FIAP GRADUAÇÃO

TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

FIAP GRADUAÇÃO

DISRUPTIVE ARCHITECTURES

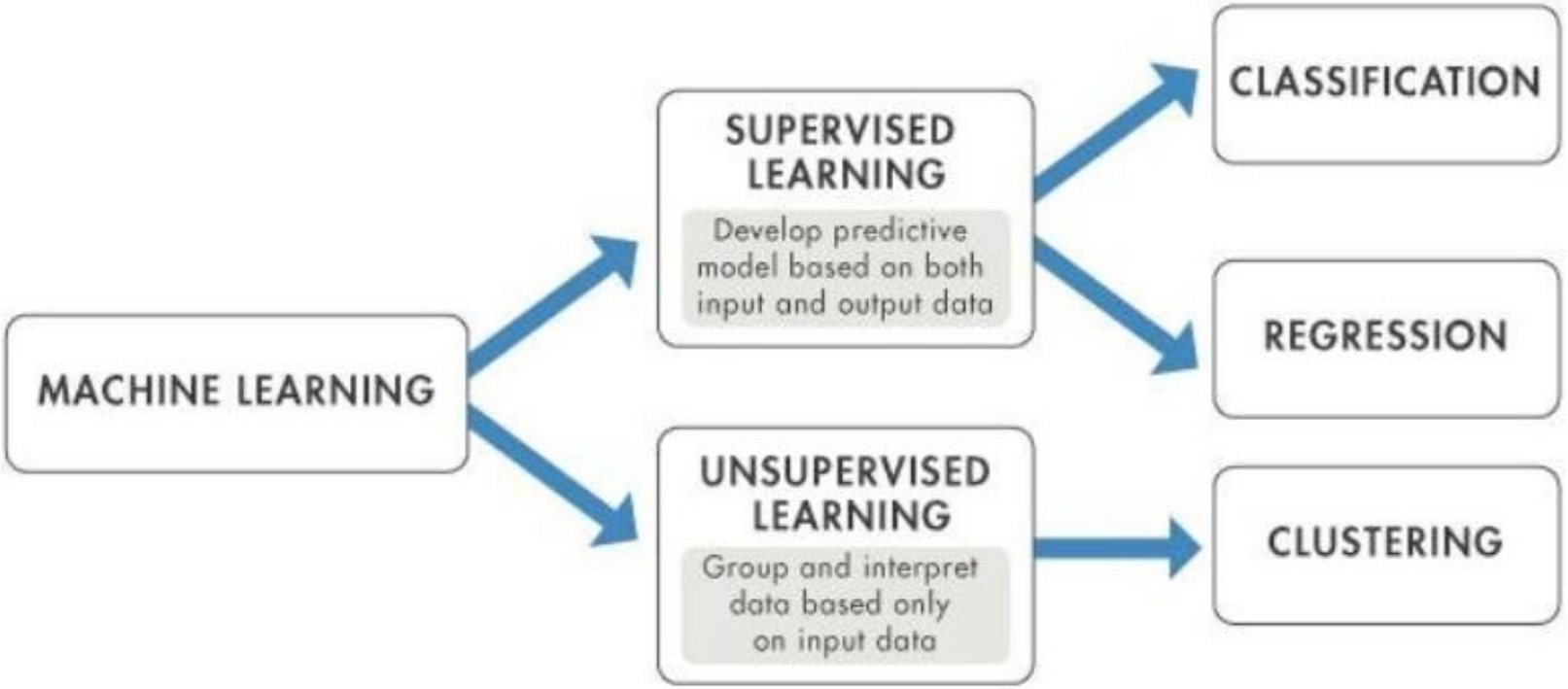
Aprendizado de Máquina Não Supervisionado – Agrupamento (Clustering)

Prof. André Tritiack
profandre.farias@fiap.com.br

2023

Aprendizado de Máquina

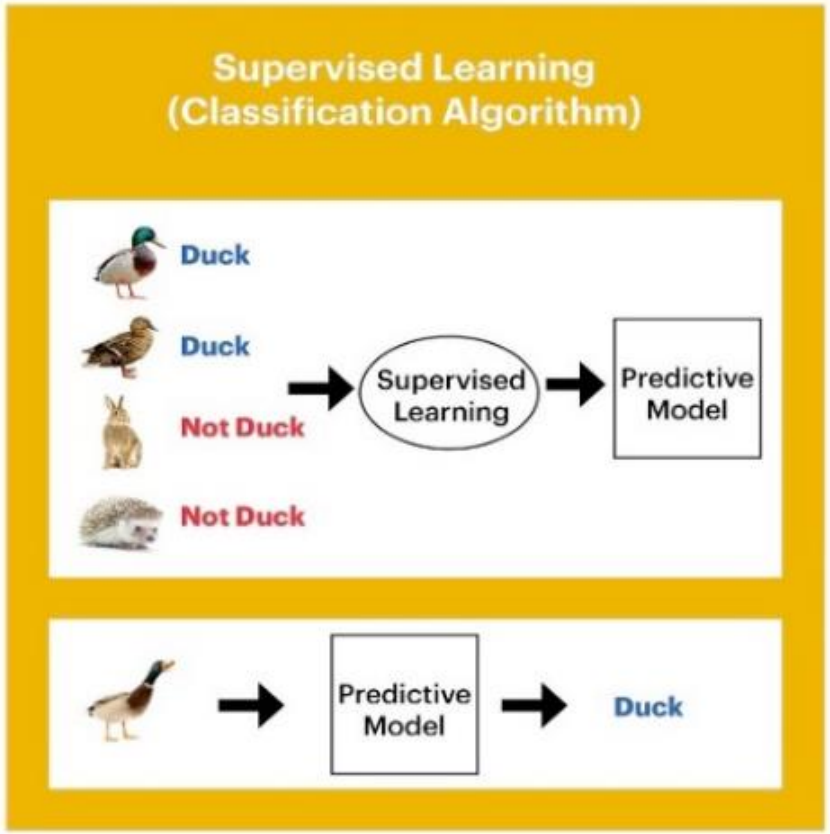
Para cada grupo de algoritmo (supervisionado, não supervisionado) existe tipos de **tarefas que os algoritmos realizam**. As três principais tarefas são Classificação, Regressão e Agrupamento.

