Introdução a Machine Learning

Lilianne M. I. Nakazono | pyladies SP | FIAP Paulista

Lilianne M. I. Nakazono

Formada em **Estatística** (IME-USP) e em **Astronomia** (IAG-USP). Doutoranda em **Astronomia** (IAG-USP) com foco em aplicações de Machine Learning e análises estatísticas. Eu procuro quasares! :)

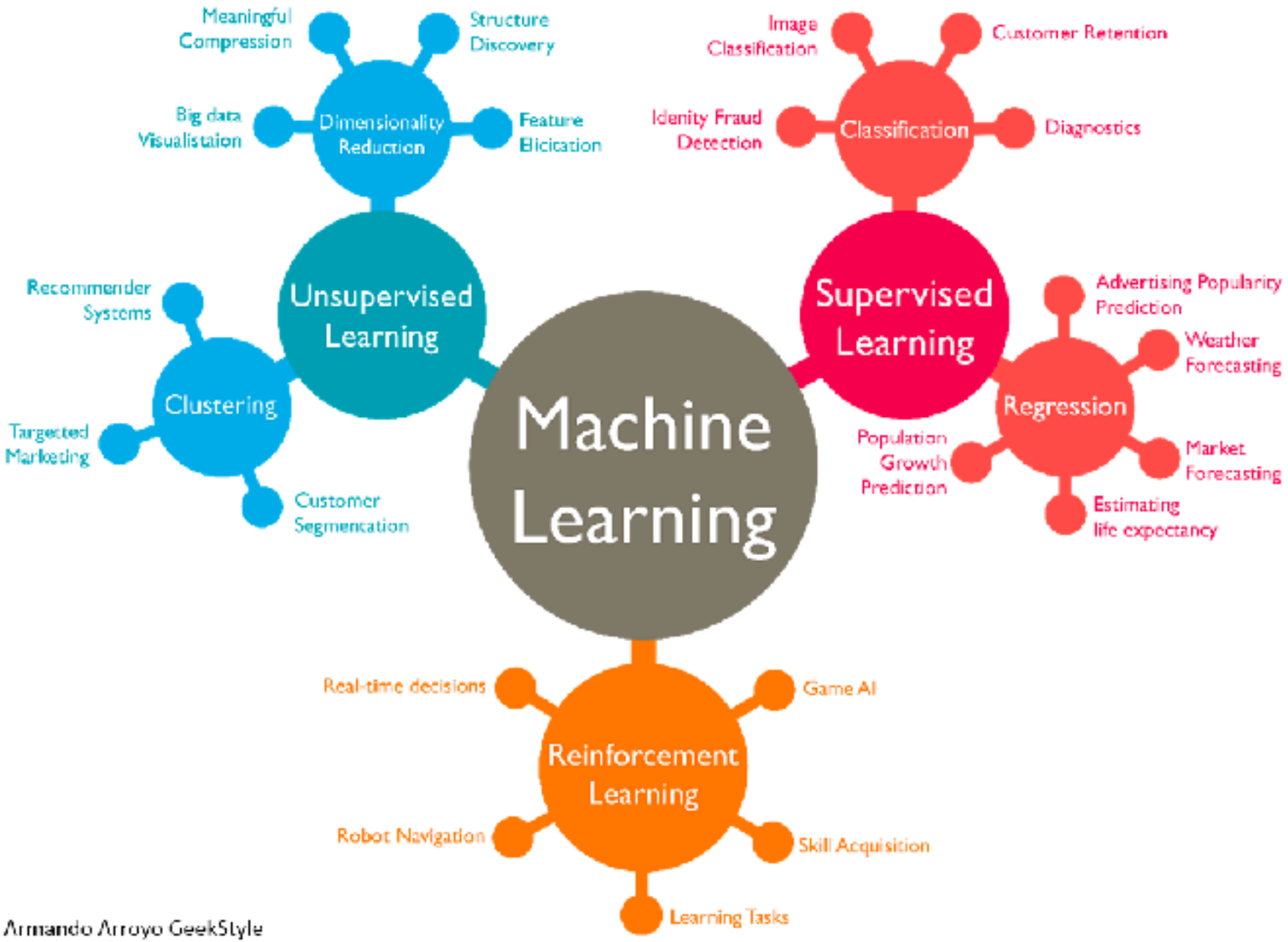
lilianne.nakazono@usp.br

github.com/marixko

@li_nkzd

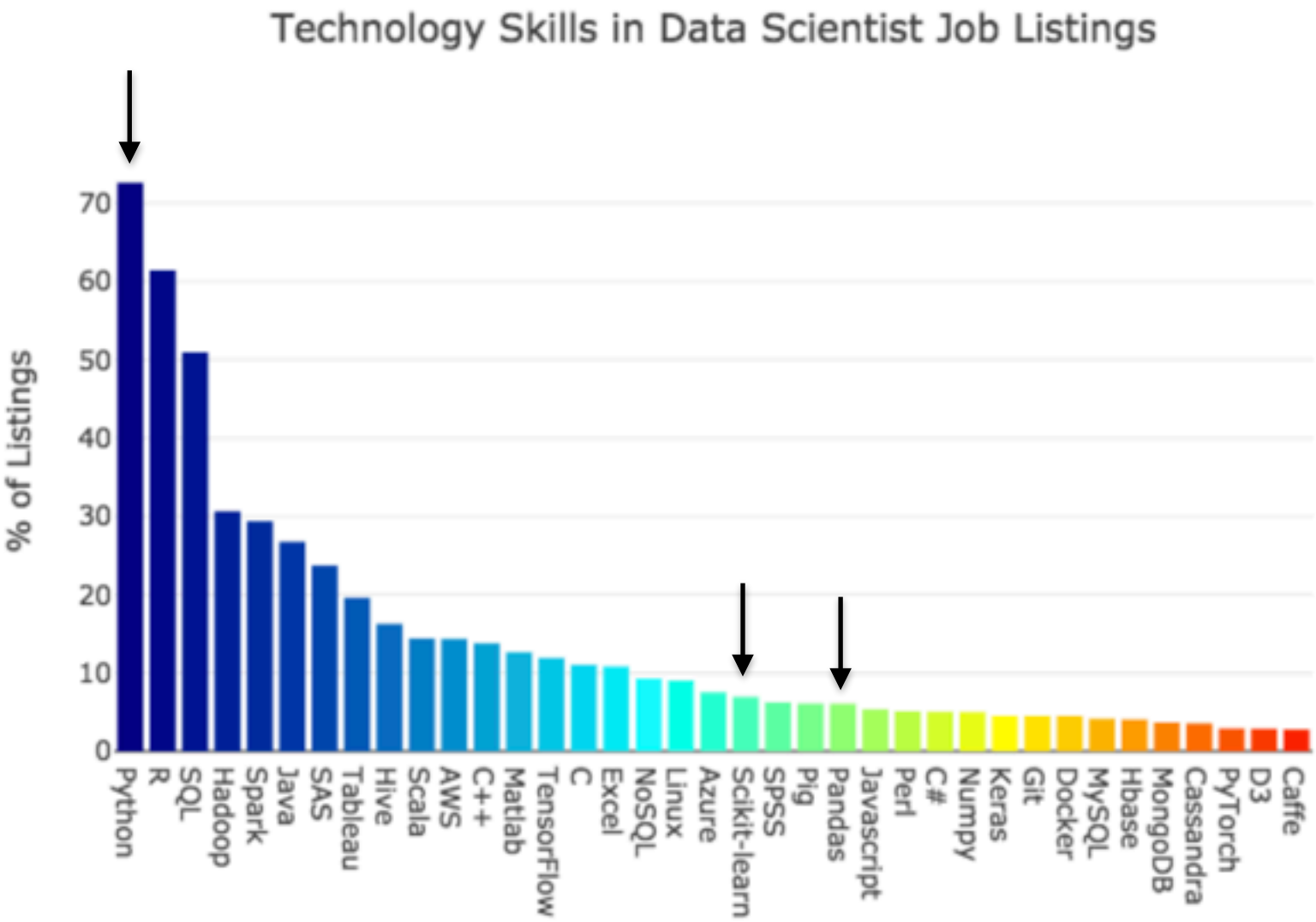
“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.”

