 1968
- SAS: 1976
- R: ~1990
- Python: ~1990

INTRODUÇÃO


Matemática
Estatística
Computação Cognitiva
Aprendizado de Máquina (Machine Learning)
Inteligência Artificial
Análise de Dados
Análise Preditiva
Aprendizado Profundo (Deep Learning)
Reconhecimento de Padrões
Mineração de Dados
Dados
Conhecimento Especializado
SESHAT Analytics

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

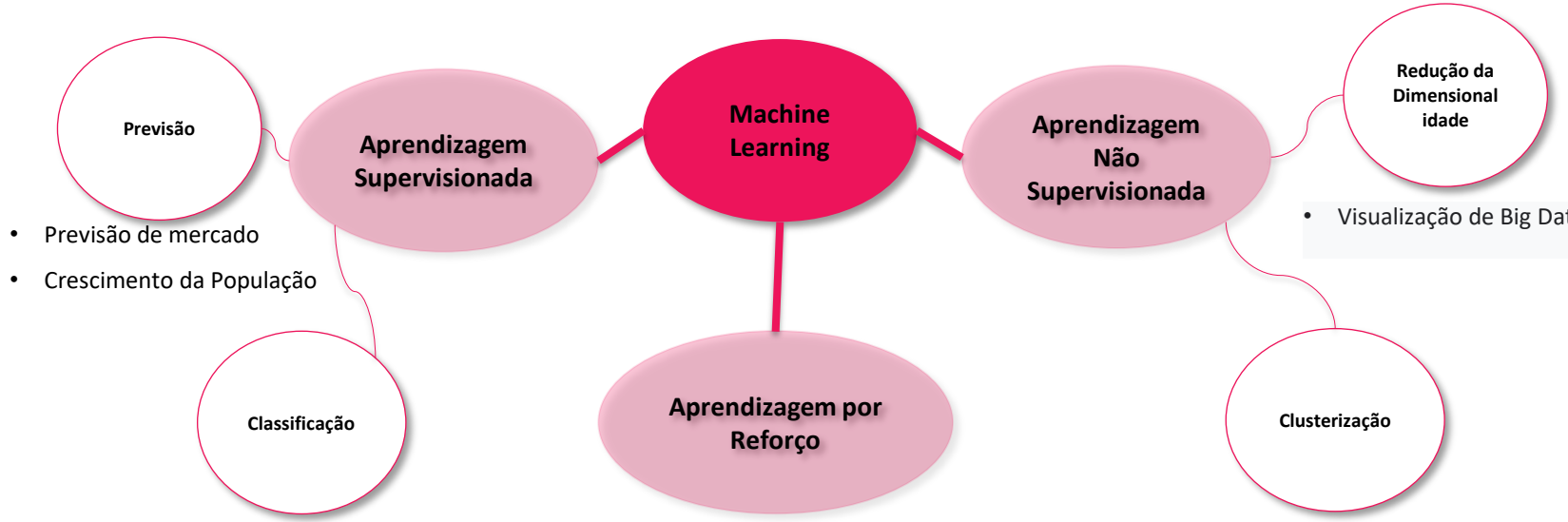
O Machine Learning (Aprendizado de Máquinas), uma linha de pesquisa dentro de Inteligência Artificial, tem por objetivo criar programas capazes de aprender uma determinada tarefa utilizando um conjunto de dados ou medida de desempenho.

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.

*O algoritmo de Machine Learning que é treinado para aprender a executar uma tarefa é conhecido como Modelo

ALGORITMOS de MACHINE LEARNING

ALGORITMOS de MACHINE LEARNING



- Previsão de mercado
- Crescimento da População

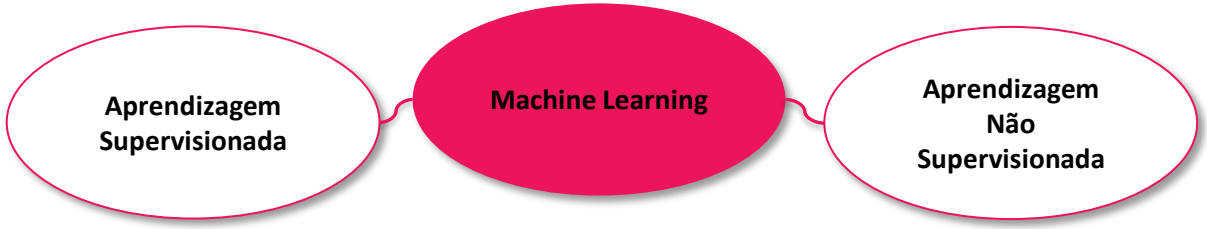
- Reconhecimento de Imagens
- Retenção de Clientes
- Modelos de Créditos
- Modelos de Fraude

- Navegação de robôs
- Tarefas de constante aprendizagem
- Aquisição de conhecimento
- Decisões em tempo real

- Segmentação de Clientes
- Sistemas de Recomendação

- Visualização de Big Data

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

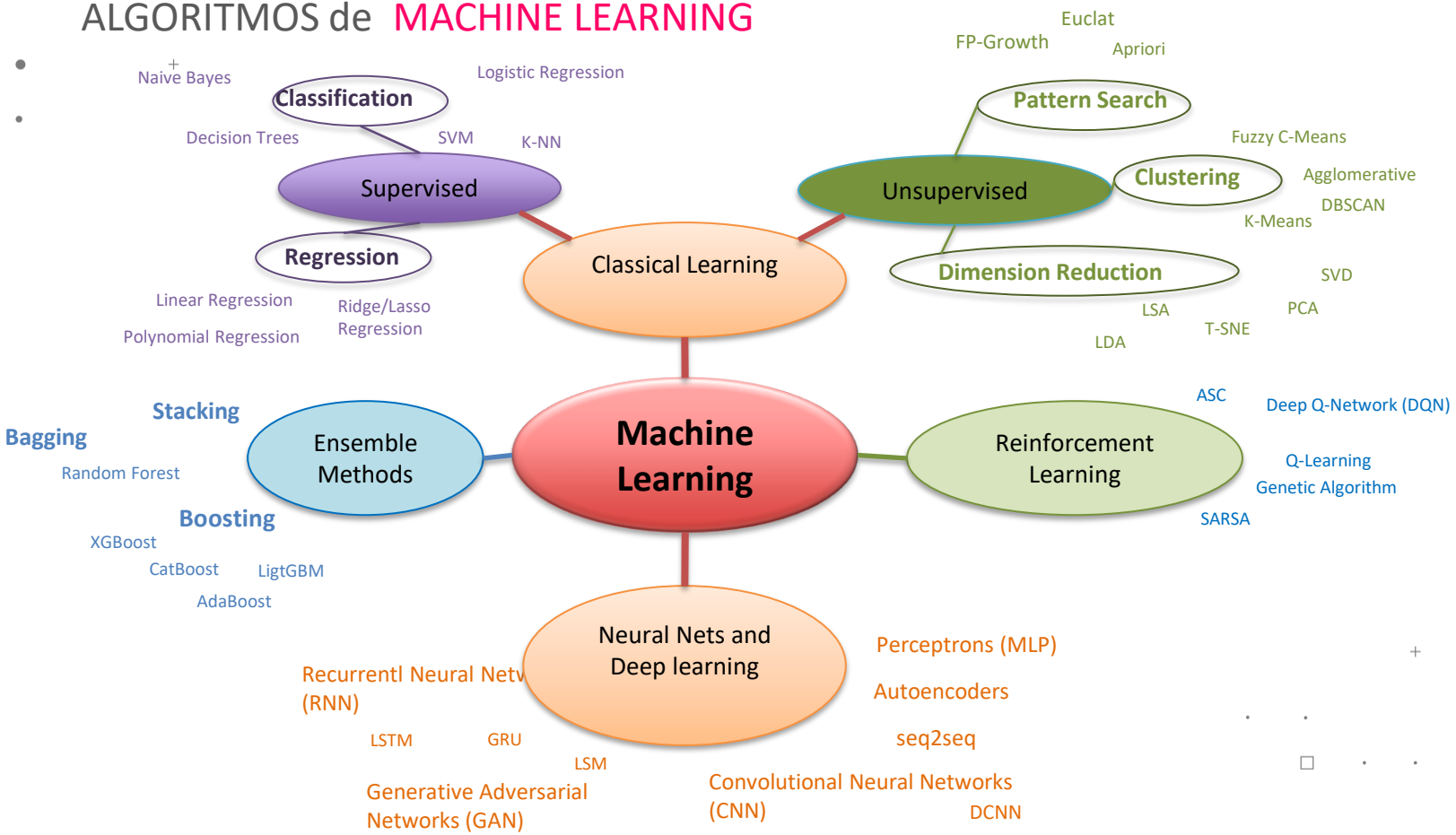
A Aprendizagem Por Reforço (ou *Reinforcement* Learning) é o treinamento de modelos de aprendizado de máquina para tomar uma sequência de decisões em um ambiente incerto e potencialmente complexo. No aprendizado por reforço, o sistema de inteligência artificial enfrenta uma situação. O computador utiliza 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.

