



Data Science
Academy

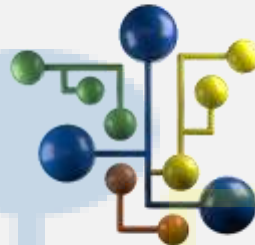
Data Science Academy gregorio@dexpertio.com.br 6004a7f1e32fc346864dec3f

Machine Learning



Data Science
Academy

Data Science Academy gregorio@dexpertio.com.br 6004a7f1e32fc346864dec3f



**Data Science
Academy**

Seja muito bem-vindo(a)!



Data Science
Academy

Data Science Academy gregorio@dexpertio.com.br 6004a7f1e32fc346864dec3f



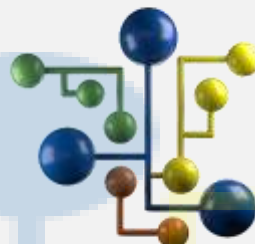
**Data Science
Academy**

K Nearest Neighbors



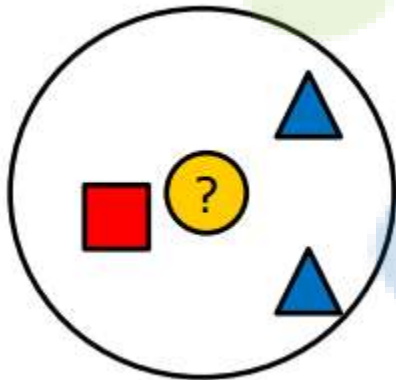
Data Science
Academy

Data Science Academy gregorio@dexpertio.com.br 6004a7f1e32fc346864dec3f



**Data Science
Academy**

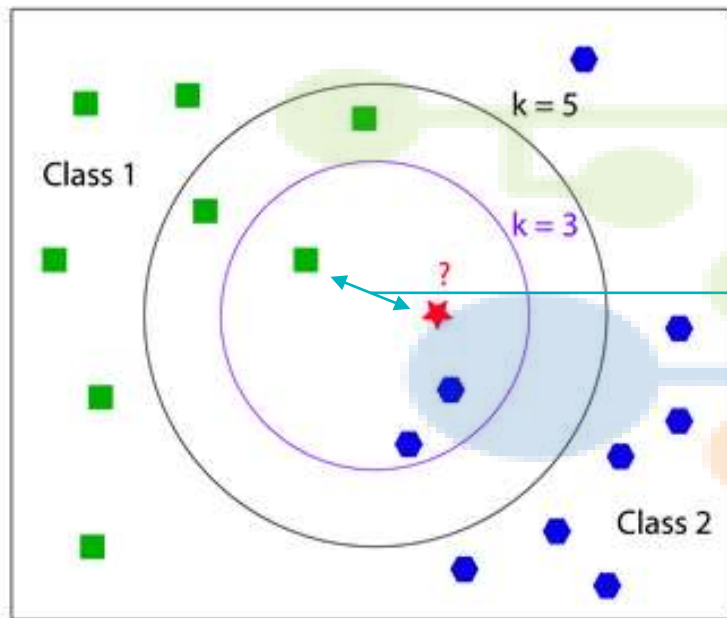
Conhecendo o Algoritmo KNN



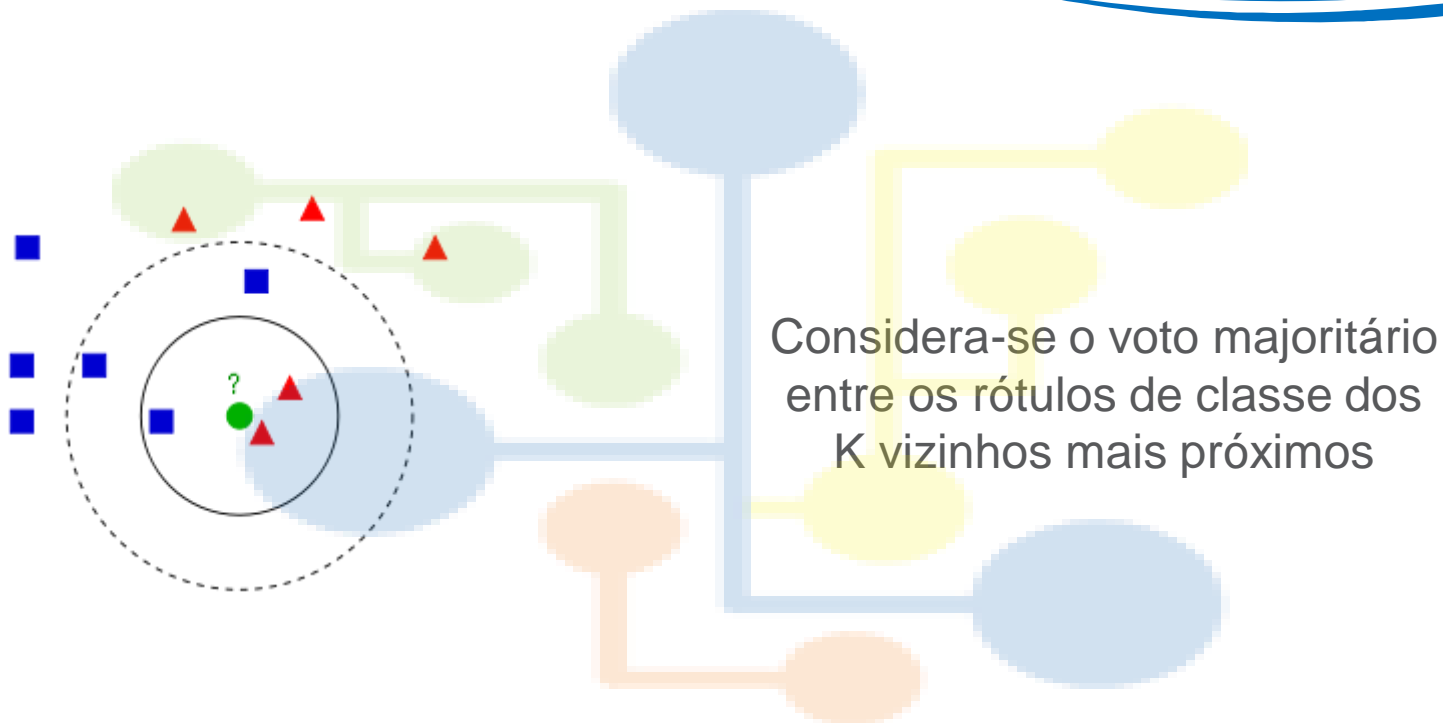
1- Dados de Treinamento

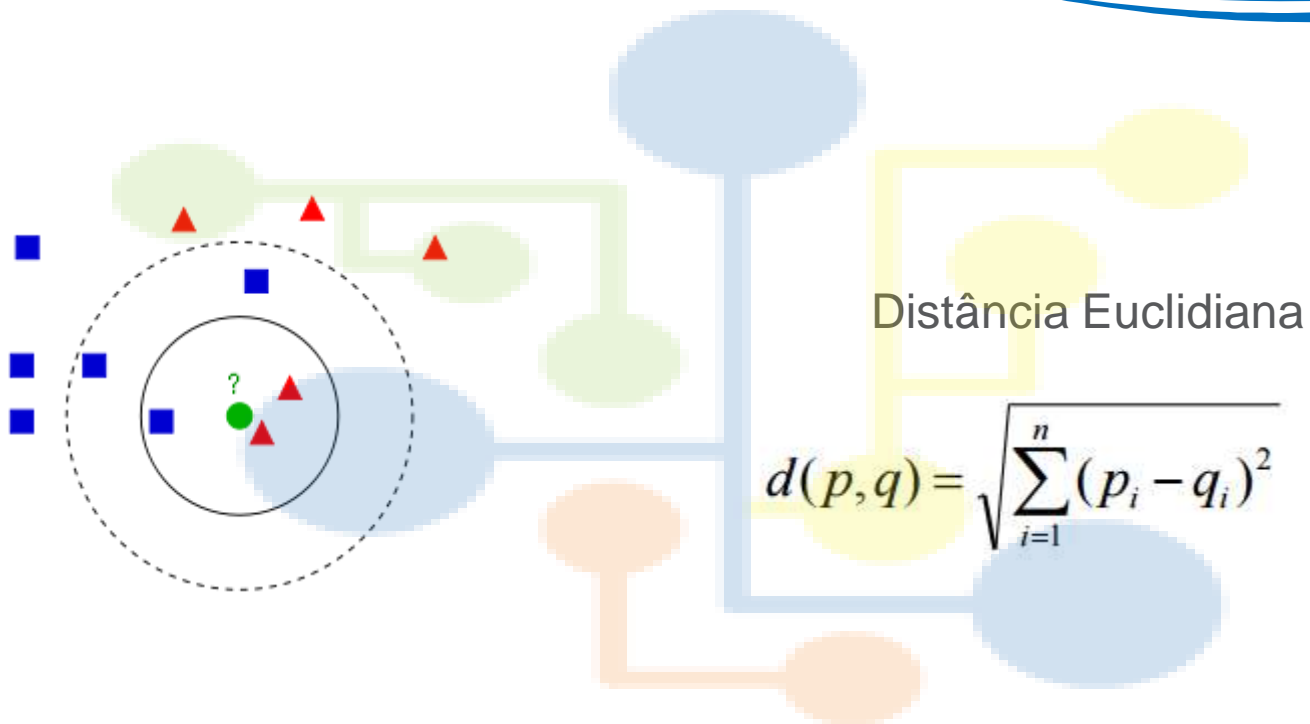
2- Definir a métrica para cálculo da distância

3- Definir o valor de K (número de vizinhos mais próximos que serão considerados pelo algoritmo)



Cálculo da distância entre
o exemplo desconhecido
e o outros exemplos do
conjunto de treinamento





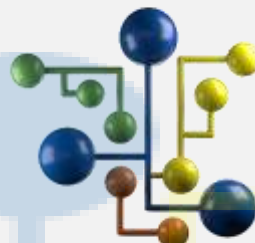




Os dados devem estar normalizados antes de aplicar o algoritmo.

