Projeto ABCIA

Módulo 02: Machine Learning

Aula 03

Prof. Dr. Luciano Ferreira Silva

UFRR | DCC DEPTO DE CIÊNCIA DA COMPUTAÇÃO | Softex | ABCIA ABC DA INTELIGÊNCIA ARTIFICIAL

Apresentação do Professor

- Graduado em Matemática (UFU) - 2003
- Mestre em Computação Gráfica (UFU) – 2006
- Doutor em Computação Gráfica (UFU) - 2009
- Certificação em Inteligência Artificial (Huawei – HCIA) - 2021

Professor do curso de Ciência da Computação/UFRR desde 2008, atuando nas disciplinas e pesquisa de **Computação Gráfica, RV e RA, Compiladores, IHM, Desenvolvimento de jogos e Visão Computacional.**

Prof. Dr. Luciano Ferreira Silva
E-mail: luciano.silva@ufrr.br

Objetivos e sua relevância

Objetivos de Aprendizagem

6. Identificar etapas básicas na construção de um modelo da realidade a partir de algoritmos de ML.

6.1. Identificar principais algoritmos existentes em função de sua aplicação para predição ou classificação.

6.2. Identificar a etapa de divisão da base, em dados de teste e de treino.

6.2.1. Identificar a finalidade da técnica de **validação cruzada**.

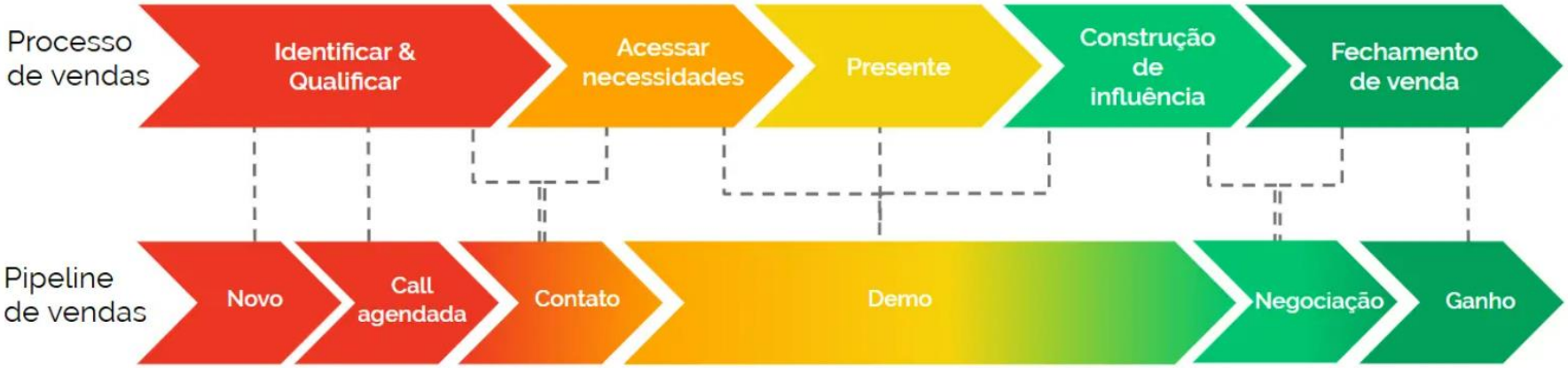
6.3. Identificar etapas de treino e teste do algoritmo.

6.4. Identificar etapa de seleção de algoritmo.

Pipeline de construção de modelos Machine Learning

Mas o que é um Pipeline?

✓O pipeline é um mapa das etapas/fases/operações que compõem um determinado processo



Pipeline ML

1. Conhecer o problema e os dados

Tipo de predição:

-Regressão

-Classificação

-Clusterização

Aprendizagem:

-Supervisionada

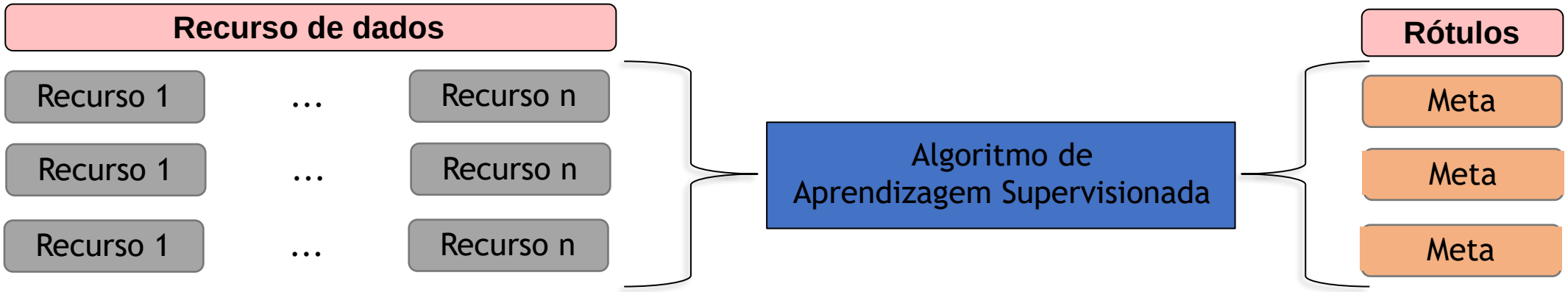
-Não supervisionada

-Semi-supervisionado

-Por reforço

Classificações da Aprendizagem

Aprendizagem Supervisionada: treinamento e aprendizagem ocorre com base em amostras de categorias conhecidas (amostras rotuladas).



Clima	Temperatura	Velocidade do vento
Ensolarado	Alta	Forte
Chuvoso	Baixa	Moderada
Ensolarado	Moderada	Fraca

Praticar esportes?
Sim
Não
Sim

Chuvoso	Moderada	Fraca

?

Classificações da Aprendizagem

Aprendizagem Supervisionada: Pode ser usada em modelos de Regressão ou Classificação.

REGRESSÃO: busca relacionar a amostra a um **Valor** por meio de função matemática.

Anos de carreira	Formação	Idade
5	Computação	37

→

Salário
Valor ?

CLASSIFICAÇÃO: busca relacionar a amostra a uma **Categoria** específica.

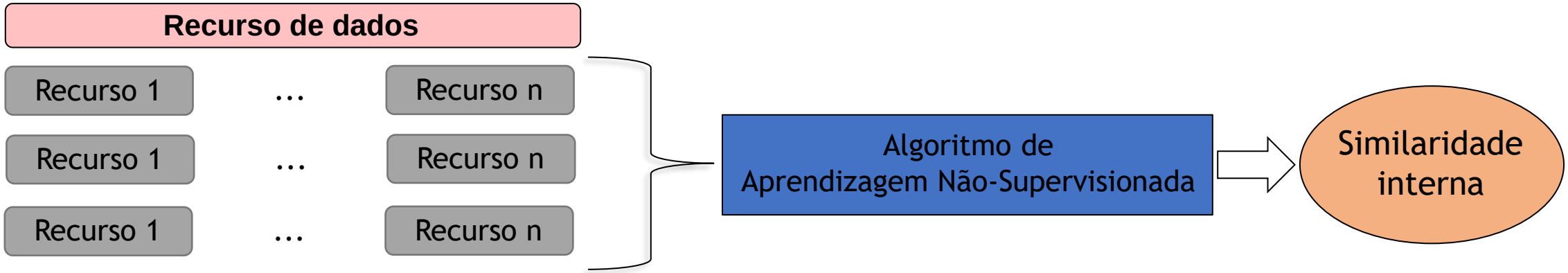
Idade do carro	Idade o motorista
5	27

→

Risco		
Baixo	**Médio**	**Alto**

Classificações da Aprendizagem

Aprendizagem Não-Supervisionada: treinamento e aprendizagem com base em amostras não rotuladas. O agrupamento é uma forma comum de aprendizagem não supervisionada.