- Tom Mitchell



Armando Arroyo GeekStyle

Quais softwares usar?



https://towardsdatascience.com/the-most-in-demand-skills-for-data-scientists-4a4a8db896db

ML está em todo lugar

Muito avanço na Ciência e Tecnologia se deve ao desenvolvimento de novos métodos de análise.

Machine Learning tem sido disruptivo em muitas áreas, por exemplo na Medicina, Biologia, Astronomia, entre outros.

Maryana Alegro, Ph.D.

“Uso deep learning pra mapear a proteína Tau, que é altamente correlacionada com declínio cognitivo em Alzheimer, em datasets de patologia de cérebros humanos inteiros.”

Mais detalhes: https://www.biorxiv.org/content/10.1101/698902v1.abstract

Dra. Maryana Alegro, pesquisadora na University of California, San Francisco (UCSF)

Hoje vocês vão fazer ciência!

HANDS-ON

Acessem:

https://github.com/marixko/workshop_pyladies

Sintam-se à vontade para codar junto ou apenas prestar atenção!

Dica:

StackOverflow será seu melhor amigo!

IRIS DATASET

Esse dataset contém medidas da pétala e da sépala de três diferentes espécies do gênero Iris: *Iris setosa*, *Iris virginica* e *Iris versicolor*.

Iris virginica

Iris setosa

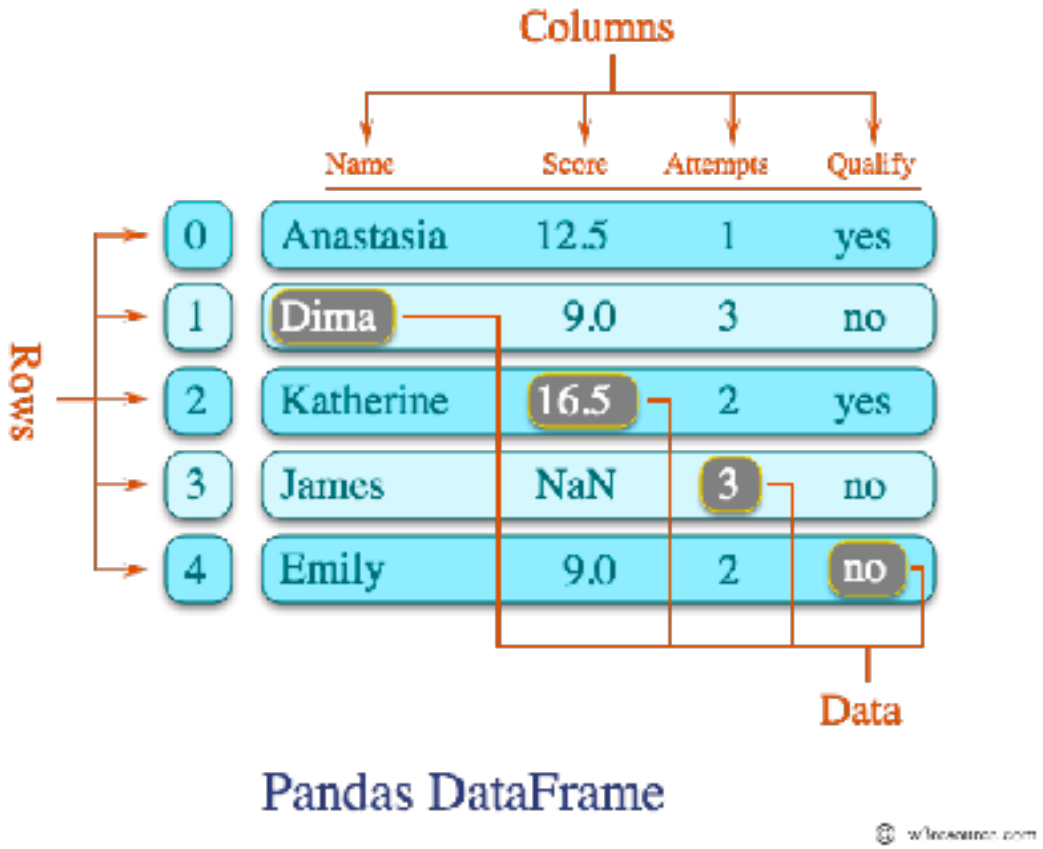
Iris versicolor

Total:
50 amostras de cada espécie

Atributos:
comprimento e largura da sépala, comprimento e largura da pétala

Dataframe?

Por definição, dataframe se refere a dados estruturados em duas dimensões, i.e. em linhas e colunas (ex: planilhas do excel)



Dataframe?

Por definição, dataframe se refere a dados estruturados em duas dimensões, i.e. em linhas e colunas (ex: planilhas do excel)

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
5	5.4	3.9	1.7	0.4
6	4.6	3.4	1.4	0.3
7	5.0	3.4	1.5	0.2
8	4.4	2.9	1.4	0.2
9	4.9	3.1	1.5	0.1
10	5.4	3.7	1.5	0.2

