FUNDAMENTOS E TÉCNICAS EM CIÊNCIAS DE DADOS

PROF. JOSENALDE OLIVEIRA

josenalde.oliveira@ufrn.br
https://github.com/josenalde/datascience

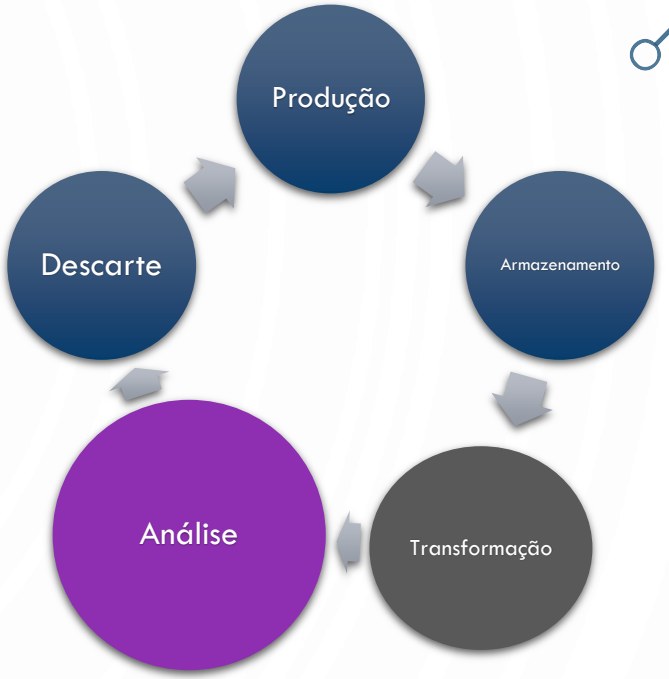
ANÁLISE E DESENVOLVIMENTO DE SISTEMAS - UFRN

ANÁLISE

No contexto big data: estuda e aplica um conjunto de ferramentas para manipular, gerir, analisar e obter informações a partir de grandes volumes de dados, em variados formatos e tipos de estruturas

No contexto geral: aplicar algum tipo de processamento/transformação em busca de conhecimento

Normalmente associada à classificação, predição, inteligência, mas pode ser dividida em Exploratória, Explícita e Implícita. Dividir auxilia na seleção da técnica, mas as mesmas podem ter uso misto.



Explícita

Folha de Pagamento
65560665401
25707968245
98130379384
33184848302
15425437382
22624531106
27758123848
99087226209
13671684577
76093580610

Prestadores de Serviços
84461614352
27632580778
89478263161
45013934591
19512753812
03753305260
64068325405
07602860004
25707968245
31477769994

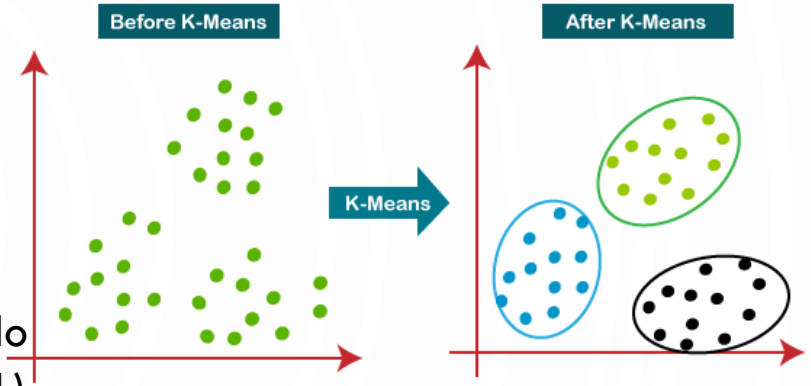
Técnicas mais simples
Informação e conhecimento explícita nos dados
Requer operação de baixa complexidade para ressaltar/destacar a informação; Destacam informações presentes; nos dados
Pode ser confundida com Exploratória, sendo aquela com o objetivo de conhecer os dados (antes de analisá-los), e a Explícita tem objetivo definido como resumir as vendas do mês, verificar notas fiscais faltantes, checar o cálculo do imposto etc.
Exemplos: filtro, consulta SQL, drill down num cubo, criação de colunas calculadas

Quais funcionários também são prestadores de serviços da empresa?