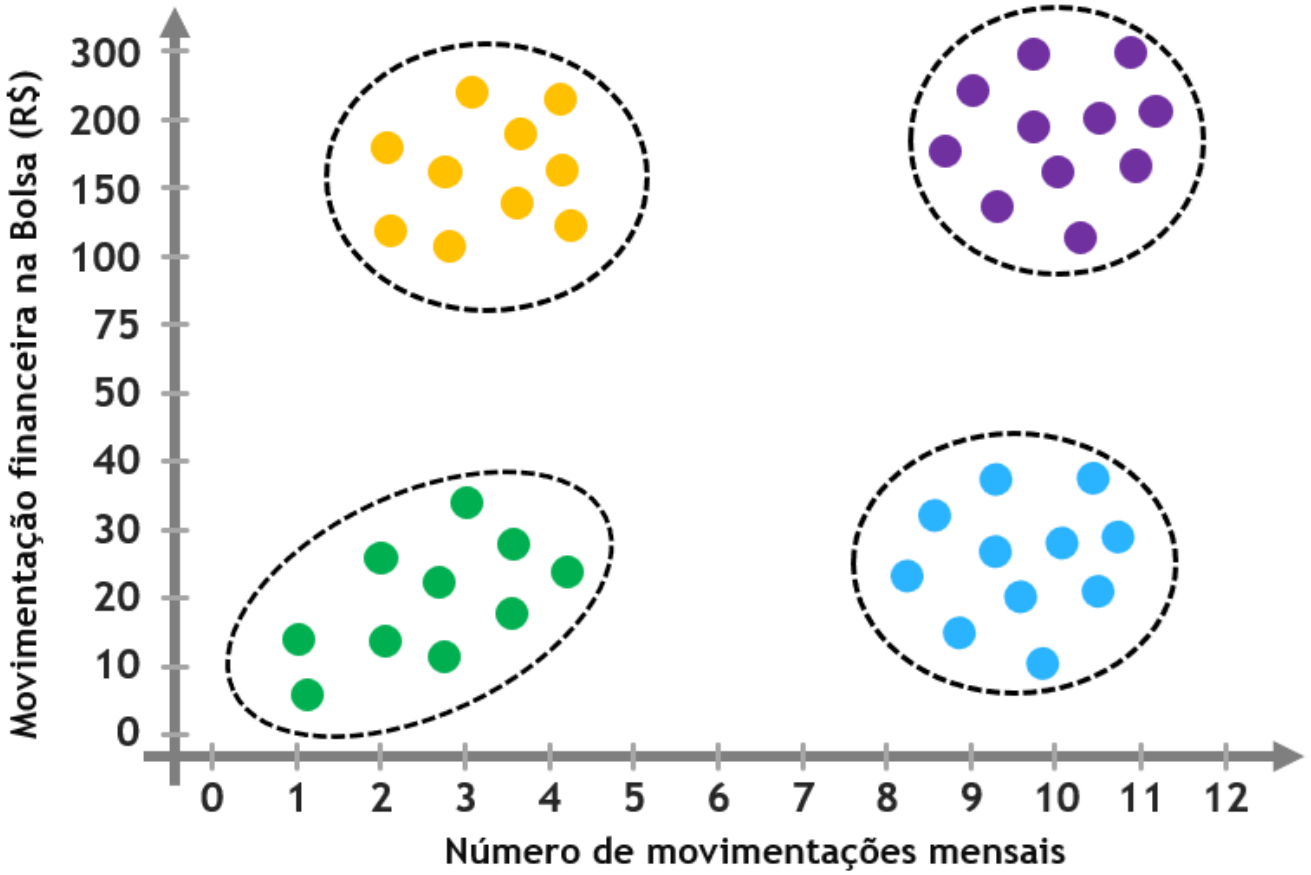


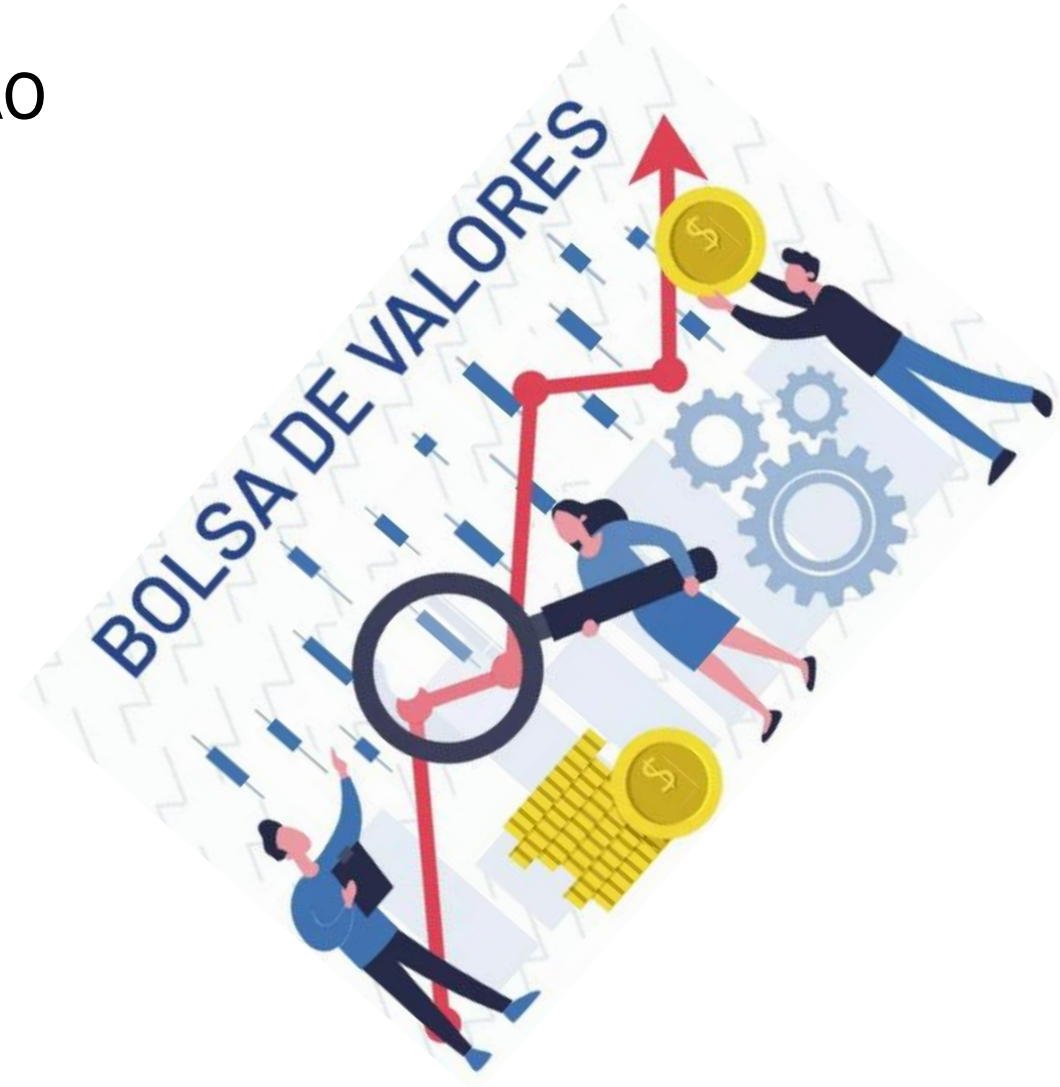
Consumo Mensal	Mercadoria	Tempo de Consumo
1000-2000	Badminton	6h - 12h
500-1000	Basquetebol	18h - 24h
1000-2000	Videogame	00:00-06:00

Categoria
Cluster 1
Cluster 2

Tipos de Predição

CLUSTERIZAÇÃO

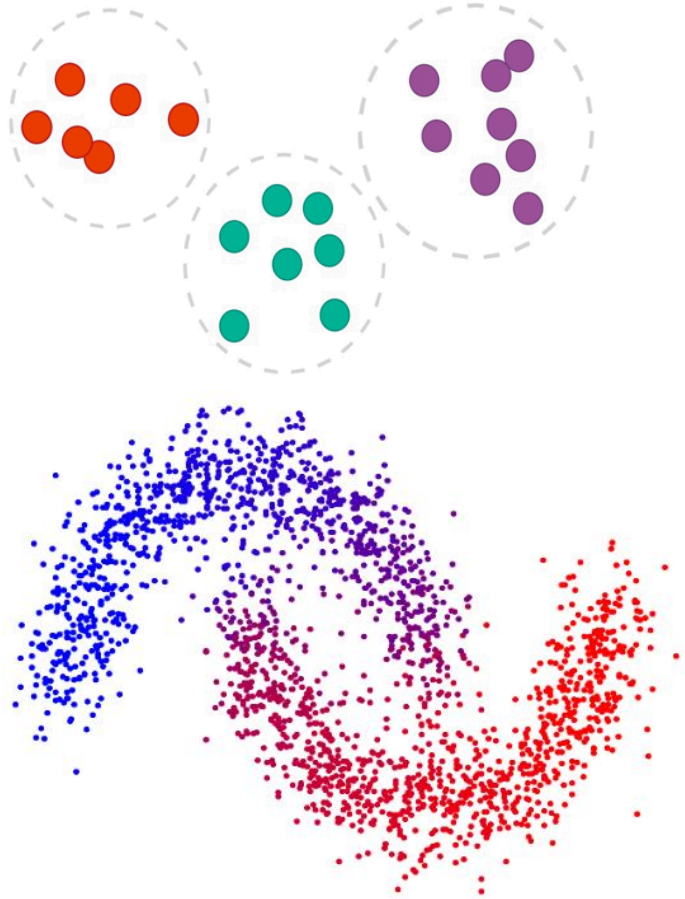




Classificações da Aprendizagem

Aprendizagem Não-Supervisionada: + EXEMPLOS

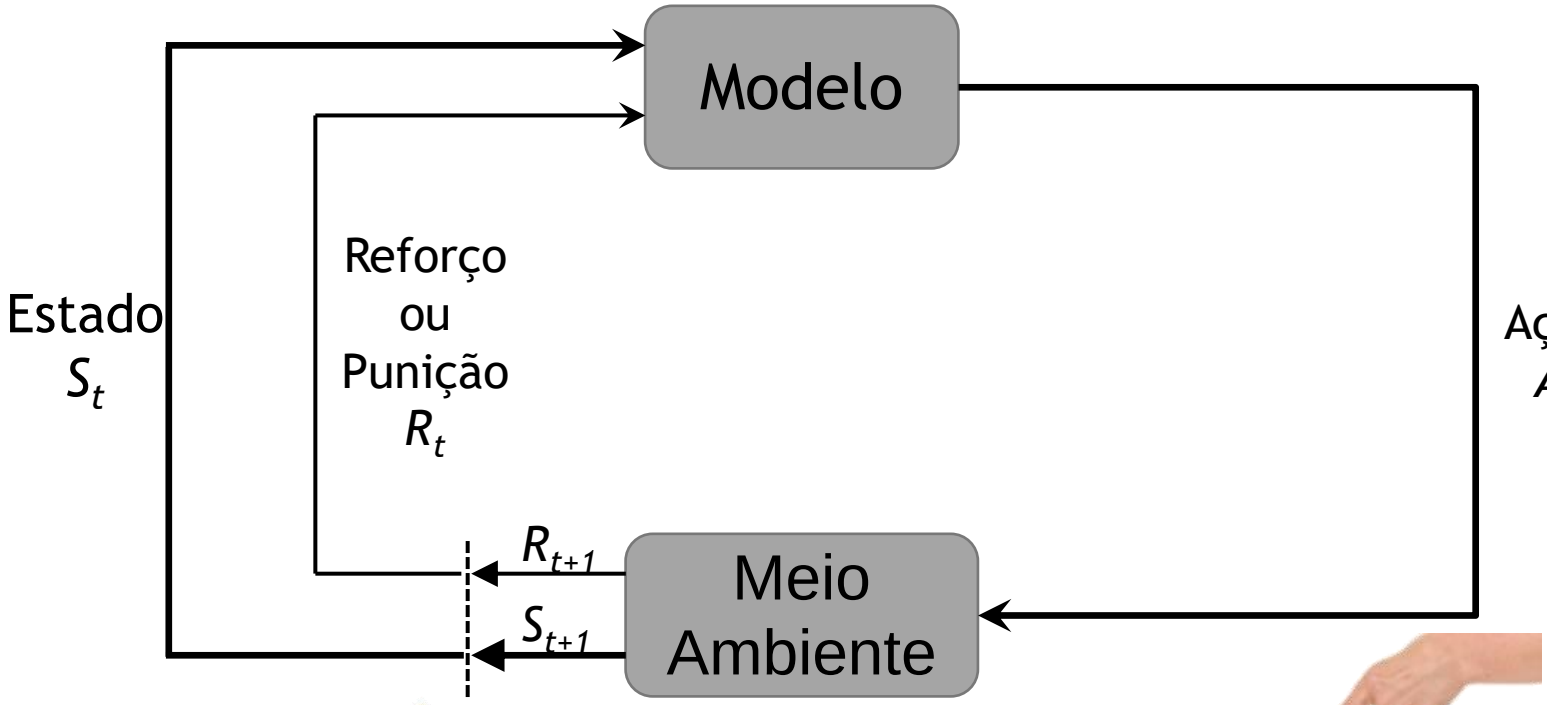
Dados	Exemplo de função do modelo
Transações bancárias	Normalidade da transação
Registros de Compras	Associação entre produtos
Registros de Compras	Perfil dos consumidores
Informações Pessoais (Idade, Sexo, etc.) + Histórico de filmes + Histórico de buscas	Perfil do “cinéfilo”:Sistema de recomendação de Filmes de Streamings (NetFlix, HBO Max, Disney+, Star+, Globoplay, etc.)