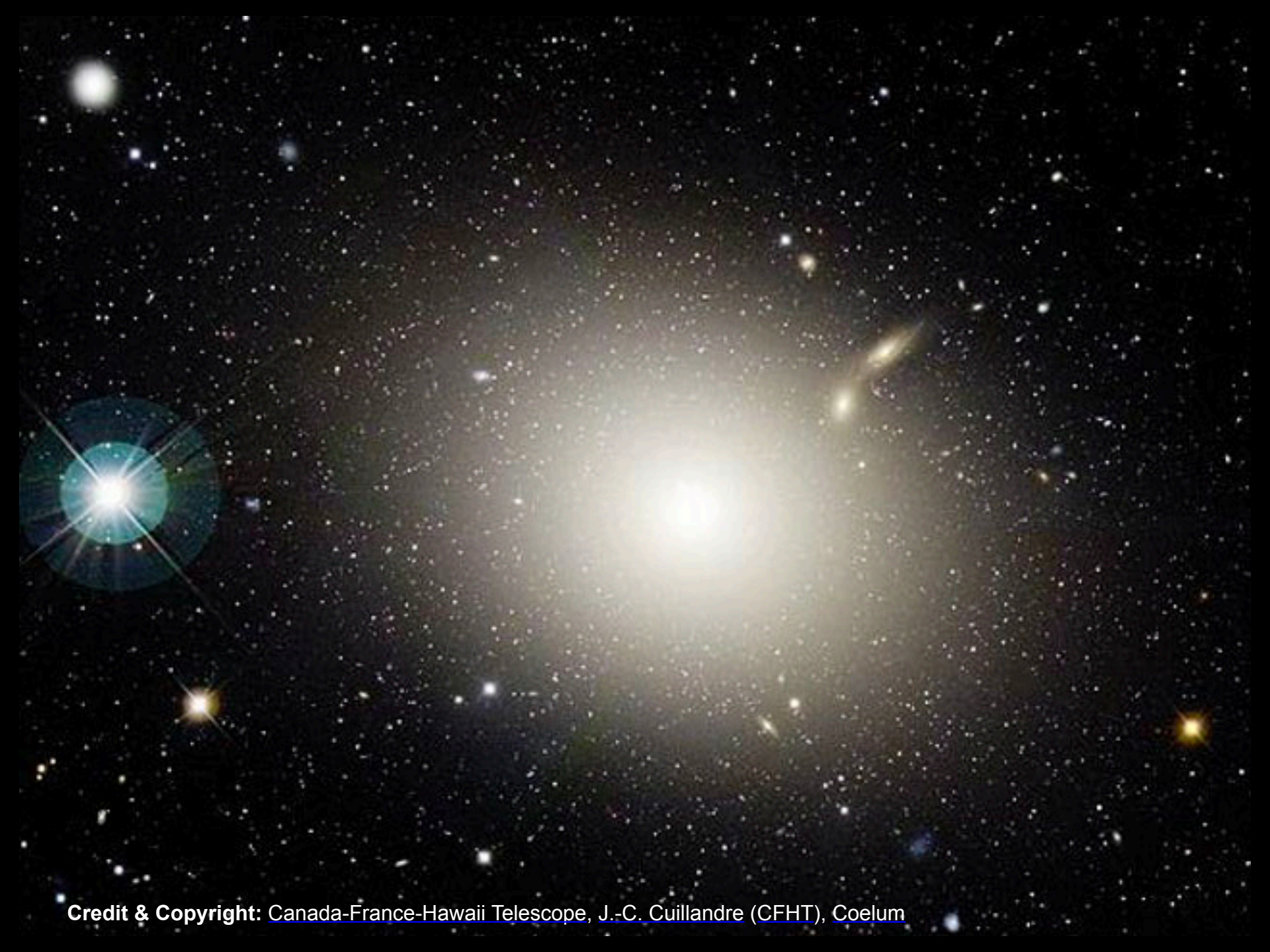
Astrônomas por um dia

Hoje em dia existem vários telescópios mapeando o céu, coletando um número gigantesco de imagens

Problema

Como classificar automaticamente estrelas e galáxias?

Image Credit: Subaru Telescope (NAOJ), Hubble Space Telescope, European Southern Observatory - **Processing & Copyright:** Robert Gendler



Credit & Copyright: Canada-France-Hawaii Telescope, J.-C. Cuillandre (CFHT), Coelum


J-PLUS
survey

Classificação estrelas e galáxias

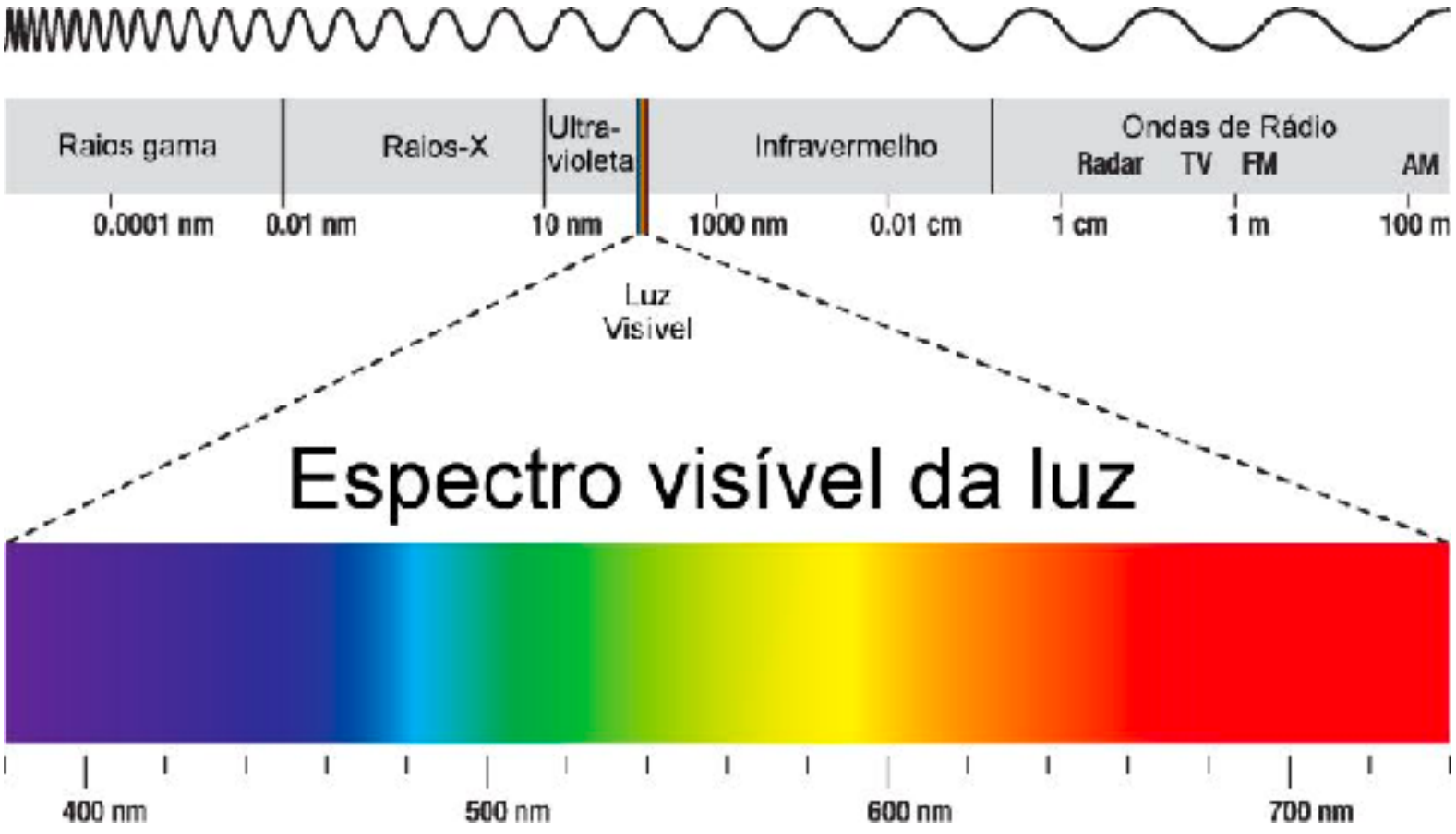
Dificuldades:

- Galáxias mais fracas e distantes podem ser facilmente confundidas como estrelas
- Resolução do telescópio e outros problemas sistemáticos

