Introdução a Ciências de Dados

Aula 1 parte 1: Introdução

Francisco A. Rodrigues
ICMC/USP
francisco@icmc.usp.br

Aula 1: Introdução

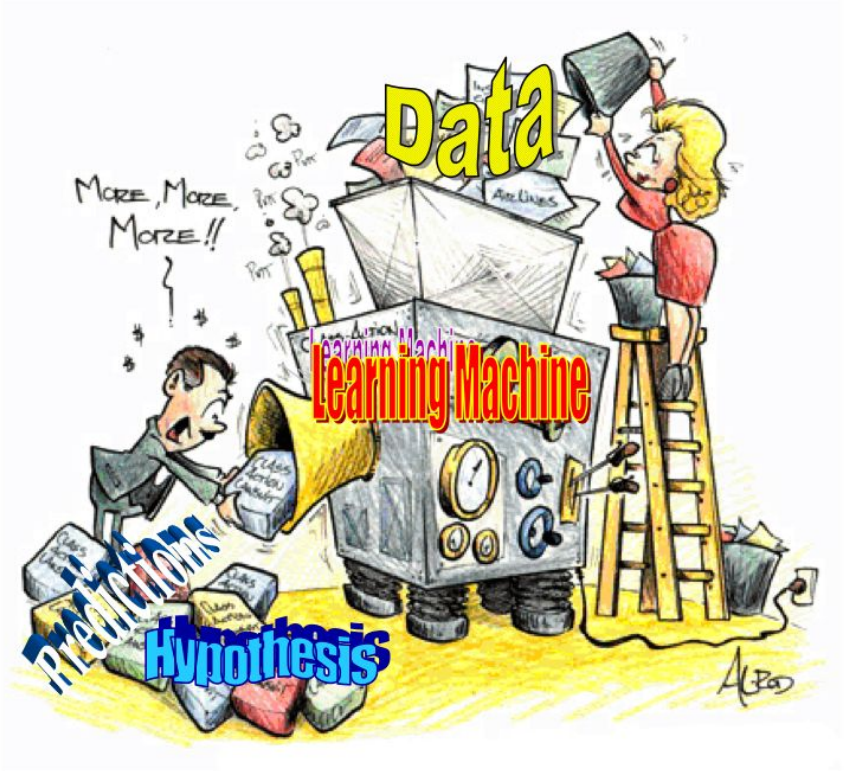
- O que é Ciência de Dados
- Problemas e Soluções em Ciência de Dados

O que é Ciência de Dados

“Ciência de dados (em inglês: data science) é uma área interdisciplinar voltada para o estudo e a análise de dados, estruturados ou não, que visa a extração de conhecimento ou insights para possíveis tomadas de decisão, de maneira similar à mineração de dados.’