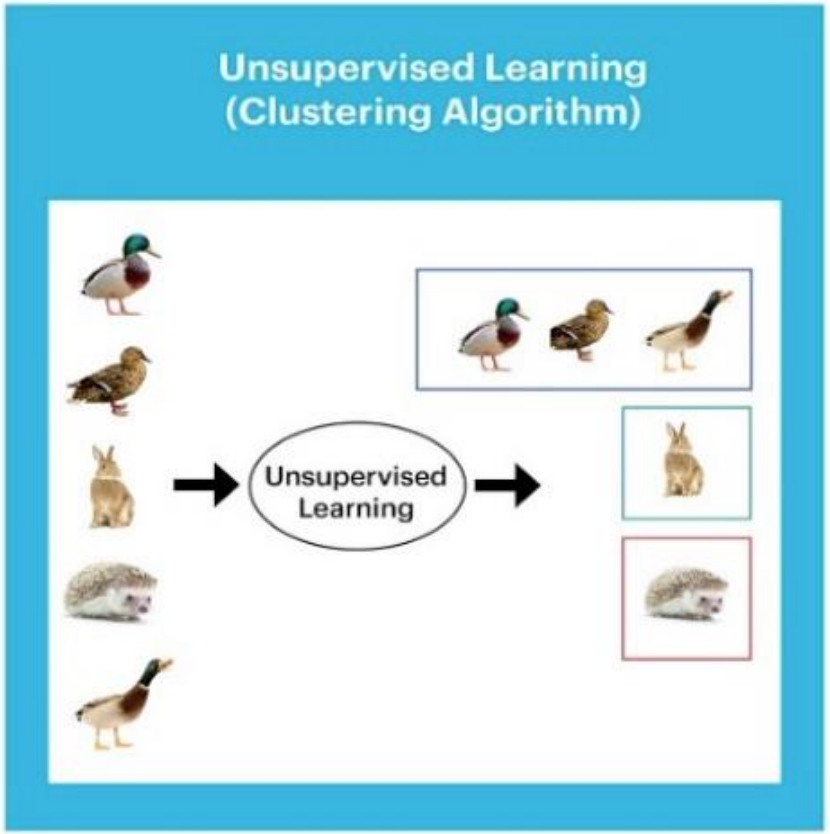


https://www.researchgate.net/profile/Rory-Bunker/publication/319937079/figure/fig1/AS:542578279366656@1506372150206/Supervised-Learning-versus-Unsupervised-Learning-Mathworks-nd.png

Aprendizado de Máquina

No aprendizado **supervisionado** temos rótulos para cada entrada de dado. No **não supervisionado**, não fornecemos nenhuma informação (rótulo) para o agrupamento.