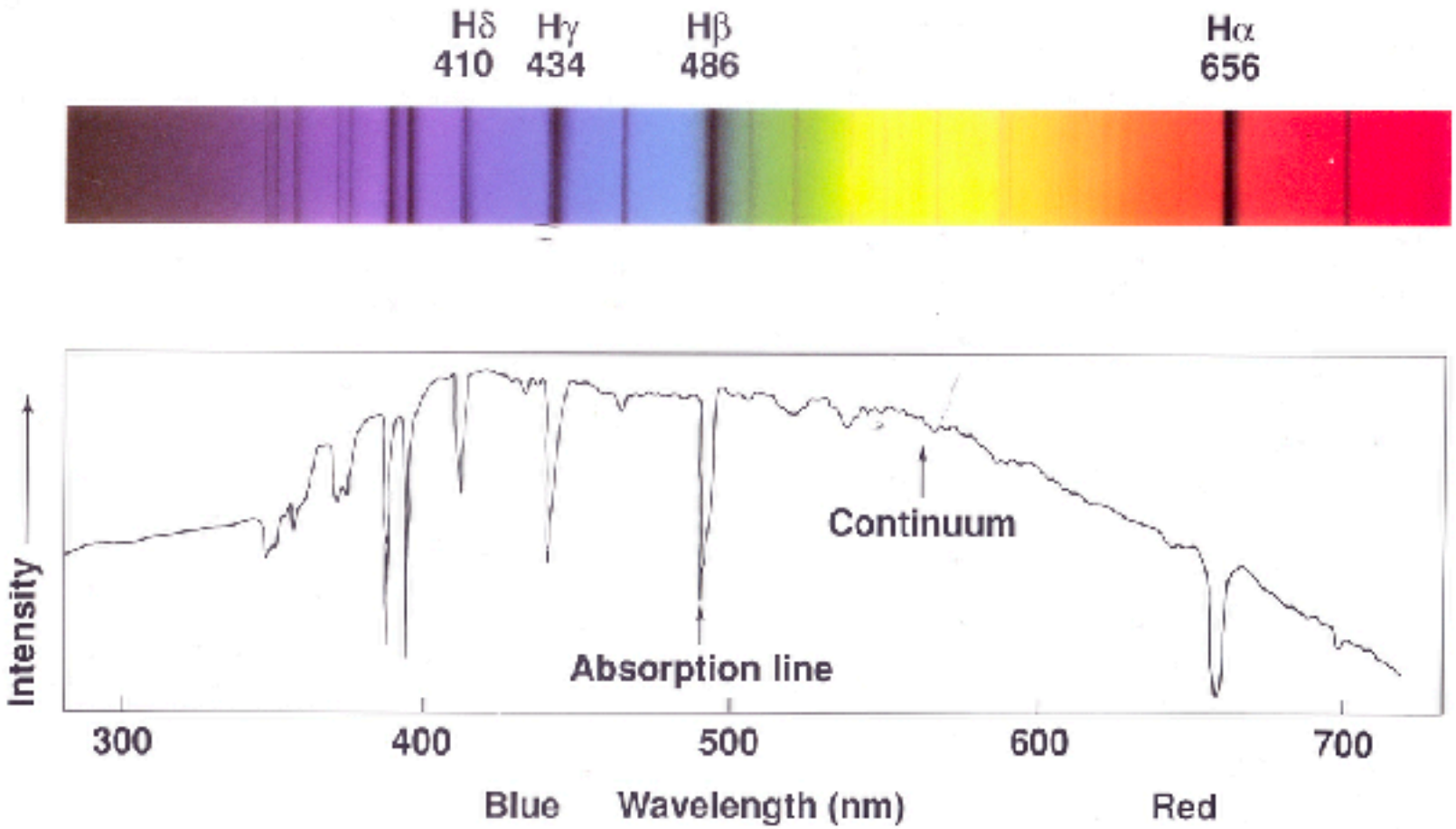
Dataset:
Galáxias e estrelas conhecidas

Atributos:
FWHM, semi-eixo maior, semi-eixo menor e distribuição de energia por comprimento de onda

Distribuição de energia



Distribuição de energia



Visual portion of stellar spectrogram

Hartmann/The Cosmic Journey, 4th ed., Fig. 16-5; The Cosmic Voyage, Fig. 16.3

Distribuição de energia

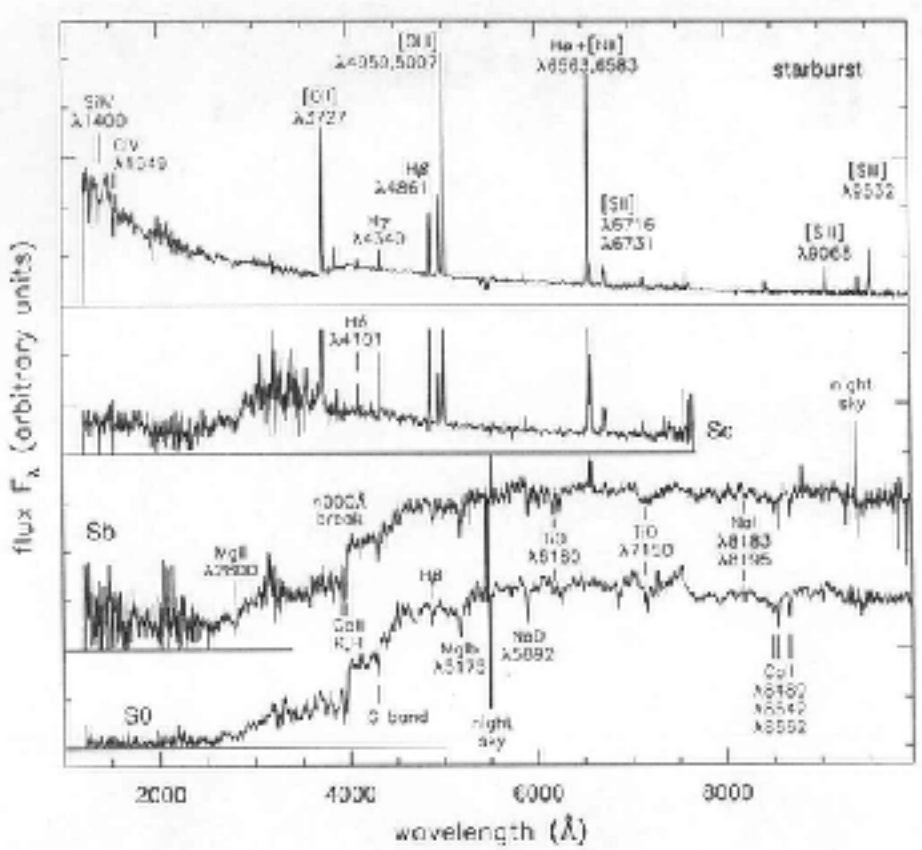
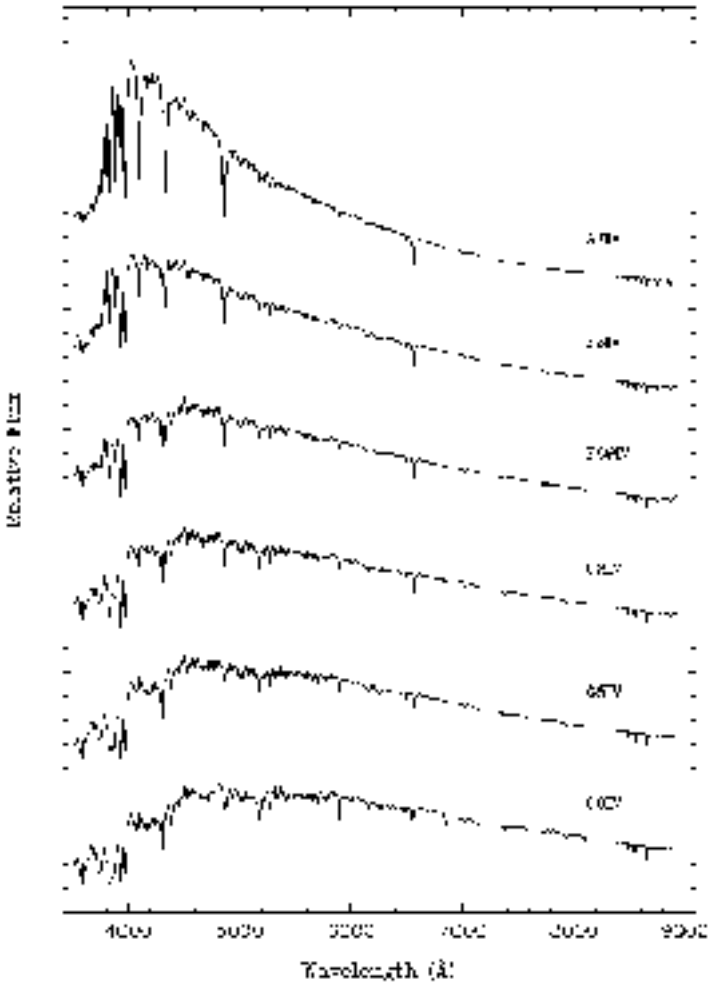