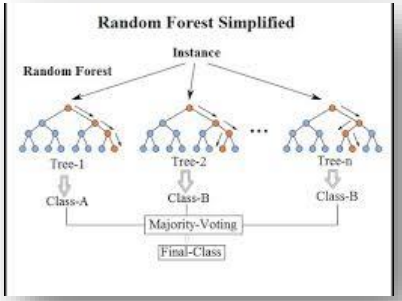
Exemplo: AlphaGo.

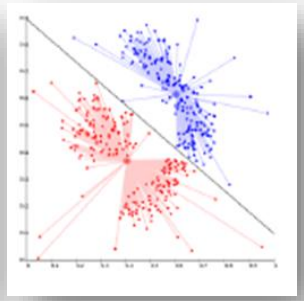
Fonte: https://www.deeplearningbook.com.br/

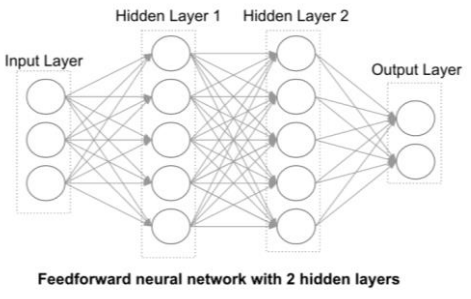
ALGORITMOS de MACHINE LEARNING



- 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







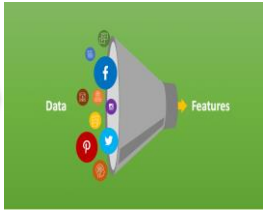
Se machine learning não se importa muito com interpretação, então se importa de fato com o quê?

➔ Performance preditiva (ou seja, acurácia das decisões)

CICLO ANALÍTICO

Entender o problema de Negócio

Coletar DADOS



Explorar/ Visualizar

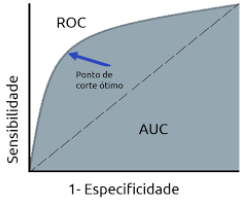
Feature Engineering

Preparar os dados

FeatureExtraction/ Selection

Machine Learning

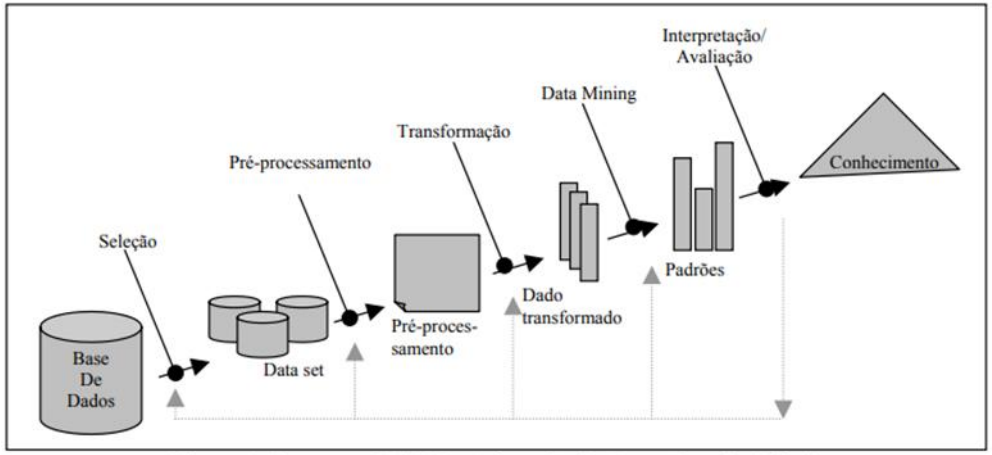
Validação / Monitoramento



Deploy / Implementar

PROCESSO KDD

KNOWLEDGE DISCOVERY IN DATABASES



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

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 nu 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.

INTRODUÇÃO

Componentes

Marketing

Decisão estratégica.

Implantação da ação.

Tecnologia

Infra estrutura.

Ferramentas.

Gerenciamento das informações.

Dados

Informações do consumidor.

Estatística /Algoritmos

Análises que suportam a tomada de decisão.

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
- ...

DATA MINING

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

VARIÁVEIS INDEPENDENTES e DEPENTENTES

Variável dependente

Mede o **fenômeno que se estuda e que se quer explicar.** São aquelas cujos efeitos são esperados de acordo com as causas. Elas se situam, habitualmente, no fim do processo causal e são sempre definidas na hipótese ou na questão de pesquisa.

Variável independente

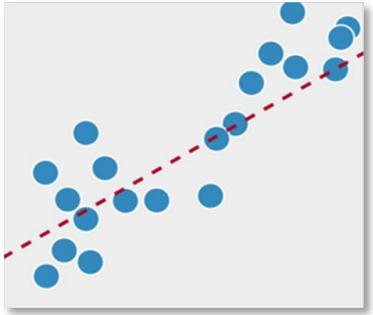
São aquelas variáveis **candidatas a explicar a(s) variável(eis) dependente(s),** cujos efeitos queremos medir. Aqui devemos ter cuidado, pois mesmo encontrando relação entre as variáveis isto, não necessariamente, significa relação causal.

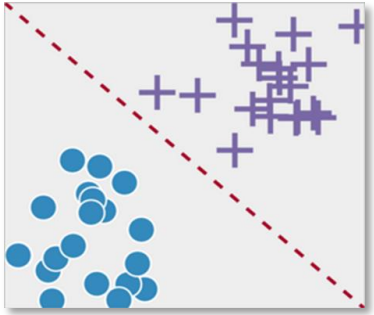
DATA MINING: MINERAÇÃO – CONSTRUÇÃO DE MODELOS

Aplicações práticas de Data Mining 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.

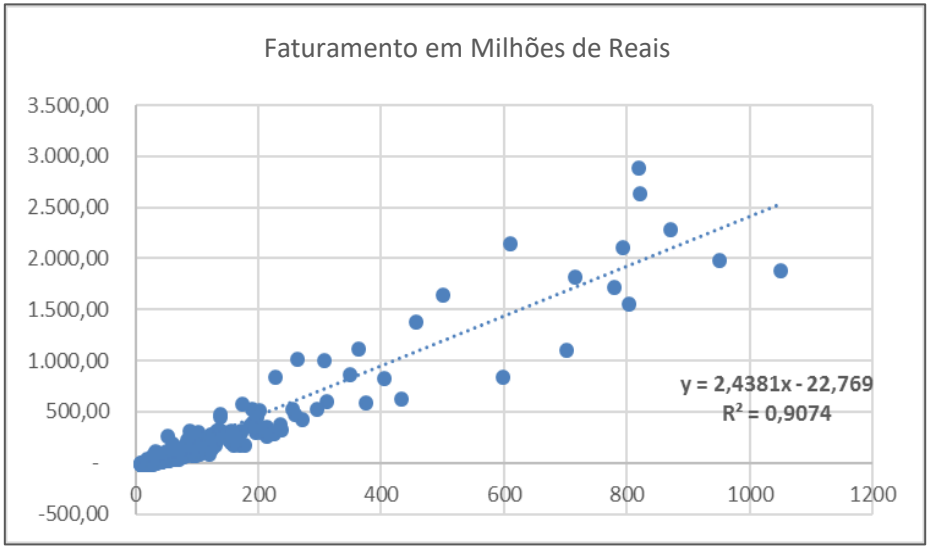




Fonte: Ronaldo Goldschmidt e Eduardo Bezerra - DataMining: conceitos, técnicas, algoritmos, orientações e aplicações. Editora Elsevier.