Classificações da Aprendizagem

Aprendizagem por Reforço:

- ✓ Utiliza tentativa e erro para encontrar uma solução para o problema.
- ✓ O modelo recebe recompensas (acertos) ou penalidades (erros) pelas ações que executa.
- ✓ Seu objetivo é maximizar a recompensa total.
- ✓ O meio ambiente pode mudar, forçando uma adaptação as novas condições.



Classificações da Aprendizagem

Aprendizagem por Reforço: muito usada para treinamento de máquinas autônomas, robôs e jogos. Exemplos:

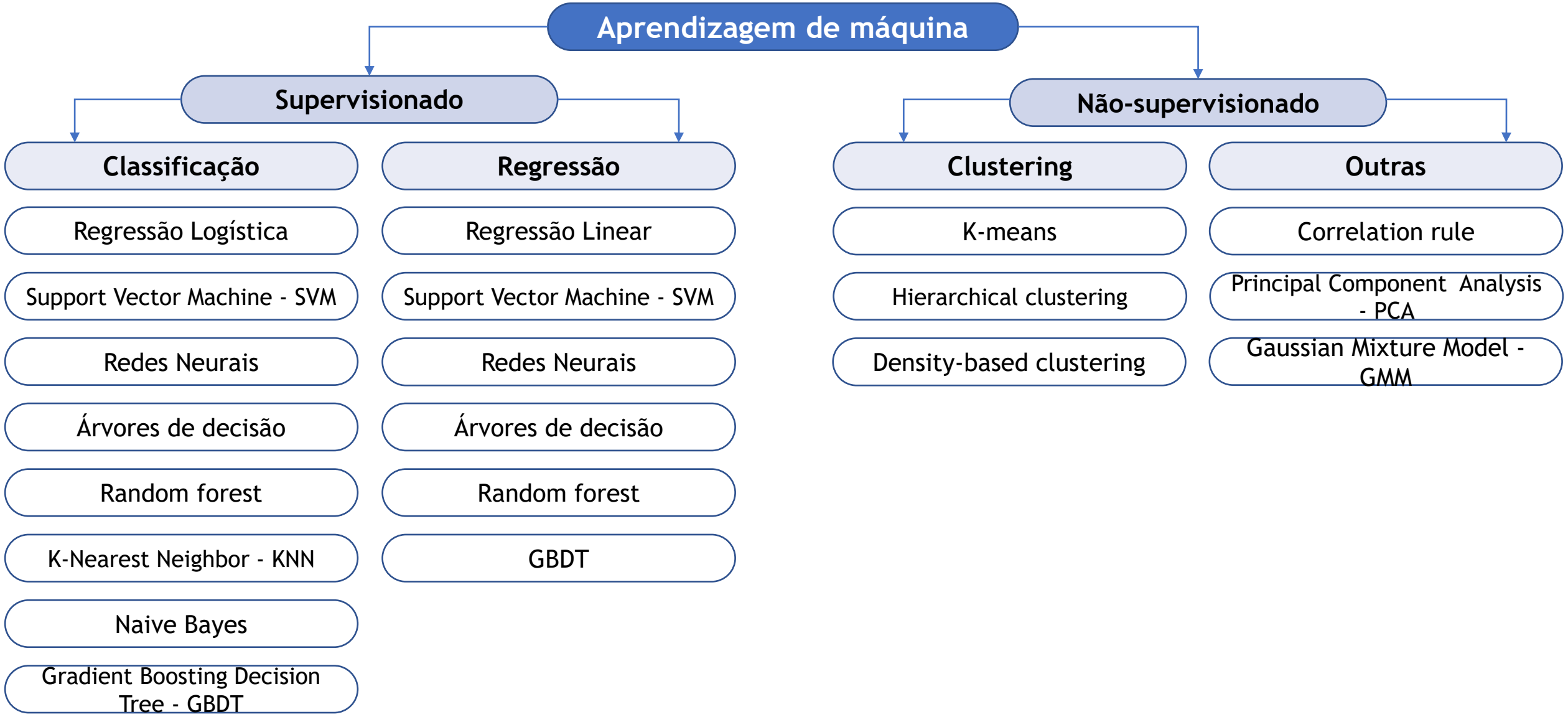
- ✓ **Jogos de xadrez:** qual a minha próxima jogada?

Kasparov perdia no xadrez para Deep Blue há 25 anos.

- ✓ **Carros autônomos:** devo frear ou acelerar quando a luz amarela começa a piscar?

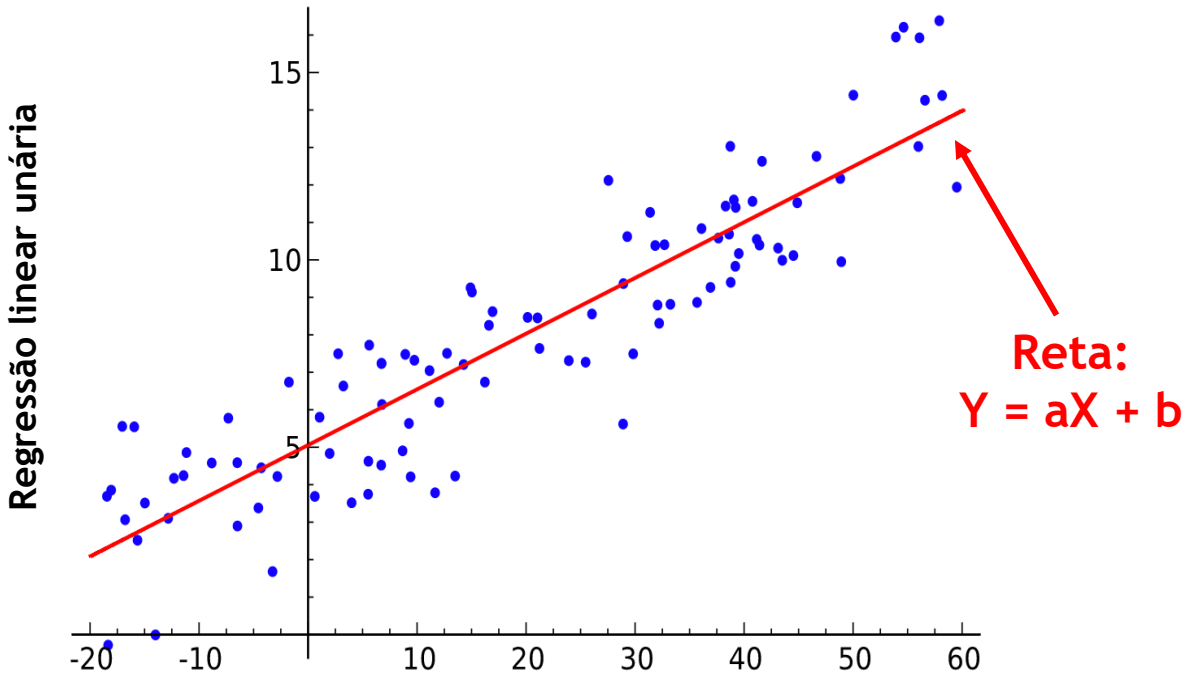
- ✓ **Robôs de limpeza:** devo continuar trabalhando ou voltar para carregar?

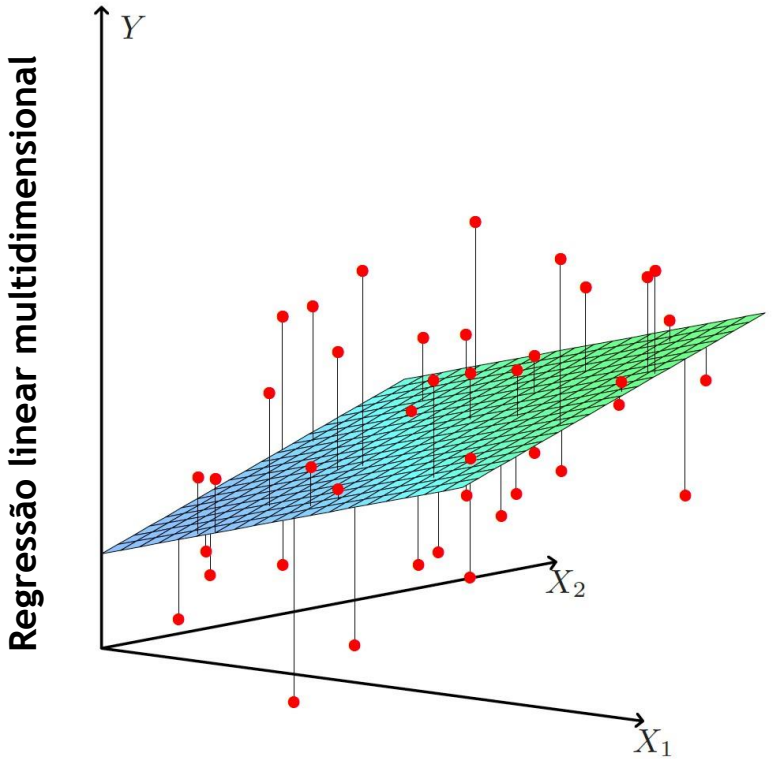
Visão geral dos algoritmos de ML



Regressão Linear

✓ Modelo estatístico cujo objetivo é indicar qual será o comportamento de uma variável dependente (Y) como uma função que contenha uma ou mais variáveis independentes (X).





