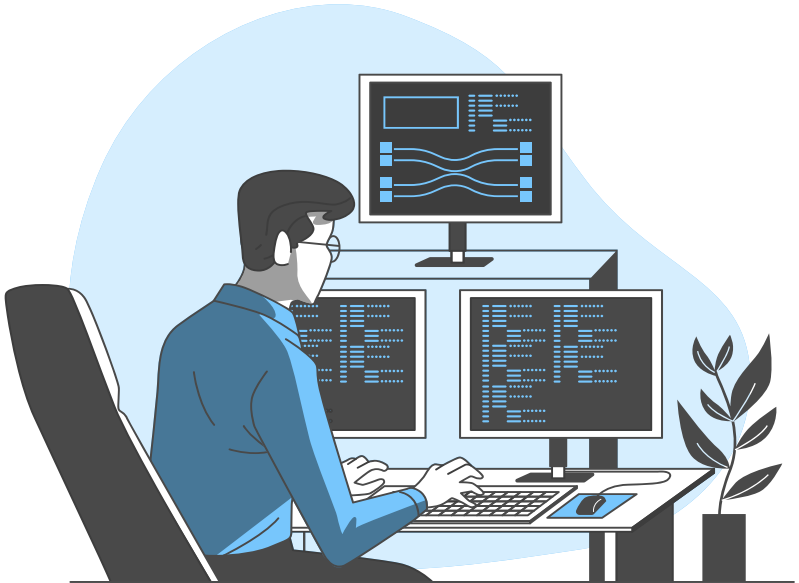
Data Science



Algoritmo KNN

K-Vizinhos mais próximos

KNN – (K Nearest Neighbor)

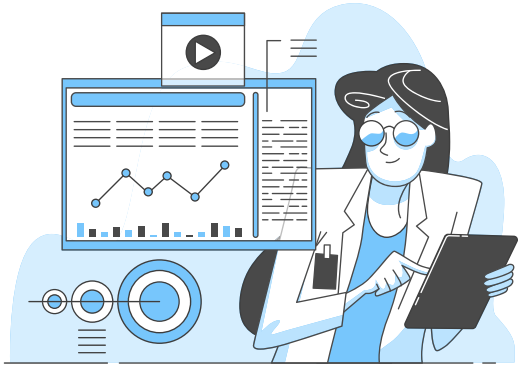
Utiliza uma técnica simples e bastante intuitiva

O aprendizado é baseado em instâncias:

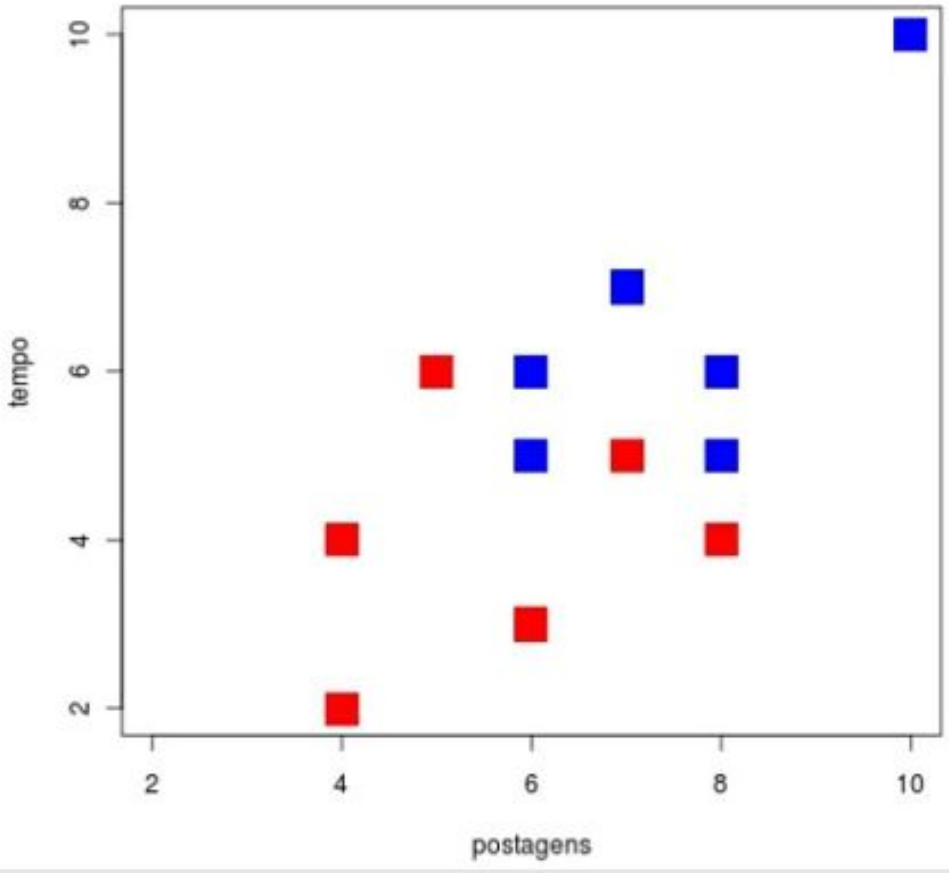
- Aprendizado: Armazena todas as instâncias de treinamento
- Classificação: Descobre a qual classe uma nova instância pertence a partir dos seus vizinhos.