TÉCNICA DE REGRESSÃO:
REGRESSÃO LINEAR SIMPLES

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



Fonte:ABRAS

TÉCNICA DE REGRESSÃO:
REGRESSÃO LINEAR MÚLTIPLA

Estimar o valor de imóveis a partir de suas características

Trecho do arquivo de dados

Ordem	ValorR$	Áream2	IA	Andar	Suítes	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

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;

Suítes [Suites]: Número de suítes;

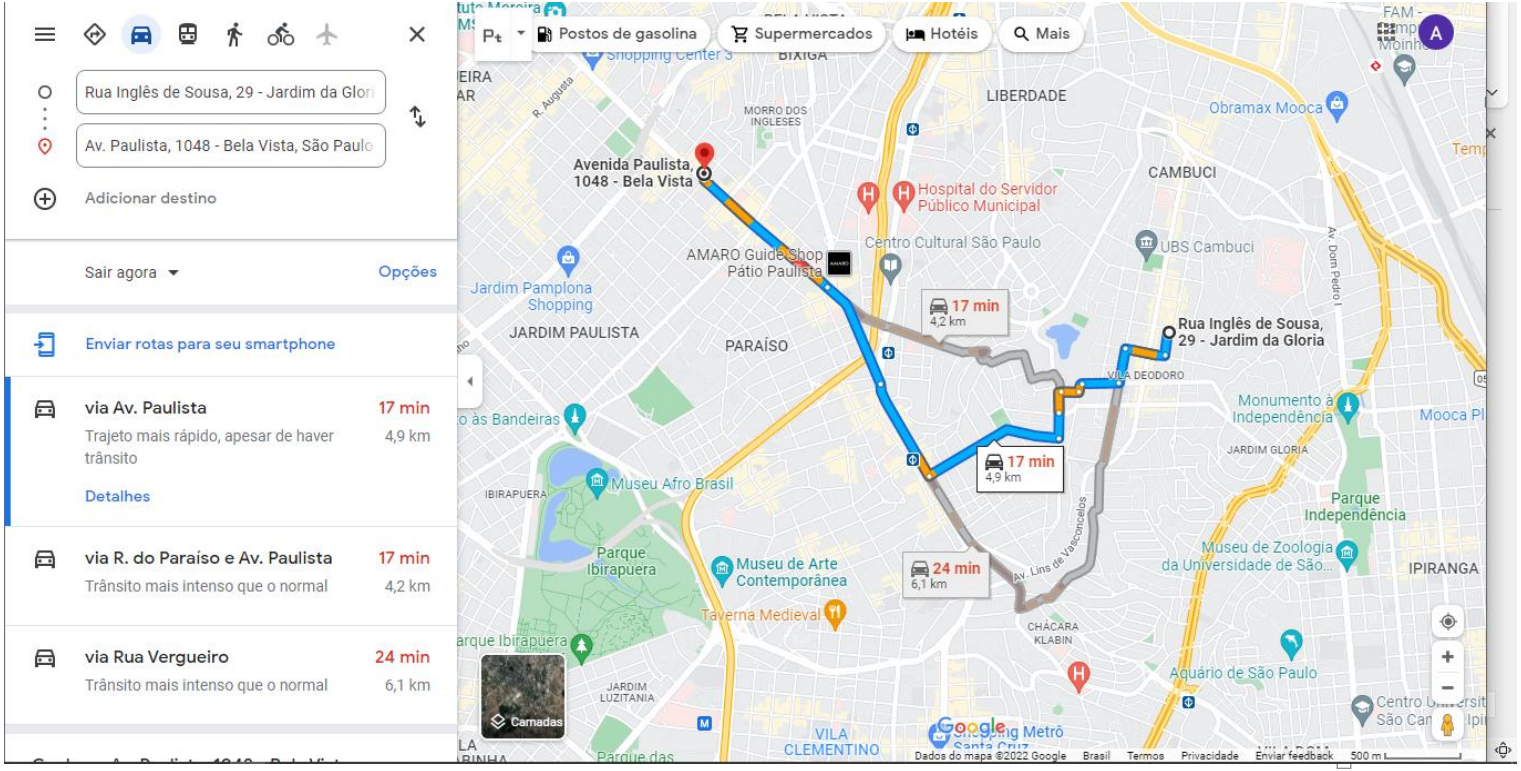
Vista Panorâmica [Vista]: A variável ambiental vista panorâmica é uma variável dicotômica: se o apartamento tiver vista panorâmica a variável vista assume valor igual a 1, se não tiver vista seu valor será 0;

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;

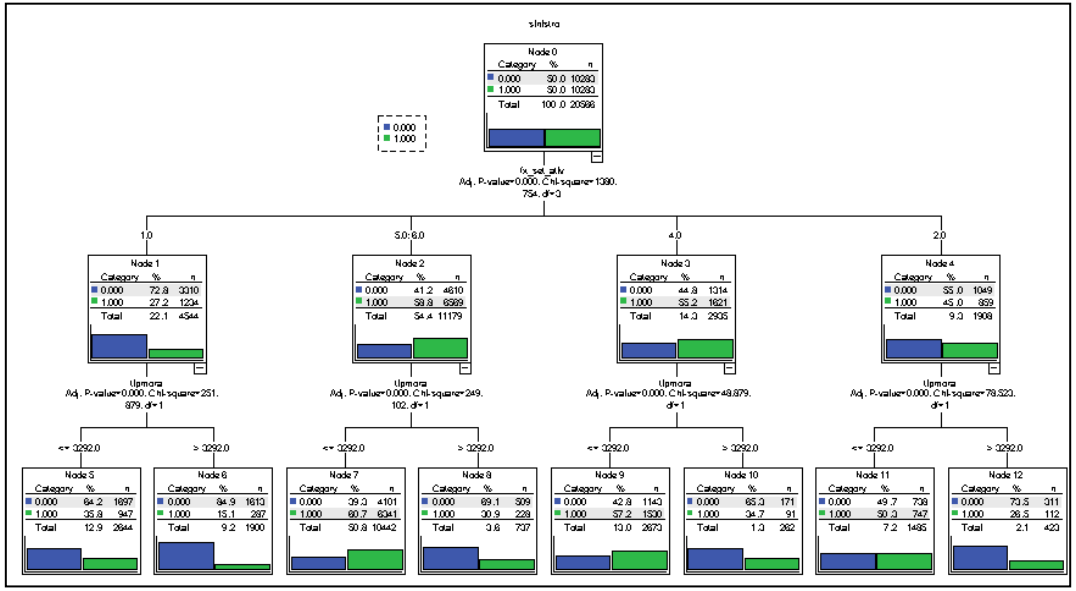
Área Verde a uma distância de 200 metros [AV 200m]: Área verde a uma distância de 200 metros assume valor igual a 1, ultrapassando 200 metros assume valor 0.