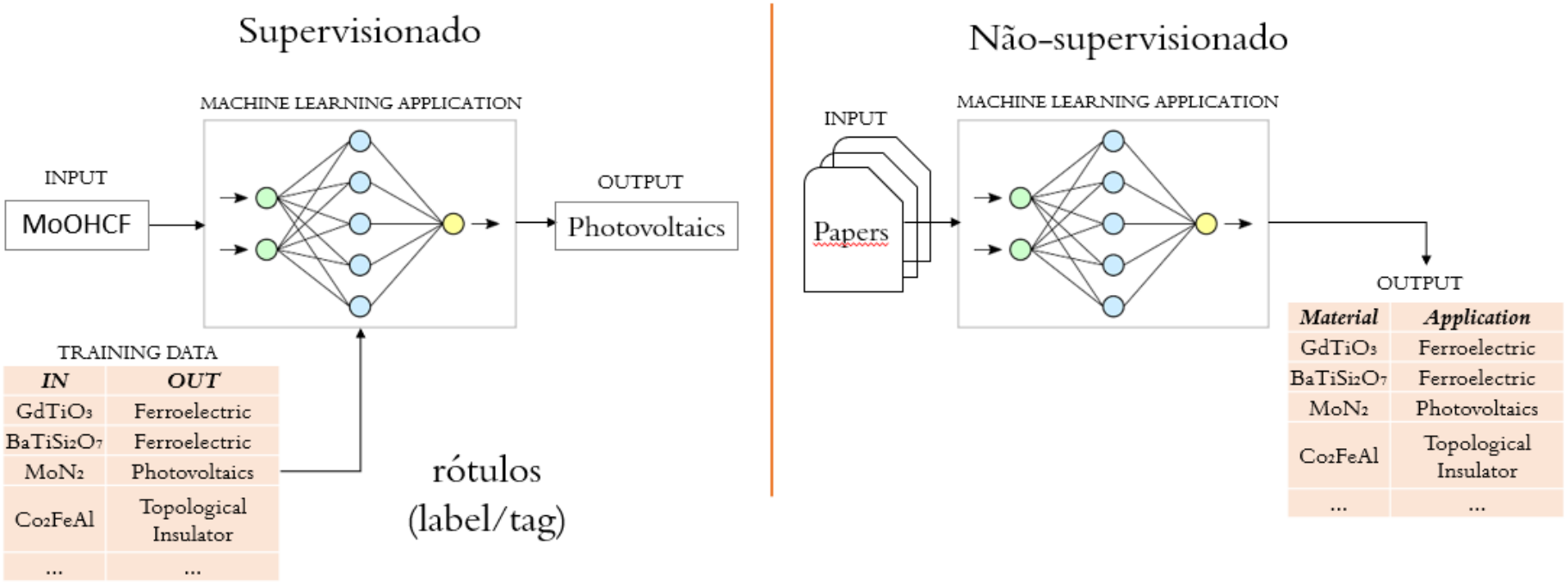


https://miro.medium.com/max/1000/0*4q_X_xQxevOQY_0u

Western Digital.

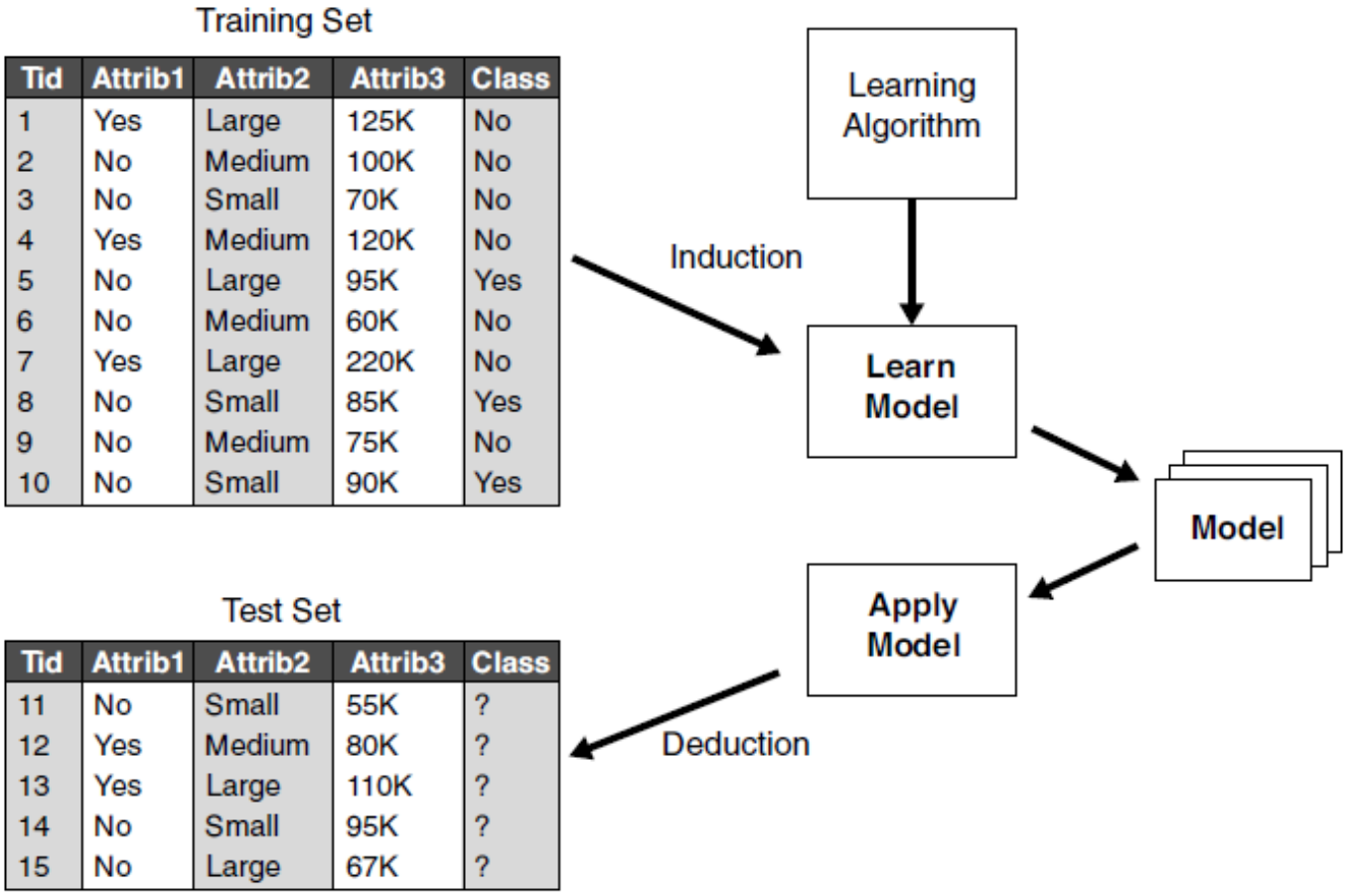
Aprendizado de Máquina

No **aprendizado supervisionado, fornecemos dados de treinamento com rótulos** (saída esperada). Uma vez que o modelo está treinado, podemos inserir entradas cuja saída são, a princípio, desconhecidas.