Figure 5.24 Spectra of galaxies from ultraviolet to near-infrared wavelengths; incompletely removed emission lines from the night sky are marked. From below: a red S0 spectrum; a bluer Sb galaxy; an Sc spectrum showing blue and near-ultraviolet light from hot young stars, and gas emission lines; a blue starburst galaxy, that has made many of its stars in the past 100 Myr – A. Kinney.



Classificação estrelas e galáxias

Dificuldades:

- Galáxias mais fracas e distantes podem ser facilmente confundidas como estrelas
- Resolução do telescópio e outros problemas sistemáticos

Dataset:
Galáxias e estrelas conhecidas

Atributos:
FWHM, semi-eixo maior, semi-eixo menor e distribuição de energia por comprimento de onda

HANDS-ON

Leia o arquivo tutorial_data.txt usando pandas e chequem a tabela com um print. Chequem que tipo de informação esse dataset. Façam um .describe().T para verificar seus dados. Usando o que viram hoje, tentem responder as seguintes perguntas:

1. Existem missing values?
2. Esse dataset tem quantas galáxias e quantas estrelas?
3. Como é a distribuição de r_auto das galáxias? E das estrelas?
4. Qual é a média e desvio padrão de r_auto das galáxias? E das estrelas?
5. Considerando apenas FWHM, A e B, faça um sns.pairplot por classe. O que você conclui?
6. Use model_selection.train_test_split do sklearn para dividir seu dataset em amostra de treinamento e de teste:

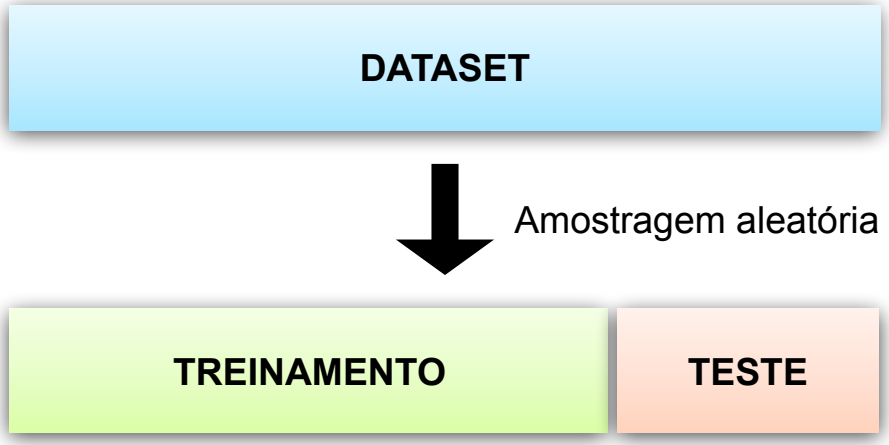
   X_train, X_test, y_train, y_test = train_test_split(*[complete!]*)

7. Use novamente o train_test_split() para separar seu X_train e y_train em treinamento e validação (Cuidado! Lembre que X_train e y_train são dados pareados)

Validação cruzada

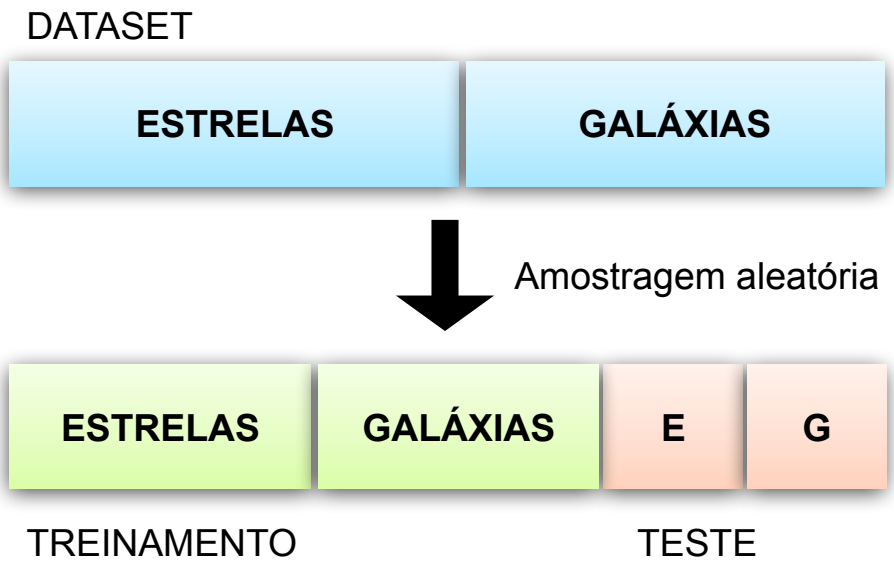
Validação cruzada

Uma forma simples de validar seu modelo: **Holdout**



Amostragem estratificada

É interessante manter a proporção de cada grupo durante a amostragem:



Matriz de confusão

Como quantificar a performance do seu modelo?

		PREDITO PELO MODELO	
		ESTRELA	**GALÁXIA**
VERDADEIRO	**ESTRELA**	VERDADEIRO POSITIVO (VP)	FALSO NEGATIVO (FN)
	GALÁXIA	FALSO POSITIVO (FP)	VERDADEIRO NEGATIVO (VN)

Matriz de confusão

Como quantificar a performance do seu modelo?