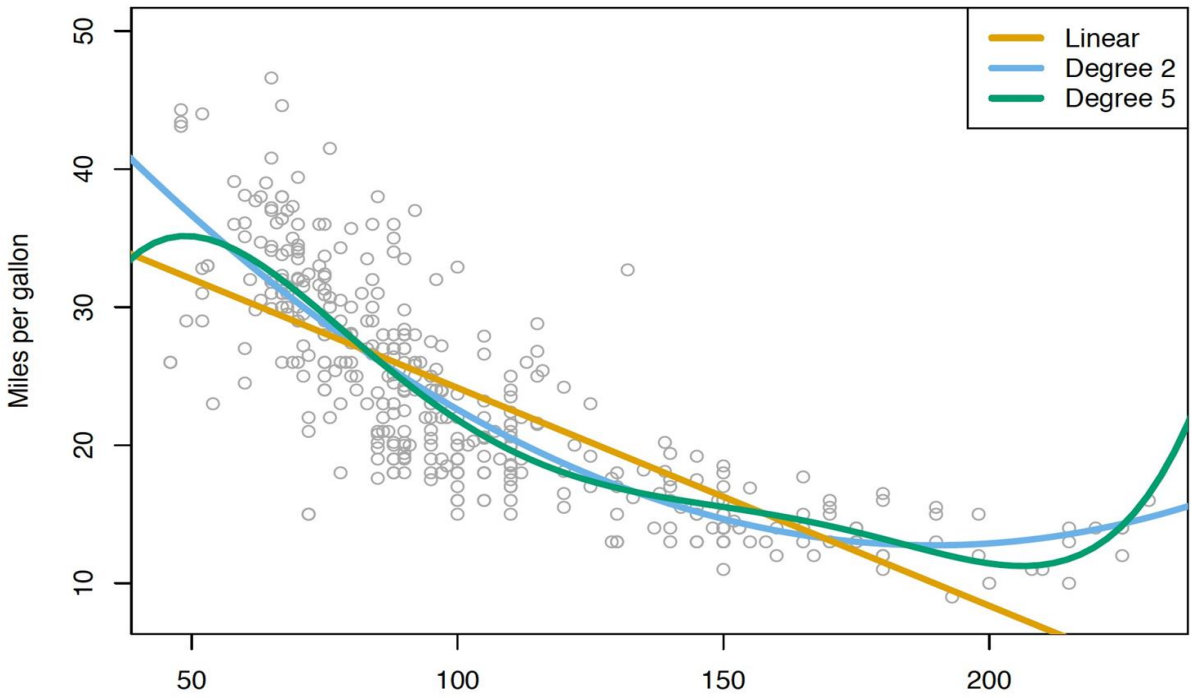
Aprendizagem	Predição
Supervisionada	Regressão

Regressão Polinomial

- ✓O modelo é uma extensão da regressão linear.
- ✓Pode ser usada quando a complexidade de um conjunto de dados excede a possibilidade de ajuste por uma linha reta.

$$h_w(x) = w_1 x + w_2 x^2 + \cdots + w_n x^n + b$$

- ✓onde, a enésima potência é uma dimensão de regressão polinomial (grau).



Comparação entre regressão linear e regressão polinomial

Aprendizagem	Predição
Supervisionada	Regressão

Regressão Logística

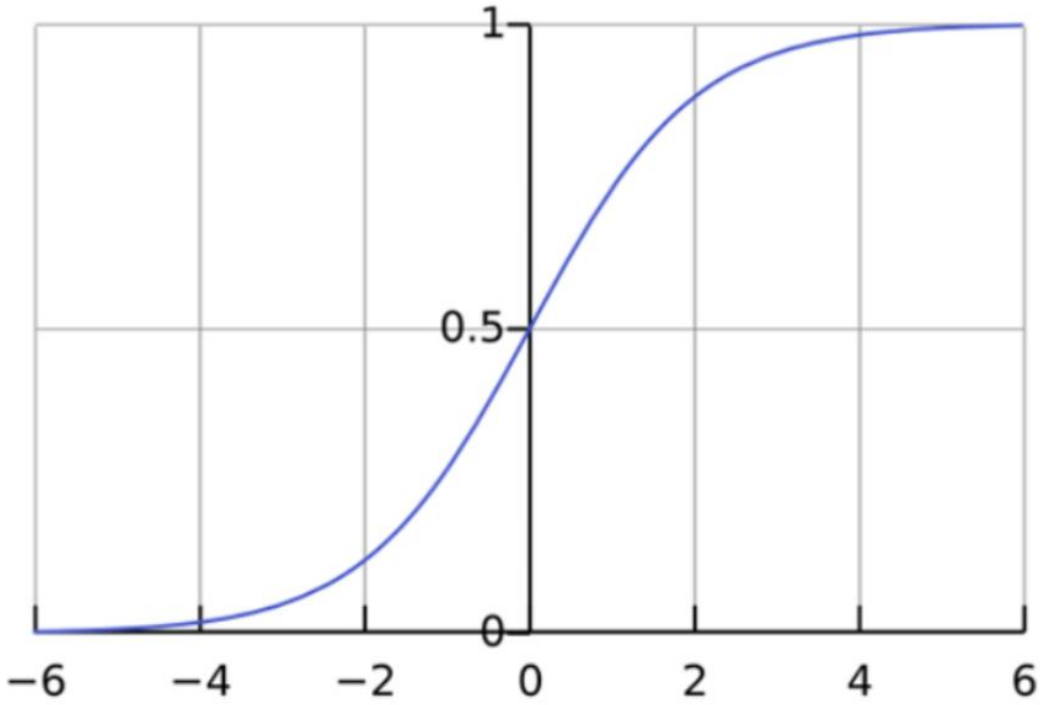
- ✓ O modelo de regressão logística é usado para resolver problemas de **classificação**.

$$P(Y=0|x) = \frac{1}{1+e^{wx+b}} \qquad P(Y=1|x) = \frac{e^{wx+b}}{1+e^{wx+b}}$$

- ✓ Se aplica apenas a problemas de **classificação binária.**
- ✓ **Observe:** $\overbrace{P(Y=0|x)}^{\text{Probabilidade Y=0}} + \overbrace{P(Y=1|x)}^{\text{Probabilidade Y=1}} = 1$
- ✓ Na prática:

Se $P(Y=0|x) > P(Y=1|x)$ classifica como Y = 0

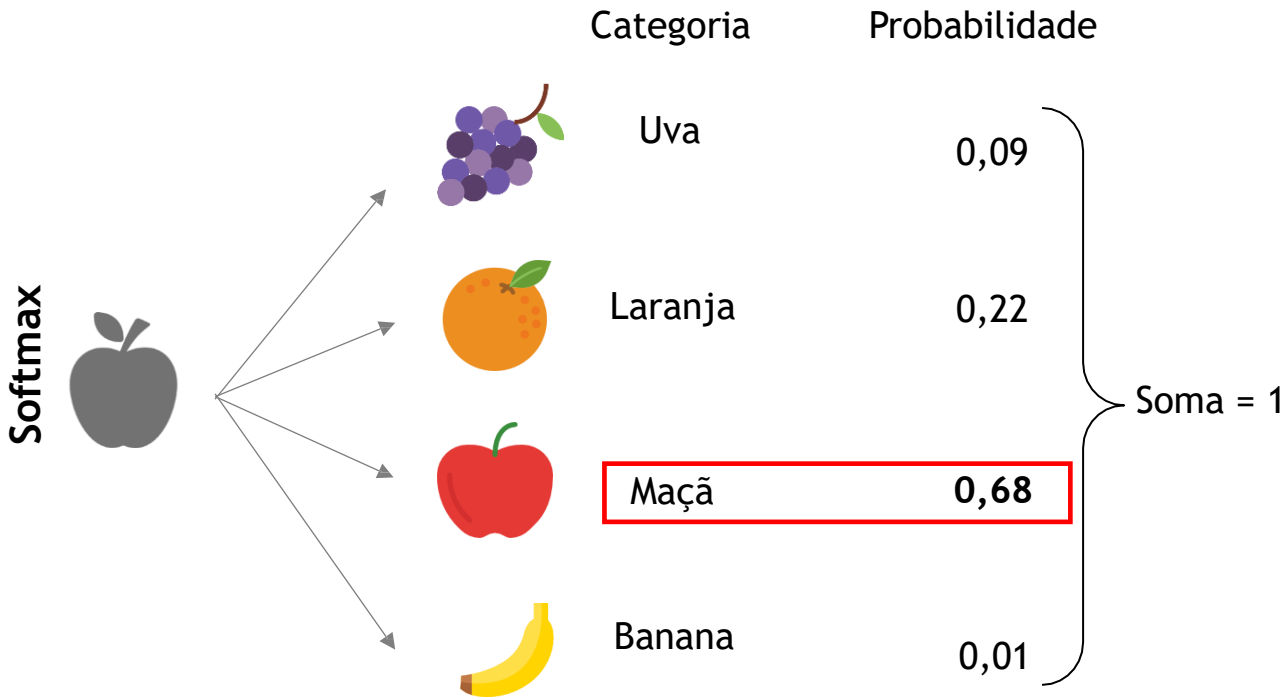
Se $P(Y=0|x) < P(Y=1|x)$ classifica como Y = 1



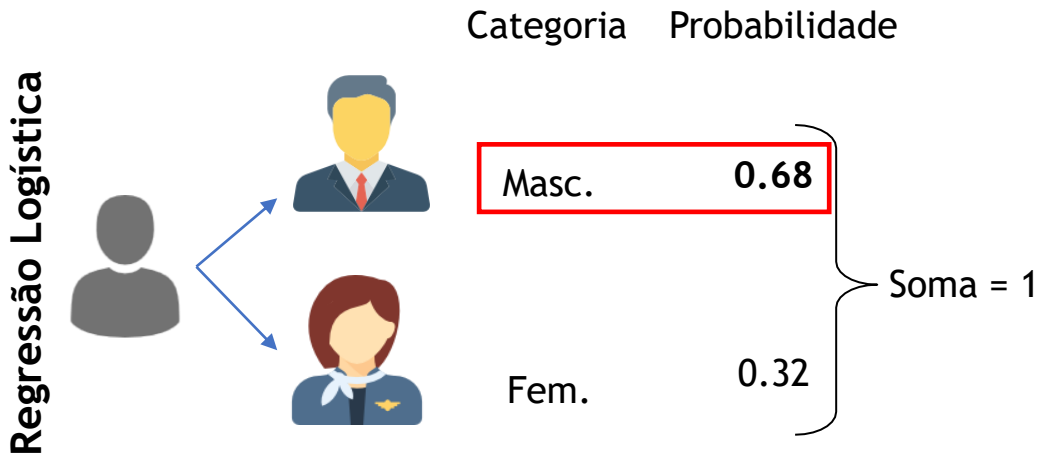
Aprendizagem	Predição
Supervisionada	Classificação

Softmax

- ✓ O modelo é uma extensão da regressão logística.
- ✓ A regressão logística se aplica apenas a problemas de classificação binária.
- ✓ Para problemas de classificação várias classes, use a função **Softmax**.