Exemplo: Base de Alunos

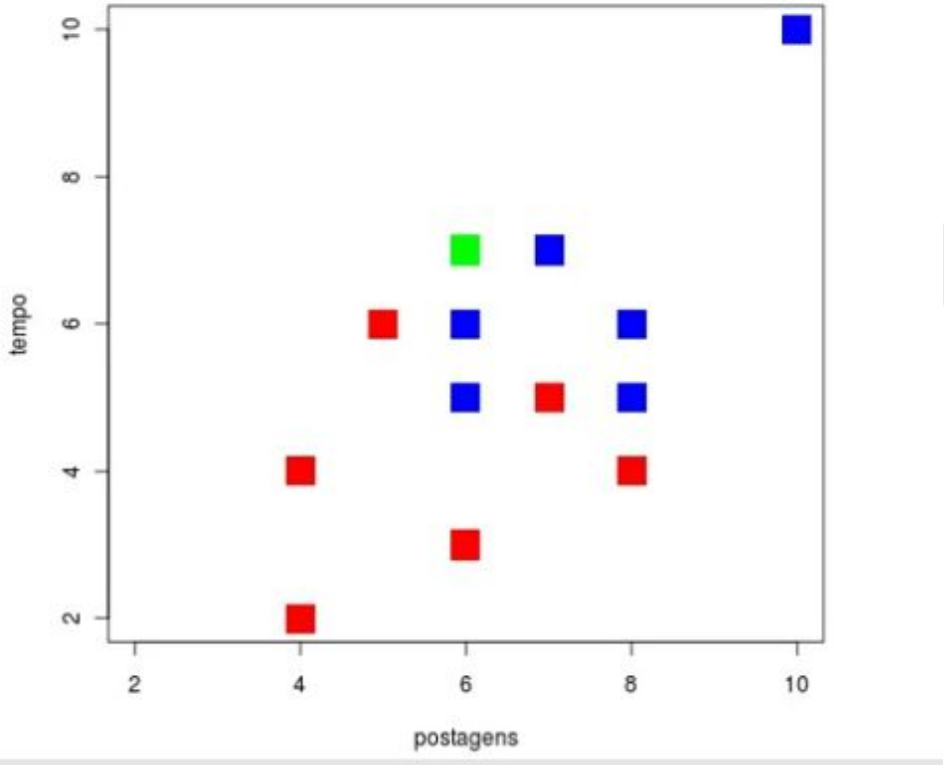
Tempo no Site - x1	Número Postagens - x2	Aprovado? - y
2	4	Não
3	6	Não
4	8	Não
4	4	Não
5	7	Não
6	6	Sim
6	5	Sim
7	7	Sim
8	5	Sim
8	6	Sim
10	10	Sim



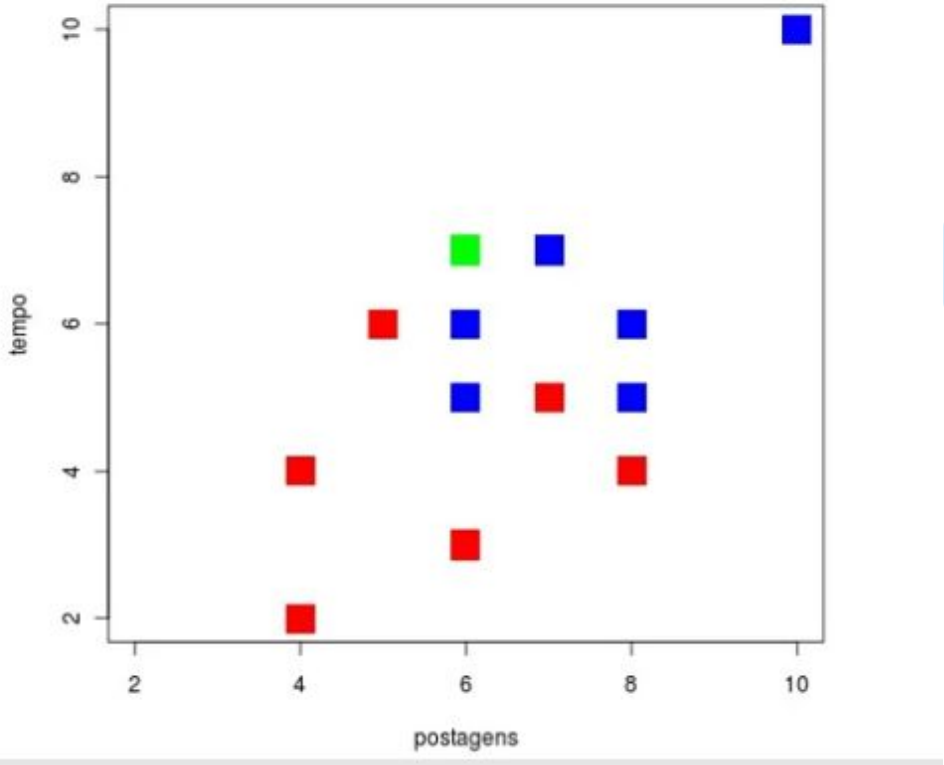
Vamos dispor em um gráfico em azul os aprovados e em vermelho os reprovados.
(comparando o tempo X postagens)



Um aluno que investiu 7 horas de estudo e teve 6 postagens (verde), será classificado como Aprovado?