- **Acurácia**

  (VP + VN) / Total

- **Precisão (+)**

  VP / (VP + FP)

- **Recall (+)**

  VP / (VP + FN)

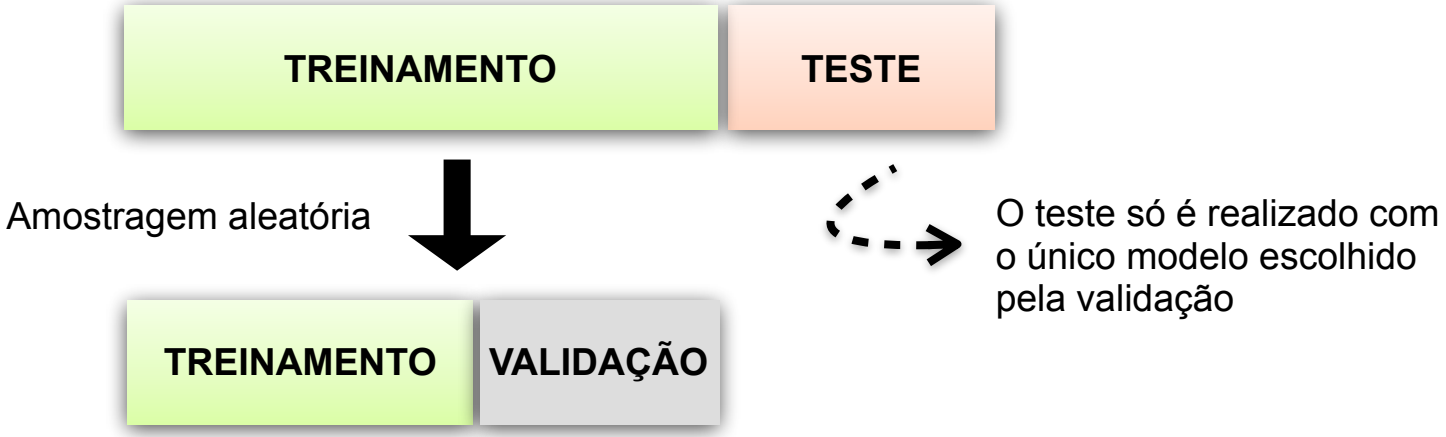
- **F-score**

  2 (Precision x Recall) / (Precision + Recall)

		PREDITO PELO MODELO	
		ESTRELA	**GALÁXIA**
VERDADEIRO	**ESTRELA**	VERDADEIRO POSITIVO (VP)	FALSO NEGATIVO (FN)
	GALÁXIA	FALSO POSITIVO (FP)	VERDADEIRO NEGATIVO (VN)

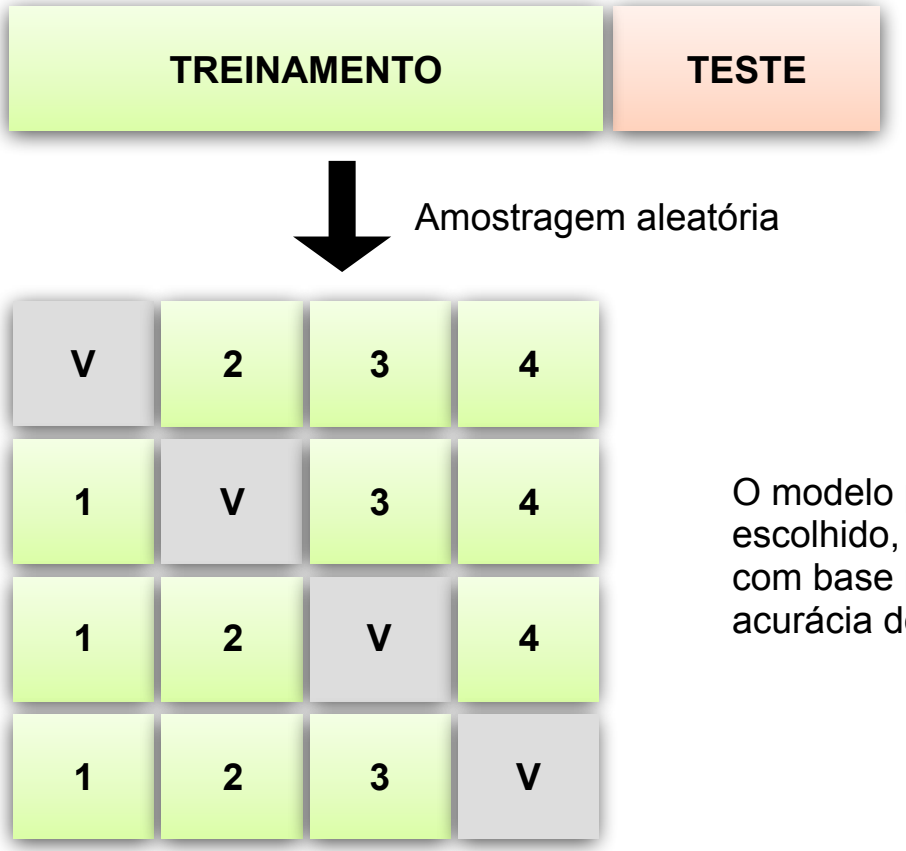
Validação cruzada

Para comparar a performance de diversos modelos (por exemplo, com diferentes parâmetros), é importante adicionar mais uma etapa: a validação



Validação cruzada

Uma forma mais robusta é o **k-fold** (ex: k = 4):