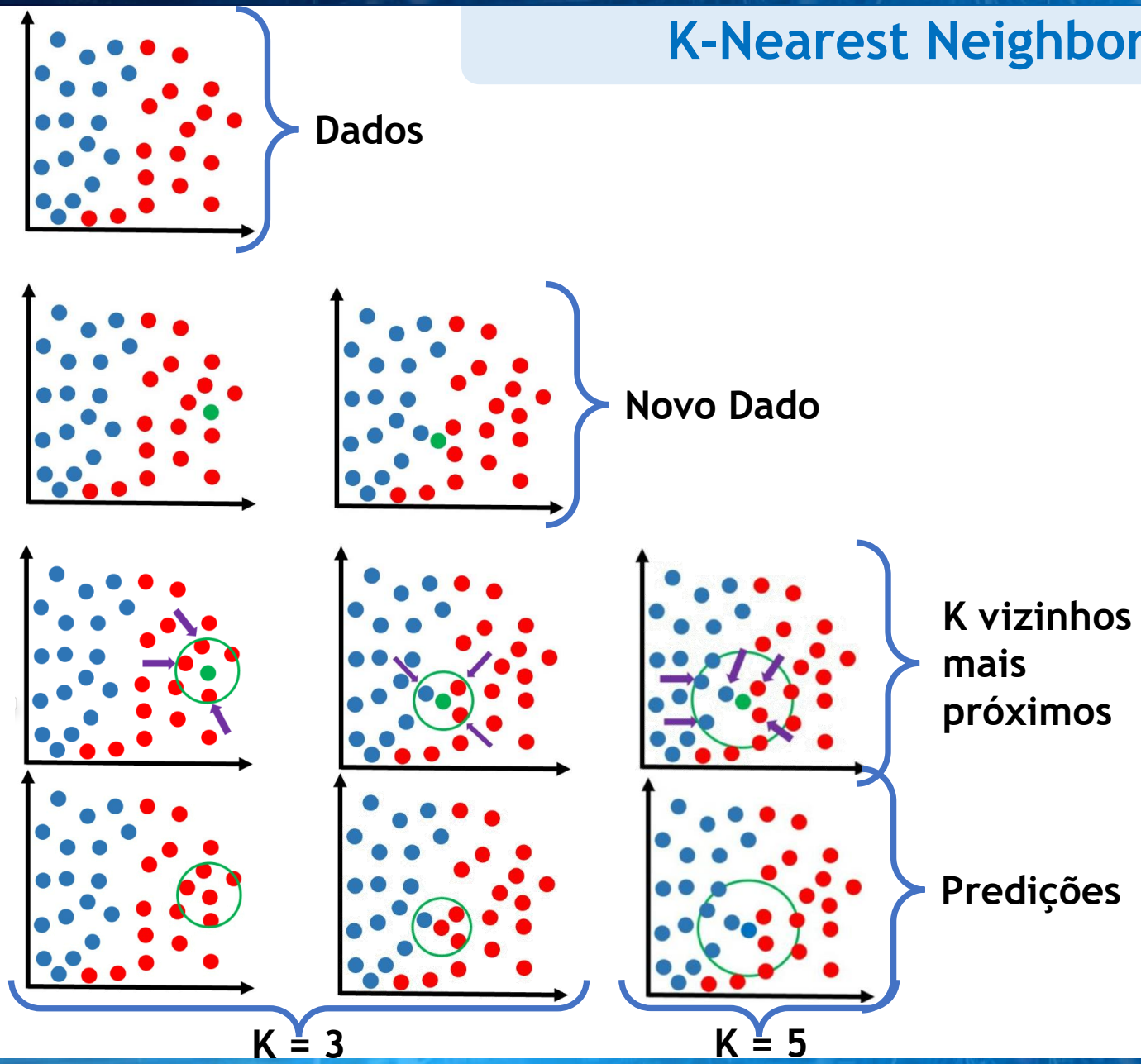
Problema de classificação multi-classe



Problema de classificação binária

Aprendizagem	Predição
Supervisionada	Classificação

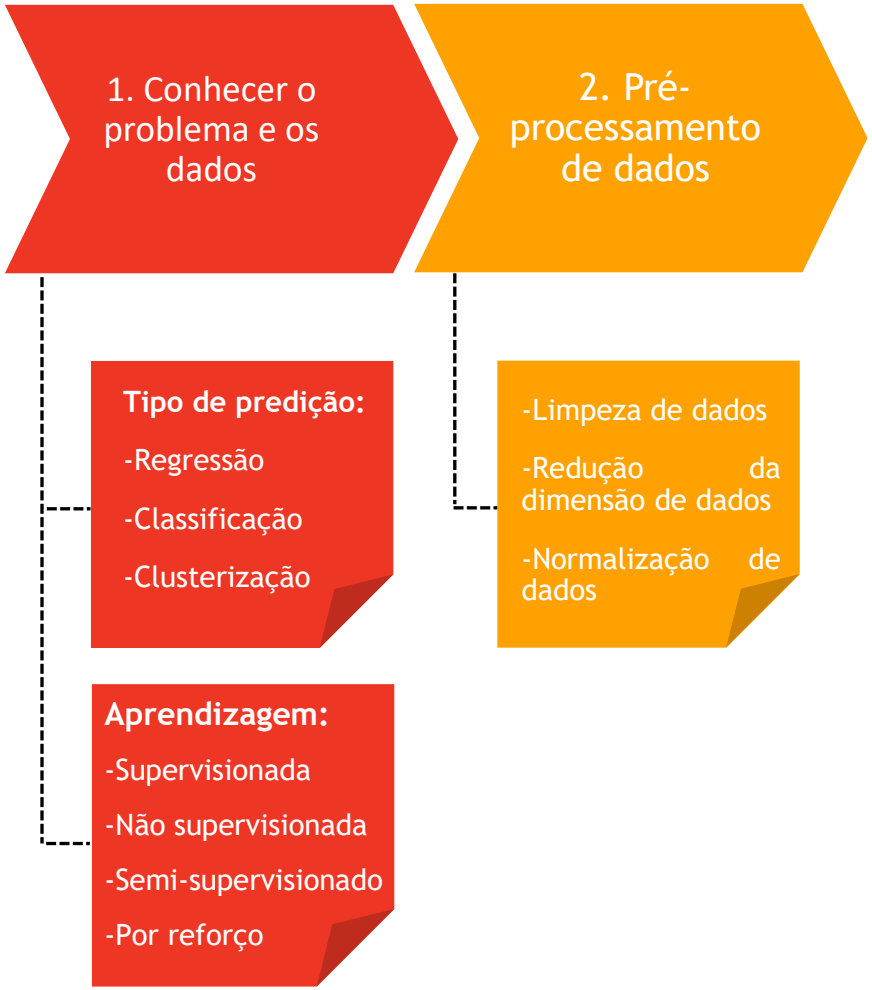
K-Nearest Neighbor - KNN



- ✓ O KNN (K-nearest neighbors, ou "K-vizinhos mais próximos") é usado para problemas de **classificação**.
- ✓ O modelo busca classificar cada amostra de um conjunto de dados avaliando sua distância em relação aos vizinhos mais próximos.
- ✓ Se os vizinhos mais próximos forem majoritariamente de uma classe, a amostra em questão será classificada nesta categoria.

Aprendizagem	Predição
Supervisionada	Classificação

Pipeline ML


1. Conhecer o problema e os dados
2. Pré-processamento de dados
Tipo de predição:
-Regressão
-Classificação
-Clusterização
Aprendizagem:
-Supervisionada
-Não supervisionada
-Semi-supervisionado
-Por reforço
-Limpeza de dados
-Redução da dimensão de dados
-Normalização de dados