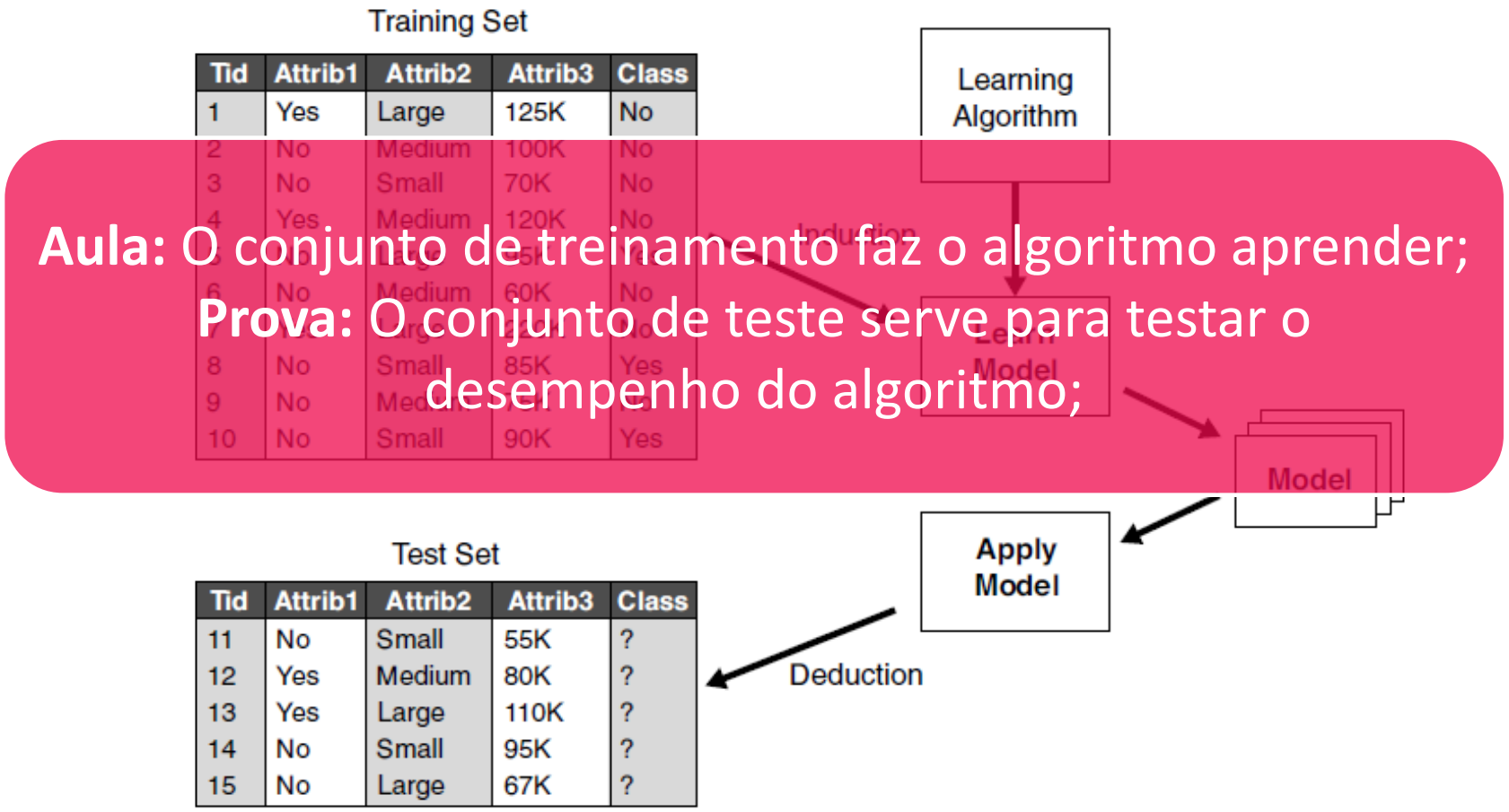
Aprendizado de Máquina - Supervisionado

No **aprendizado supervisionado** iremos separar nosso conjuntos de dados em dois grupos, o **conjunto de treinamento** e o **conjunto de teste**.



Aprendizado de Máquina - Supervisionado

No **aprendizado supervisionado** iremos separar nosso conjuntos de dados em dois grupos, o **conjunto de treinamento** e o **conjunto de teste**.



Aula: O conjunto de treinamento faz o algoritmo aprender;

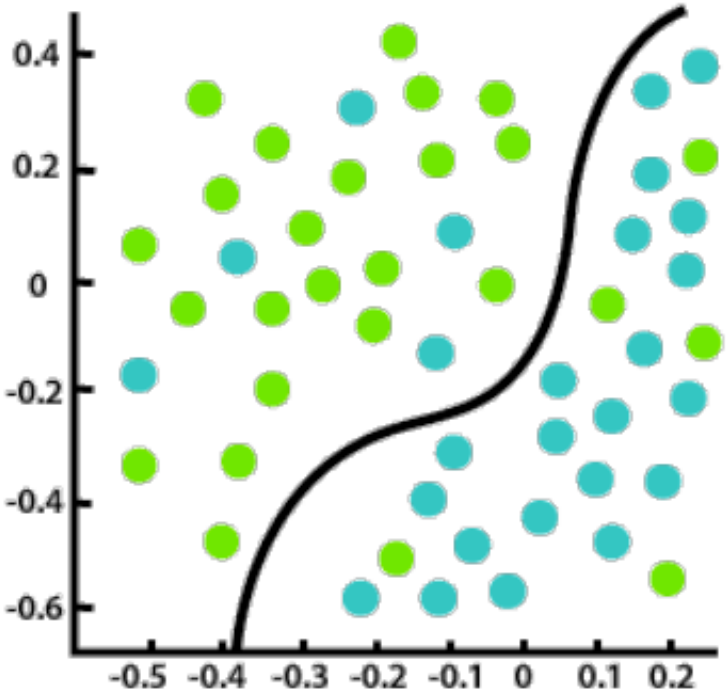
Prova: O conjunto de teste serve para testar o desempenho do algoritmo;