WIKIPÉDIA
A enciclopédia livre

O que é Ciência de Dados



O que é Ciência de Dados

Ciência de Dados

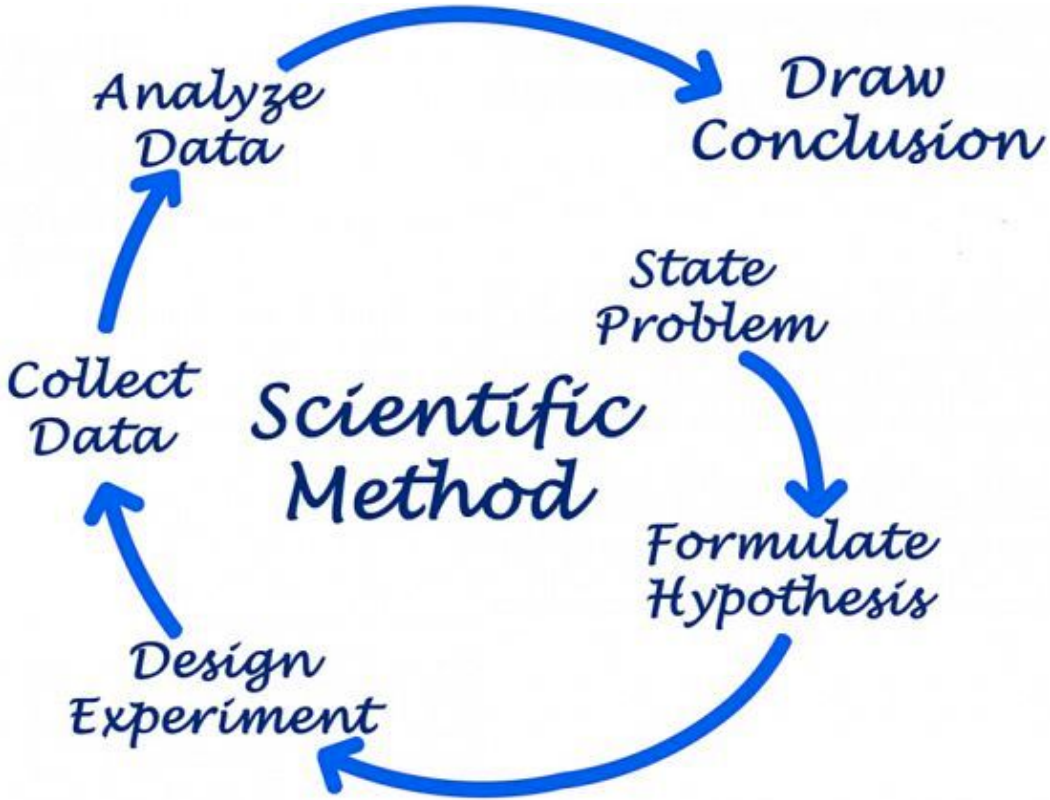
=

Ciência

+

Dados

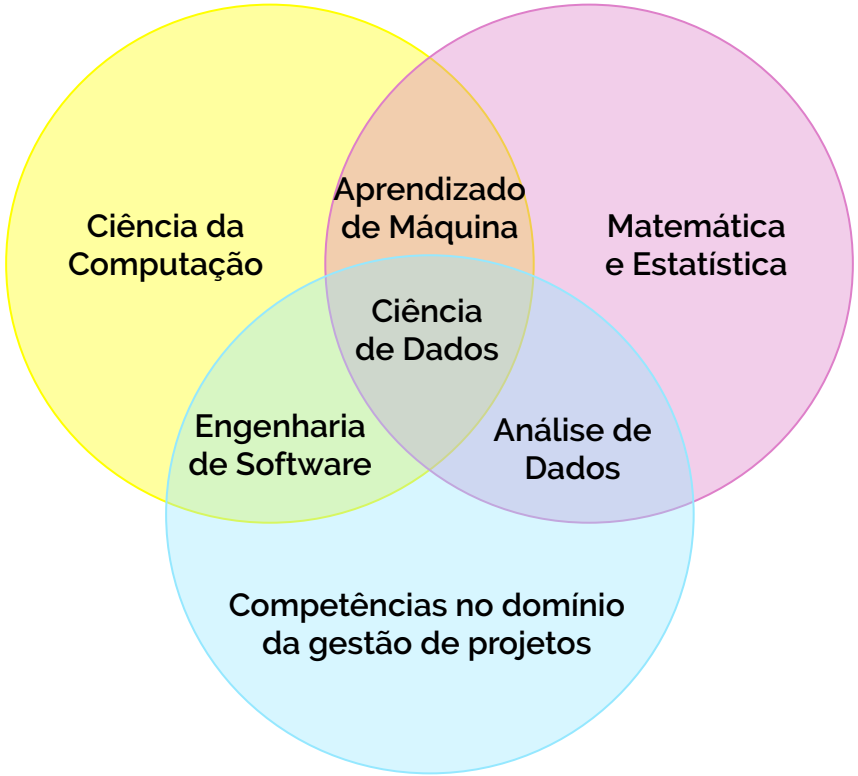
Método científico



Método Científico

USP ICMC USP SÃO CARLOS CeMEAI CEPID - Centro de Ciências Matemáticas Aplicadas à Indústria

O que é Ciência de Dados



Problemas em Ciência de Dados

Classificação de documentos:

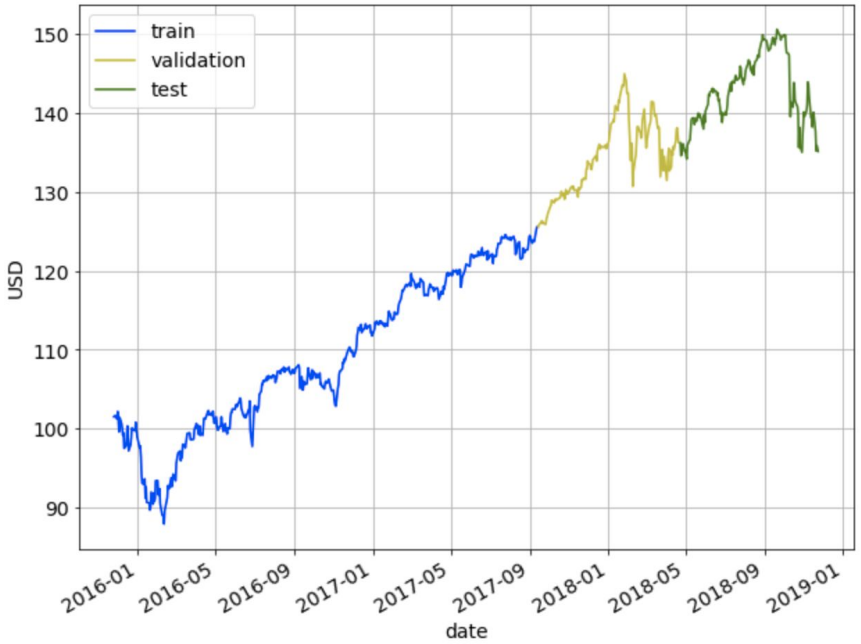
Sports
Science
News

Problemas em Ciência de Dados

Agrupar imagens similares:

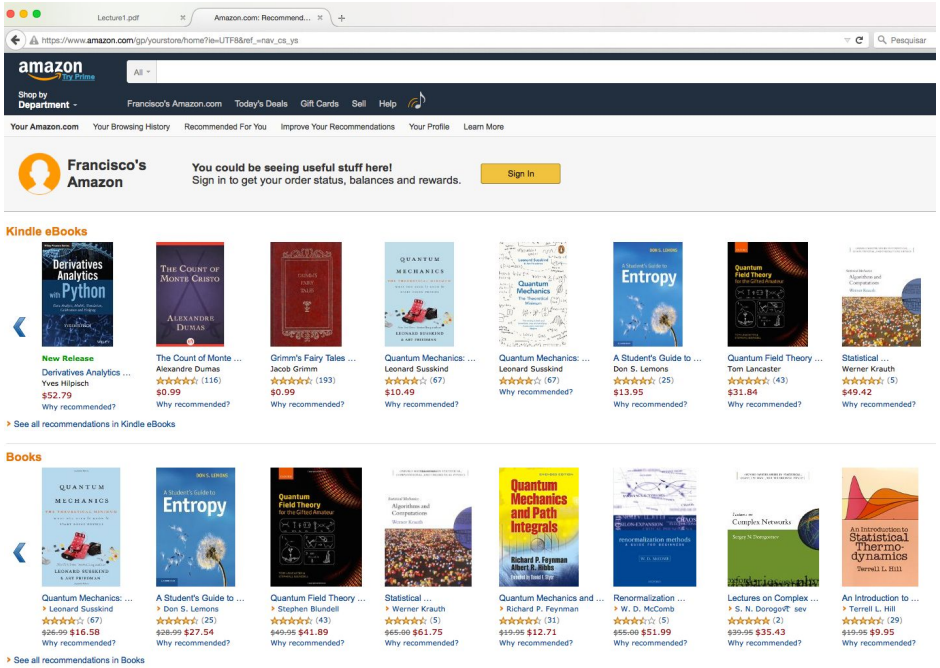
Problemas em Ciência de Dados

Mercado de ações:



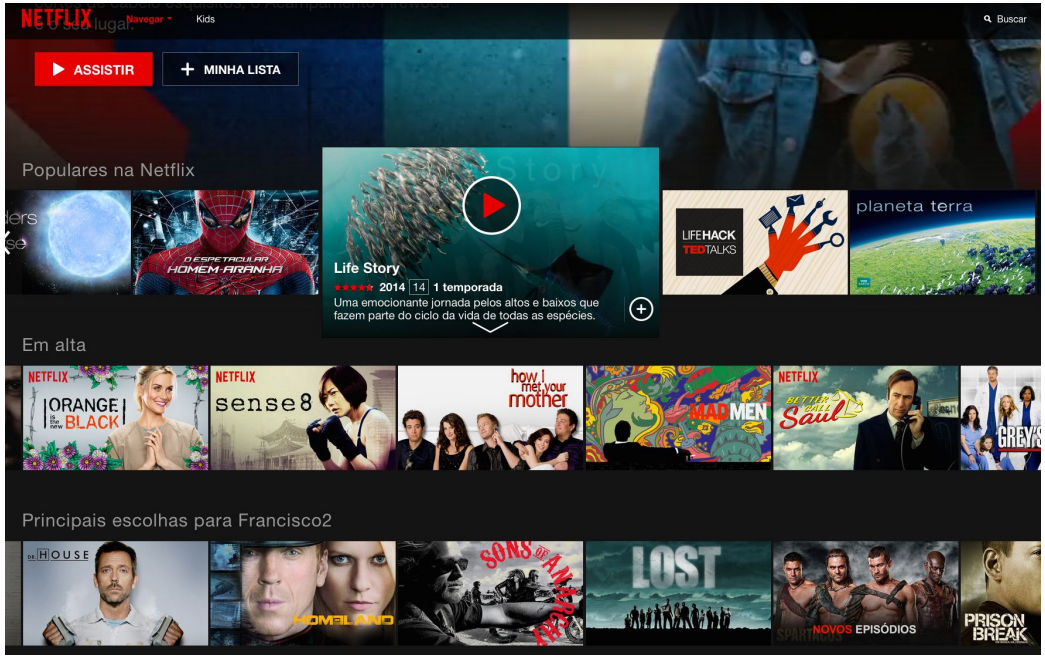
Problemas em Ciência de Dados

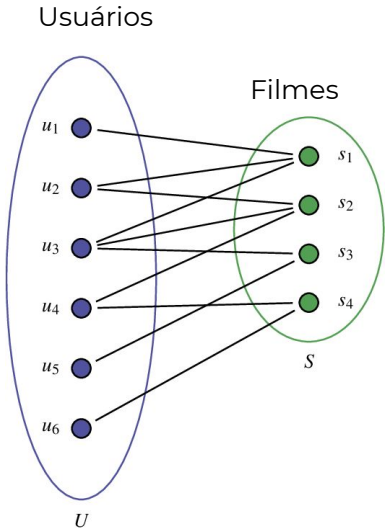
Recomendação



Problemas em Ciência de Dados

Recomendação





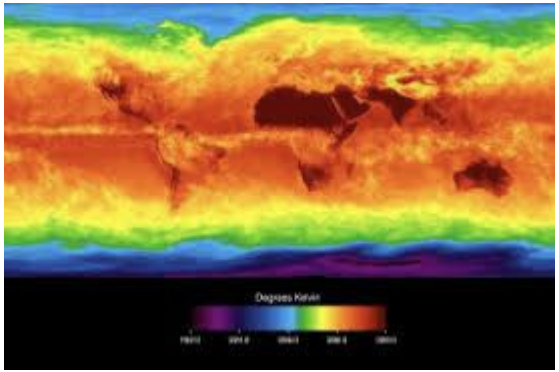
Problemas em Ciência de Dados

Diagnóstico médico:

pixabay.com

Problemas em Ciência de Dados

Predição de secas:



pixabay.com

Problemas em Ciência de Dados

Agricultura:

pixabay.com

Problemas em Ciência de Dados

Operações Bancárias:

pixabay.com

Problemas em Ciência de Dados

Detecção de Fraudes:

pixabay.com

Problemas em Ciência de Dados

E-commerce
- Identificação de clientes
- Recomendação de produtos
- Análise de avaliações de produtos

Bancos
- Detecção de fraudes
- Modelagem de risco de crédito
- Mercado futuro

Medicina
- Análise de dados médicos
- Descoberta de novas drogas
- Bioinformática

Transporte
- Carros autônomos
- Sistema de monitoramento
- Segurança

Finanças
- Segmentação de usuários
- Decisões estratégicas
- Análise de risco

Agricultura
- Uso de pesticidas
- Previsão das safras
- Planejamento de lavouras

Tarefas

- Classificação
- Regressão
- Agrupamento
- Regras de associação
- Visualização

USP ICMC USP SÃO CARLOS CeMEAI CEPID - Centro de Ciências Matemáticas Aplicadas à Indústria

Aprendizado supervisionado

Modelos preditivos: função que, dado um conjunto de exemplos rotulados, constrói um estimador.

$$y = f(X, \theta) + \epsilon$$

Classificação

- Rótulos nominais (conjunto discreto e não ordenado de valores)
  - Ex. {doente, saudável}, {bom pagador, mau pagador}, {iris setosa, iris versicolor, iris virginica}
- Estimador é chamado **classificador**.

Regressão

- Rótulos contínuos (conjunto infinito ordenado de valores)
  - Ex. peso, temperatura, vazão de água.
- Estimador é chamado **regressor**.

Estimadores podem ser vistos como funções.