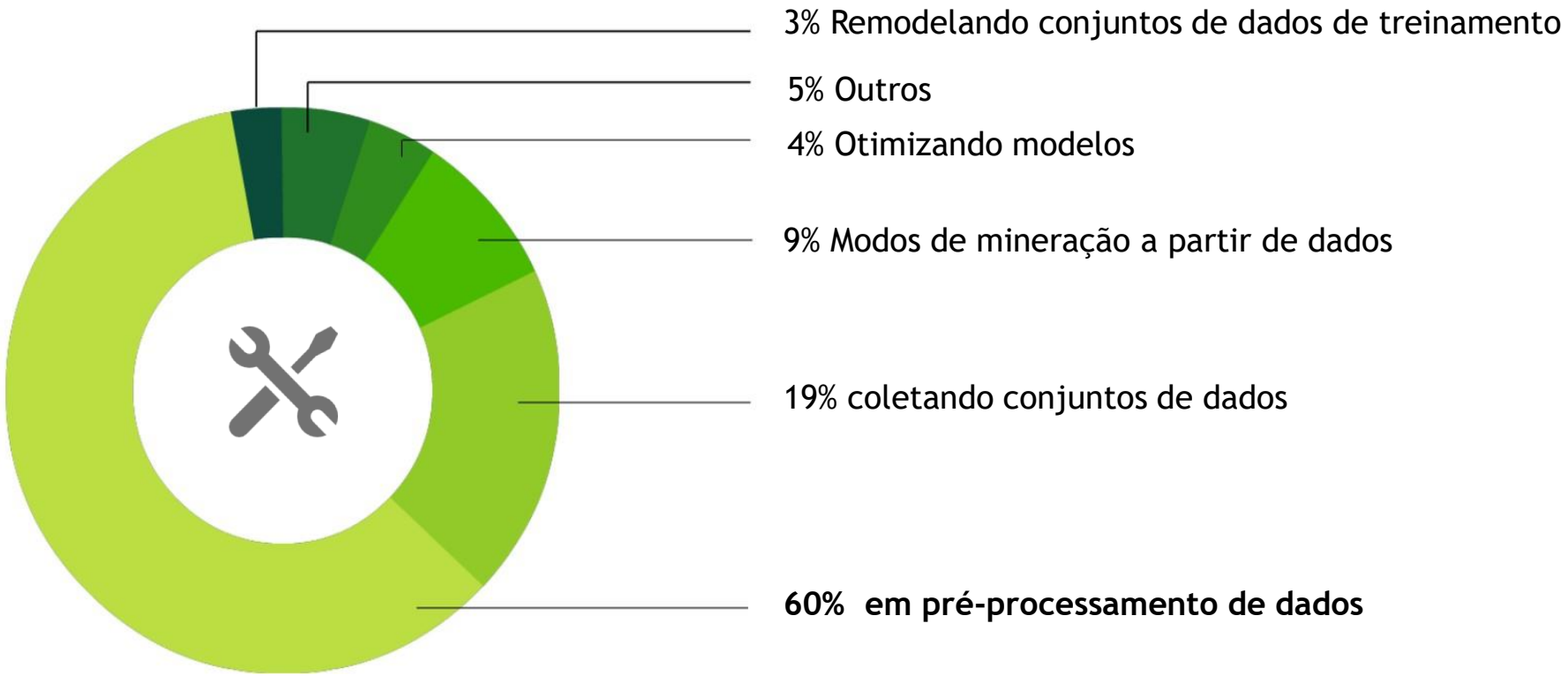
Visão geral dos dados

✓Exemplo de um conjunto típico de dados.

		Recurso 1	Recurso 2	Recurso 3	Rótulo
	No.	Área	Distritos escolares	Direção	Preço da casa
Conjunto de treinamento	1	100	8	Sul	1000
	2	120	9	Sudoeste	1300
	3	60	6	Norte	700
	4	80	9	Sudeste	1100
Conjunto de teste	5	95	3	Sul	850

Pré-processamento de dados

- ✓Sem bons dados, não existe um bom modelo.
- ✓Estatísticas sobre o trabalho de cientistas de dados em aprendizagem de máquina



CrowdFlower Data Science Report 2016

Pré-processamento de dados

#	Identidade	Nome	Aniversário	Gênero	IsTeacher	#Students	País	Cidade
1	111	João	31/12/1990	M	0	0	Irlanda	Dublin
2	222	Mery	15/10/1978	F	1	15	Islândia	
3	333	Alice	19/04/2000	F	0	0	Espanha	Madrid
4	444	Mara	11/01/1997	M	0	0	França	Paris
5	555	Alex	15/03/2000	UMA	1	23	Alemanha	Berlim
6	555	Peter	01/12/83	M	1	10	Itália	Roma
7	777	Calvin	05/05/1995	M	0	0	Itália	Itália
8	888	Roxane	08/03/1948	F	0	0	Portugal	Lisboa
9	999	Anne	09/05/1992	F	0	5	Suíça	Genebra
10	6666	Maria	15/07/92	F	1	250	Brasil	Brasília
11	101010	Paulo	14/11/1992	M	1	26	Ytali	Roma

Item duplicado inválido

Pode ser desnecessária: Reduzir dimensão

Valor ausente

Valor inválido

Valor que deve estar em outra coluna

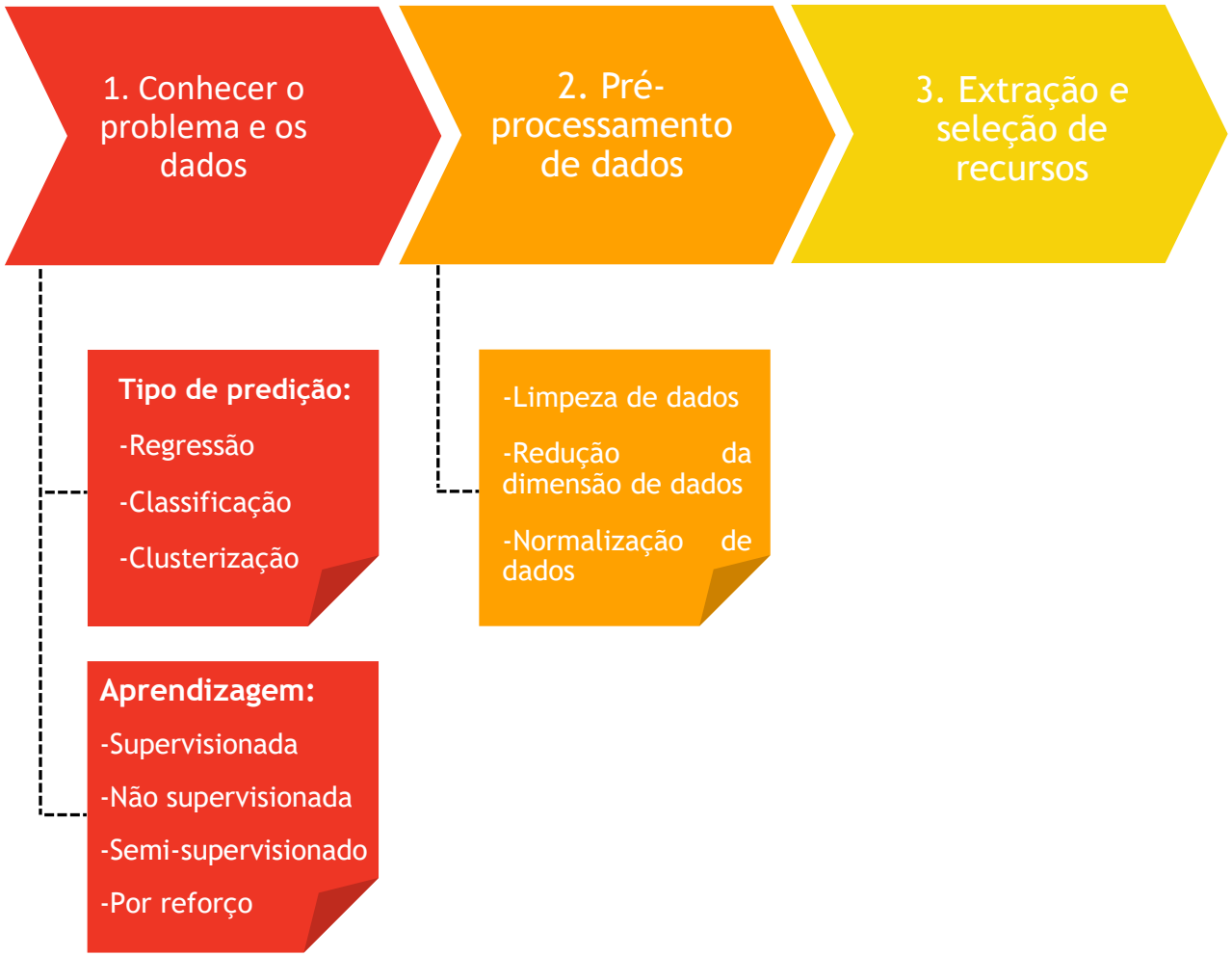
Erro ortográfico

Formato incorreto

Dependência de atributo

Valor discrepante: Normalizar

Pipeline ML


1. Conhecer o problema e os dados
2. Pré-processamento de dados
3. Extração e seleção de recursos
Tipo de predição:
-Regressão
-Classificação
-Clusterização
Aprendizagem:
-Supervisionada
-Não supervisionada
-Semi-supervisionado
-Por reforço
-Limpeza de dados
-Redução da dimensão de dados
-Normalização de dados

Extração e seleção de recursos

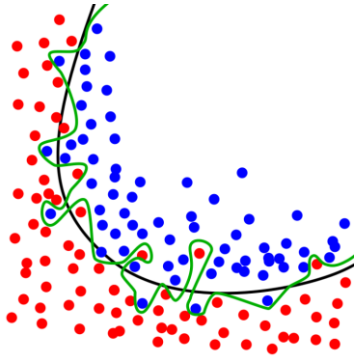
✓Geralmente, um conjunto de dados tem muitos recursos, alguns dos quais podem ser redundantes ou irrelevantes para o valor a ser previsto.