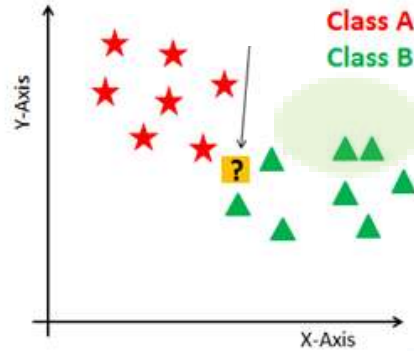


Data Science Academy gregorio@dexpertio.com.br 6004a7f1e32fc346864dec3f

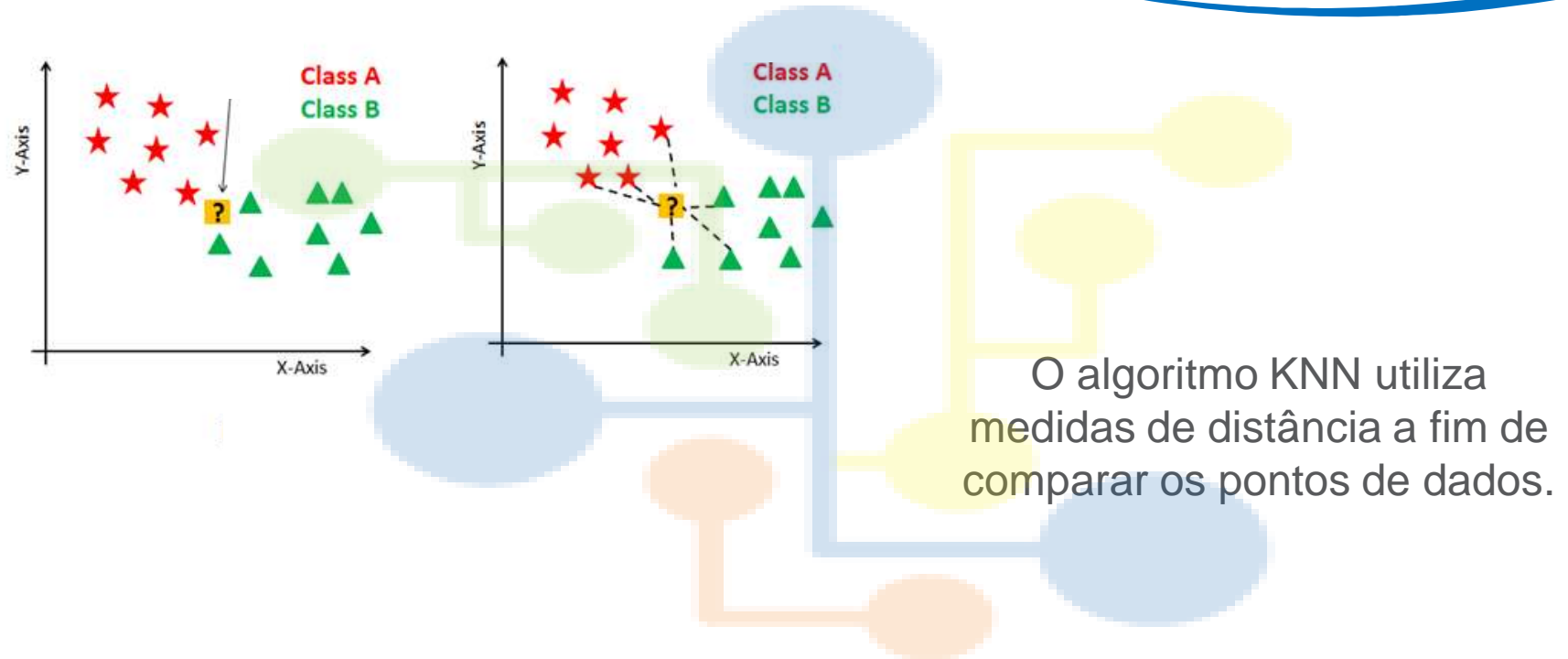


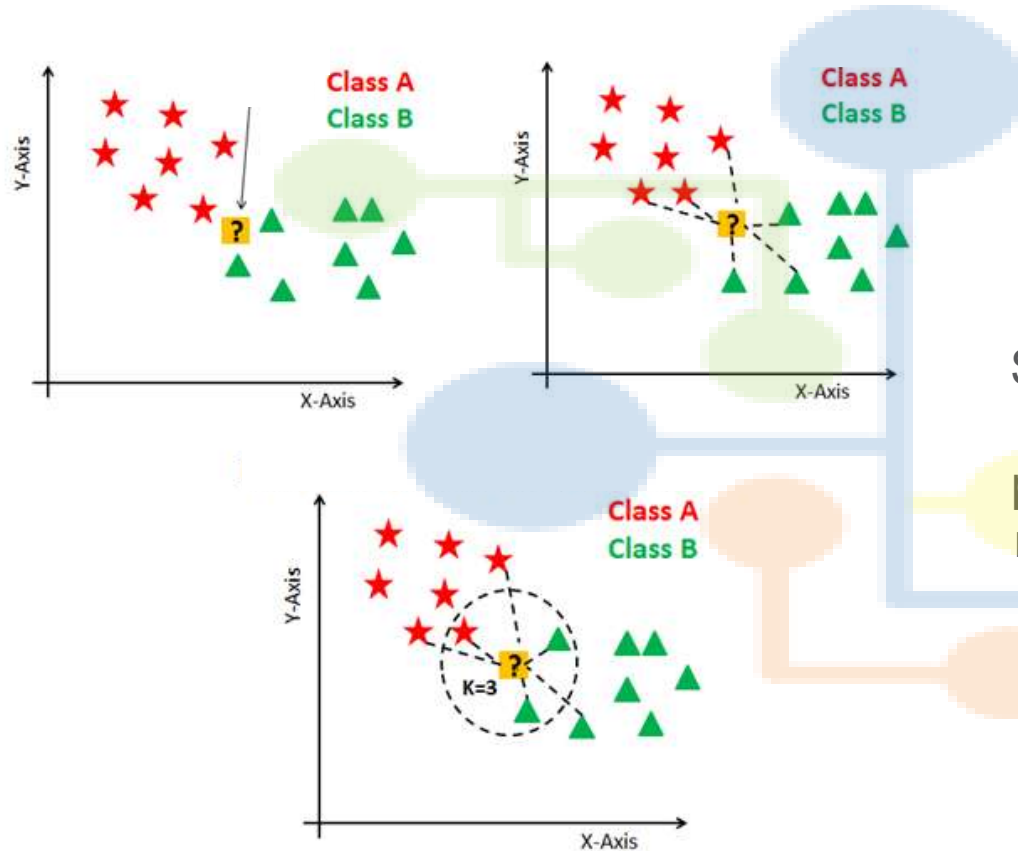
**Data Science
Academy**

KNN e Estrutura de Células de Voronoi

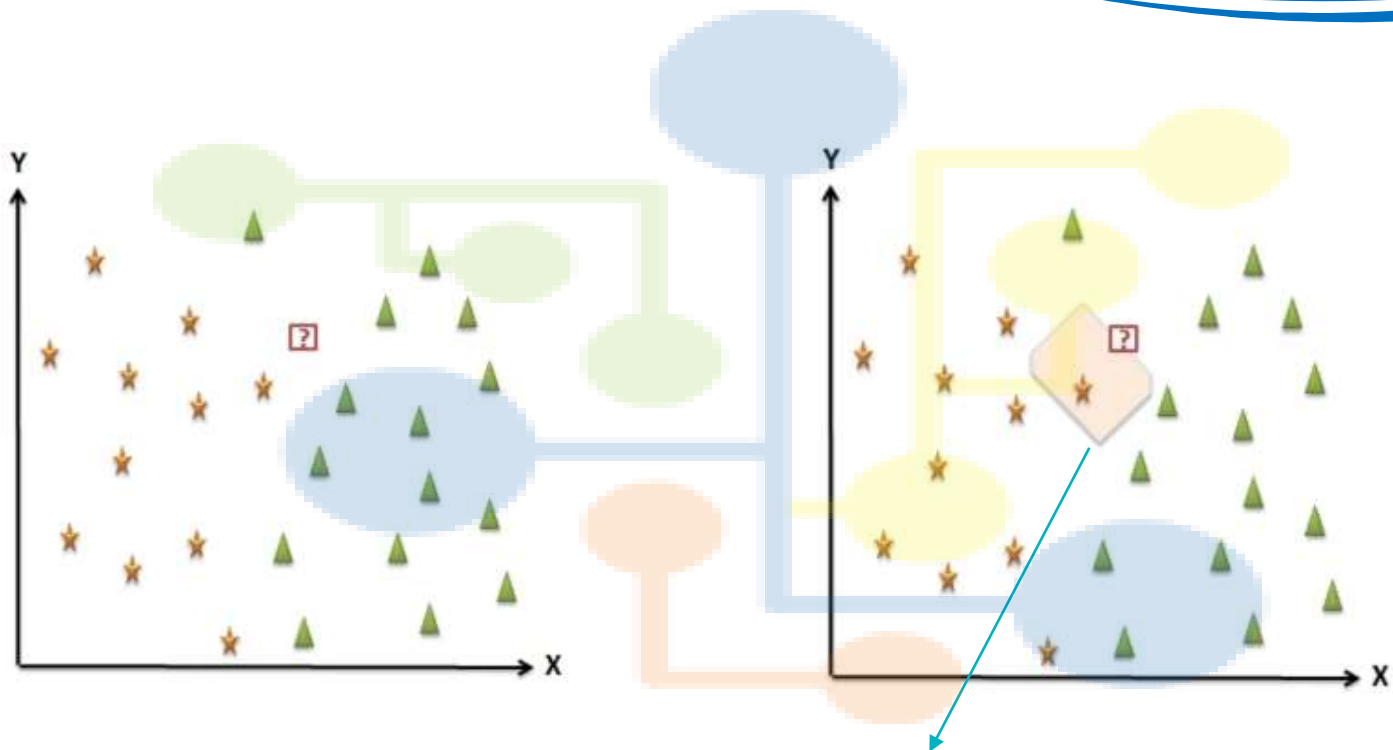


A pergunta que queremos responder é: a qual dos 2 grupos o ponto amarelo com a interrogação pertence?