TÉCNICA DE REGRESSÃO:

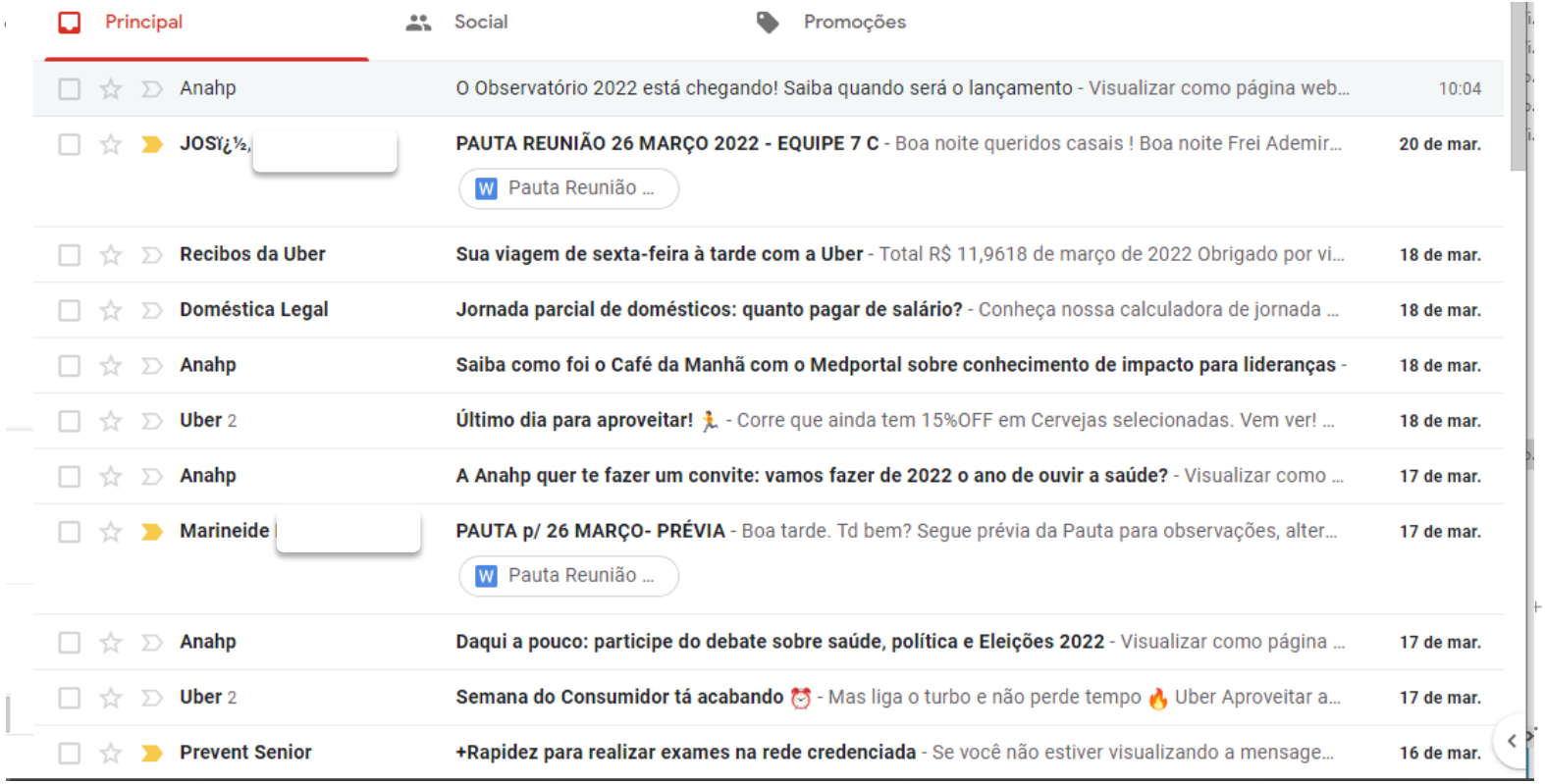


TÉCNICA DE CLASSIFICAÇÃO:
ÁRVORE DE DECISÃO

Exemplo de Modelo



TÉCNICA DE CLASSIFICAÇÃO:



TÉCNICA DE CLASSIFICAÇÃO:
REGRESSÃO LOGÍSTICA

Exemplo de 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

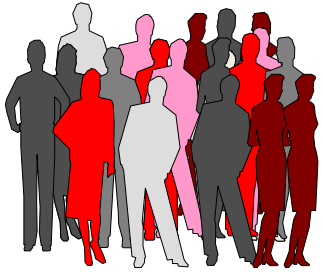
Clientes com o evento(*)

(*) Evento (exemplos)

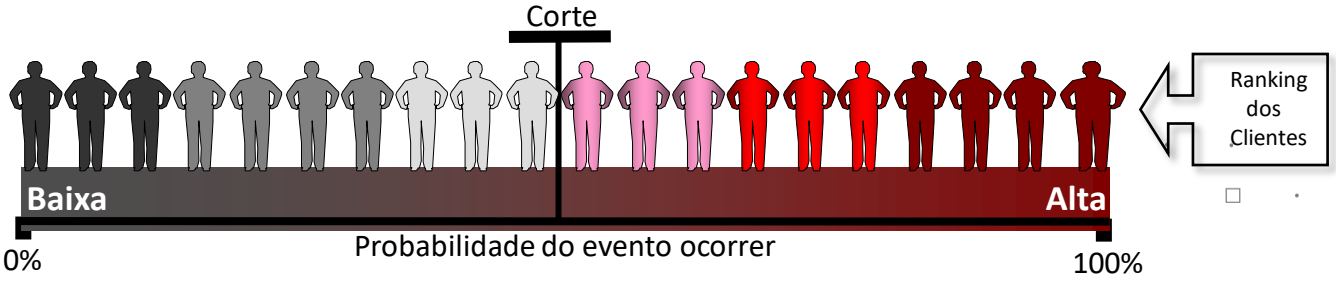
- Aquisição
- Cancelamento
- Pagamento

Modelo

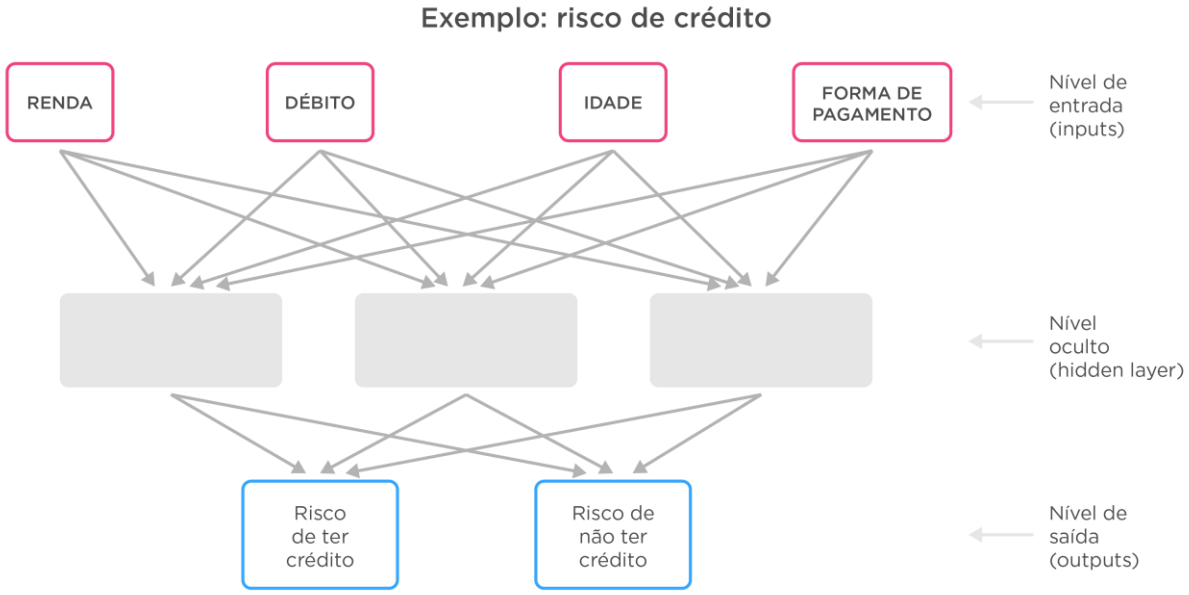
Pontuação dos Clientes com Probabilidade