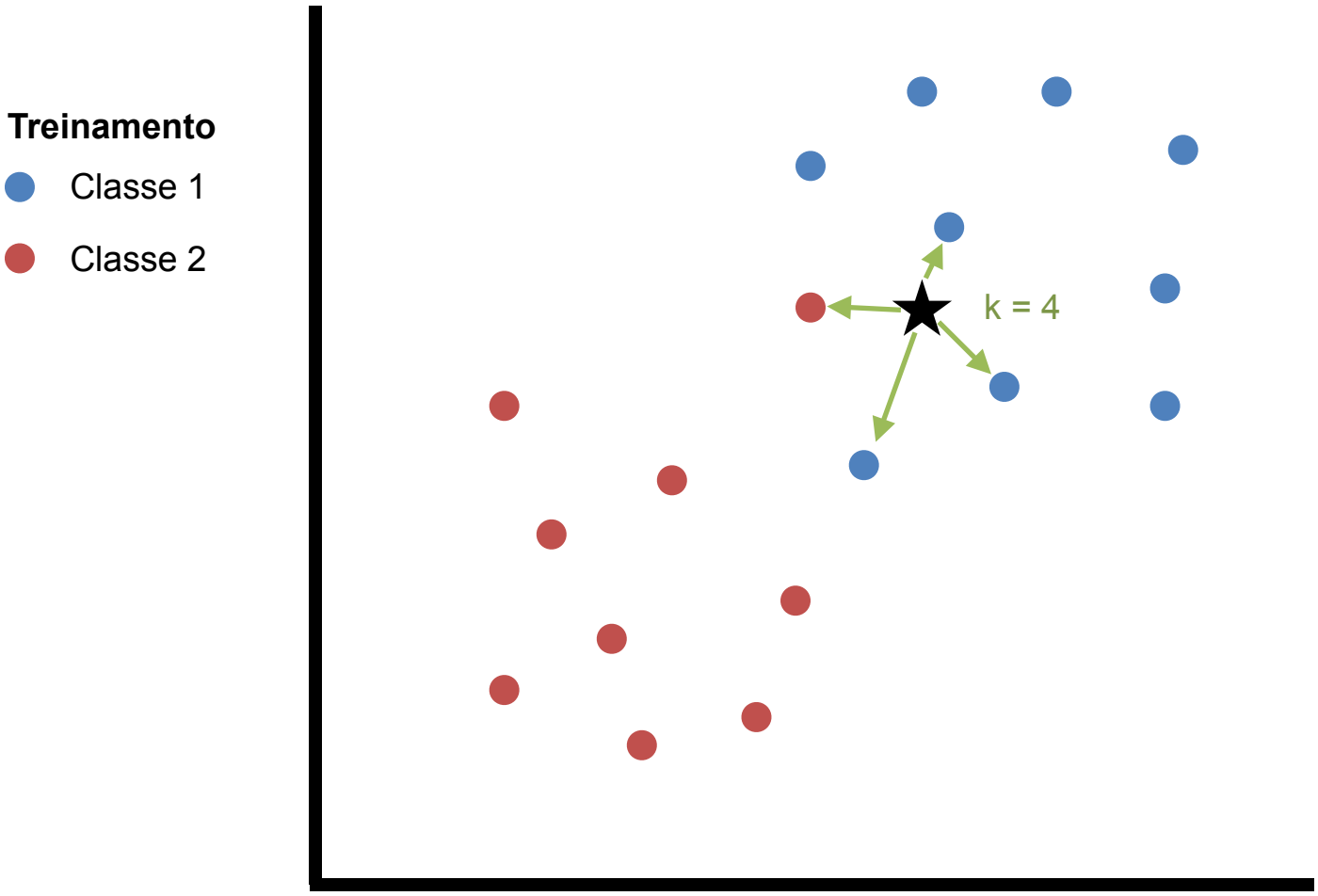
O modelo pode ser escolhido, por exemplo, com base na média da acurácia dos 4 fits

Estratégia

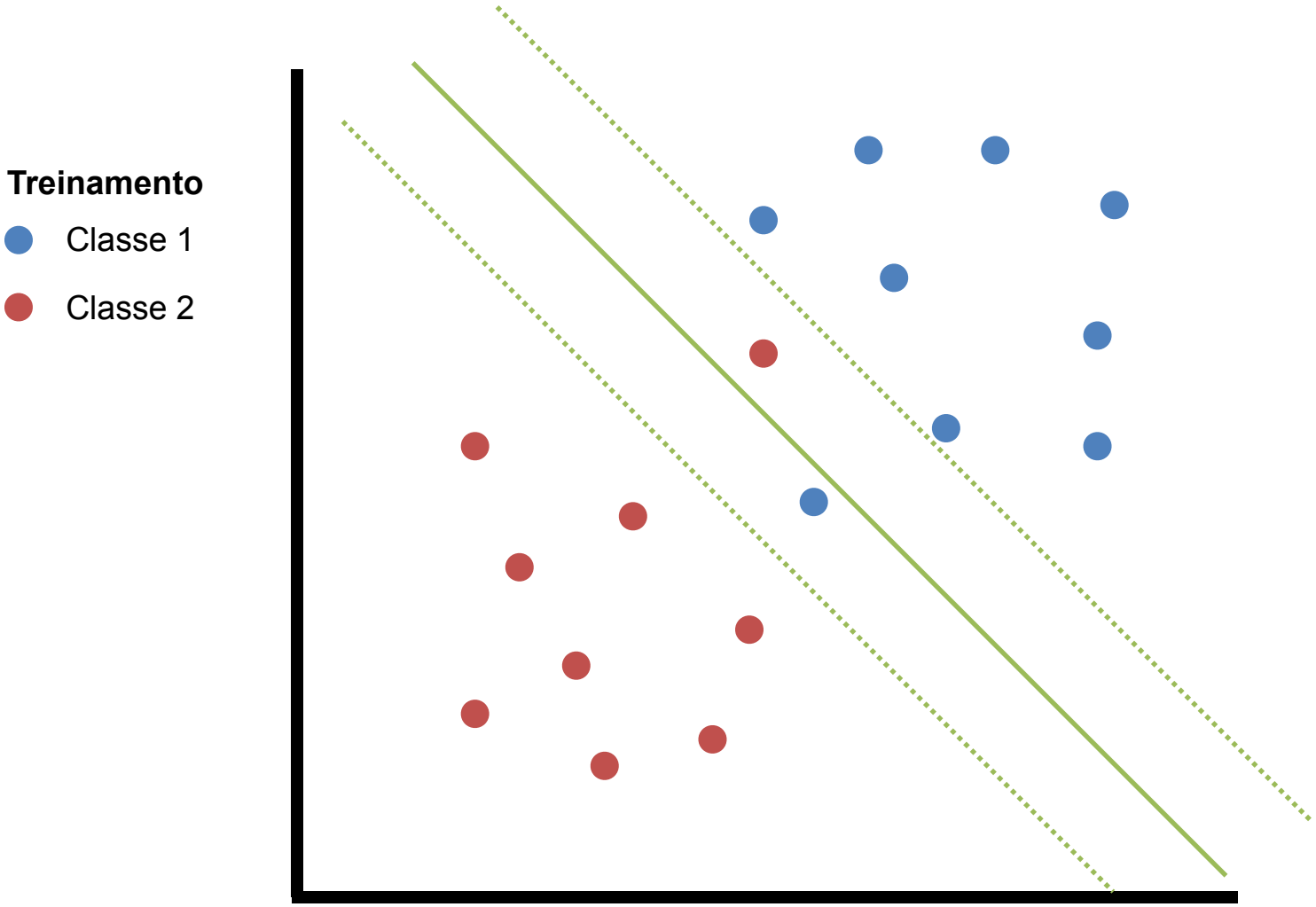
1. Divida seu dataset em amostra de treinamento e de teste de forma aleatória (e estratificada, se for o caso)
2. Caso vá testar diversos modelos, separe uma parte da sua amostra de treinamento para validação
3. Escolha alguma(s) métrica(s) para decidir qual modelo teve a melhor performance. A escolha da métrica deve fazer sentido com o contexto do seu problema
4. Após escolher o melhor modelo, faça o teste final. É deste teste que você terá uma estimativa mais realista do quão assertivas serão suas predições
5. Treine novamente seu modelo escolhido com todo seu dataset
6. Faça suas predições!