✓A seleção de recursos é necessária nos seguintes aspectos:

✓Simplifica os modelos para torná-los fáceis de serem interpretados pelos usuários

✓Reduz o tempo de treinamento

✓Evita a explosão da dimensão



✓Melhora a generalização do modelo e evita overfitting

Pipeline ML


1. Conhecer o problema e os dados
2. Pré-processamento de dados
3. Extração e seleção de recursos
4. Treinamento de modelo
Tipo de predição:
-Regressão
-Classificação
-Clusterização
Aprendizagem:
-Supervisionada
-Não supervisionada
-Semi-supervisionado
-Por reforço
-Limpeza de dados
-Redução da dimensão de dados
-Normalização de dados

Treinamento do modelo

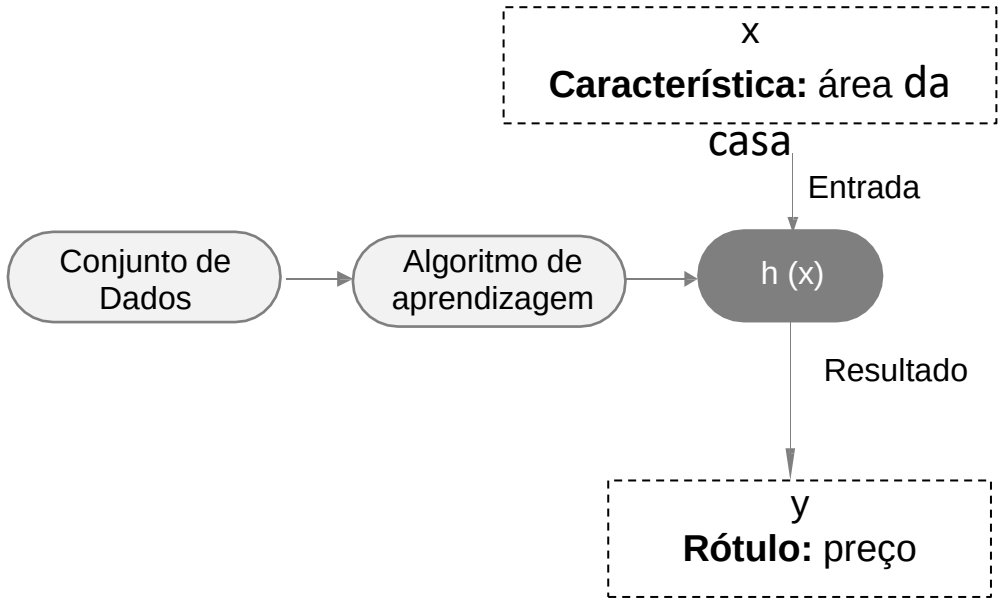
✓ Suponha que haja um conjunto de dados contendo as áreas das casas e os preços de 21.613 unidades habitacionais vendidas em uma cidade. Com base nesses dados, podemos prever os preços de outras casas da cidade.

Área da Casa	Preço
1.180	221.900
2.570	538.000
770	180.000
1.960	604.000
1.680	510.000
5.420	1.225.000
1.715	257.500
1.060	291.850
1.160	468.000
1.430	310.000
1.370	400.000
1.810	530.000
…	…

Conjunto de Dados de Treinamento

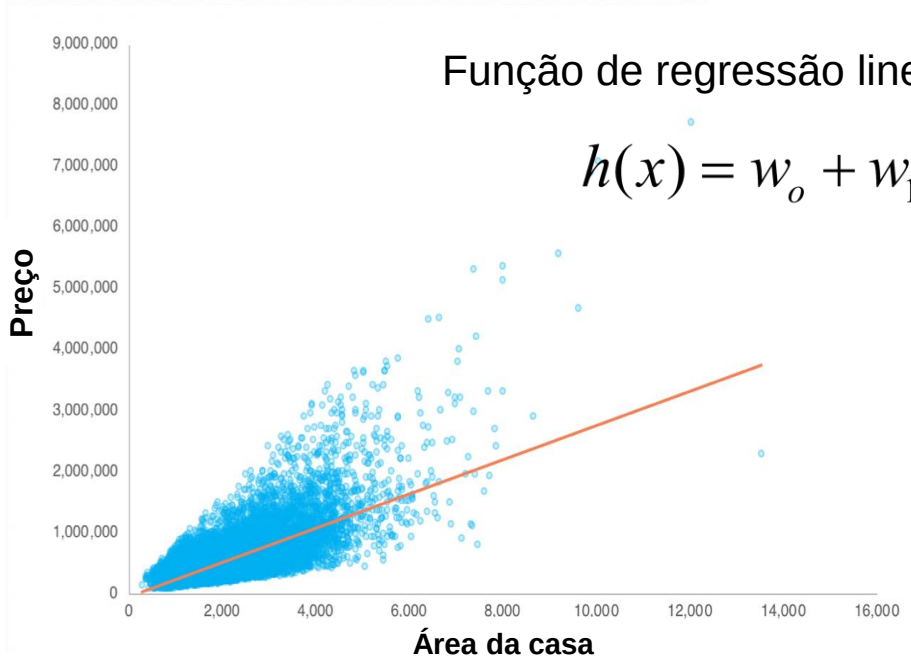
Treinamento do modelo

- ✓ Como os dados são rotulados usaremos **aprendizagem supervisionada.** Considerando as características do problema podemos usar **regressão linear.**
- ✓ Nosso objetivo é construir uma função de modelo **h (x)** que se aproxime da função que expressa a distribuição verdadeira do conjunto de dados.



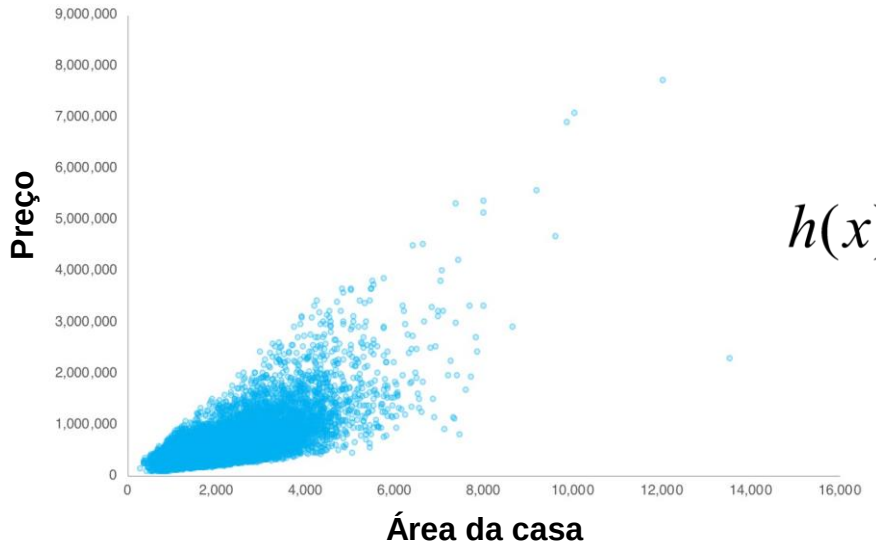
Função de regressão linear unária

$$h(x) = w_o + w_1 x$$

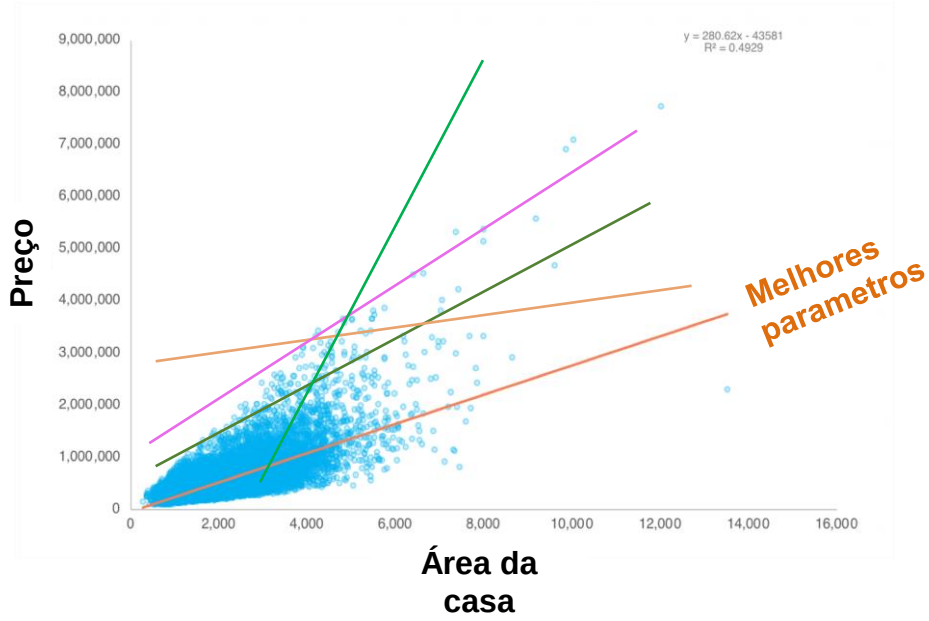


Treinamento do modelo

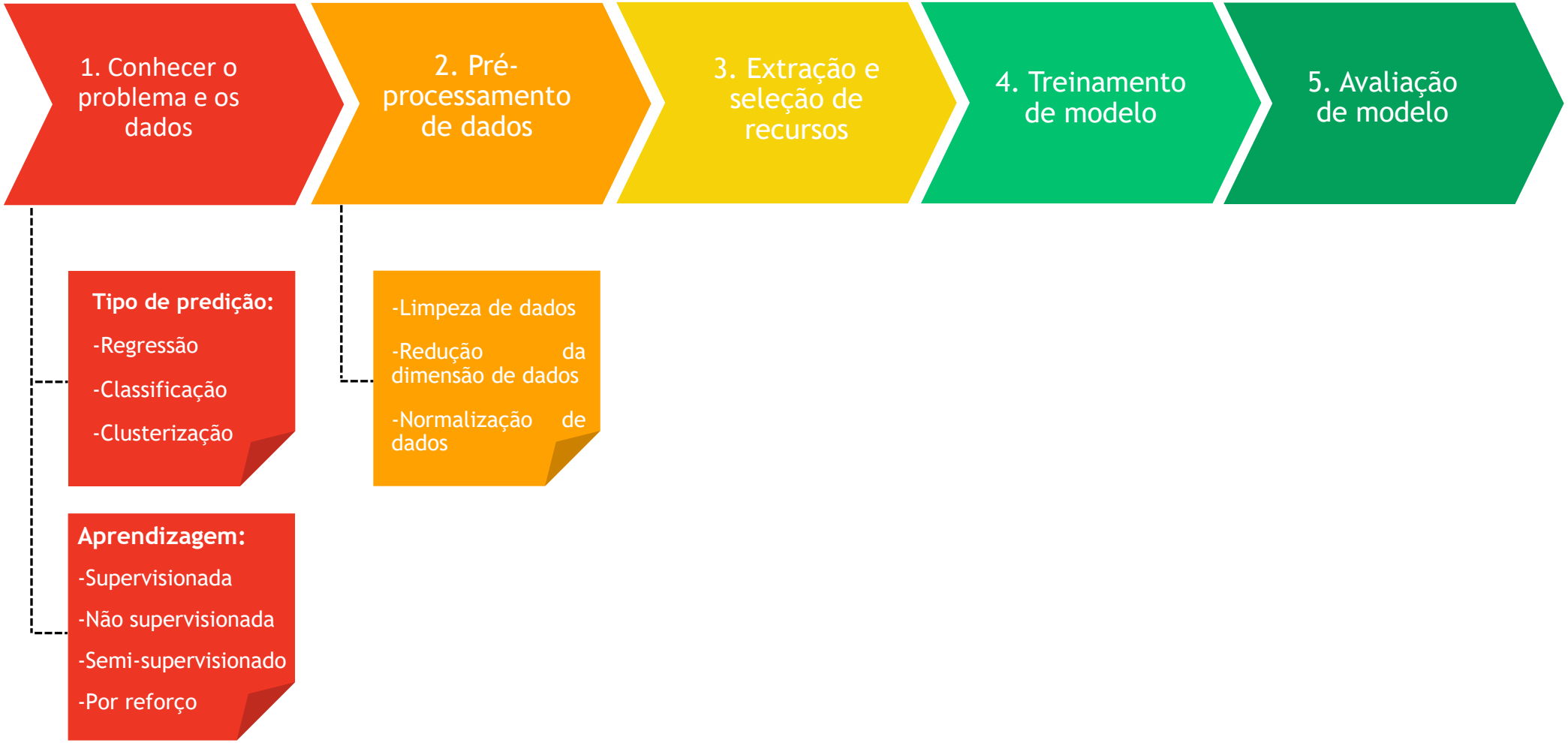
- ✓ A Regressão Linear visa encontrar uma linha reta que melhor se adapte ao conjunto de dados.
- ✓ Em resumo: treinar um modelo de Regressão Linear significa aprender/descobrir os melhores parâmetros w_0 e w_1.
- ✓ A ideia de encontrar os melhores parâmetros é aplicável a outros modelos de ML.