Regra de Decisão



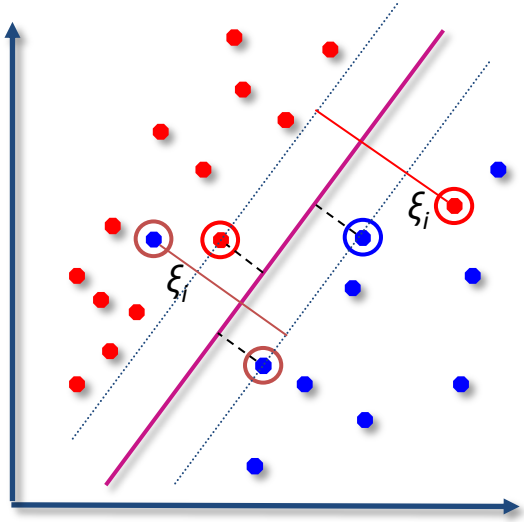
TÉCNICA DE CLASSIFICAÇÃO:
REDES NEURAIS

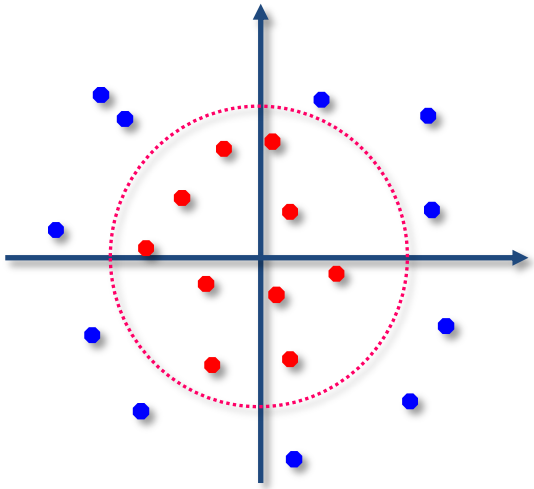


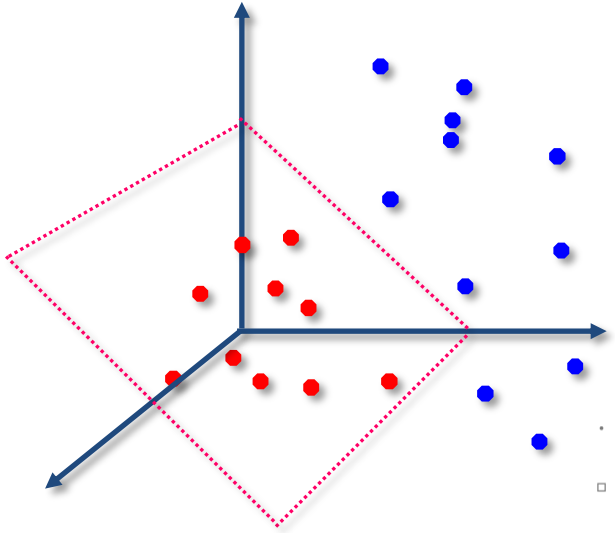
As redes neurais usam dados de entrada.
Atribui pesos nas conexões entre os atributos (neurônios).
E obtém um resultado (risco de ter ou não crédito) - nível de saída.

TÉCNICA DE CLASSIFICAÇÃO:

SVM – Support Vector Machine



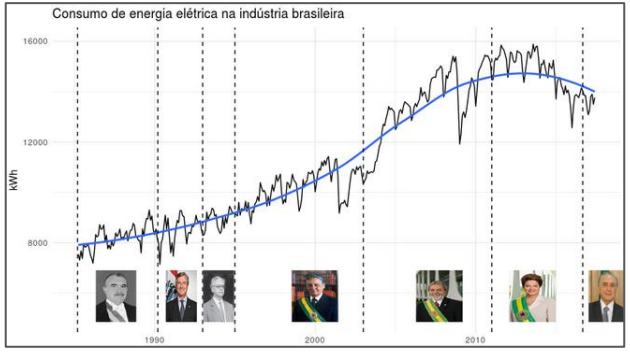




DATA MINING

Aplicações práticas de Data Mining 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.

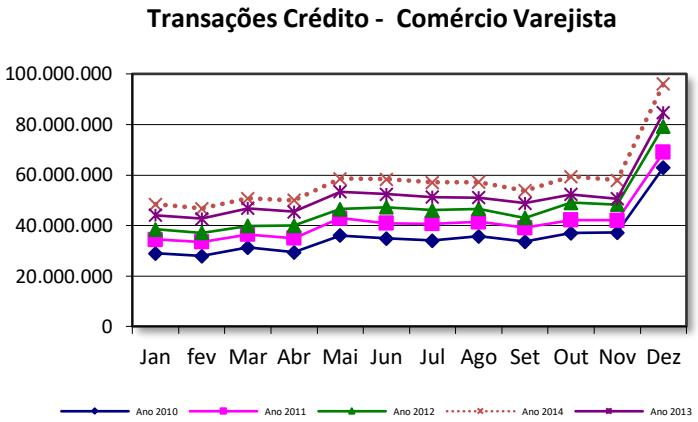


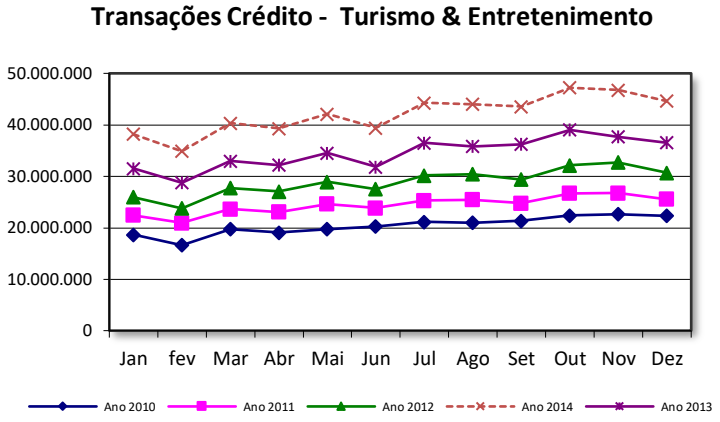
Fonte: Ronaldo Goldschmidt e Eduardo Bezerra - DataMining: conceitos, técnicas, algoritmos, orientações e aplicações. *Editora Elsevier.*

MODELOS DE SÉRIES TEMPORAIS

Previsão

Quantidade de transações mensais com cartões de crédito





Fonte:ABECS

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:

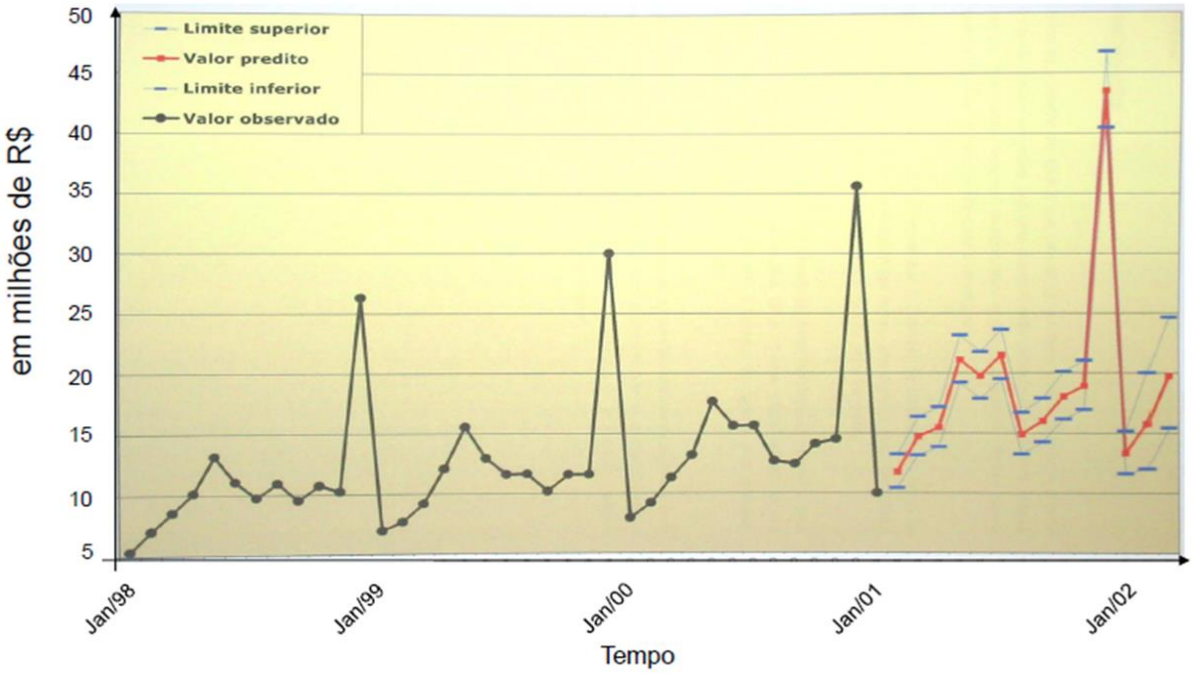
Período	Proporção de vendas
17/01 a 23/01	34.1
24/01 a 30/01	27.9
31/01 a 06/02	26.7
07/02 a 13/02	15.4
14/02 a 20/02	37.0
21/02 a 27/02	25.0
28/02 a 06/03	46.7

