Escola Politécnica de Pernambuco

Especialização em Ciência de Dados e Analytics

Introdução à Ciência de Dados
Aula 6

Prof. Dr. Alexandre Maciel

alexandre.maciel@upe.br

ANÁLISE PREDITIVA DE DADOS

“... é a arte de se obter informação a partir de dados coletados e a utilizar para prever padrões de comportamento e tendências.”

Covington

DEFINIÇÕES

Aprendizado de Máquina

é um **campo da IA** que, por meio de algoritmos, fornece aos computadores a capacidade de **identificar padrões** de dados em massa para fazer **previsões** (análise preditiva).

(Provost & Fawcett).

APRENDIZADO DE MÁQUINA

CLASSIFICAÇÃO

- Tarefa supervisionada de aprender uma **função alvo f(X)** que mapeie cada conjunto de **atributos X** para um dos rótulos de **classes Y** pré-determinados.

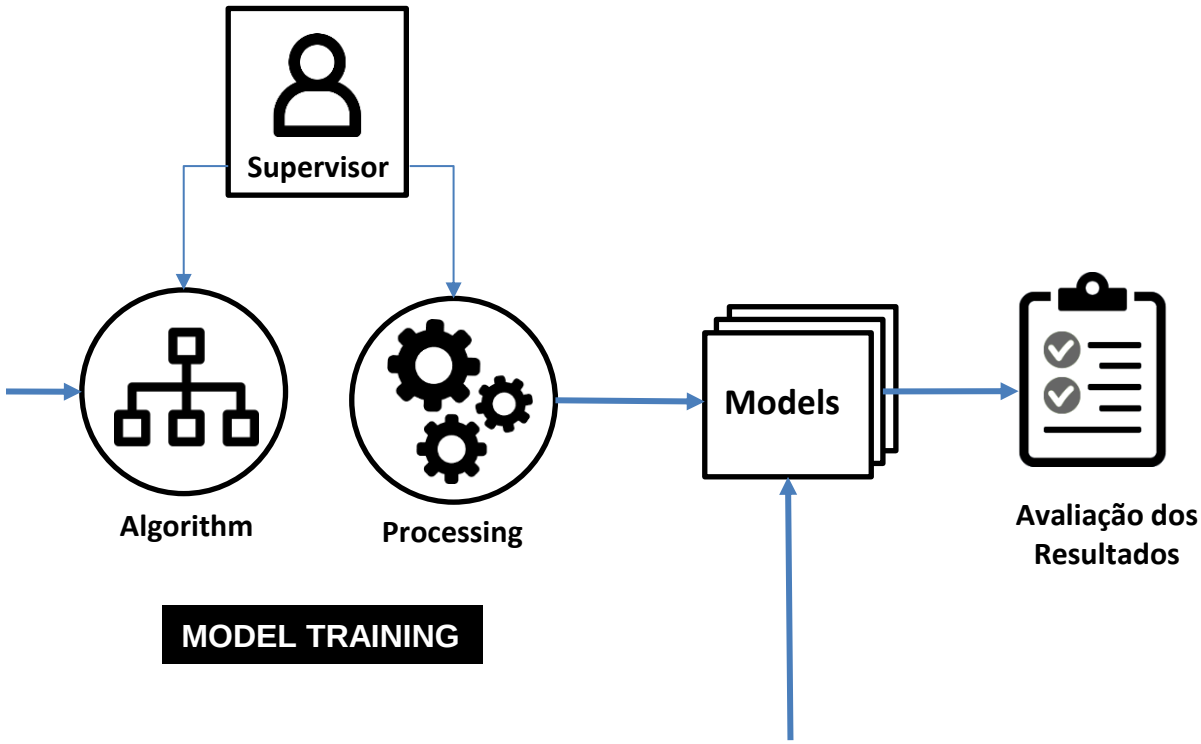
- O objetivo é aproximar a função alvo para que, sempre que forem apresentados **novos dados de entrada** (x), a máquina possa **prever as variáveis de saída** (Y) para esses dados.

- Isso requer que o algoritmo de aprendizagem **generalize** a partir dos dados de treinamento para situações invisíveis de uma forma "razoável“ (erro de generalização).

CONSTRUÇÃO DOS MODELOS

TRAINING DATASET

BI-RADS	Idade	Forma	Margem	Densidade	Severidade
5.0	67.0	3.0	5.0	3.0	1
4.0	43.0	1.0	1.0	NaN	1
5.0	58.0	4.0	5.0	3.0	1
4.0	28.0	1.0	1.0	3.0	0
5.0	74.0	1.0	5.0	NaN	1
4.0	65.0	1.0	NaN	3.0	0
4.0	70.0	NaN	NaN	3.0	0
5.0	42.0	1.0	NaN	3.0	0
5.0	57.0	1.0	5.0	3.0	1
5.0	60.0	NaN	5.0	1.0	1

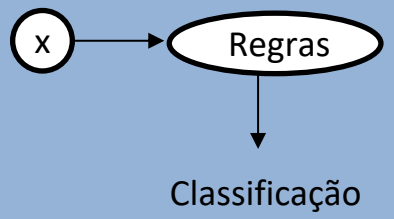


BI-RADS	Idade	Forma	Margem	Densidade	Severidade
5.0	67.0	3.0	5.0	3.0	?
4.0	43.0	1.0	1.0	NaN	?
5.0	58.0	4.0	5.0	3.0	?

