SCC0275 - Introdução à Ciência de Dados

Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo

Professor: Tiago Santana de Nazare
E-mail: tiagosn@alumni.usp.br

1º semestre de 2023

Me apresentando...

Tiago Santana de Nazare (tiagosn@alumni.usp.br)

Graduação:

Ciências de Computação ICMC-USP (2010-2015)

Doutorado:

Ciências de Computação ICMC-USP (2015-2020)

Profissional:

Cientista de Dados (2017 - ...)

Professor (2018 - ...)

Avaliação

Termos 3 trabalhos durante a disciplina

- Pelo menos 2 semanas para a entrega
- Questões para serem respondidas
- Código em Python
- Presença: entregar pelo menos 2 trabalhos

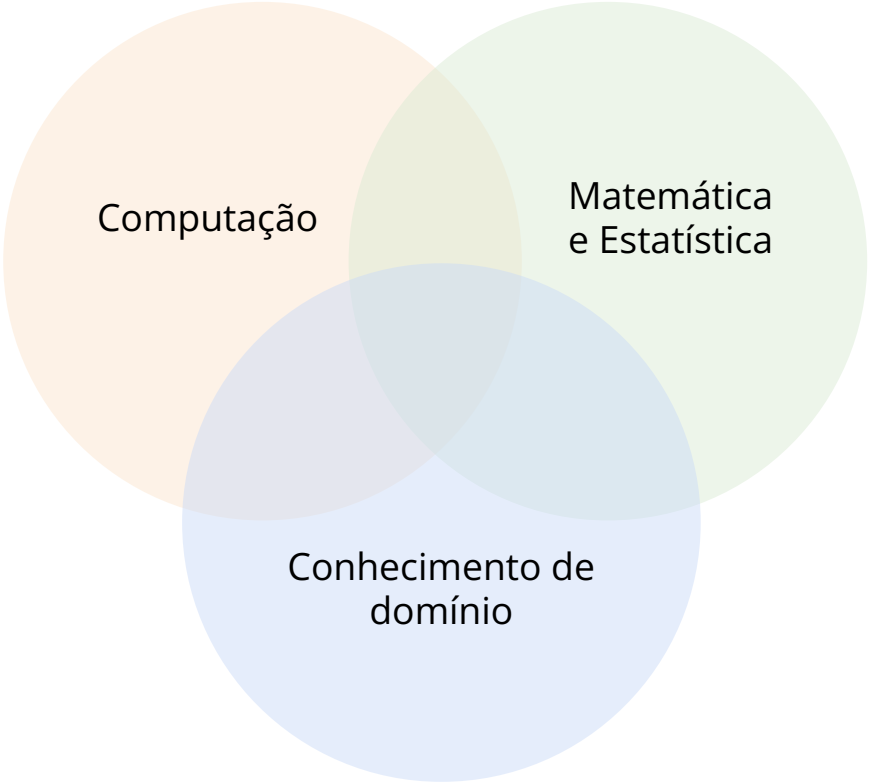
Nota final (NF):

$$NF = \frac{T_1 + T_2 + T_3}{3}$$

Ciência de dados

É uma área interdisciplinar que usa **dados** para:

- Descobrir problemas em um processo
- Monitorar um sistema
- Encontrar padrões
- Gerar previsões
- ...



Vamos usar o Python nessa disciplina...

Link: https://colab.research.google.com
Vantagens: não precisa instalar nada, na nuvem
Desvantagens: não tem muito recurso computacional

Link: https://hub.docker.com/r/jupyter/scipy-notebook
Vantagens: já vem com muitos pacotes instalados
Desvantagens: tem que instalar e configurar o Docker

Link: https://www.anaconda.com (**instalar versão 64 bits**)
Vantagens: já vem com muitos pacotes instalados
Desvantagens: tem que instalar e configurar (mas é fácil)

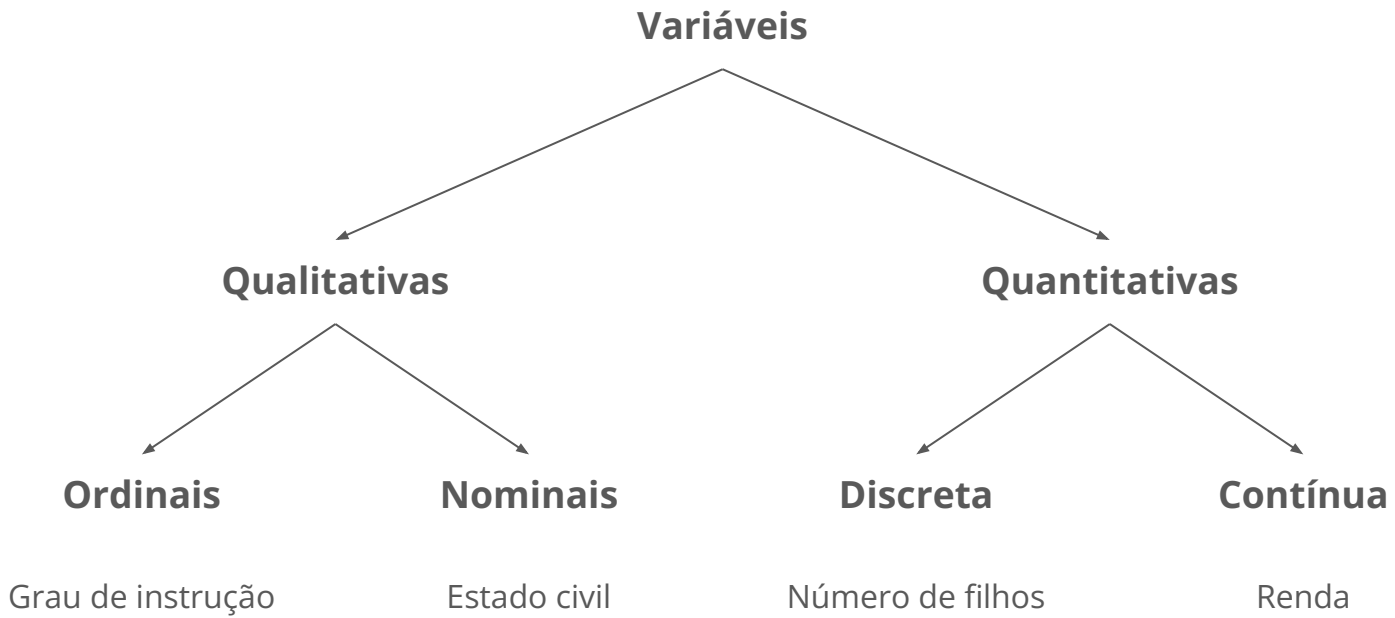
Link: http://tiny.cc/if8wtz
Vantagens: Python portable com pacotes instalados
Desvantagens: só tem para Windows