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

MODELOS DE SÉRIES TEMPORAIS

Previsão de 12 meses para o faturamento mensal - Varejo (Vestuário) -

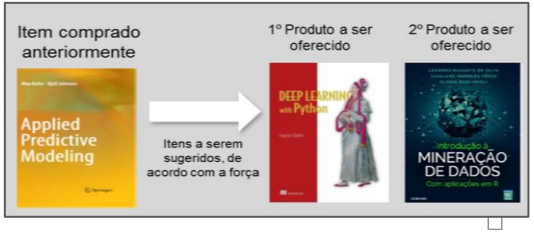


DATA MINING

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

- **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

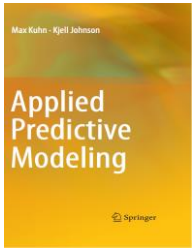
Quais associações são significativas?

Exemplo Associações significativas

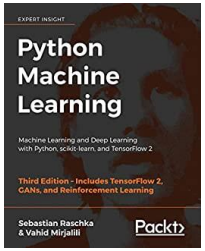
Item comprado anteriormente

Itens a serem sugeridos de acordo com a força

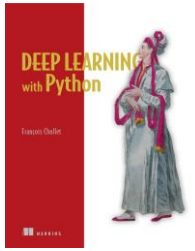
1° produto



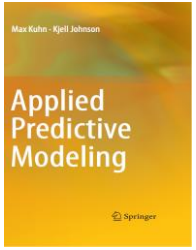
2° produto



3° produto

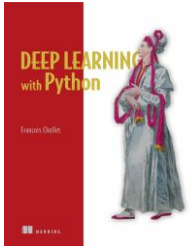


Item comprado anteriormente



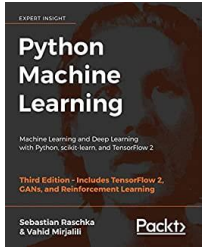
Itens a serem sugeridos de acordo com a força

1° produto



2° produto

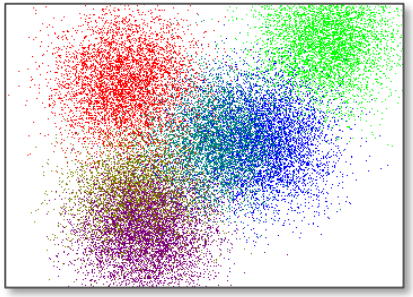
3° produto

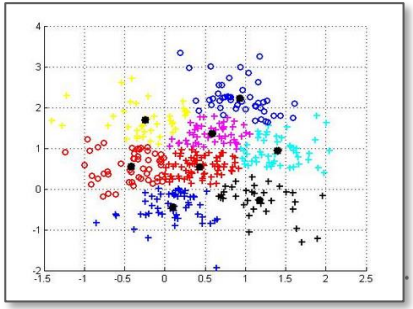


DATA MINING

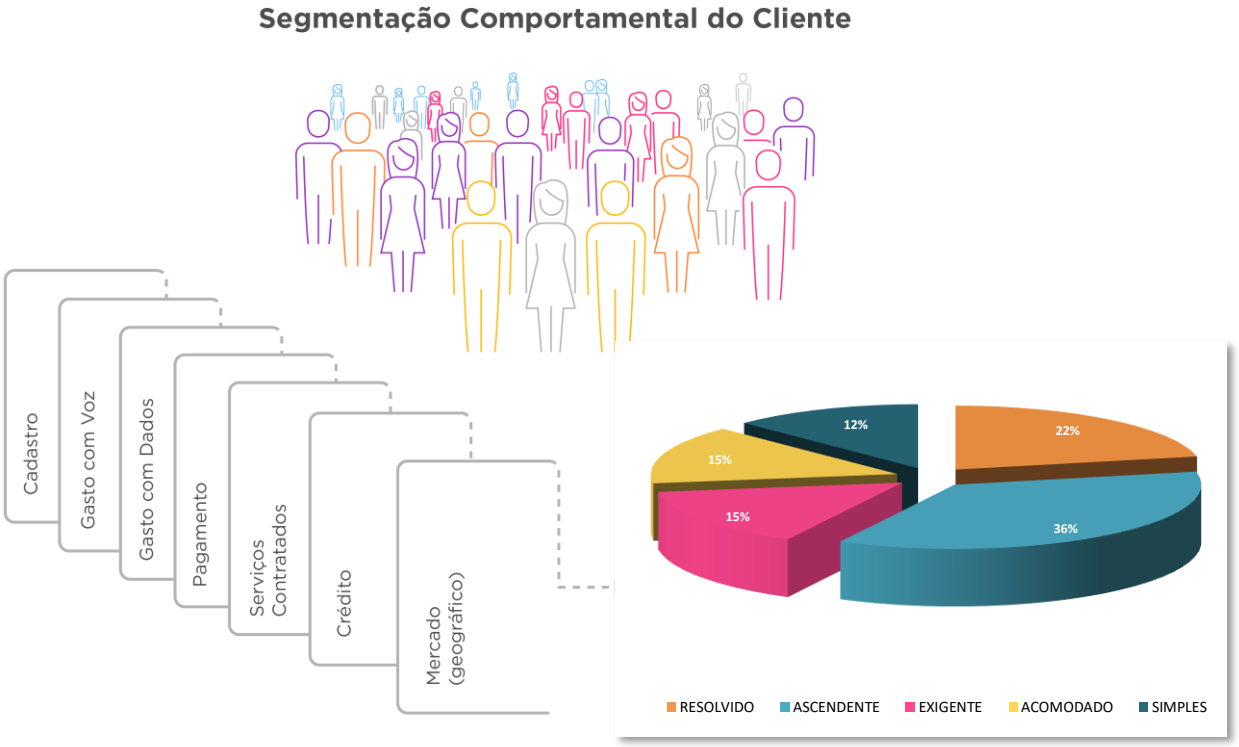
Aplicações práticas de Data Mining podem ser categorizadas de acordo com a tarefa que se pretende resolver.

- 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.





TÉCNICA DE AGRUPAMENTO:
ANÁLISE DE CLUSTERS



DATA MINING

Aplicações práticas de Data Mining 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.

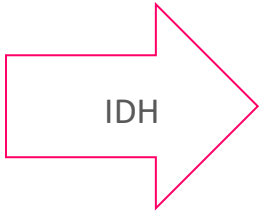
Exemplo: Construção do indicador de satisfação

- 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 dos 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

- 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.

TÉCNICA DE SUMARIZAÇÃO :

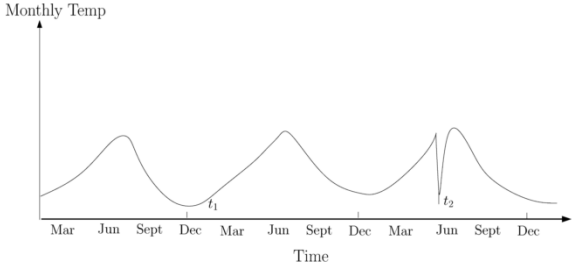
Exemplo: Construção do IDH-Municipal

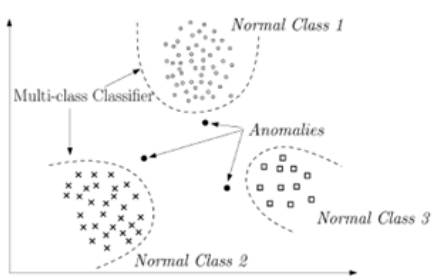
MUNICÍPIO	UF	Esp_Vida	Tx_alfab	Tx_freq_esc	rendacapita	IDH_M	Class_UF	Class_BR
São Caetano do Sul	SP	78,18	97,01	98,57	834,00	0,919	1	1
Águas de São Pedro	SP	77,44	97,06	85,75	954,65	0,908	2	2
Santos	SP	72,27	96,44	92,62	729,62	0,871	3	6
Vinhedo	SP	74,87	94,08	79,73	627,47	0,857	4	15
Jundiaí	SP	73,94	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

DATA MINING

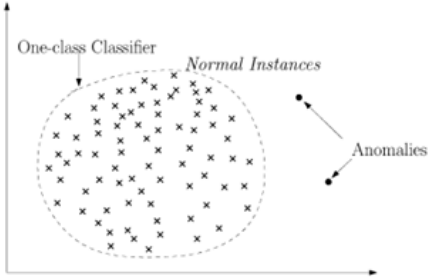
Aplicações práticas de Data Mining podem ser categorizadas de acordo com a tarefa que se pretende resolver.