Classificação

Definição formal: Dado um conjunto de observações:

$$D = \{\mathbf{X}, \mathbf{y}, i = 1,\ldots,N\}$$

- **f** representa uma função desconhecida (função objetivo).
- Essa função mapeia as entradas nas saídas correspondentes.
- O algoritmo preditivo aprende a aproximação, que permite estimar valores de **f** para novos valores de **X**.

$$y_i = f(X_i, \theta) + \epsilon_i$$

Classificação

$$y_i \in \{C_1, C_2, \ldots, C_n\}$$

Classificação

Classificação de documentos:

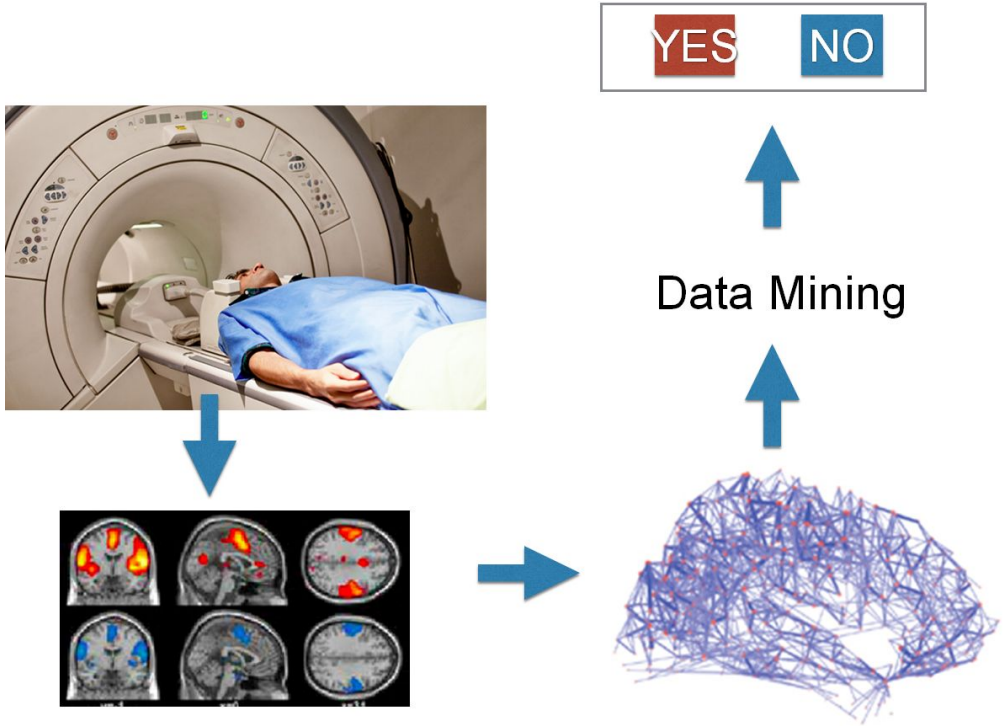
Sports
Science
News

Classificação

Identificação de Fake News

Classificação

Exemplo



Classificação

Produtos Bancários:

pixabay.com

Regressão

Definição formal: Dado um conjunto de observações:

$$D = \{\mathbf{X}, \mathbf{y}, i = 1,\ldots,N\}$$

- **f** representa uma função desconhecida (função objetivo).
- Essa função mapeia as entradas nas saídas correspondentes.
- O algoritmo preditivo aprende a aproximação, que permite estimar valores de **f** para novos valores de **X**.