Primeiro passo de qualquer projeto/estudo

Carregar corretamente os dados e entender como eles estão organizados

sepal_length_cm	sepal_width_cm	petal_length_cm	petal_width_cm	type
5.7	2.9	4.2	1.3	Iris-versicolor
6.7	2.5	5.8	1.8	Iris-virginica
5.1	3.4	1.5	0.2	Iris-setosa
4.4	3.0	1.3	0.2	Iris-setosa
6.9	3.2	5.7	2.3	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica
5.5	2.6	4.4	1.2	Iris-versicolor
5.1	3.8	1.5	0.3	Iris-setosa
4.9	2.5	4.5	1.7	Iris-virginica

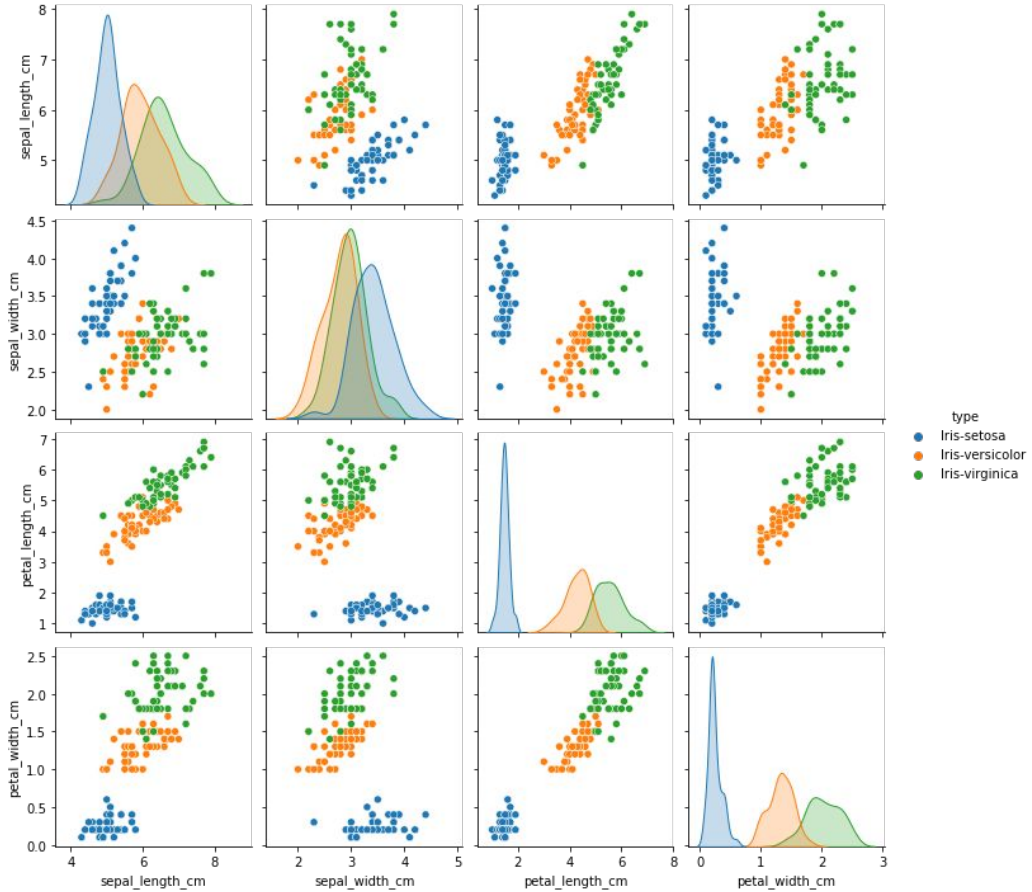
Entender quais tipos de dados nós temos



Tentar entender a distribuição dos dados