$$h(x) = w_o + w_1 x$$



Pipeline ML


1. Conhecer o problema e os dados
2. Pré-processamento de dados
3. Extração e seleção de recursos
4. Treinamento de modelo
5. Avaliação de modelo
Tipo de predição:
-Regressão
-Classificação
-Clusterização
Aprendizagem:
-Supervisionada
-Não supervisionada
-Semi-supervisionado
-Por reforço
-Limpeza de dados
-Redução da dimensão de dados
-Normalização de dados

Avaliação do modelo

- ✓Analisar a **capacidade de generalização** do modelo:
  - ✓O objetivo da aprendizagem de máquina é que o modelo obtido após o aprendizado tenha um bom desempenho em novas amostras, não apenas em amostras usadas para treinamento.
  - ✓A capacidade de aplicar um modelo a novas amostras é chamada de generalização ou robustez.
  - ✓Por isso dividimos os dados em Conjunto de dados treinamento e Conjunto de dados de testes

Validação Cruzada

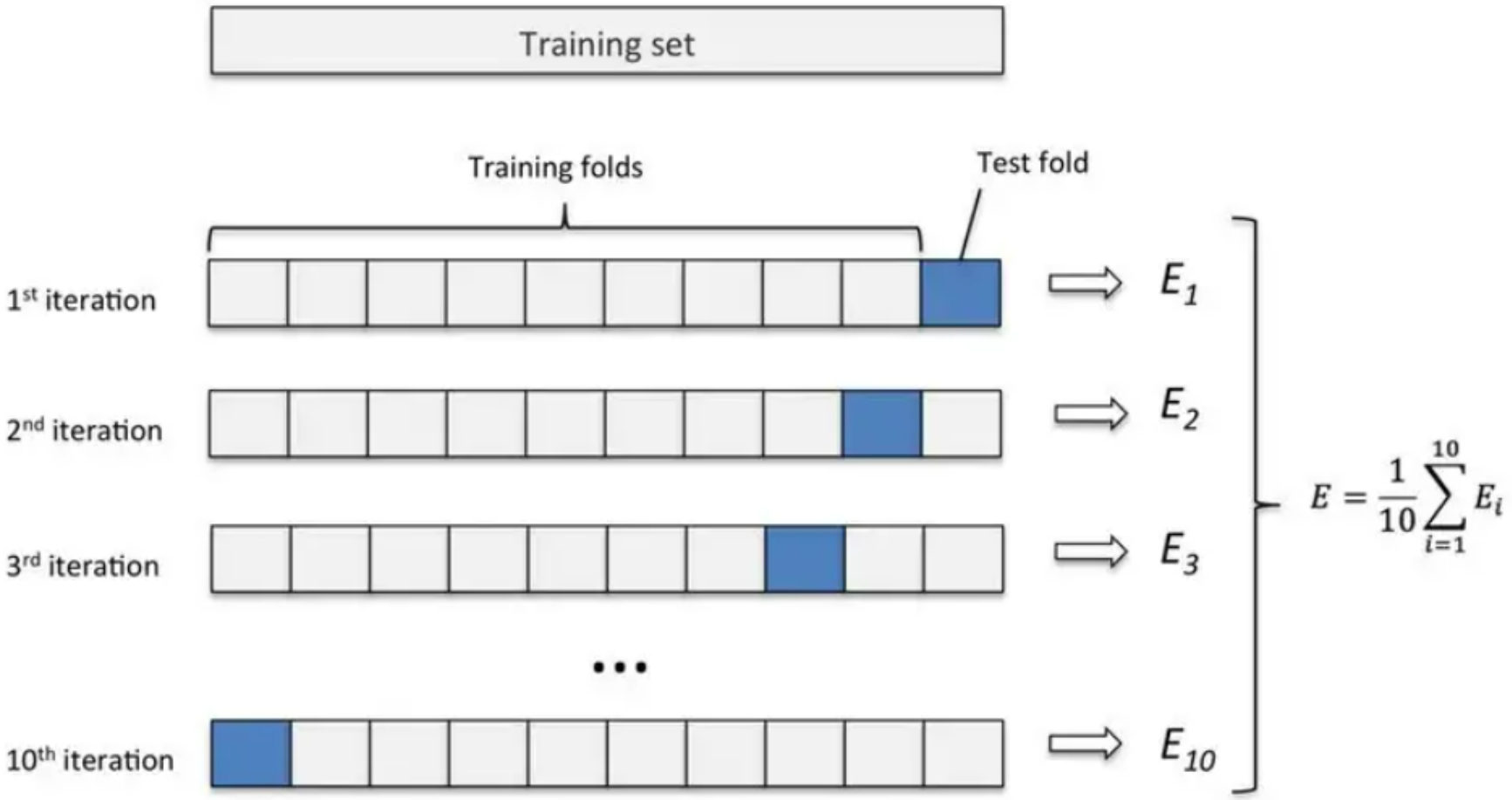
- ✓ Observaram acima que treinamos e testemos um modelo da seguinte forma:

- ✓ Nessas condições, é comum trabalharmos como:
  - ✓ Conjunto de treinamento (80%) e Conjunto de teste (20%)

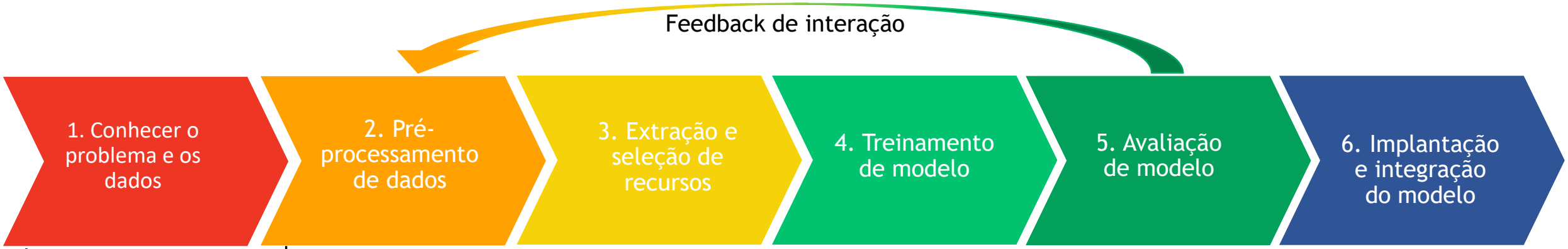
- ✓ Mas observe que, nessa partição dos dados, podemos criar cenários que não representação o conjunto todo?
- ✓ Exemplo: conjunto muito diferente → má avaliação do modelo. Ou seja, a análise da capacidade de generalização foi comprometida!

Validação Cruzada

✓ Para tanto, temos a Validação cruzada (k-fold cross-validation). O Conjunto de dados (dataset) é dividido aleatoriamente em “K” grupos.



Pipeline ML



Tipo de predição:

-Regressão

-Classificação

-Clusterização

Aprendizagem:

-Supervisionada

-Não supervisionada

-Semi-supervisionado

-Por reforço

-Limpeza de dados

-Redução da dimensão de dados

-Normalização de dados

Praticando exercícios

Exemplo de Questão

1) Qual o tipo de predição realizada por um método que prêve se a receita de uma loja no próximo trimestre será 200-300 ou 300-400, com base em histórico de dados rotulados.

a) Regressão
b) Classificação
c) Por regra
d) Clusterização

Exemplo de Questão

2) Qual das seguintes afirmações é verdadeira sobre o aprendizado não-supervisionado?

a) O algoritmo não-supervisionado processa apenas "recursos" e não processa rótulos
b) O algoritmo Hierarchical clustering não é aprendizado não-supervisionado
c) Os algoritmos K-means e SVM pertencem ao aprendizado não-supervisionado
d) Nenhuma das acima

Exemplo de Questão

3) Qual dos seguintes algoritmos não é aprendizado supervisionado?

a) Regressão linear
b) árvore de decisão
c) KNN
d) K-means

Exemplo de Questão

4) Quais são os tipos comuns de dados sujos (Dirty data) ?

a) valor malformado
b) Valor duplicado
c) valor logicamente errado
d) valor ausente

Exemplo de Questão

5) A validação cruzada (k-fold cross-validation) refere-se à divisão do conjunto de dados de teste em K sub-conjuntos de dados .

a) Verdadeiro
b) Falso

UFRR

DEPTO DE CIÊNCIA DA COMPUTAÇÃO

Softex

ABCIA
ABC DA INTELIGÊNCIA ARTIFICIAL

Até a próxima aula!

Continue praticando :)