Aprendizado de Máquina - Supervisionado

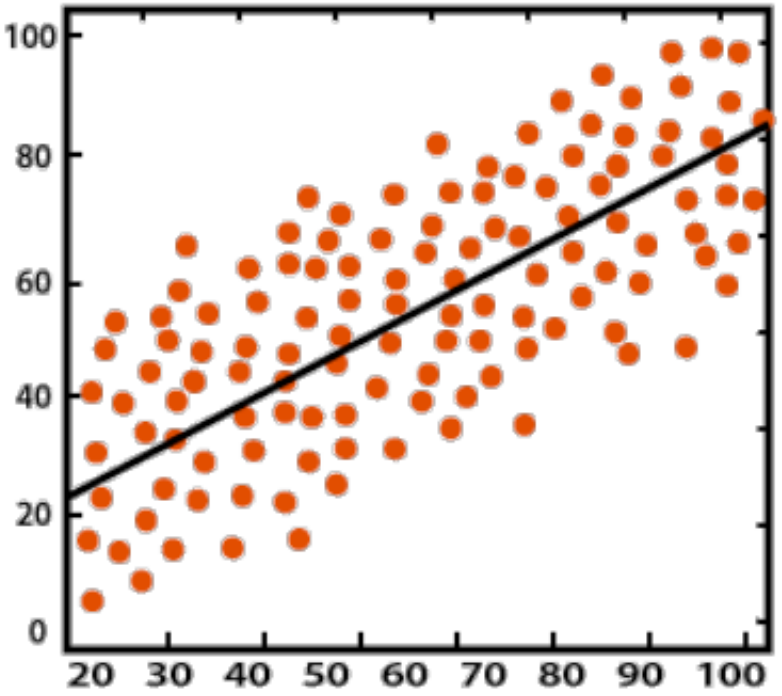
FIAP

Dentro do aprendizado supervisionado temos as tarefas de Regressão e Classificação:

Classificação: predição de classe



Regressão: predição de valor

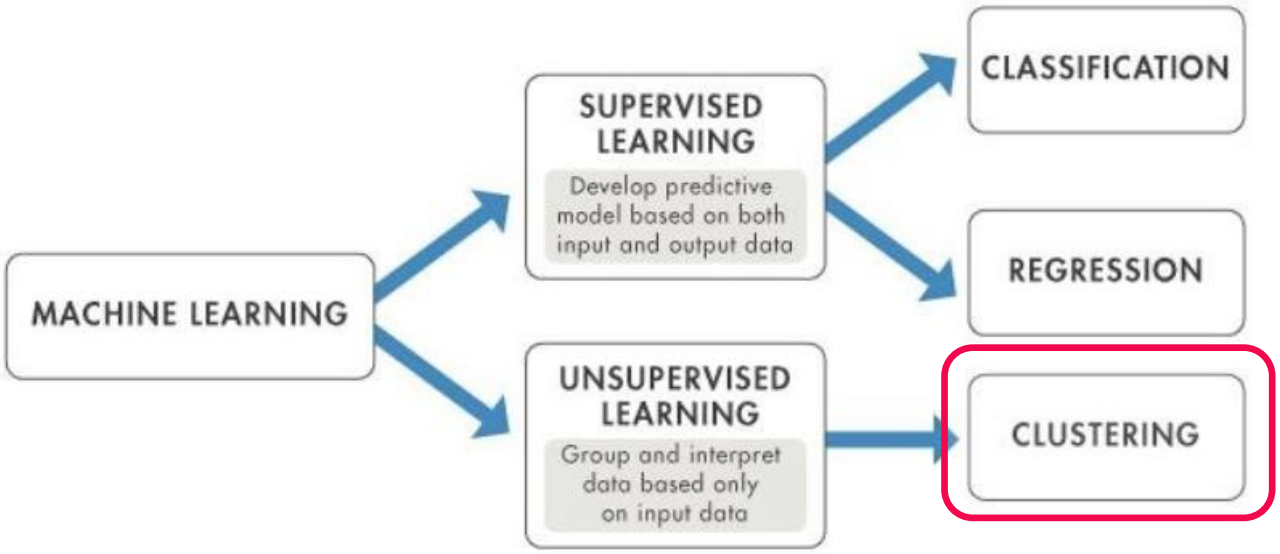


https://static.javatpoint.com/tutorial/machine-learning/images/regression-vs-classification-in-machine-learning.png

Fluxo de trabalho em ML

Aprendizado não supervisionado

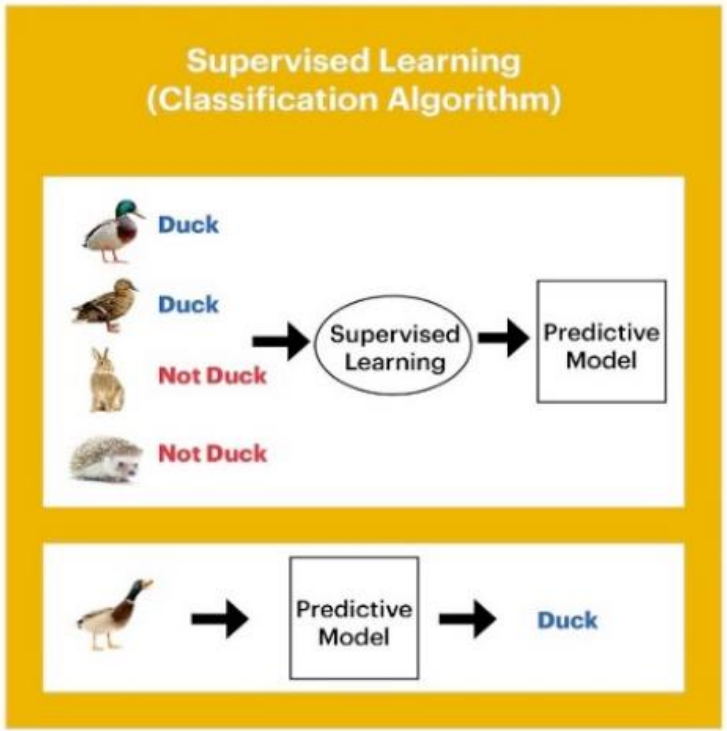
- No aprendizado não supervisionado **não sabemos inicialmente os rótulos das classes**.
- Com essas técnicas gostaríamos de resolver problemas de:
  - **Agrupamento** (ou clusterização)
  - **Regras de Associação**
  - **Redução de Dimensionalidade**
  - **Detecção de Outliers**