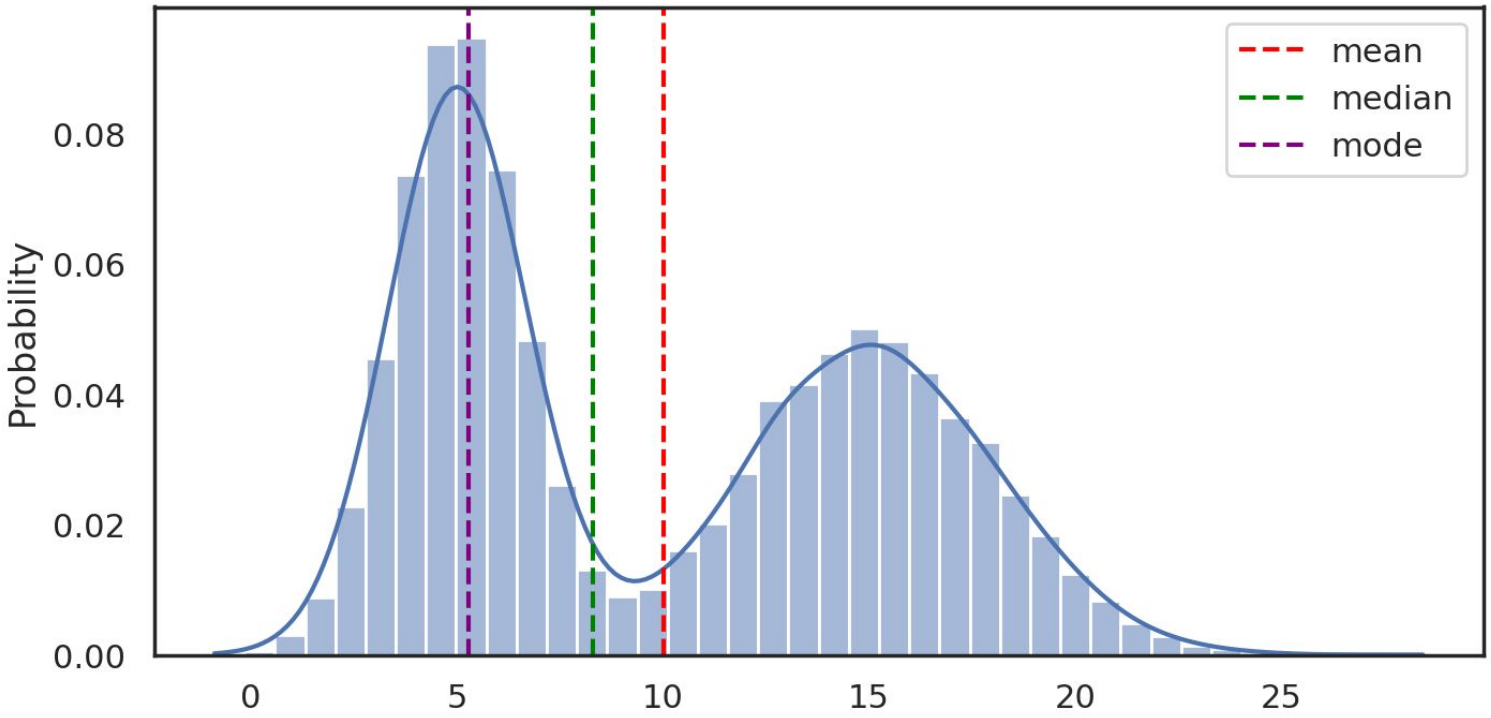
Podemos usar:

- Histograma
- Distribuição
- Scatter plot



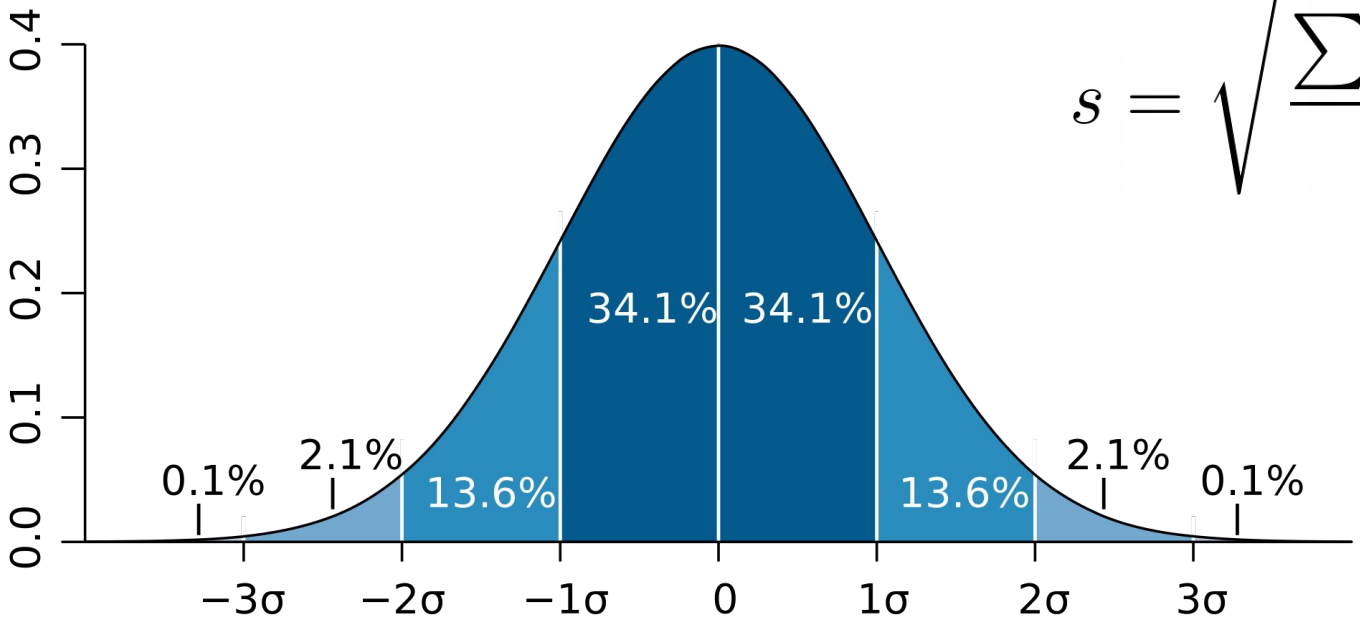
Centralidade

Medidas de centralidade: Média, Mediana e Moda



Dispersão

Medidas de dispersão: variância e desvio padrão



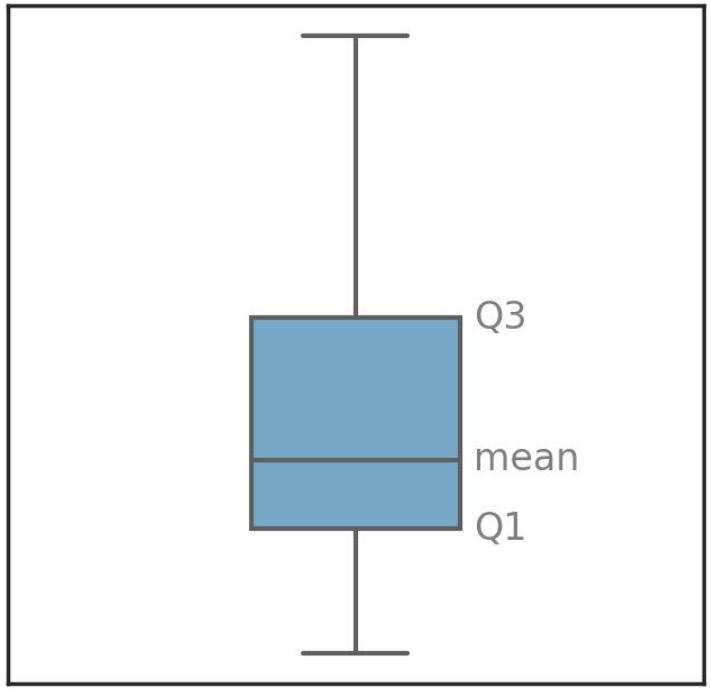
$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$

https://en.wikipedia.org/wiki/Standard_deviation

Box-plot

Sumário de uma distribuição

- Mediana
- Q1: primeiro quartil (25% percentil)
- Q3: primeiro quartil (75% percentil)
- IQR = Q3 - Q1