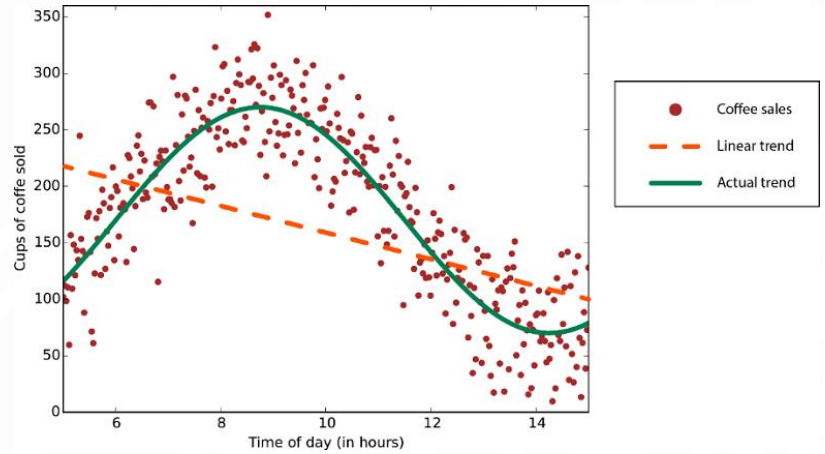
ANÁLISE

Implícita

A informação não está disponível claramente no conjunto de dados
Mesmo olhando de várias formas, filtre, selecione ou faça algum cálculo
Necessita função mais específica, como aprendizagem de máquina (ML) ou uma lei estatística



Quem vem solicitar crédito será um bom pagador?
Quem foi mal pagador no passado (histórico) será no futuro?



Dado um conjunto de características (features) de um cliente,
pode-se atribuir pesos a estas features e prever, com certa margem de erro, se será ou não bom pagador (Ex. Naïve Bayes, KNN, SVM etc.) - classificar

Pode-se também agrupar itens com features comuns (recursos), de acordo com determinadas regras (clusterização)

Pode-se desenvolver equações de predição numérica ou categórica

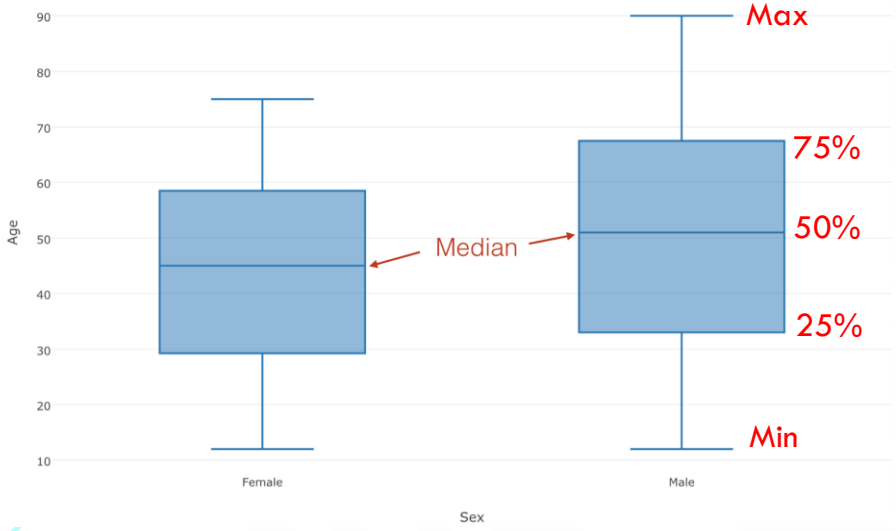
Feature STORE: featurestore.org
https://www.cienciaedados.com/o-que-sao-feature-stores-e-por-que-sao-essenciais-na-escalabilidade-em-data-science/