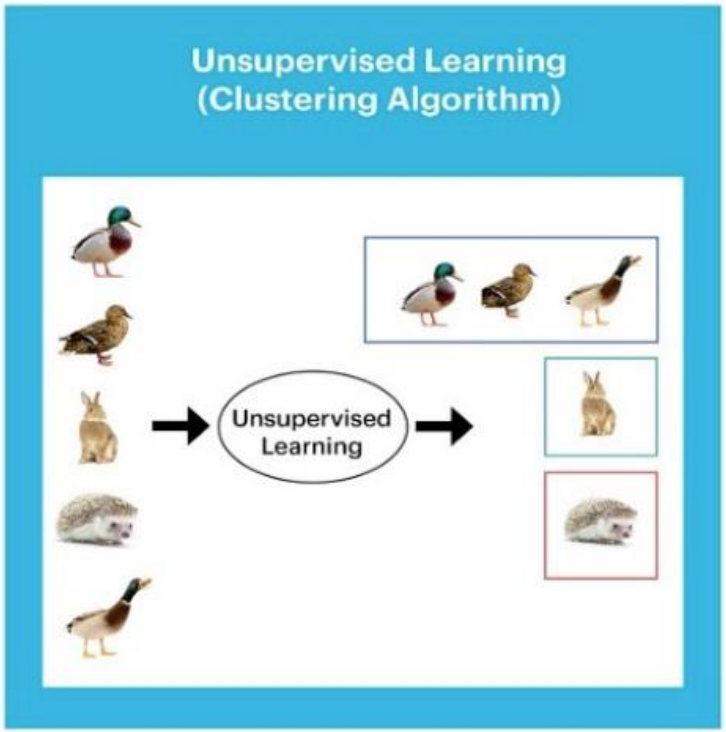


https://www.researchgate.net/profile/Rory-Bunker/publication/319937079/figure/fig1/AS:542578279366656@1506372150206/Supervised-Learning-versus-Unsupervised-Learning-Mathworks-nd.png

Aprendizado não supervisionado

- Na **classificação** nosso atributo alvo é um objeto (string) previamente conhecido (dado rotulado);
- Na **clusterização** nosso atributo alvo é um número (int) previamente desconhecido (dado não rotulado);





https://miro.medium.com/max/1000/0*4q_X_xQxevOQY_0u

Western Digital.

Agrupamento

Agrupamento é uma técnica não supervisionada!

$$f(x_1, x_2, \ldots, x_N)$$

Atributos descritivos

Grupo alvo desconhecido

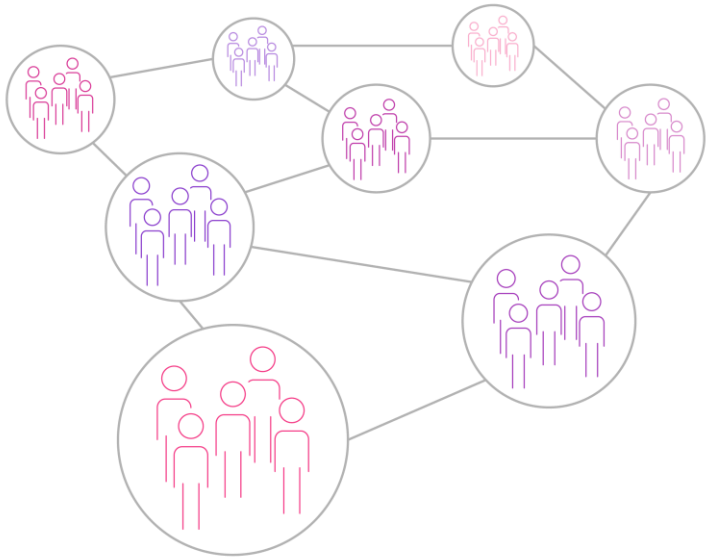
Índice da linha	x1	x2	...	xn	$\hat{y}$
1	548.4	-9789		0.4875	?
2	689.4	-10235		-0.358	?
3	3154.8	-1031858	...	-0.1458	?
...	...			...	
k	803.54	-20000		1.054	?

Queremos criar essa coluna!