Correlação

$$s_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{N-1}$$

$$s_x = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$

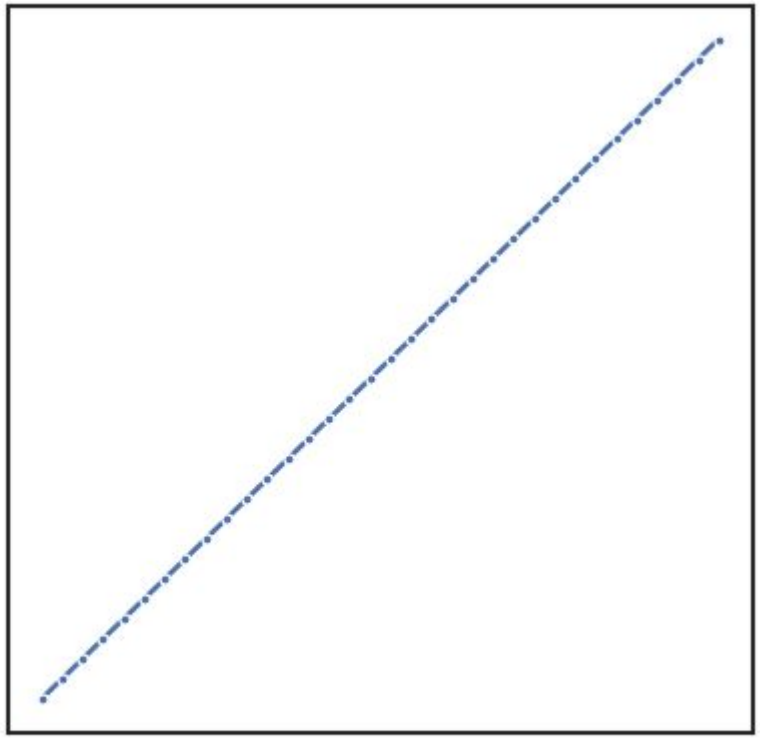
Correlação de Pearson

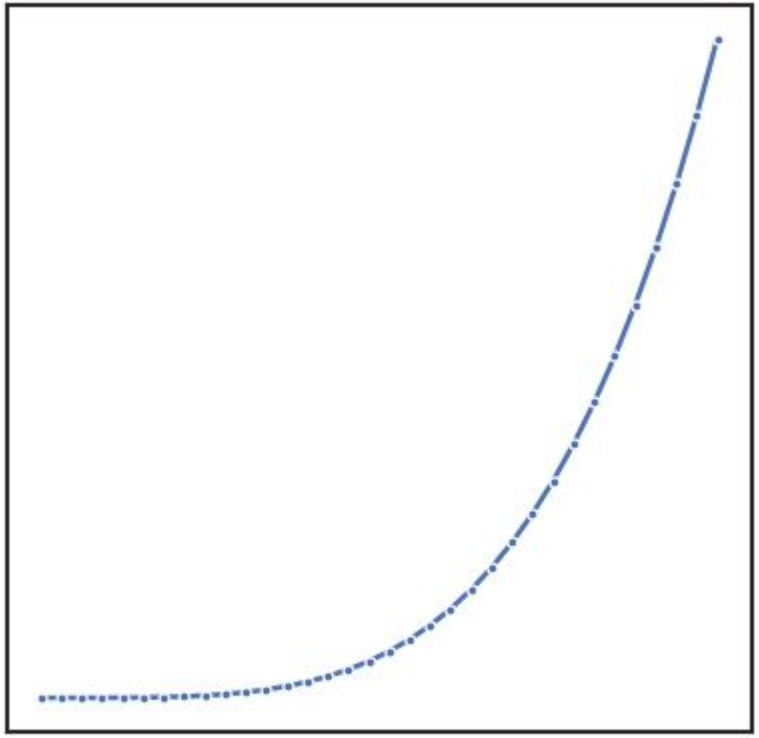
$$\rho_{x,y} = \frac{s_{x,y}}{s_x s_y}$$

Correlação de Spearman

$$\rho_{\mathrm{rank}(x),\mathrm{rank}(y)} = \frac{s_{\mathrm{rank}(x),\mathrm{rank}(y)}}{s_{\mathrm{rank}}(x)s_{\mathrm{rank}}(y)}$$

Correlação - alguns exemplos





Problemas nas bases de dados - Missing values

Algumas variáveis podem não estar presentes para algumas amostras

Como resolver?

1. Remover as linhas
2. Remover as colunas
3. Usar média, moda, mediana
4. "Sinalizar" o missing (outra var ou valor)
5. Usar um modelo para inferir

Col1	Col2	Col3
1	0.42	5.2
3		2.1
2	0.35	

Problemas nas bases de dados - Outliers

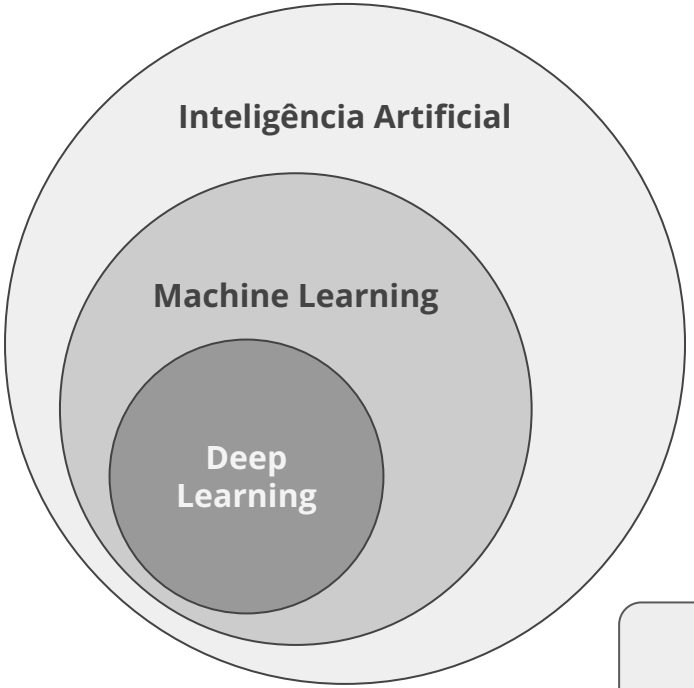
Valores muito fora do "normal"