Um aluno que investiu 7 horas de estudo e teve 6 postagens (verde), será classificado como Aprovado?

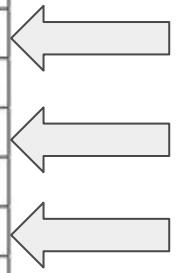


Passo 1: Calcular a distância do novo registro a cada um dos registros existentes.

- Dados dois pontos, $A(x_1^A ... x_n^A)$ e $B(x_1^B ... x_n^B)$;
- Distância euclidiana quadrada: $d(A, B) = \sum_{i=1}^{n}(x_i^A - x_i^B)^2$.

x_1 : **utilização**	x_2 : **postagens**	**Distância para o** $(6, 7)$
2	4	$(2-6)^2 + (4-7)^2 = 25$
3	6	10
4	8	5
4	4	13
5	7	1
6	5	4
6	6	1
6	5	4
7	7	1
8	5	8
8	6	5
10	10	25

x_1 : utilização	x_2 : postagens	Distância para o $(6, 7)$
2	4	$(2-6)^2 + (4-7)^2 = 25$
3	6	10
4	8	5
4	4	13
5	7	1
6	5	4
6	6	1
6	5	4
7	7	1
8	5	8
8	6	5
10	10	25



- Considerando k=3

Obviamente, nem tudo são flores...

Normalização

Precisamos normalizar os dados.

...

Empates

Como devemos tratar eventuais empates?

...

Vizinhos

Qual o melhor número de K a ser utilizado?

...

Vamos trabalhar com um Dataset chamado Irís

Fonte do Dataset: https://www.kaggle.com/datasets/uciml/iris

O conjunto de dados consiste em 150 registros, sendo 50 amostras de cada uma das três espécies de Iris (Iris setosa, Iris virginica e Iris versicolor). Quatro variáveis foram medidas em cada amostra: o comprimento e a largura das sépalas e pétalas, em centímetros.

As colunas disponíveis no dataset são:

* Id
* SepalLengthCm - Comprimento da Sépala
* SepalWidthCm - Largura da Sépala
* PetalLengthCm - Comprimento da Pétala
* PetalWidthCm - Largura da Pétala
* Species (São 3 espécies: 0-Iris Setosa, 1-Iris Versicolour ou 2- Iris Virginica)

Íris Dataset

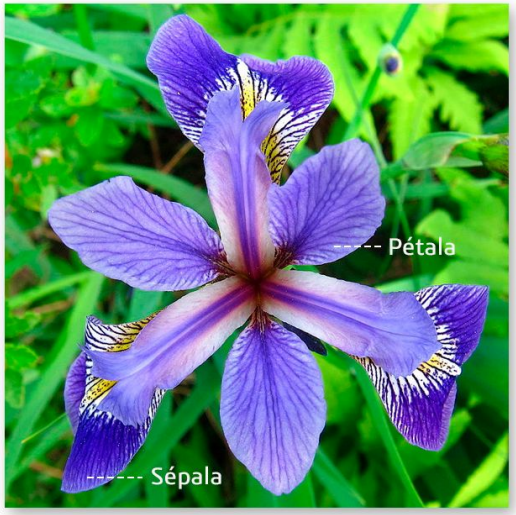
Base de dados das Flores de Íris

Iris flower dataset

Setosa

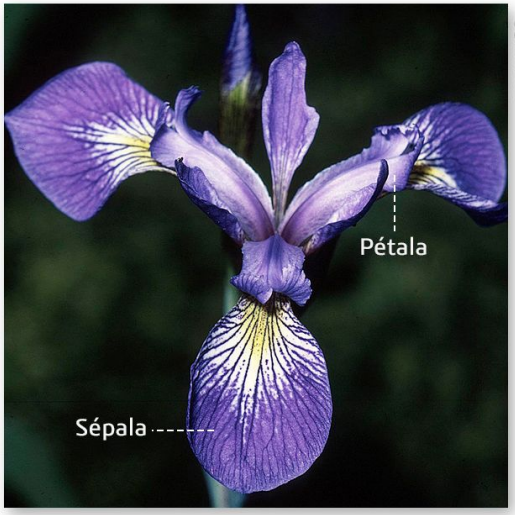
Tiia Monto, CC BY-SA 4.0, via Wikimedia Commons

Versicolor



Charles de Mille-Isles from Mille-Isles, Canada, CC BY 2.0, via Wikimedia Commons

Virginica



Robert H. Mohlenbrock. Courtesy of USDA NRCS, Public domain, via Wikimedia Commons

Implementando o Knn na prática!

O objetivo é dado a largura e comprimento da Sépala e da Pétala, iremos predizer qual tipo de Íris é.