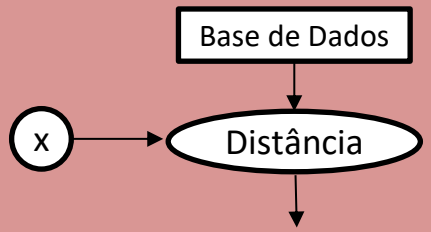
TEST DATASET

MÉTODOS DE CLASSIFICAÇÃO

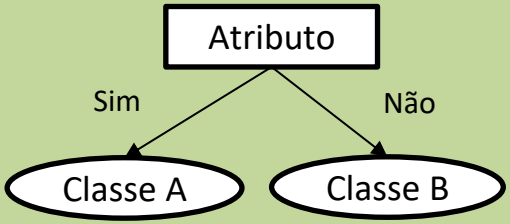
Baseados em conhecimento



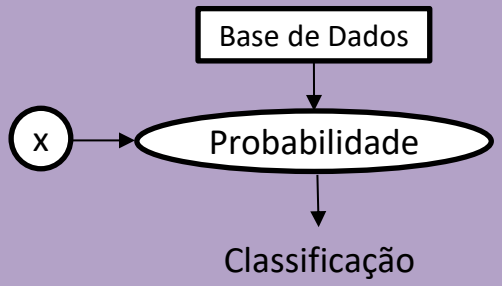
Baseados em distância



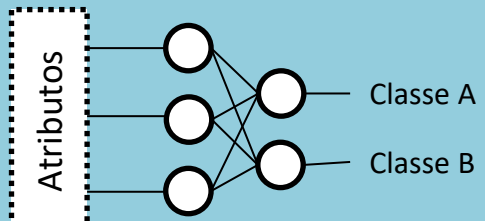
Baseados em árvore



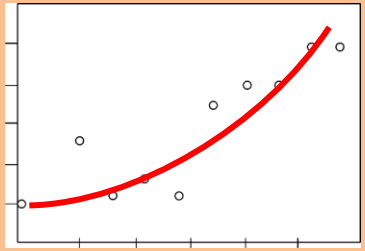
Probabilísticos



Conexionistas



Baseados em função



AVALIAÇÃO DA CLASSIFICAÇÃO

- **Matriz de Confusão**

  - Matriz que relaciona classes originais com as preditas.

		Classe predita	
		Positiva	**Negativa**
Classe original	**Positiva**	VP	FN
	Negativa	FP	VN

- **Acurácia**

  - Percentual de classificações corretas.

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN}$$

- **Precision**

  - Dentre todas as classificações positivas, quantas estão corretas?

$$Precision = \frac{VP}{VP + FP}$$

- **Recall**

  - Dentre todas as situações de classe positivas, quantas estão correta?

$$Recall = \frac{VP}{VP + FN}$$

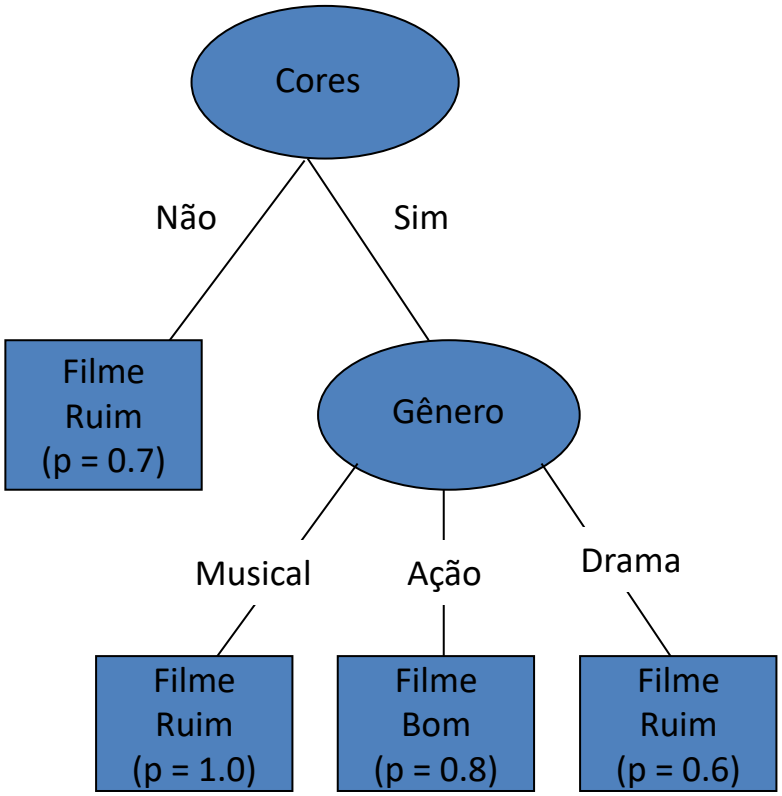
- **F1-Score**

  - Média harmônica entre Precison e Recall.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