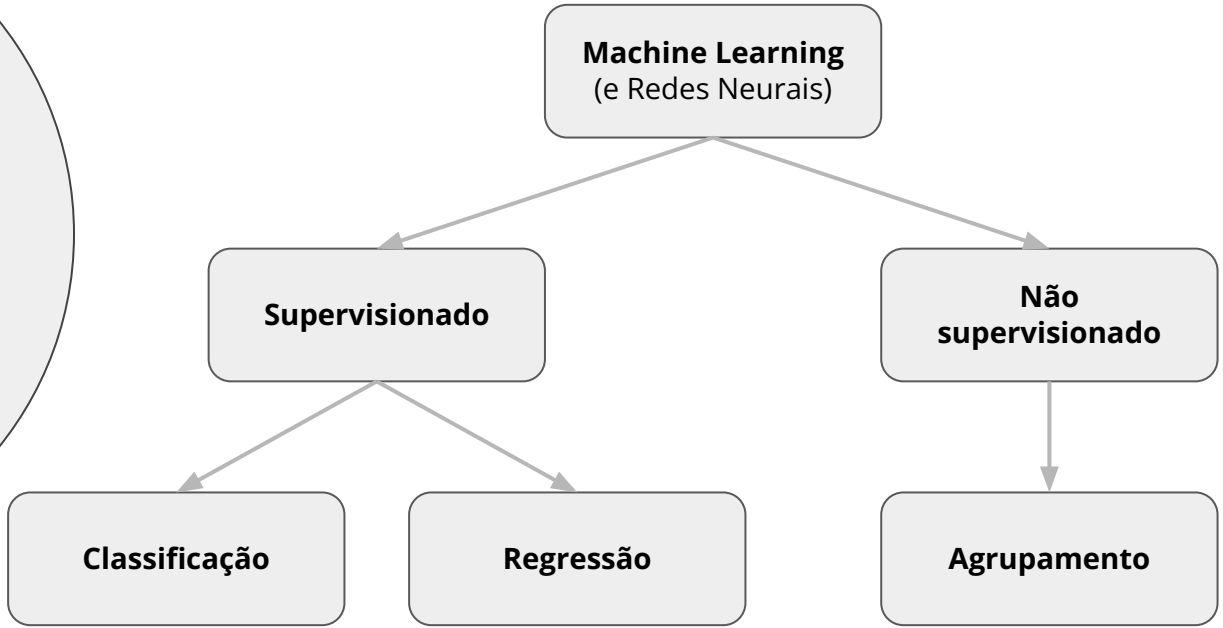
Como resolver?

1. Descartar linhas (remover outliers)
2. Detectar anomalias
3. Detectar novidades

Col1	Col2	Col3
1	0.42	5.2
3	0.21	2.1
2	0.35	35000

Contexto





https://www.venturus.org.br/en/machine-learning-for-laymen/

Aprendizado Supervisionado

- Objetivo: criar um modelo para prever uma variável futura (variável resposta) usando algumas variáveis conhecidas (variáveis explicativas)
- Requerimentos:
  - dados para treinar o modelo e avaliá-lo → variável resposta já obtida
  - uma métrica de sucesso para avaliar o resultado do modelo criado
- Tipos de problemas que resolvemos:
  - Classificação: quando a variável resposta é categórica (discreta)
  - Regressão: quando a variável resposta é contínua

Detectar fraudes em cartões de crédito
(problema de classificação)

Prover o valor de venda de uma casa
(problema de regressão)

Aprendizado Não-supervisionado

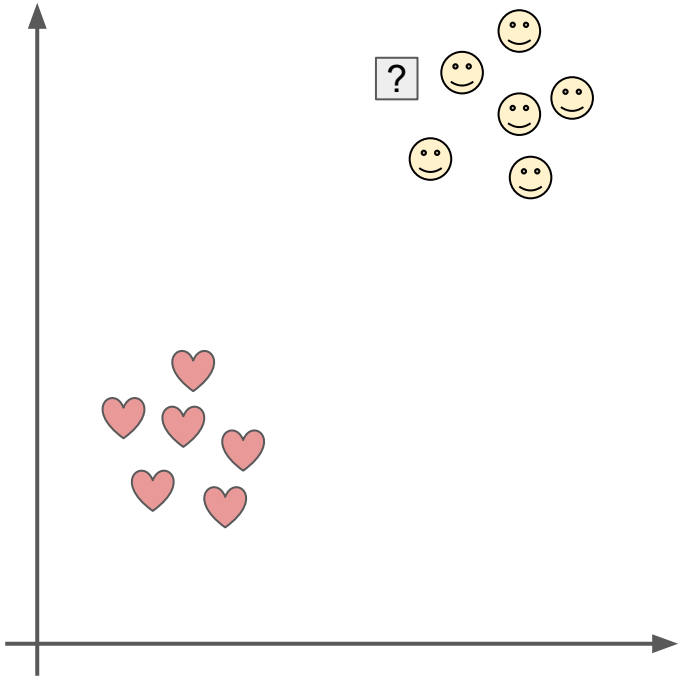
- Objetivo: Descobrir subgrupos (exemplos similares) em um conjunto de dados

- Requerimentos:
  - dados de treinamento para treinar o modelo e avaliá-lo
  - **NÃO** precisa de uma variável resposta