ANÁLISE

Exploratória

John Wilder Tukey: Exploratory Data Analysis, 1977

- Conhecer os dados que serão analisados
  - como estão distribuídos
  - quais suas médias, mediana (describe())
  - desvio padrão, amplitude
  - como estão relacionados
  - existem valores anormais (outliers)?
  - baseada em técnicas **quantitativas e visuais**
    - gráficos de dispersão (relacionar variáveis numéricas)
    - boxplot (diagrama de caixa)
    - histoarama (freauência do dado)/nuvens de palavras



pandas

Dataframes

Vamos utilizar essencialmente DataFrames Pandas
Arrays NumPy
Visualização de dados com MatPlotLib e Seaborn

index labels

column names

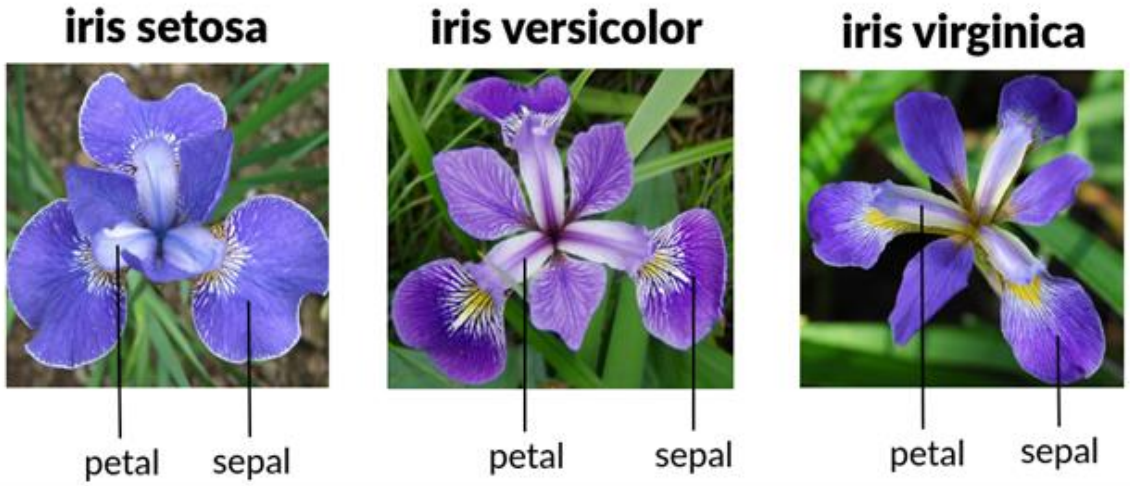
	Mountain	Height (m)	Range	Coordinates	Parent mountain	First ascent	Ascents bef. 2004	Failed attempts bef. 2004
0	Mount Everest / Sagarmatha / Chomolungma	8848	Mahalangur Himalaya	27°59′17″N 86°55′31″E	NaN	1953	>>145	121.0
1	K2 / Qogir / Godwin Austen	8611	Baltoro Karakoram	35°52′53″N 76°30′48″E	Mount Everest	1954	45	44.0
2	Kangchenjunga	8586	Kangchenjunga Himalaya	27°42′12″N 88°08′51″E	Mount Everest	1955	38	24.0
3	Lhotse	8516	Mahalangur Himalaya	27°57′42″N 86°55′59″E	Mount Everest	1956	26	26.0
4	Makalu	8485	Mahalangur Himalaya	27°53′23″N 87°05′20″E	Mount Everest	1955	45	52.0
5	Cho Oyu	8188	Mahalangur Himalaya	28°05′39″N 86°39′39″E	Mount Everest	1954	79	28.0
6	Dhaulagiri I	8167	Dhaulagiri Himalaya	28°41′48″N 83°29′35″E	K2	1960	51	39.0
7	Manaslu	8163	Manaslu Himalaya	28°33′00″N 84°33′35″E	Cho Oyu	1956	49	45.0
8	Nanga Parbat	8126	Nanga Parbat Himalaya	35°14′14″N 74°35′21″E	Dhaulagiri	1953	52	67.0
9	Annapurna I	8091	Annapurna Himalaya	28°35′44″N 83°49′13″E	Cho Oyu	1950	36	47.0

data

ANÁLISE

Exploratória + Limpeza

- Vamos aplicar algumas técnicas para análise exploratória, no dataset didático das espécies de flores Iris, com suas características (features) de comprimento (length) e largura (width) das respectivas pétalas e sépalas – *nb_aed_iris*



https://www.kaggle.com/arshid/iris-flower-dataset