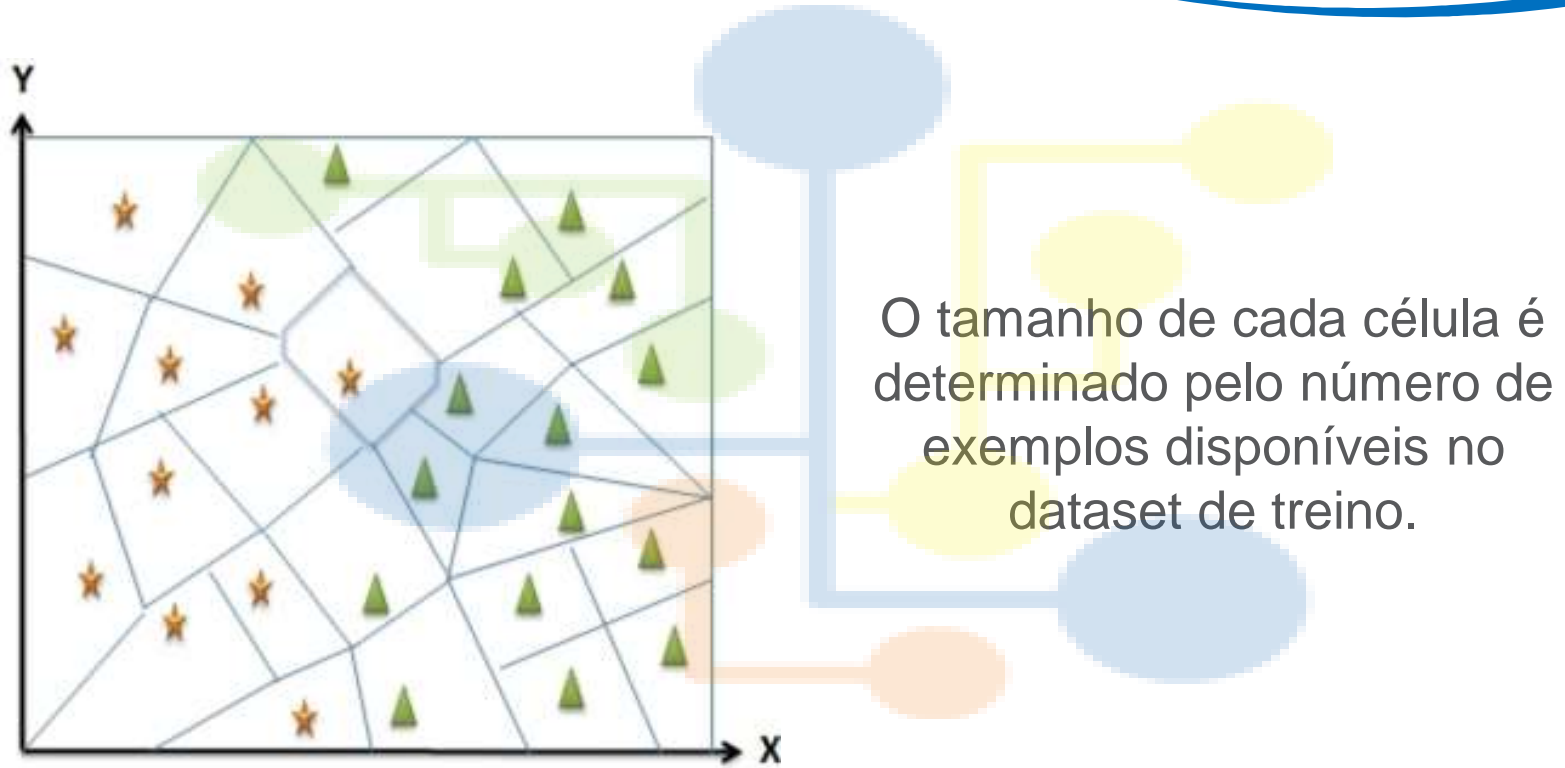


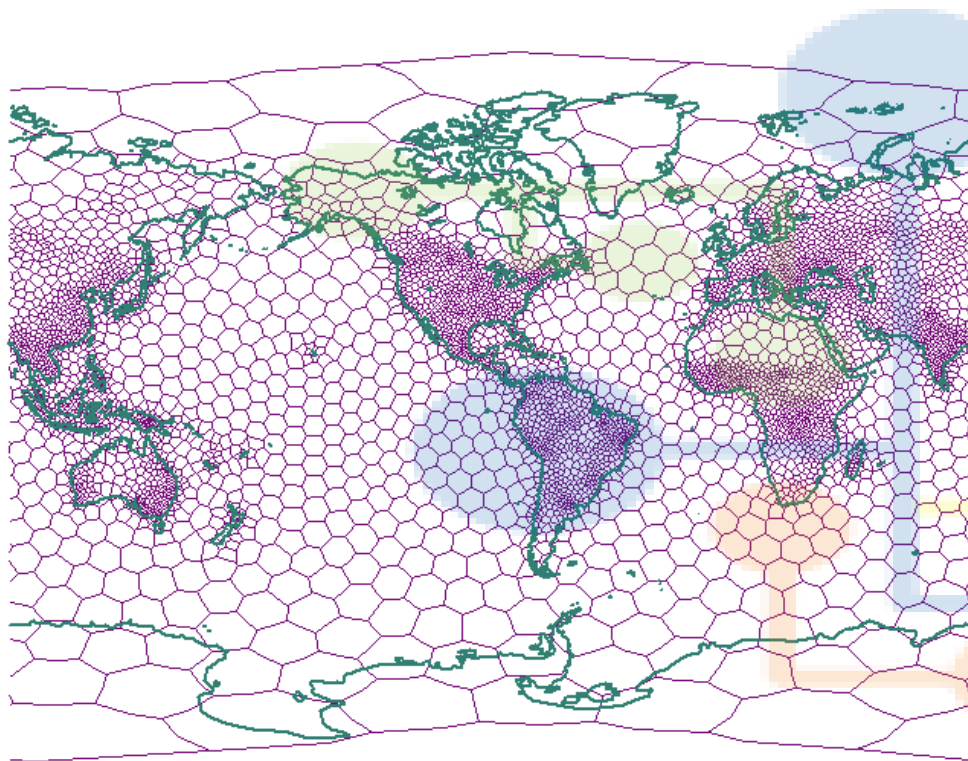


Se usamos um valor de K igual a 3, o KNN vai comparar o novo ponto de dado com os 3 vizinhos mais próximos em cada classe e então fazer uma votação.