Agrupamento

GERAÇÃO DE GRUPOS OU CLUSTERS

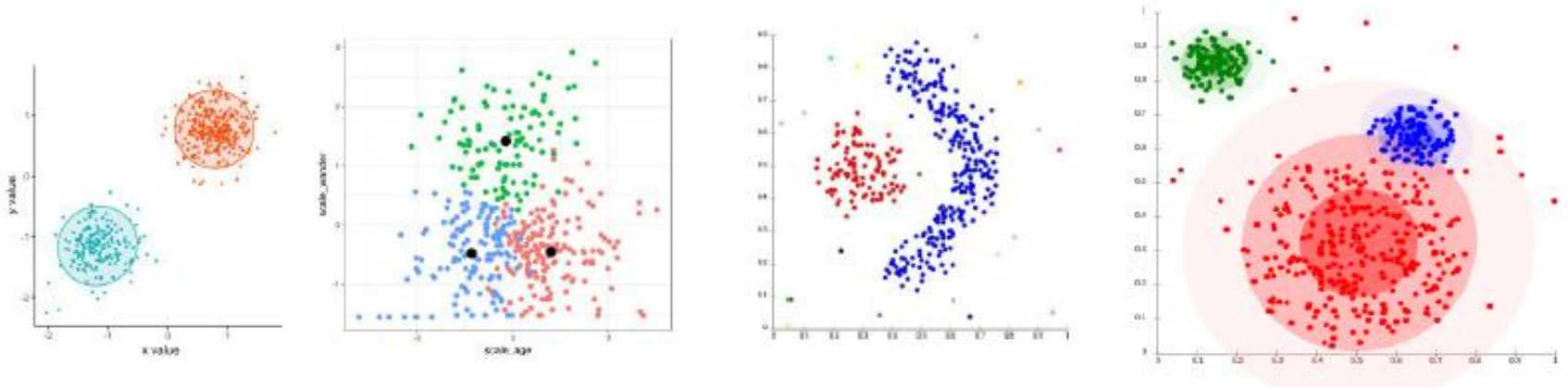
- Os grupos são formados de maneira a maximizar a similaridade entre os elementos de um grupo (similaridade intragrupo) e minimizar a similaridade entre elementos de grupos diferentes (similaridade intergrupos).
- Aprendizado não supervisionado.



ALGORITMOS DE CLUSTERIZAÇÃO

Agrupamento

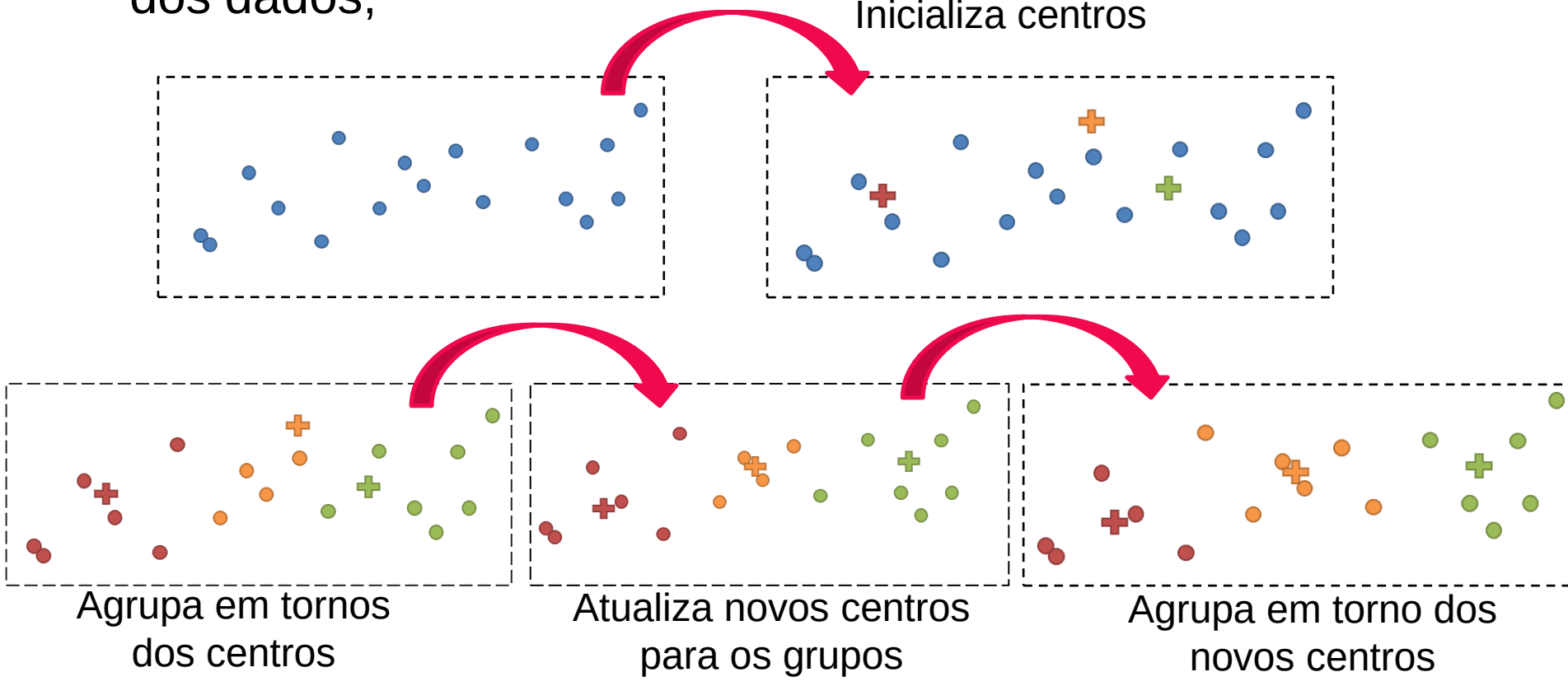
- Grupos podem:
  - Ter diferentes tamanhos, formas e densidades;
  - Formar uma hierarquia;
  - Ter sobreposição ou serem disjuntos