- 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).





(a) Multi-class Anomaly Detection



(b) One-class Anomaly Detection

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 Estrutural

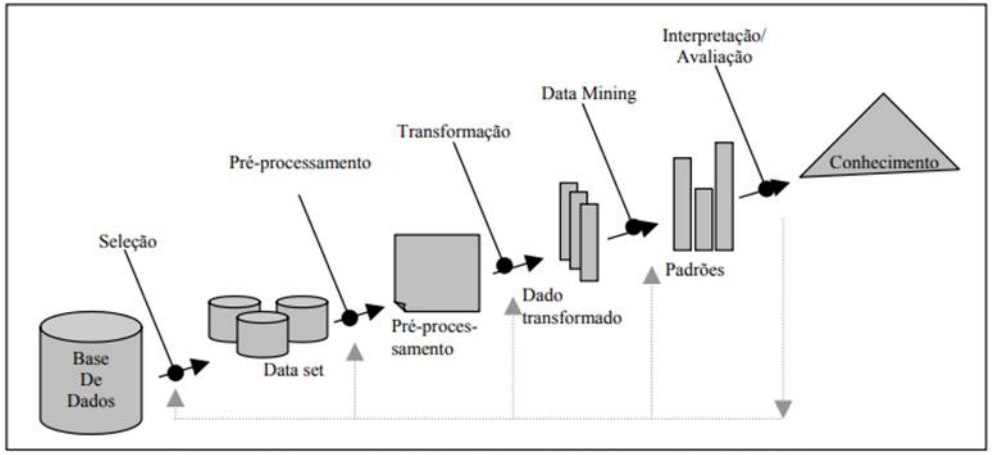
- Técnicas de Interdependência.
- Técnicas Multivariadas que procuram agrupar dados com base em semelhança, permitindo assim a interpretação das estruturas dos dados. Não há distinção entre variáveis dependentes e independentes.

Aprendizado Supervisionado

Aprendizado Não Supervisionado

PROCESSO KDD

KNOWLEDGE DISCOVERY IN DATABASES



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

PROCESSO 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.

Visão das Etapas de KDD (Adaptado de Fayyad, Piatetsky-Schapiro, Smyth, 1996).

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 incluam 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.

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.

SOFTWARE ESTATÍSTICO

- SAS
- SPSS
- Minitab
- STATISTICA
- STATA
- R
- Mplus
- Python
- KNIME
- WEKA

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 Estrutural

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

Aprendizado Supervisionado

Aprendizado Não Supervisionado

ANÁLISE ESTATÍSTICA MULTIVARIADA

NOÇÕES GERAIS

ANÁLISE MULTIVARIADA

O que são dados multivariados?

- Amostra de indivíduos selecionados aleatoriamente: pessoas residentes em uma cidade, municípios do Brasil, domicílios etc.
- Em cada indivíduo são observadas diversas dimensões (variáveis): sexo, idade, número de mensagens enviadas, horários em que costuma ligar, reações a diferentes tipos de ações, avaliações da empresa em diferentes aspectos etc.
- Em função de essas variáveis serem medidas no mesmo indivíduo, existirão, provavelmente, relações de interdependência e correlações entre elas.

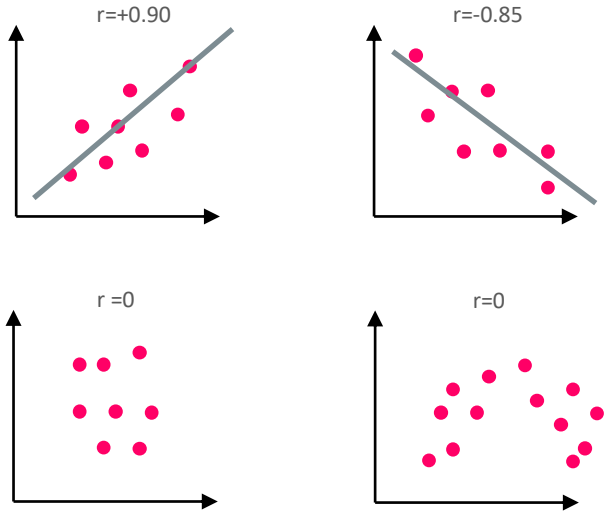
Como analisar estas informações?

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.

Apoio

RELEMBRANDO...

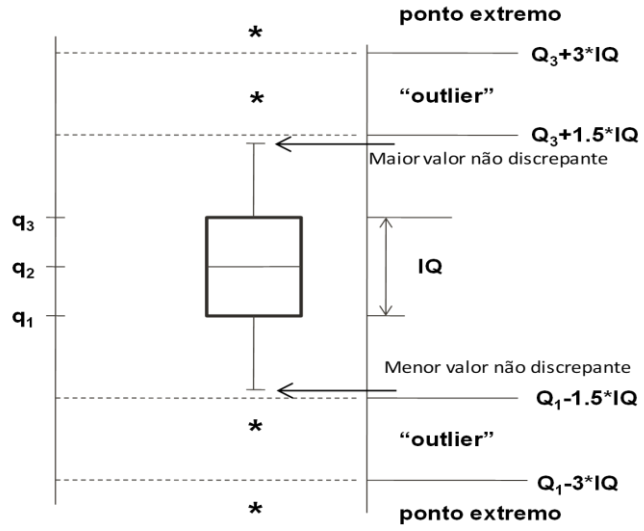
ANÁLISE EXPLORATÓRIA DOS DADOS

Outliers

Observações que apresentam um grande afastamento das restantes ou são inconsistentes com elas.

- Dado incorreto
- População diferente
- Dado correto – Evento raro

Representação Gráfica na Análise dos Dados

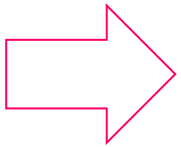


CUIDADO

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.

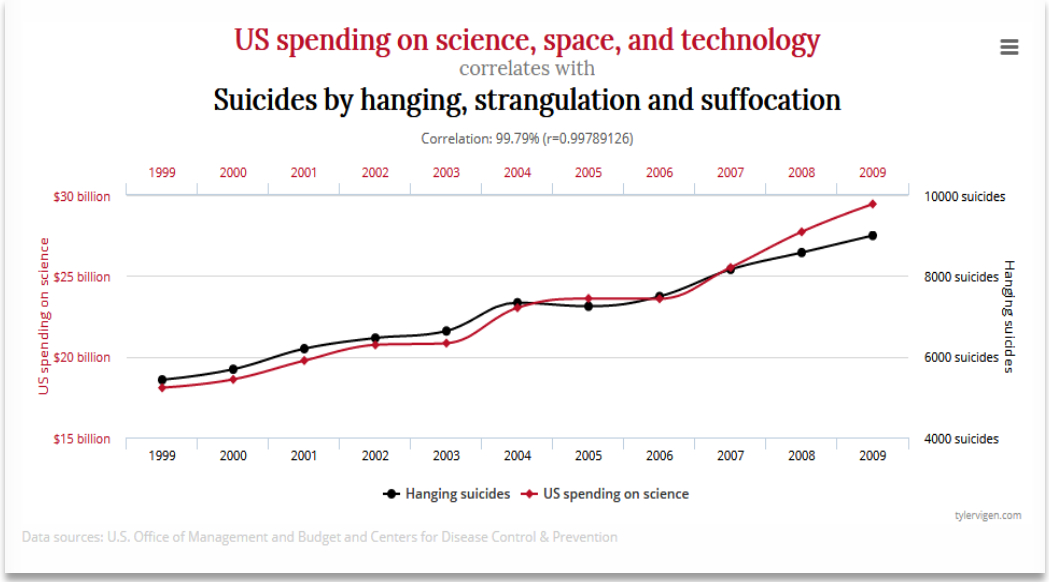
ASSOCIAÇÕES ESPÚRIAS

- Associação entre dois fatores e quando queremos saber se um causa o outro.
- Big data muitos resultados estatisticamente significativos que não fazem sentido causal.
- Variável de confusão quando há muitas variáveis na análise.