Algoritmos de classificação supervisionada

k-Nearest Neighbours (kNN)

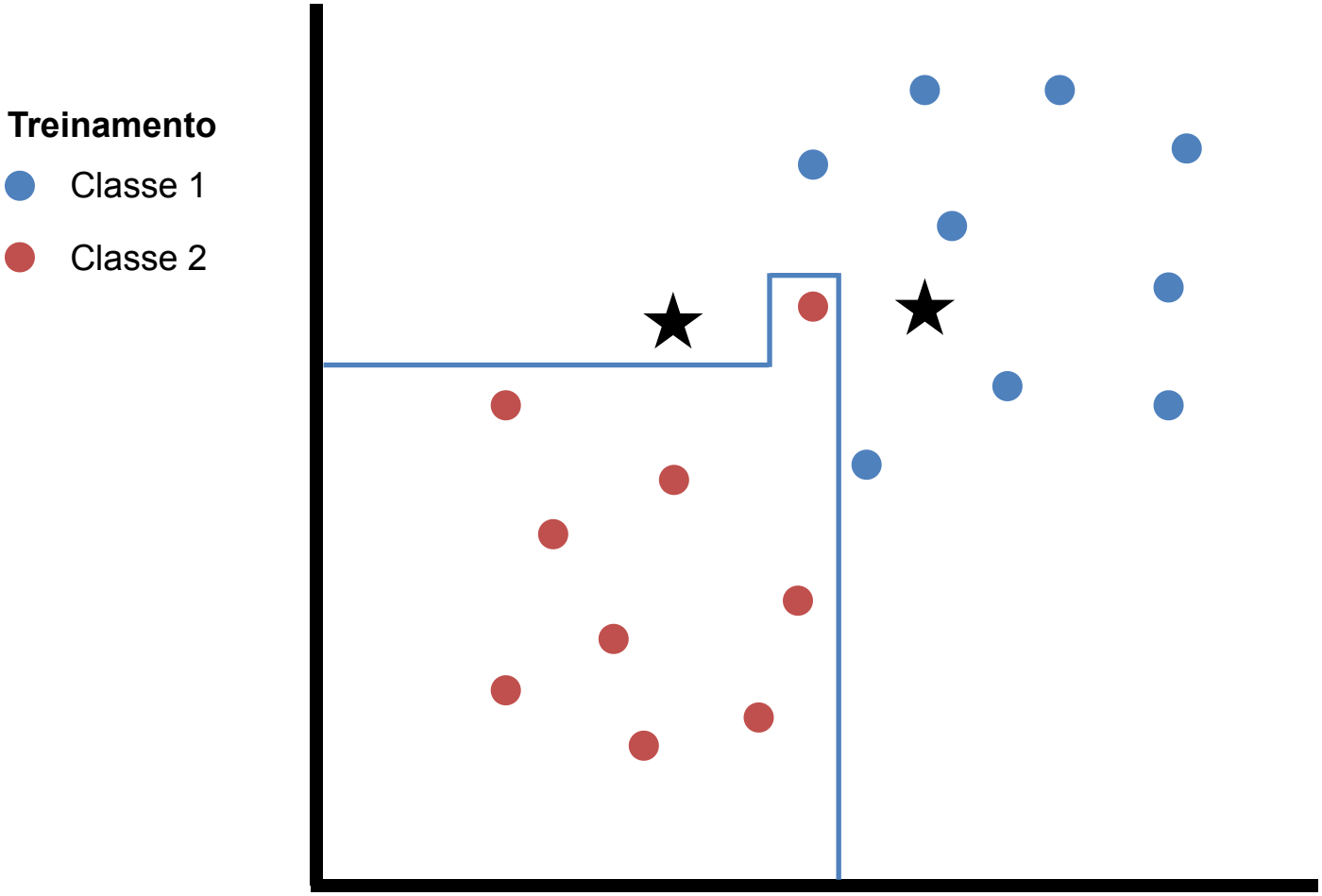


Support Vector Machine (SVM)



A separação não precisa ser necessariamente linear
Para tanto, modifica-se o kernel

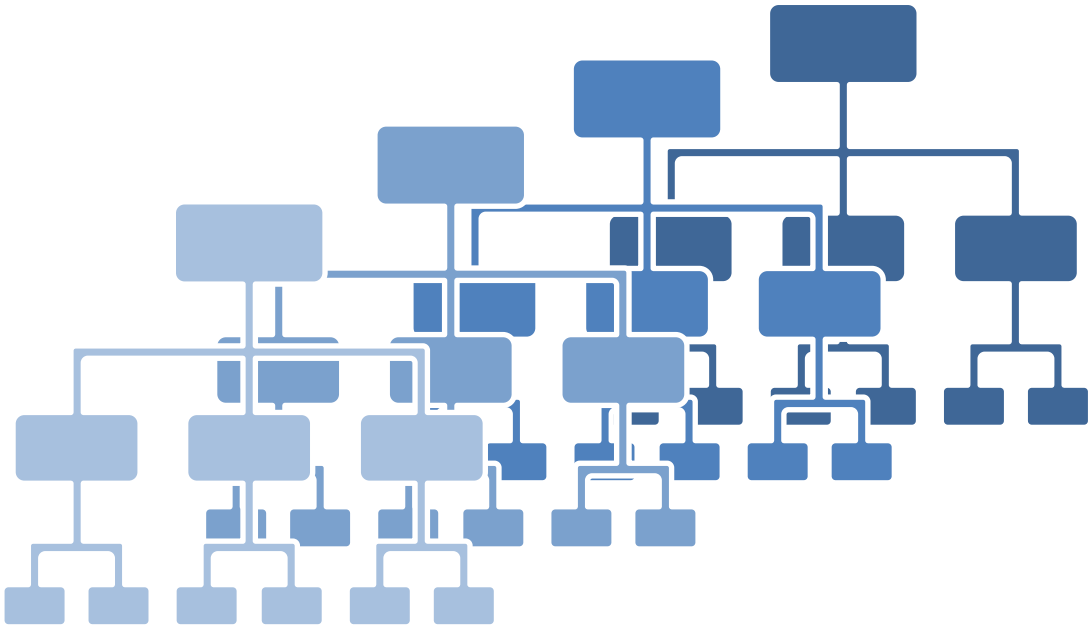
Decision Trees



Random Forest

Várias árvores de decisão construídas de forma independente e aleatoriamente* constituem um Random Forest

A classificação é dada pelo voto majoritário de todas as árvores



HANDS-ON

1. Use model_selection.train_test_split do sklearn para dividir seu dataset em amostra de treinamento e de teste:

   X_train, X_test, y_train, y_test = train_test_split(*[complete!]*)

2. Use novamente o train_test_split() para separar seu X_train e y_train em treinamento e validação (Cuidado! Lembre que X_train e y_train são dados pareados)

HANDS-ON

1. Escolha pelo menos um algoritmo:
   tree.DecisionTreeClassifier()
   ensemble.RandomForestClassifier()
   neighbours.KNeighborsClassifier()
   svm.SVC()

2. Treinem alguns modelos variando os parâmetros. Use sua amostra de validação para avaliar a performance de cada modelo. Qual deu o melhor resultado?

Considerações Finais

Hoje tentei passar pra vocês um básico de Machine Learning. Aqui eu foquei no raciocínio que se deve ter do momento que vocês recebem um dataset até à validação de modelo.

Nada do que passei é regra absoluta para tudo. Existem ótimas discussões pela internet (e.g. towardsdatascience, reddit, medium) e que recomendo para iniciar a aprofundar no assunto.

Lembrem-se que Machine Learning é uma área em rápido desenvolvimento. O aprendizado da máquina pode ser feito em um curto período de tempo, mas o **nosso** aprendizado é contínuo.