ÁRVORE DE DECISÃO

- O nó mais elevado da árvore é conhecido como ***raiz***.
- Cada ***nó interno*** contém um teste sobre os valores de um dado atributo.
- Cada ***ramo*** representa um resultado do teste.
- Os ***nós folhas*** da árvore representam as classes.
- Comumente acompanhado de um ***grau de confiança***.



AGRUPAMENTO

- Tarefa de aprendizado de máquina não supervisionada.

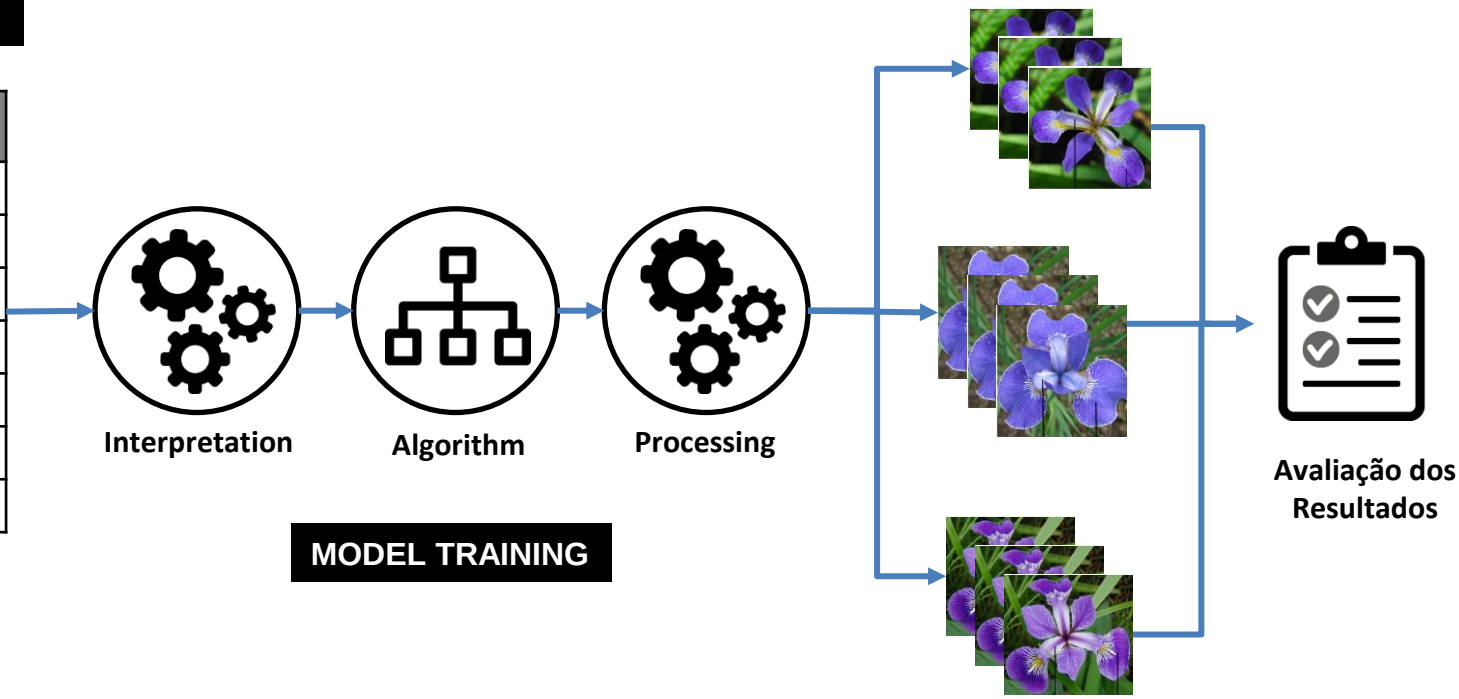
- Organização de um conjunto de objetos (*representados por vetores de características*) em grupos **baseada na similaridade** entre eles.

- Particionar um conjunto de dados em subconjuntos de forma que os objetos de cada grupo (idealmente) **compartilhem características comuns**.