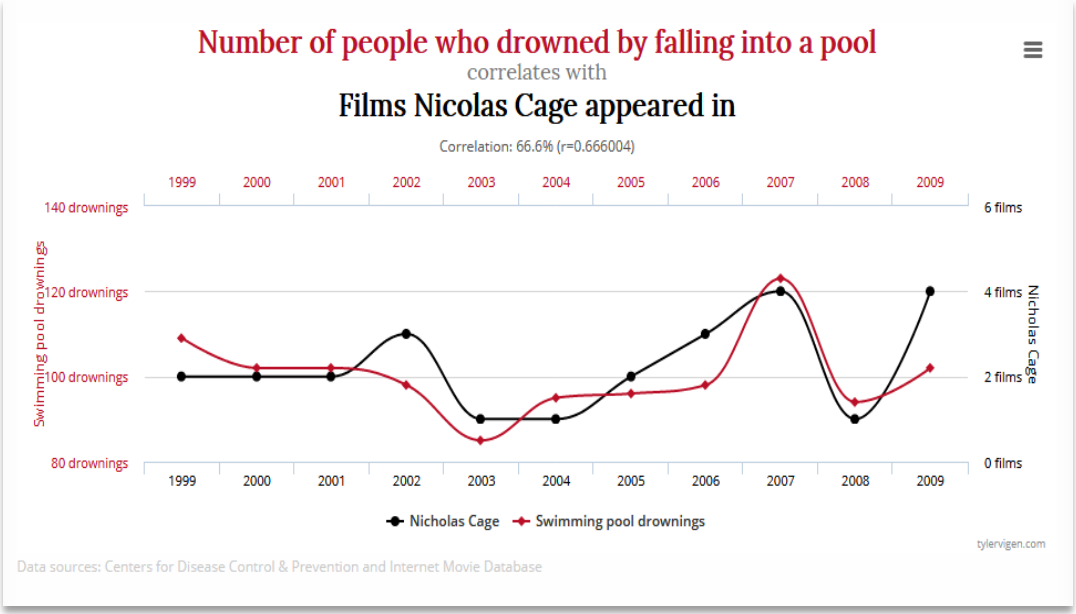
Uma relação estatística existente entre duas variáveis, mas onde não existe nenhuma relação causa-efeito entre elas. Essa relação estatística pode ocorrer por pura coincidência ou por causa de uma terceira variável.

ASSOCIAÇÕES ESPÚRIAS



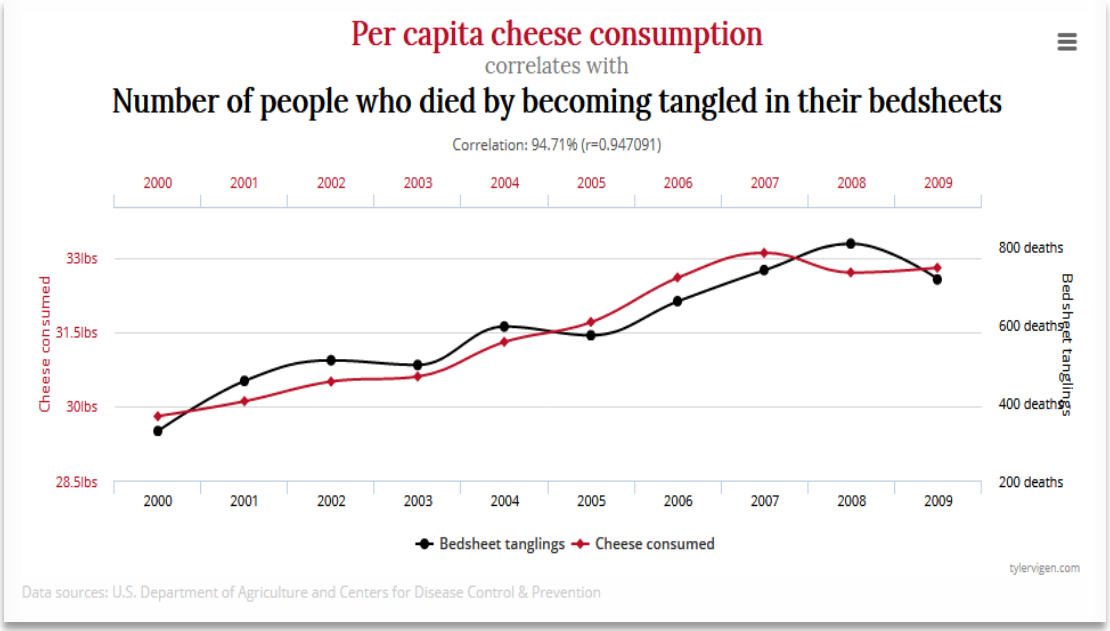
Escritório de Administração e Orçamento dos EUA e Centros de controle e prevenção de doenças

ASSOCIAÇÕES ESPÚRIAS



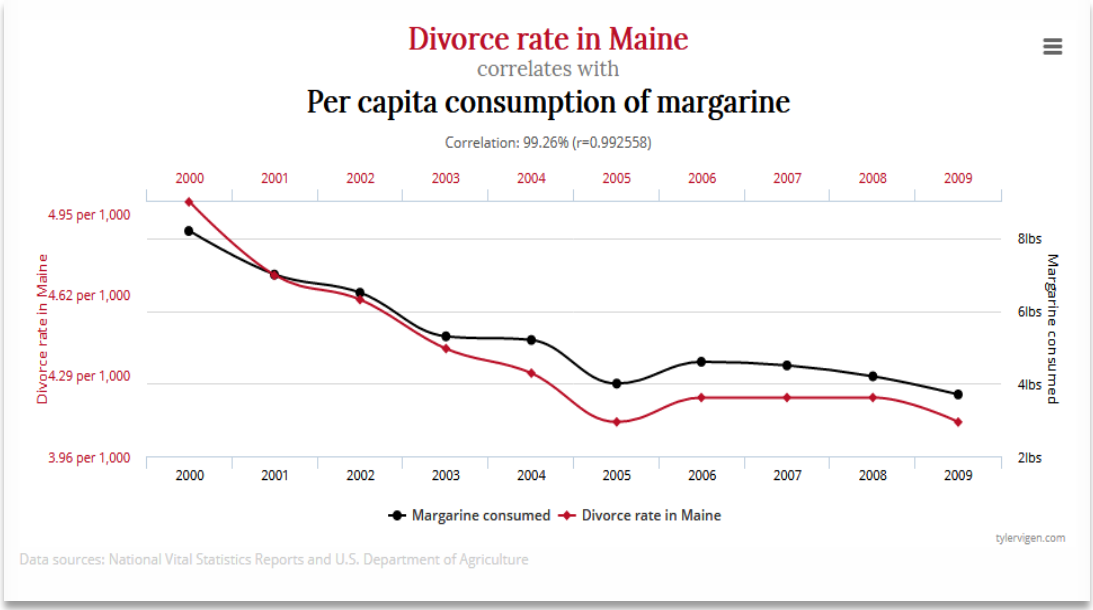
Data sources: Centers for Disease Control & Prevention and Internet Movie Database

ASSOCIAÇÕES ESPÚRIAS



Departamento de Agricultura dos EUA e Centros de Controle e Prevenção de Doenças

ASSOCIAÇÕES ESPÚRIAS



Relatórios Nacionais de Estatísticas Vitais e Departamento de Agricultura dos EUA

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
- TORGO, 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