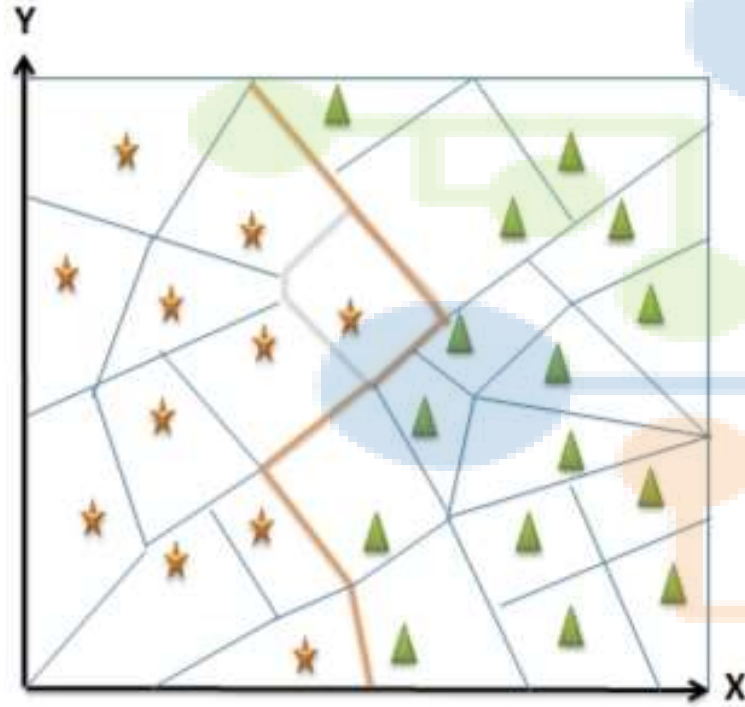


Célula Voronoi





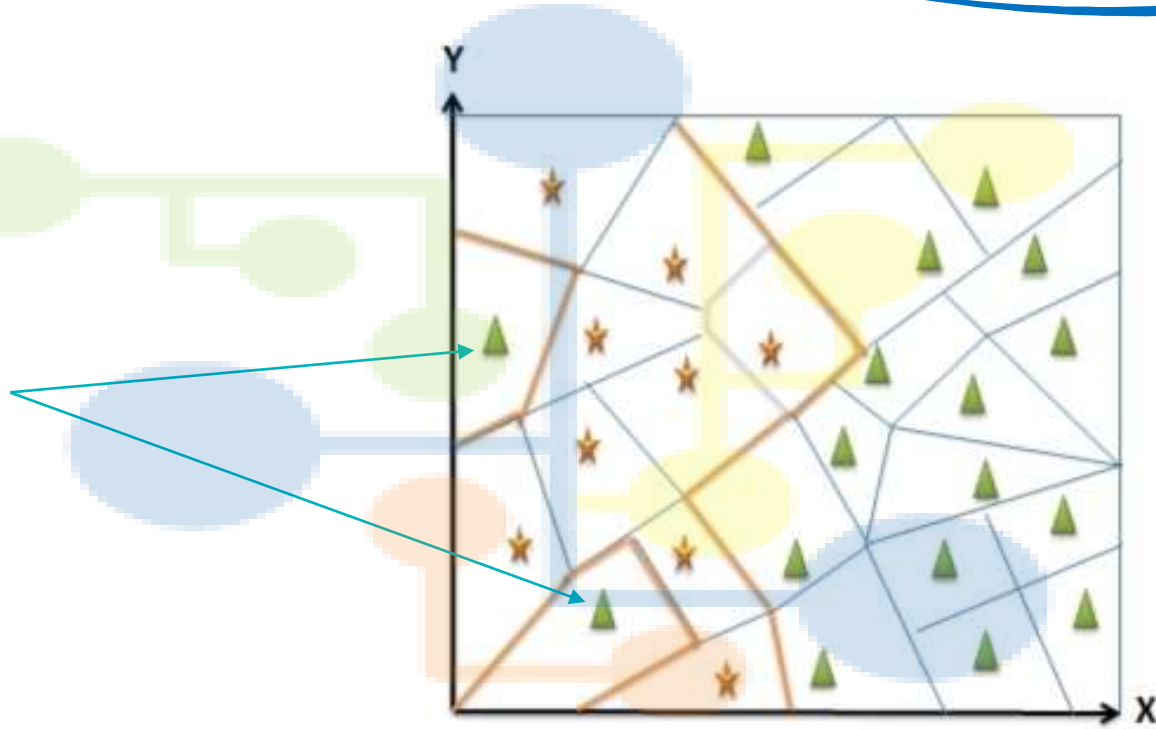
*Estrutura de célula de
Voronoi*



Um aspecto interessante
dessa separação por células
é que existe uma fronteira
que forma a separação entre
as classes de dados.