CONSTRUÇÃO DOS MODELOS

UNLABELED DATASET

sepal. length	sepal. width	petal. length	petal. width
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
5.6	3.0	4.5	1.5
6.4	3.2	4.5	1.5
7.7	3.8	6.7	2.2
7.7	2.6	6.9	2.3
6.7	2.5	5.8	1.8

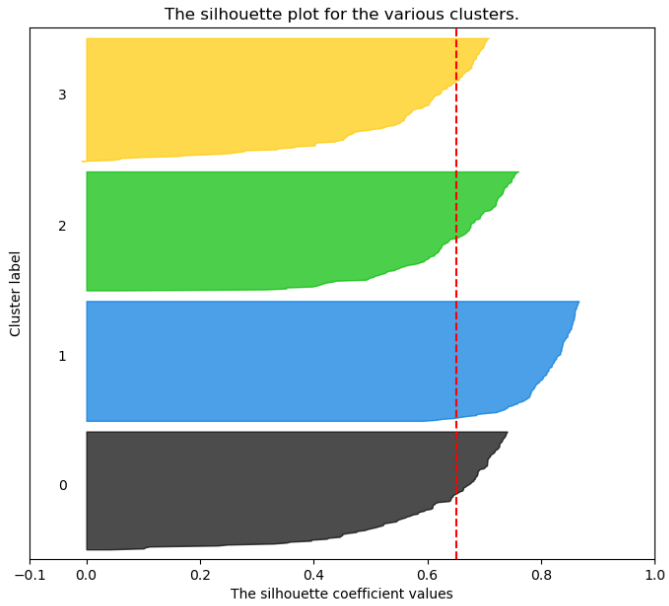


AVALIAÇÃO DO AGRUPAMENTO

- **Índice da Silhueta**
  - Medida de quão semelhante um objeto é ao seu próprio cluster (coesão) em comparação com outros clusters (separação).

$$SIL(g) = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{|g_i|}\sum_{j=1}^{|g_i|}\frac{b(j)-a(j)}{\max\{a(i),b(j)\}}$$

onde
- g é o agrupamento resultante;
- k é o número de grupos;
- $|g_i|$ é o número de objetos no i-ésimo grupo;
- $a(j)$ é a distância média do j-ésimo objeto do grupo g_i aos objetos do mesmo grupo; e
- $b(j)$ é a menor distância média do j-ésimo do objeto do grupo g_i aos objetos dos outros grupos.



MÉTODOS DE AGRUPAMENTO

Particionais

- Constrói *k* partições dos dados.

- Partição inicial + algoritmo de realocação iterativa.

- Hard ou Soft (Fuzzy).

Hierárquicos

- Aglomerativos: cada objeto um grupo -> similares unem-se no mesmos grupos.

- Divisivos: todos objetos no mesmo grupo -> grupos menores

Densidade

- Num grupos definidos automaticamente

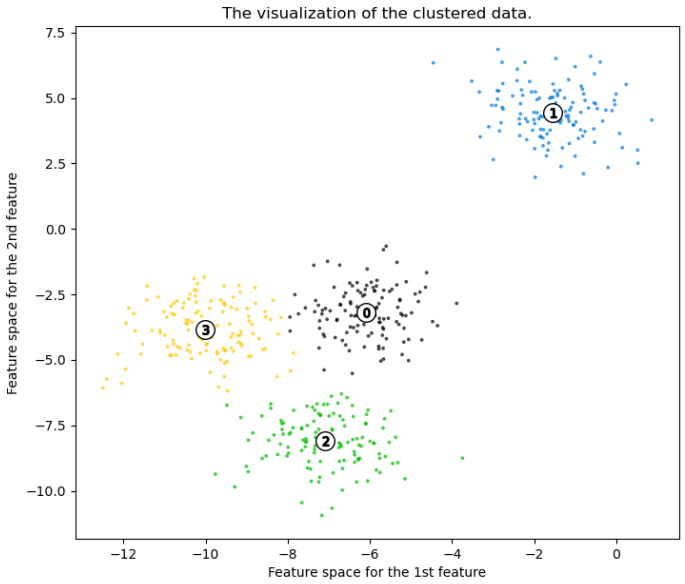
- Raio de vizinhança e número mínimo de pontos

K-MEANS

- Toma como entrada o parâmetro *k* grupos desejados e particiona o conjunto de dados em *k* grupos com *n* objetos.

- Busca similaridade intragrupos e dissimilaridade intergrupos.

- No particionamento cada objeto pertence ao grupo do centroide mais próximo a ele.



FERRAMENTAS

Linguagens

IDEs

Analytics

"Point and Click"

Infraestrutura e Armazenamento

ATIVIDADE 6

1. Com a base de dados do seu projeto execute os algoritmos
   - Árvore de Decisão
   - K-means.

2. Entregue os slides (+2) de sua apresentação.

3. Faça o upload no Google Classroom.