Segmentar grupos de clientes

Sistemas de recomendação de produtos

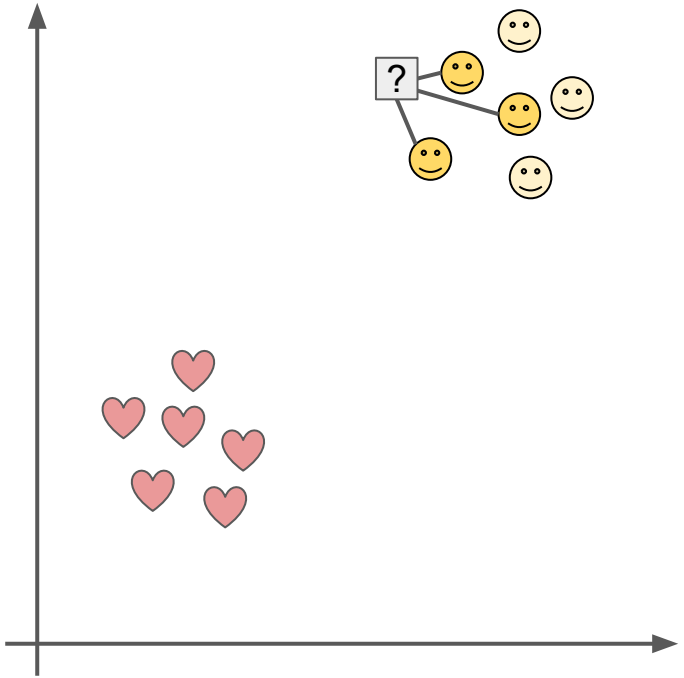
K-Nearest Neighbors (kNN) - Classificação



Intuitivamente falando...

- Qual é a classe do novo elemento?
- Porque?

K-Nearest Neighbors (kNN) - Classificação



Hiperparâmetros:

- k → número de vizinhos mais próximos a serem olhados

Treinamento:

- Salva a base de treinamento → lazy learner

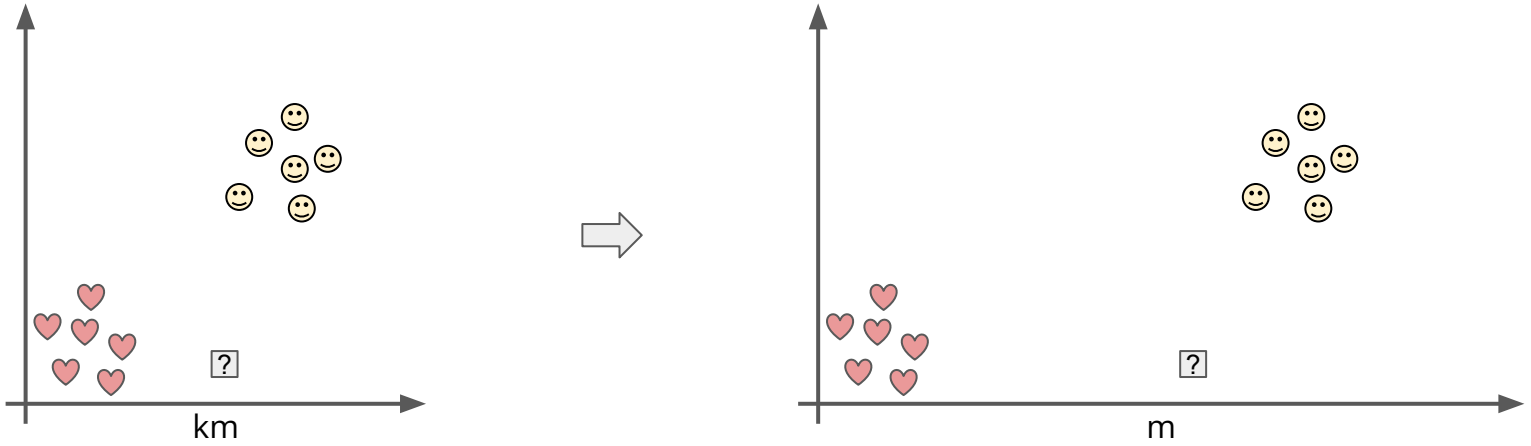
Inferência (previsão para novos dados):

- Encontrar os **k** vizinhos mais próximos na base de treino
- Atribui a classe da maioria dos vizinhos ao novo exemplo
- Proporção dos vizinhos de cada classe pode ser usada para computar probabilidades

Problemas do kNN

1. Custo computacional: comparar com a base toda (índices ajudam)

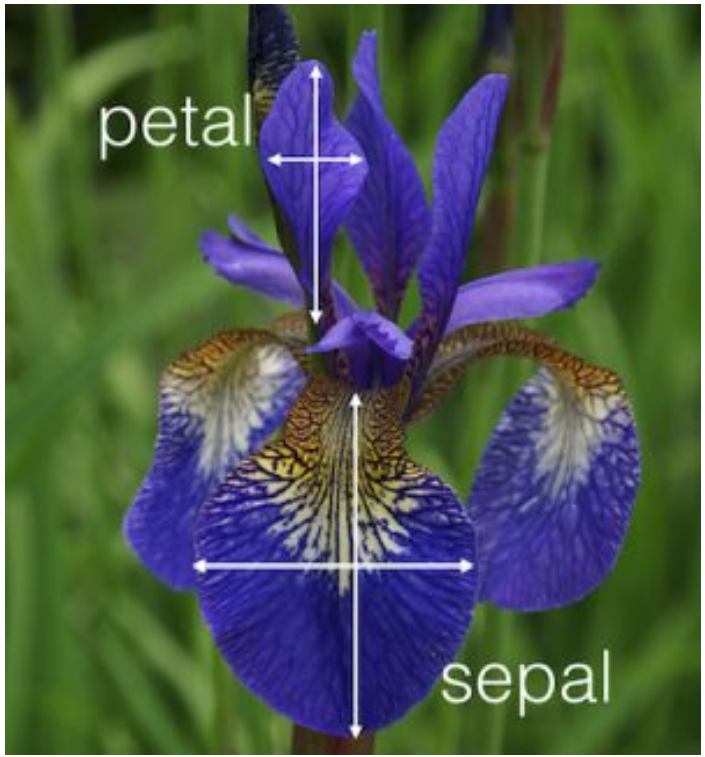
2. Variáveis com diferentes escalas (ou irrelevantes) podem atrapalhar
   - Transformar todas as variáveis para a mesma escala
   - Remover variáveis irrelevantes