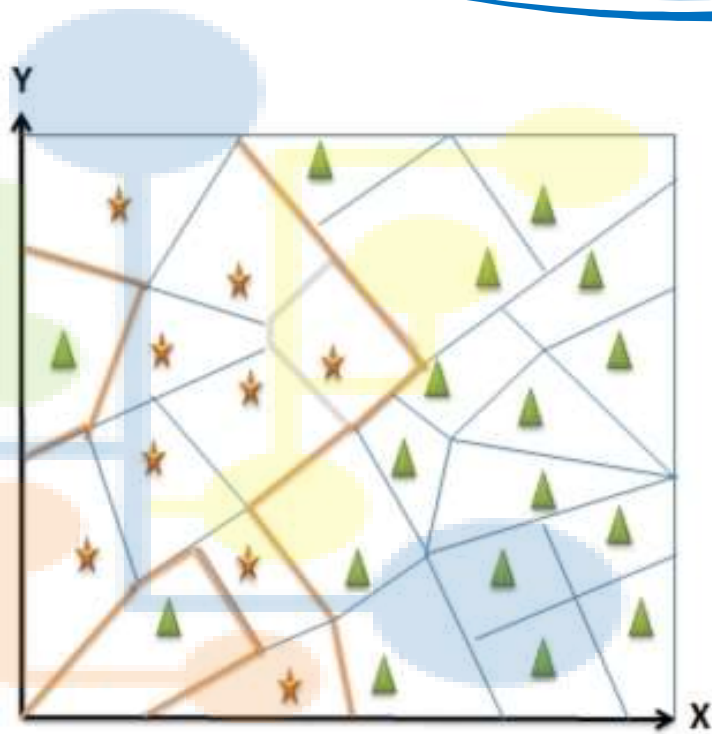


Outliers



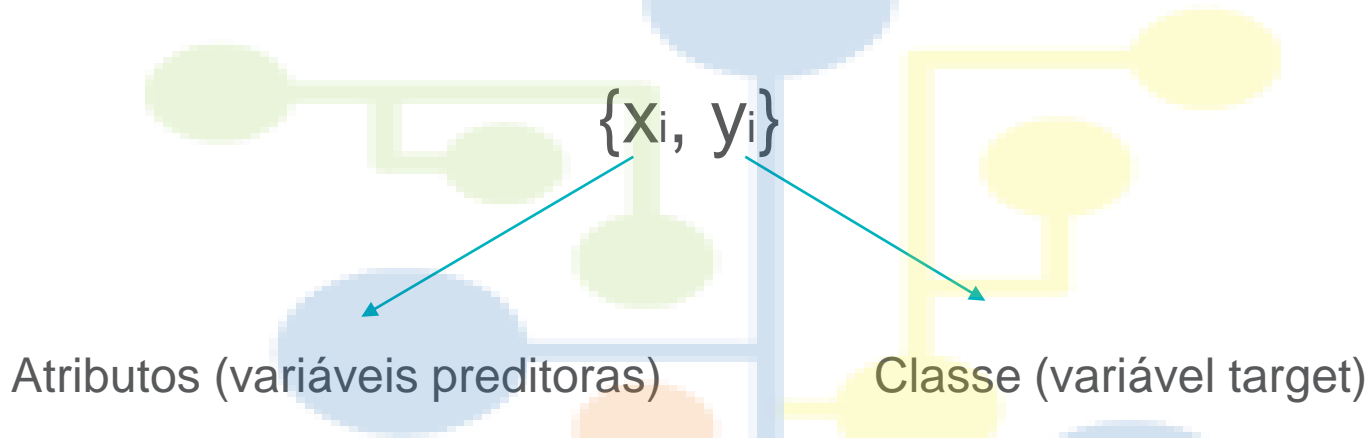


k-vizinhos mais próximos



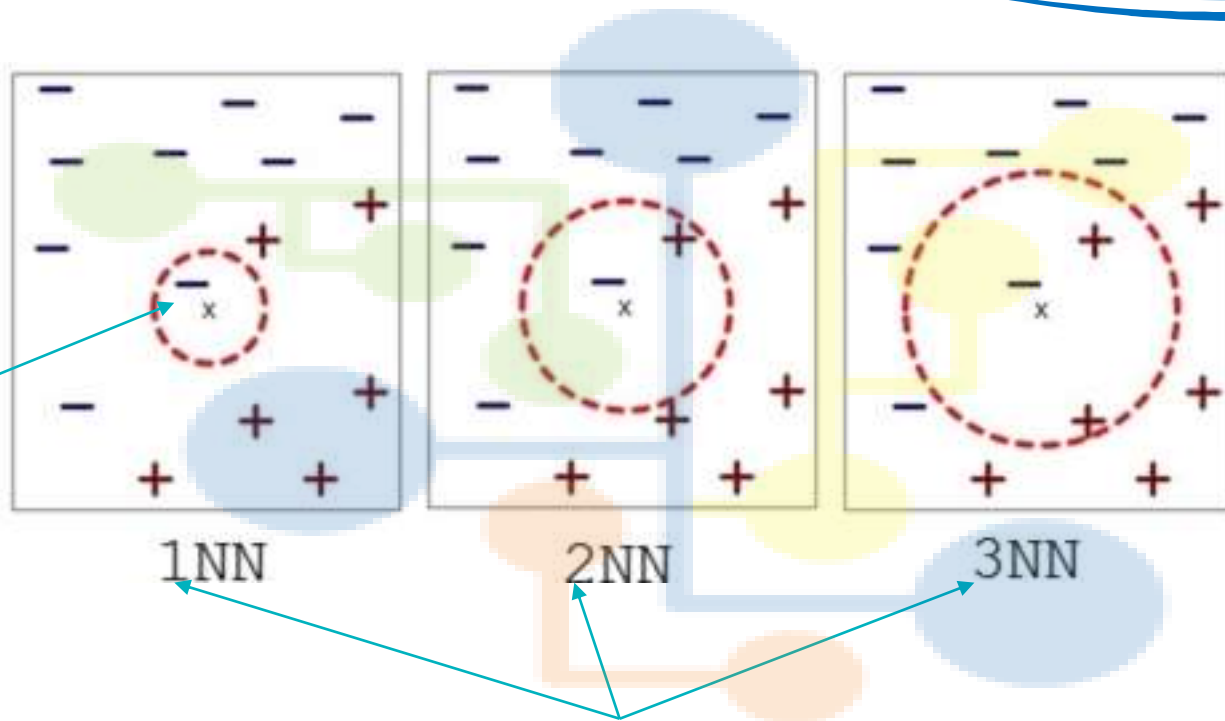


Como funciona o KNN



Como classificar um novo ponto de dado X

- 1- A distância é computada entre X e X_i para cada valor de X_i .
- 2- É escolhido o k-vizinho mais próximo X_{in} e sua respectiva classe.
- 3- Retorna-se o valor de y mais frequente na lista $y_{i1}, y_{i2}, \dots, y_{in}$.