- Existem muitos algoritmos diferentes para fazer agrupamento. Alguns deles são baseados em:
  - **Ligações** como **Agrupamento Hierárquico**;
  - **Densidade** como o **DBSCAN**;
  - **Partições** como o **K-Means**;
  - **Grid** como o STING e o WaveCluster
  - **Modelos** como o SOM, redes neurais e **Mistura Gaussiana**;

Baseado em Partição – k-Means

- O k-Means é um dos métodos mais antigos (referências originais datam de 1956, 1965 e 1967) e mais utilizados;
- Ele é simples e intuitivo, baseado na ideia de se quebrar o espaço multidimensional em partições a partir do centroide dos dados;



k-Means – Algoritmo

O pseudocódigo do k-Means pode ser sumarizado como:

1. Escolher aleatoriamente k centros para os clusters;
2. Atribuir cada objeto para o cluster de centro mais próximo segundo alguma métrica de distância (ex: euclidiana);
3. Mover cada centro para a média (centroide) dos objetos do cluster correspondente;
4. Repetir os passos 2 e 3 até que algum critério de convergência seja atendido (ex: número máximo de interação, limiar mínimo de mudança nos centroides).

k-Means – Algoritmo / Métrica

- Precisamos usar uma métrica de distância entre os centroides e os pontos de dados;
- Podemos usar diferentes métricas. A mais comum é a distância Euclidiana:

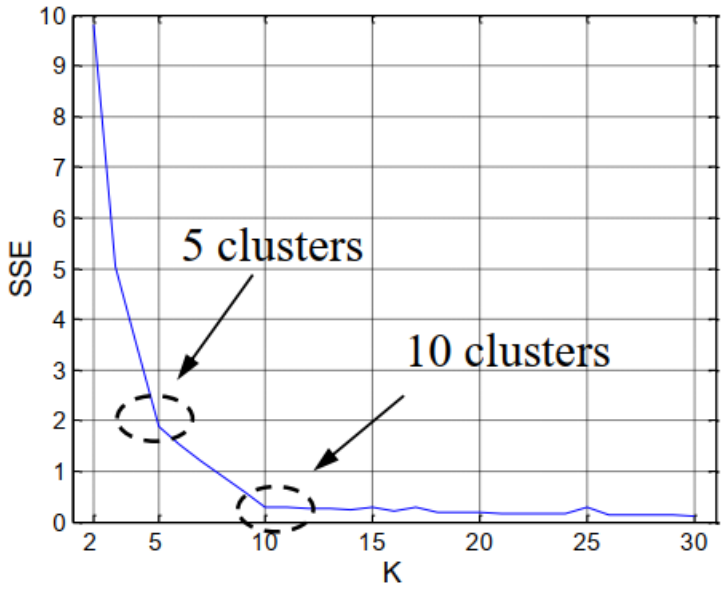
$$d(A,B) = \sqrt{\sum_{i=1}^{n} \left(A_{x_i} - B_{x_i}\right)^2}$$

identifier	class name	args	distance function
"euclidean"	EuclideanDistance	•	`sqrt(sum((x - y)^2))`
"manhattan"	ManhattanDistance	•	`sum(|x - y|)`

Outras métricas:
https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.DistanceMetric.html

k-Means – Algoritmo / Hiperparâmetro

- O k-Means tem o hiperparâmetro k que é o número de grupos;
- Como saber qual é o melhor número de k?
- Podemos usar a Soma dos Erros Quadráticos (SSE) em relação ao centroide para encontrar o “joelho” da curva de otimização:



$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2$$

- $dist$ é a distância euclidiana;
- c_i é o centro do i-ésimo agrupamento;
- x são os dados pertencentes ao i-ésimo agrupamento.

k-Means – Vantagens e Desvantagens

FIAP

Vantagens:

- Implementação simplificada.
- Facilidade em lidar com qualquer medida de similaridade e por consequência, qualquer tipo de atributo.

Desvantagens:

- Dificuldade na definição do valor de “k”.
- Suscetível a outliers e a ausência de normalização.

MÉTRICAS DE DESEMPENHO DE AGRUPAMENTO

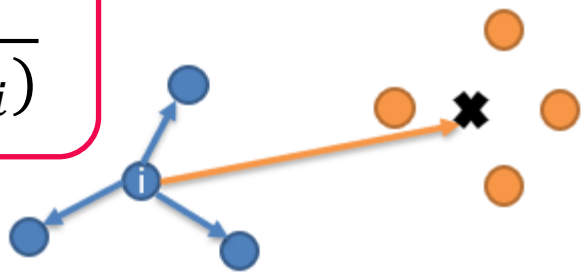
Silhouette

O *Silhouette Score* é uma métrica que avalia o "formato" dos clusters obtidos.

Ele é obtido calculando a distância média entre um dado de um agrupamento com todos os outros dados do mesmo cluster (a) e com a média desse mesmo dado com todos os dados do agrupamento mais próximo.

Essa métrica é definida como:

$$Silhouette(i) = \frac{b_i - a_i}{\max(a_i, b_i)}$$



b: distância entre i e o agrupamento mais próximo

a: distância média entre i e todos os pontos de seu agrupamento

Silhouette

A média do *Silhouette Score* de todos os dados nos define o quão bom é o nosso agrupamento:

$$-1 \leq Silhouette \leq 1$$

Dados foram atribuídos aos grupos de forma errada

Agrupamentos bem separados e é possível distingui-los

Agrupamentos com problemas

Agrupamentos ok

FIAP GRADUAÇÃO

Colaboração e adaptação: Prof. André Tritiack