$$y_i = f(X_i, \theta) + \epsilon_i$$

Regressão

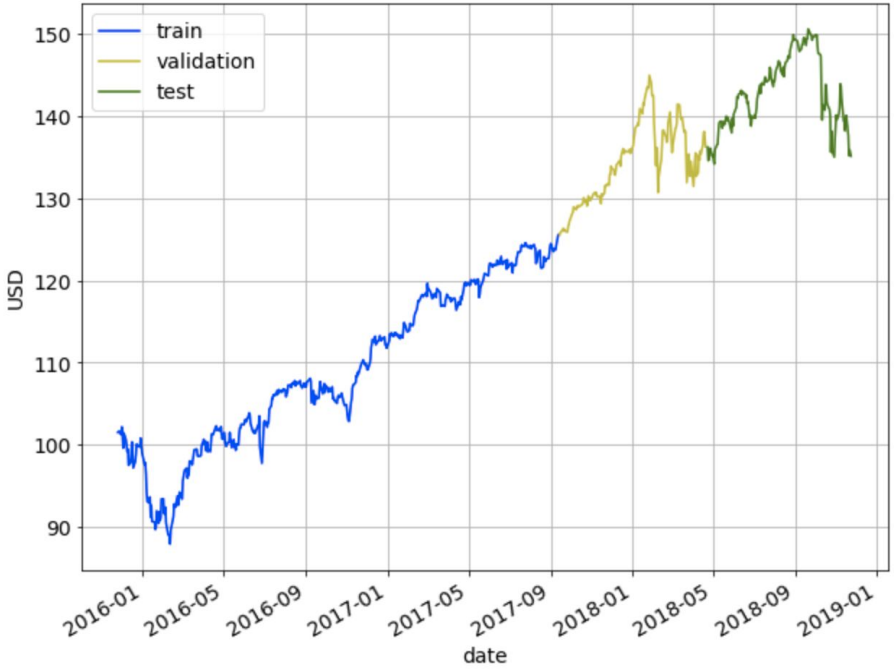
$$y_i \in \mathbb{R}$$

Regressão

Exemplo:

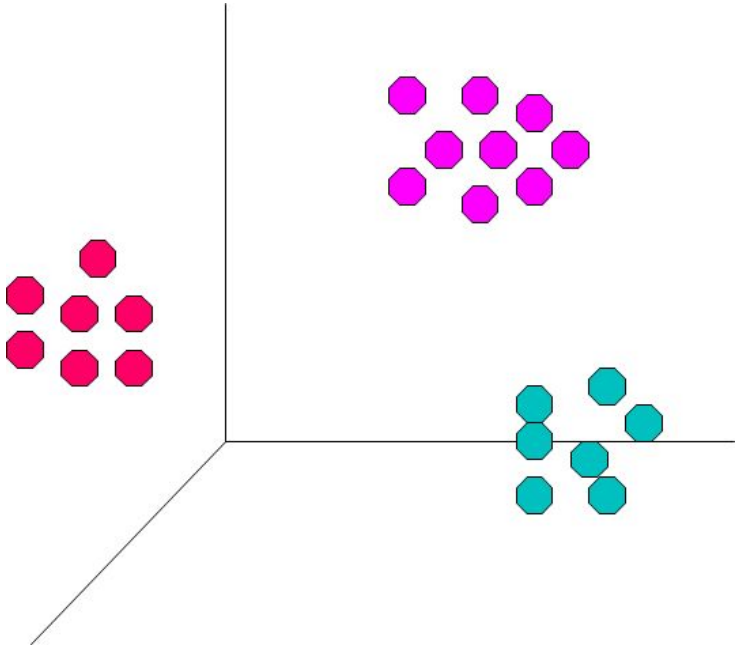
Regressão

Exemplo: Previsão de séries temporais



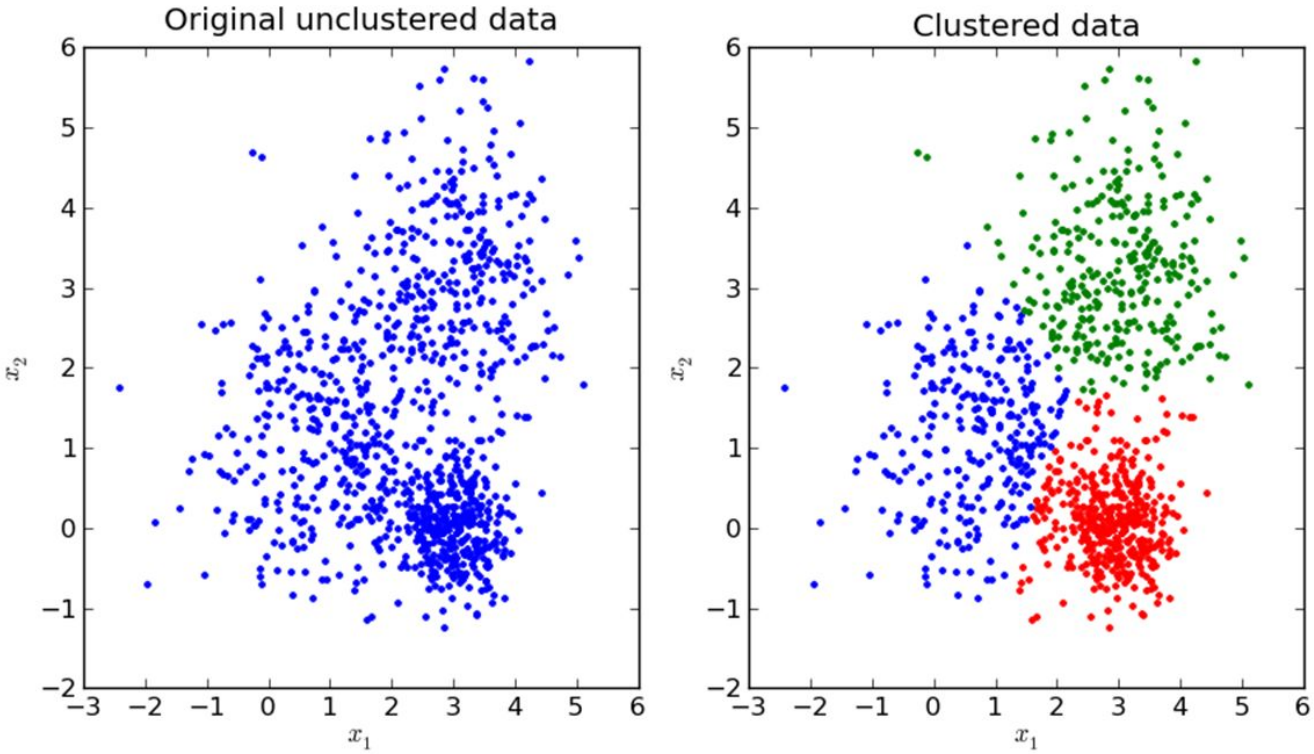
Agrupamento

Objetivo: Agrupar as observações de forma que a similaridade entre objetos no mesmo grupo seja máxima e a entre grupos seja mínima.



Agrupamento

Exemplo:



Agrupamento

Agrupar imagens similares:

Regras de associação

Dado um conjunto de transações, onde cada uma contém um número de itens de uma dada coleção,
produzir regras de dependência para predizer um item baseado na ocorrência de outros itens.

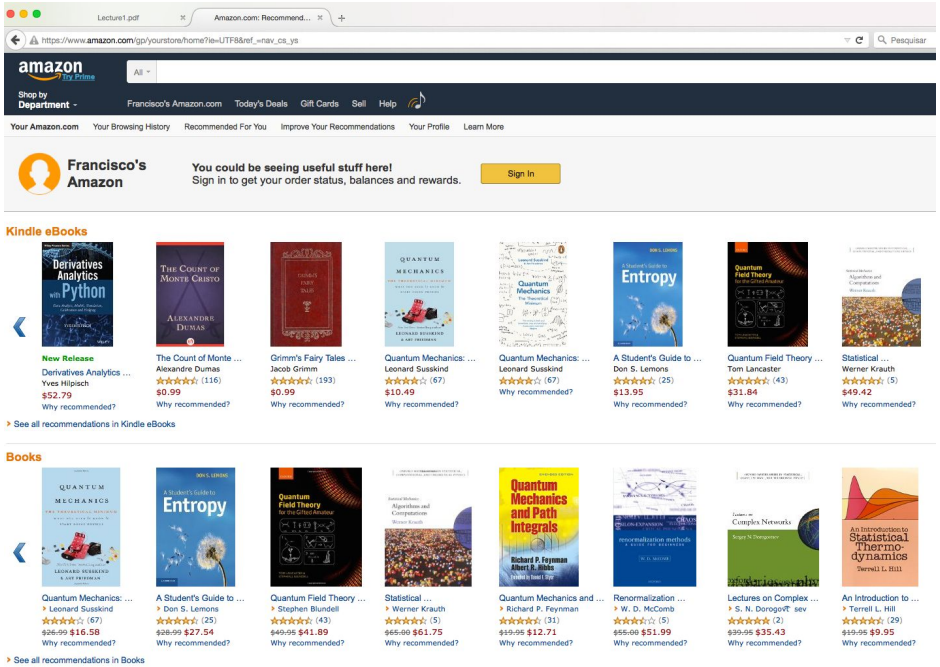
ID	Itens
1	Pão, Café, Leite
2	Cerveja, Pão
3	Cerveja, Café, Fralda, Leite
4	Cerveja, Pão, Fralda, Leite
5	Fralda, Leite, Café

Regras Descobertas:

{Leite} -> {Café}
{Fralda, Leite} -> {Cerveja}

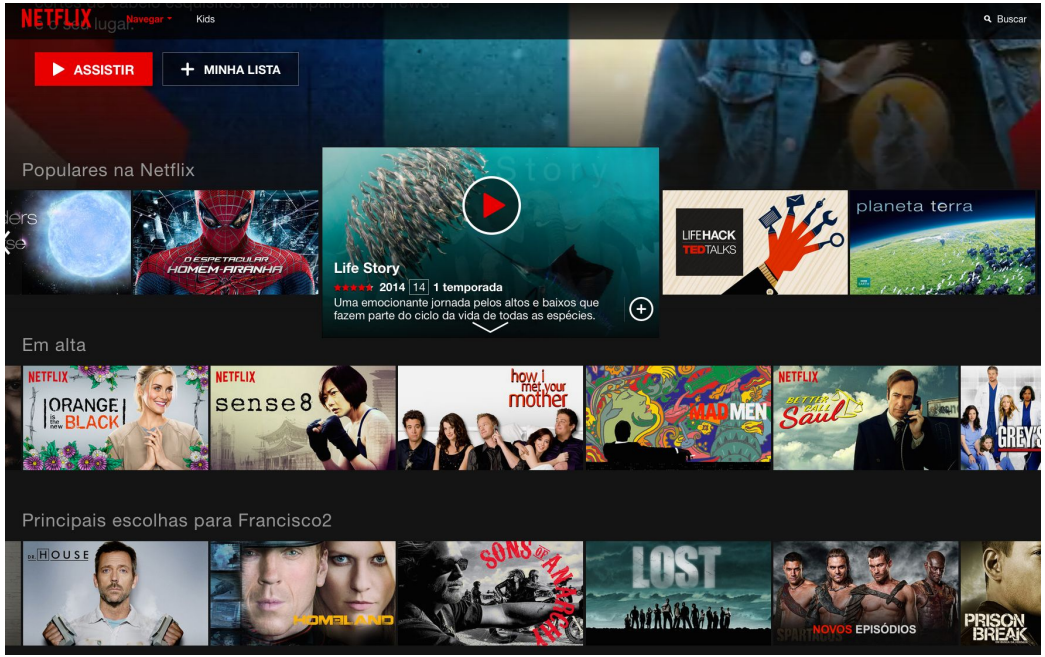
Regras de Associação

Recomendação



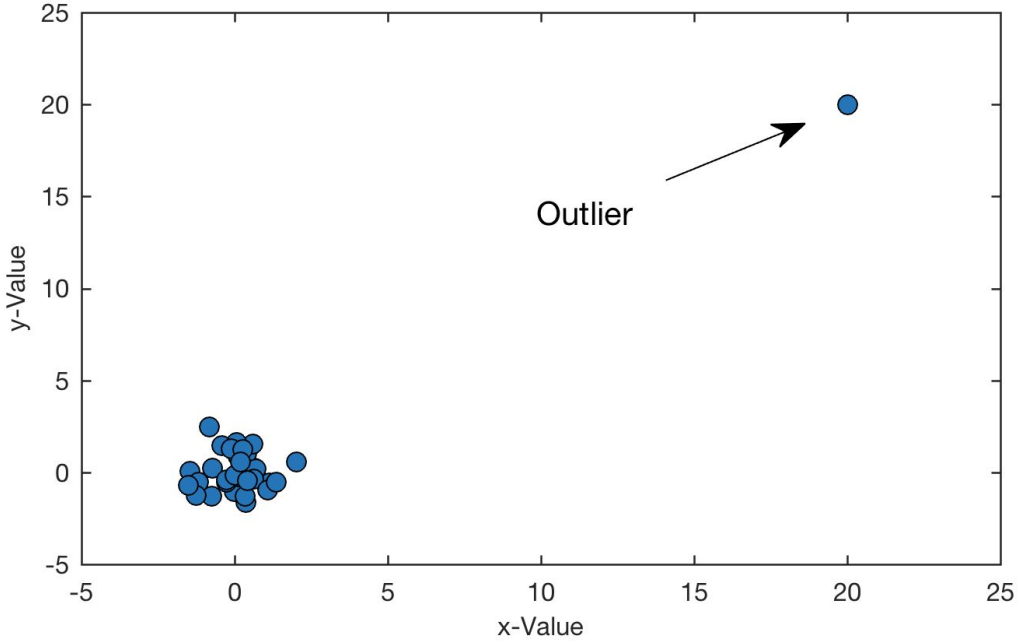
Regras de Associação

Recomendação



Detecção de outliers

Os outliers são dados que se diferenciam drasticamente de todos os outros, são pontos fora da curva.



Detecção de Outliers

Detecção de Fraudes:

pixabay.com

Outras tarefas

- Visualização
- Classificação semi-supervisionada
- Aprendizado por reforço

Sumário

- O que é Ciência de Dados
- Problemas e Soluções em Ciência de Dados

Leitura Complementar

- Tan, Steinbach, Karpatne, Kumar, **Introduction to Data Mining**, Pearson, 2013 (capítulo 1)