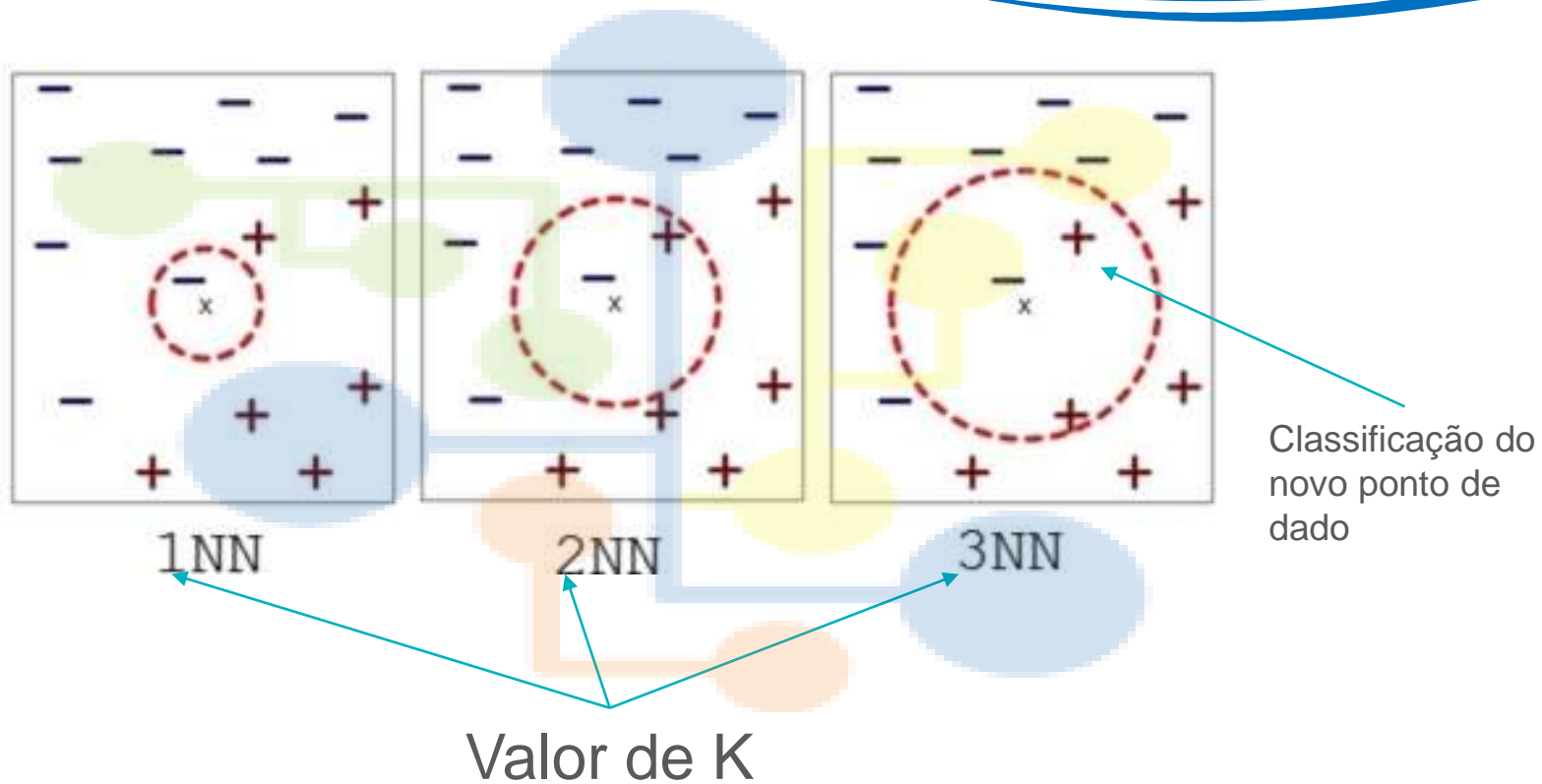
Classificação do
novo ponto de
dado

1NN

2NN

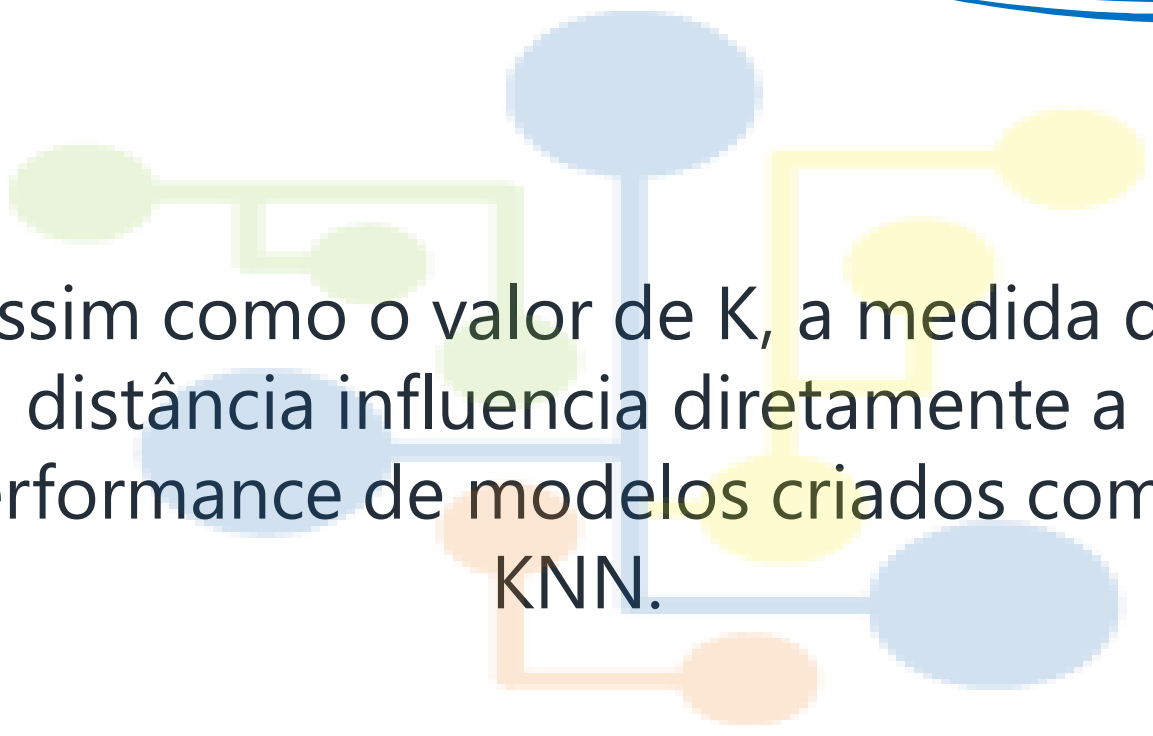
3NN

Valor de K





Existem diversas medidas de distância disponíveis e vamos discutir aqui algumas das mais comuns. O principal propósito da medida de distância é identificar os dados que são similares e que não são similares.

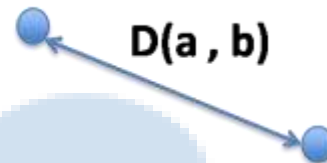
A decorative background diagram consisting of several colored circles (blue, green, yellow, orange) connected by lines, forming a network-like structure.