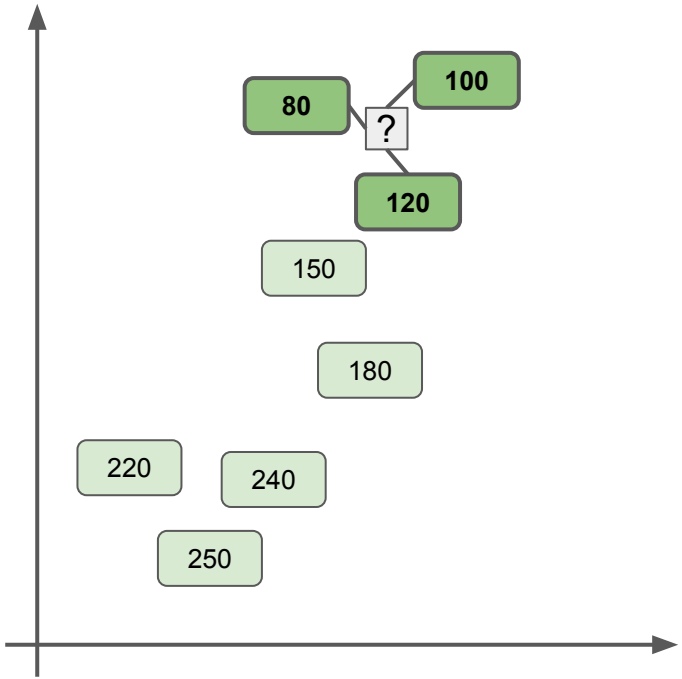
Prática no Python... junto com os conceitos

Notebook para estudo:

- knn_classification.ipynb



K-Nearest Neighbors (kNN) - Regressão



Hiperparâmetros:

- k → número de vizinhos mais próximos a serem olhados

Treinamento:

- Salva a base de treinamento → lazy learner

Inferência (previsão para novos dados):

- Encontrar os **k** vizinhos mais próximos na base de treino
- Atribui a média dos valores da variável resposta dos vizinhos ao novo exemplo

Como medir a qualidade do modelo supervisionado

1. Usar dados não vistos pelo modelos
   - Base de treinamento: experimentamos com vários modelos e hiperparâmetros
     - podemos subdividir em treino e validação, para otimizar os hiperparâmetros
   - Base de teste: fica reservada até que o melhor modelo seja escolhido

2. Escolher uma métrica apropriada ao problema
   - Classificação:
     - Acurácia: % de acertos na base onde o modelo é avaliado $Acc = \frac{\#\text{acertos}}{\#\text{total de exemplos}}$
     - AUC: Area Under The Curve → dados desbalanceados
   - Regressão:
     - MSE, MAE, ...
     - nem sempre o resultado tem tanto significado prático
     - é comum fazer outros estudos para determinar a qualidade do modelo

Vamos praticar um pouco...

Classificação de dados médicos

Notebook para estudo:

- knn_classification.ipynb

Previsão do valor de imóveis

Notebook para estudo:

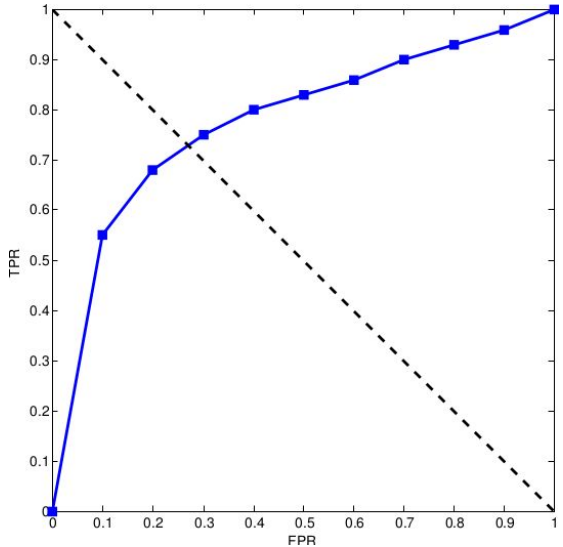
- knn_regression.ipynb

Problemas com dados desbalanceados

Um modelo com a acurácia alta é sempre bom?

- Não, ele pode sempre classificar novos dados com a classe majoritária

AUC pode ajudar nesses casos:



$$\begin{cases} \mathrm{TPR} = \dfrac{\mathrm{TP}}{\mathrm{P}} = \dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \\ \mathrm{FPR} = \dfrac{\mathrm{FP}}{\mathrm{N}} = \dfrac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \end{cases}$$

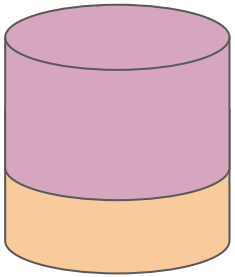
Como escolher os valores dos hiperparâmetros

Testar várias combinações

- Grid-search, Randomized search, ...
- Só usar os dados de treinamento para isso (**teste fica guardado**)

Como usar os dados de treinamento:

Hold-out



K-fold cross validation

Fold 1	Fold 2	Fold 3
Fold 1	Fold 2	Fold 3
Fold 1	Fold 2	Fold 3

Dados de treinamento
Dados de validação (avaliar o modelo)

K-means - Agrupamento

Inicialização:

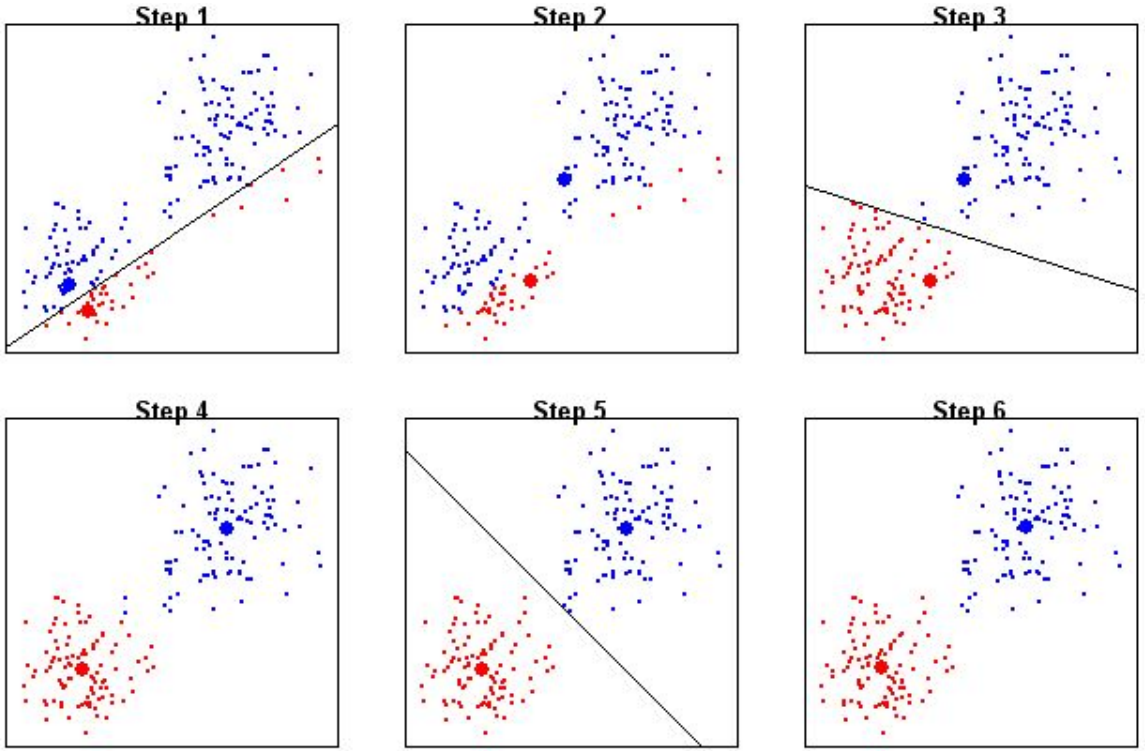
- Criar **k** centróides aleatórios

Iteração:

- Associar os exemplos ao centróide mais próximo
- Atualizar os centróides → média dos elementos associados a ele

Outras versões:

- K -medoids



https://animoidin.wordpress.com/2018/07/20/k-means-clustering/

K-means - Limitações

Escolher o valor de k

- Podemos usar a silhueta

Grupos com formatos não hiper esféricos ou com outliers

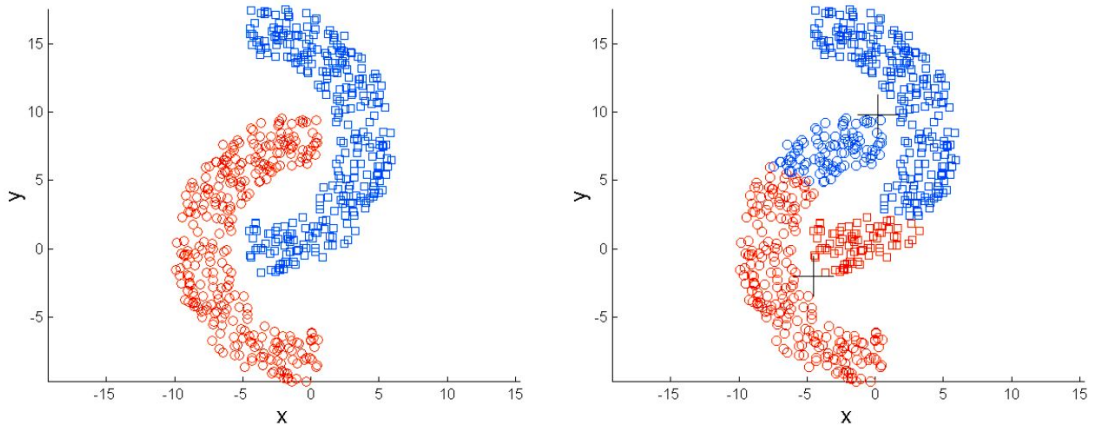
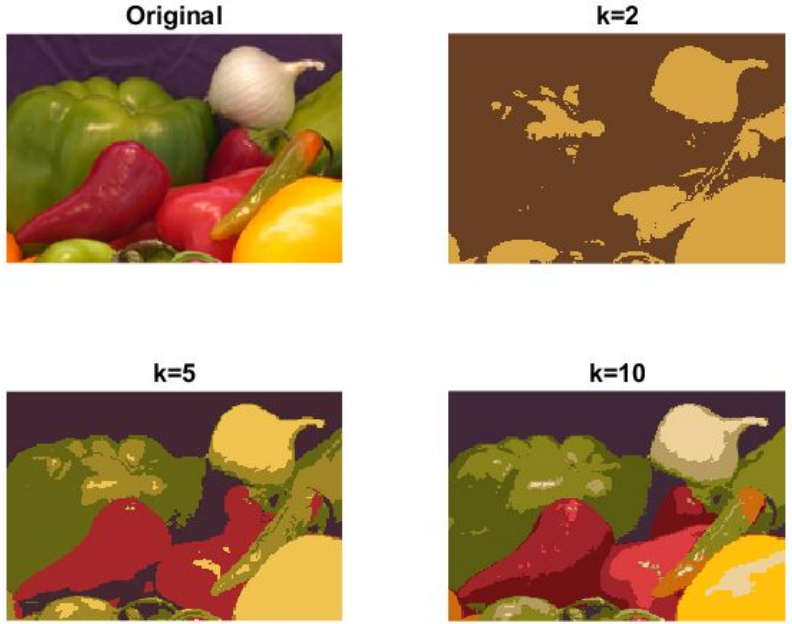


Imagem das notas de aula do Prof. Dr. Rodrigo Mello

Quantização de imagens

Notebook para estudo:

- kmeans_img_quantization.ipynb



https://www.ics.uci.edu/~yug10/uci/cs216/hw3/solution.md.html