Assim como o valor de K , a medida de distância influencia diretamente a performance de modelos criados com o KNN.



Distância Euclidiana

$$D(a, b) = \sqrt{\sum_{i=1}^n (b_i - a_i)^2}$$





Distância de Hamming

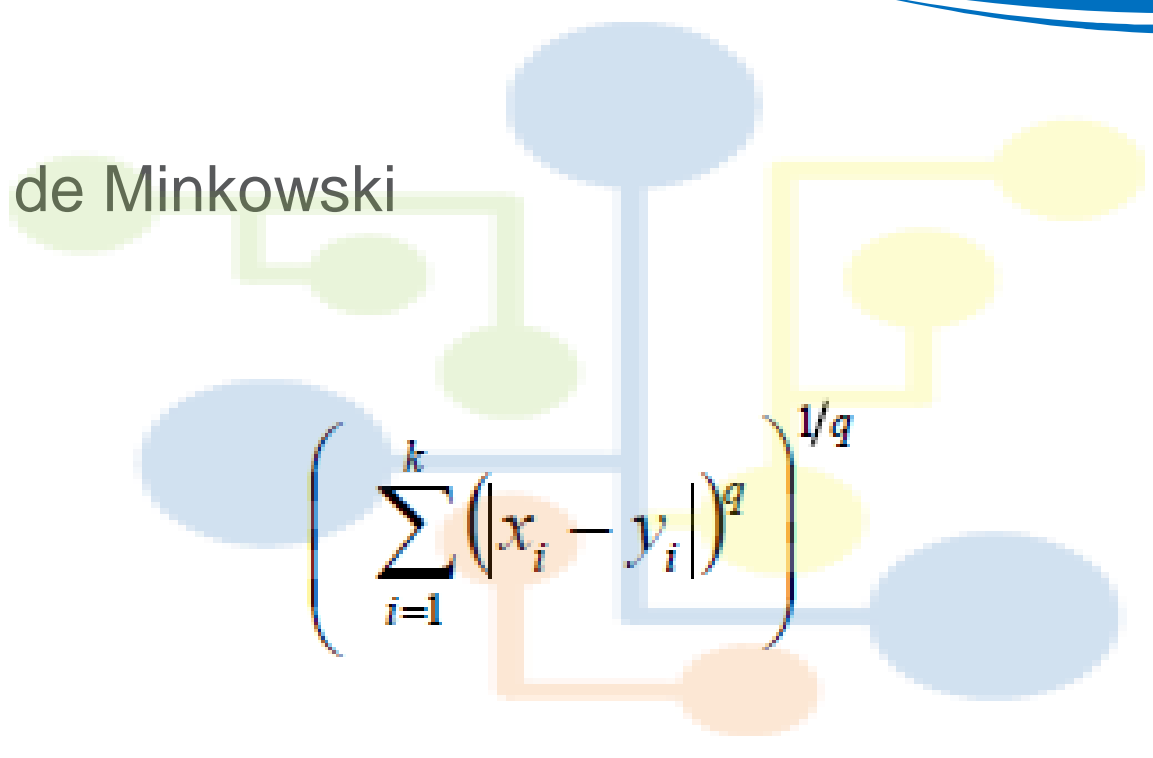
$$D_H = \sum_{i=1}^k |x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

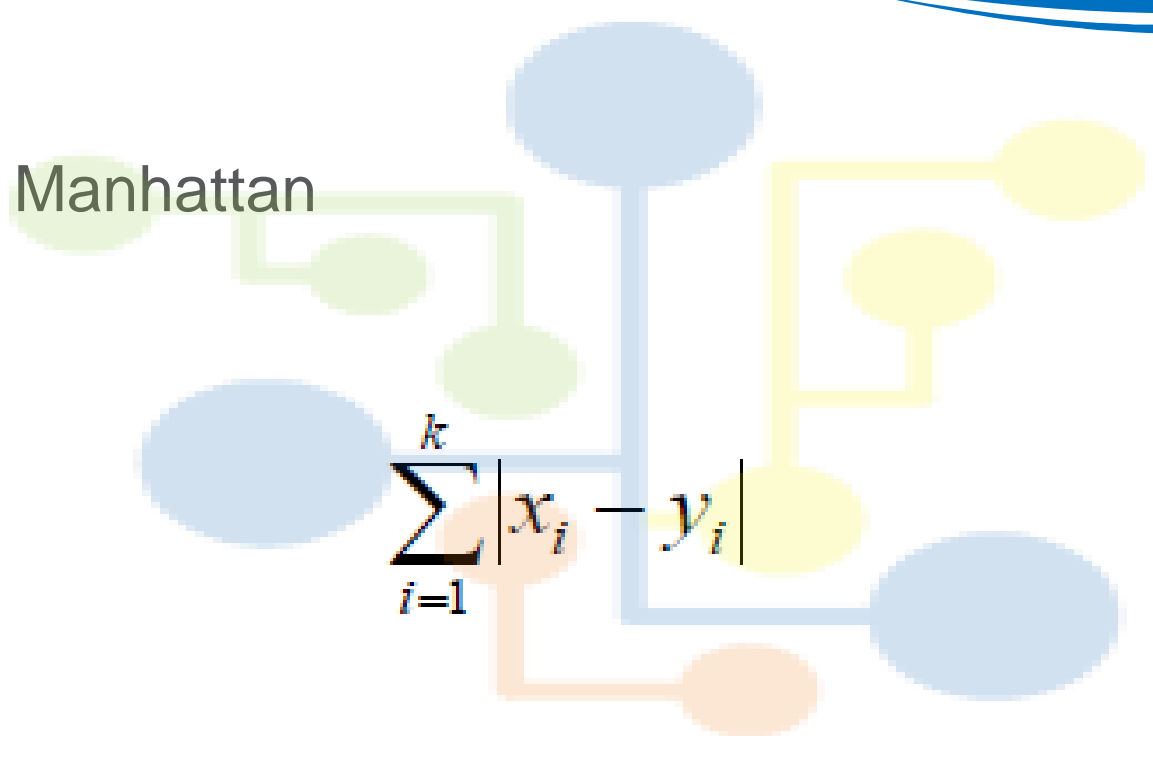


Distância de Minkowski

A decorative background diagram consisting of several colored circles (blue, green, yellow, orange) connected by lines of the same color, forming a network-like structure.
$$\left(\sum_{i=1}^k \left(|x_i - y_i| \right)^q \right)^{1/q}$$

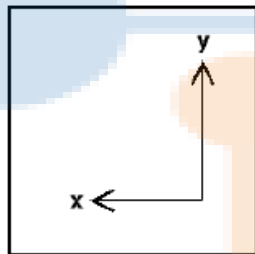


Distância Manhattan

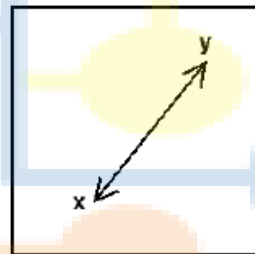
A diagram illustrating the Manhattan distance concept. It features several nodes of different colors (blue, green, yellow, orange) connected by lines. The lines are colored to match the nodes they connect, forming a network. The equation for Manhattan distance is overlaid on the diagram.
$$\sum_{i=1}^k |x_i - y_i|$$



As distâncias Manhattan e Euclidiana são as mais comuns.



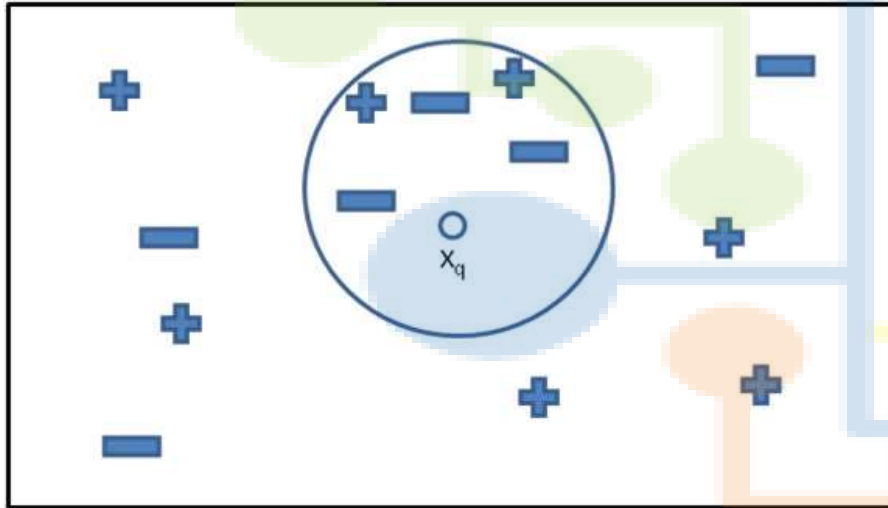
Manhattan



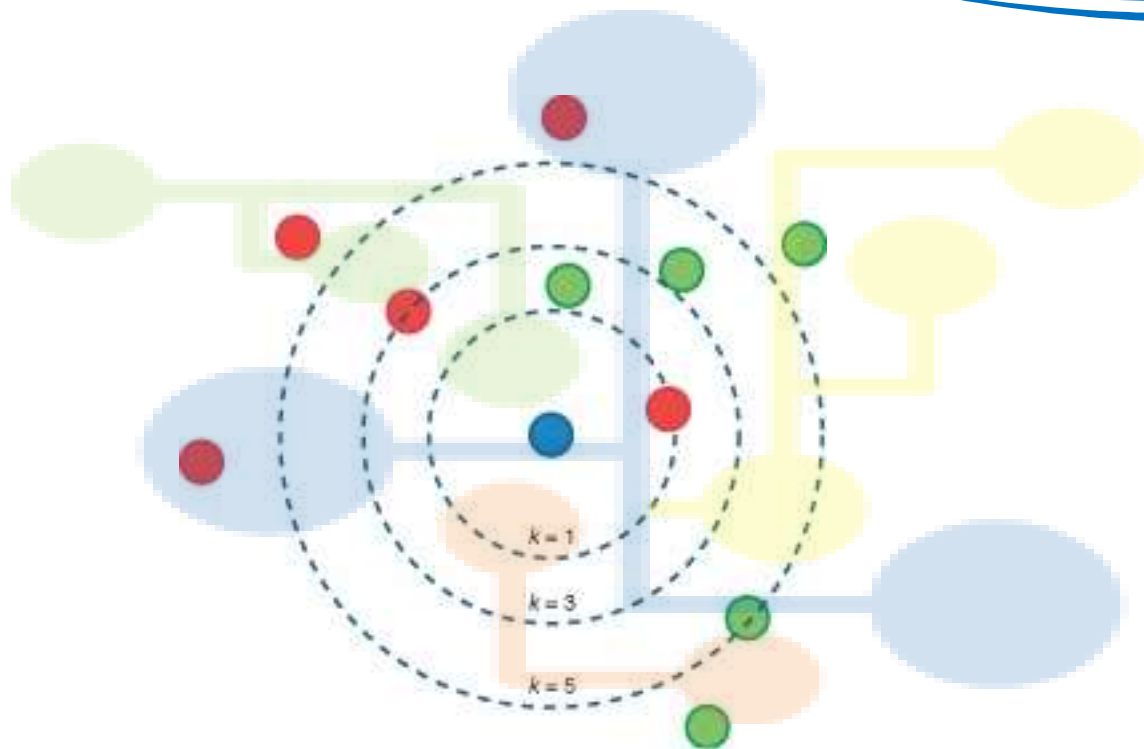
Euclidean



A precisão da classificação utilizando o algoritmo KNN depende fortemente do modelo de dados. Na maioria das vezes os atributos precisam ser normalizados para evitar que as medidas de distância sejam dominadas por um único atributo.



O algoritmo KNN é uma
variação do algoritmo NN.





Data Science
Academy

Data Science Academy gregorio@dexpertio.com.br 6004a7f1e32fc346864dec3f



Data Science Academy

KNN

Vantagens e Desvantagens





Vantagens

- Rápido treinamento
- Capaz de aprender funções complexas
- Não perde/desperdiça informação
- Bastante flexível
- Em alguns casos pode apresentar bons resultados





Desvantagens

- Classificar um exemplo desconhecido pode ser um processo computacionalmente complexo, pois requer um cálculo de distância para cada exemplo de treinamento
- Lento para realizar uma consulta
- A precisão da classificação pode ser severamente degradada pela presença de ruído ou características irrelevantes
- Não constrói um modelo de classificação





Data Science
Academy

Data Science Academy gregorio@dexpertio.com.br 6004a7f1e32fc346864dec3f



Continue Trilhando uma Excelente Jornada de Aprendizagem!

Muito Obrigado!