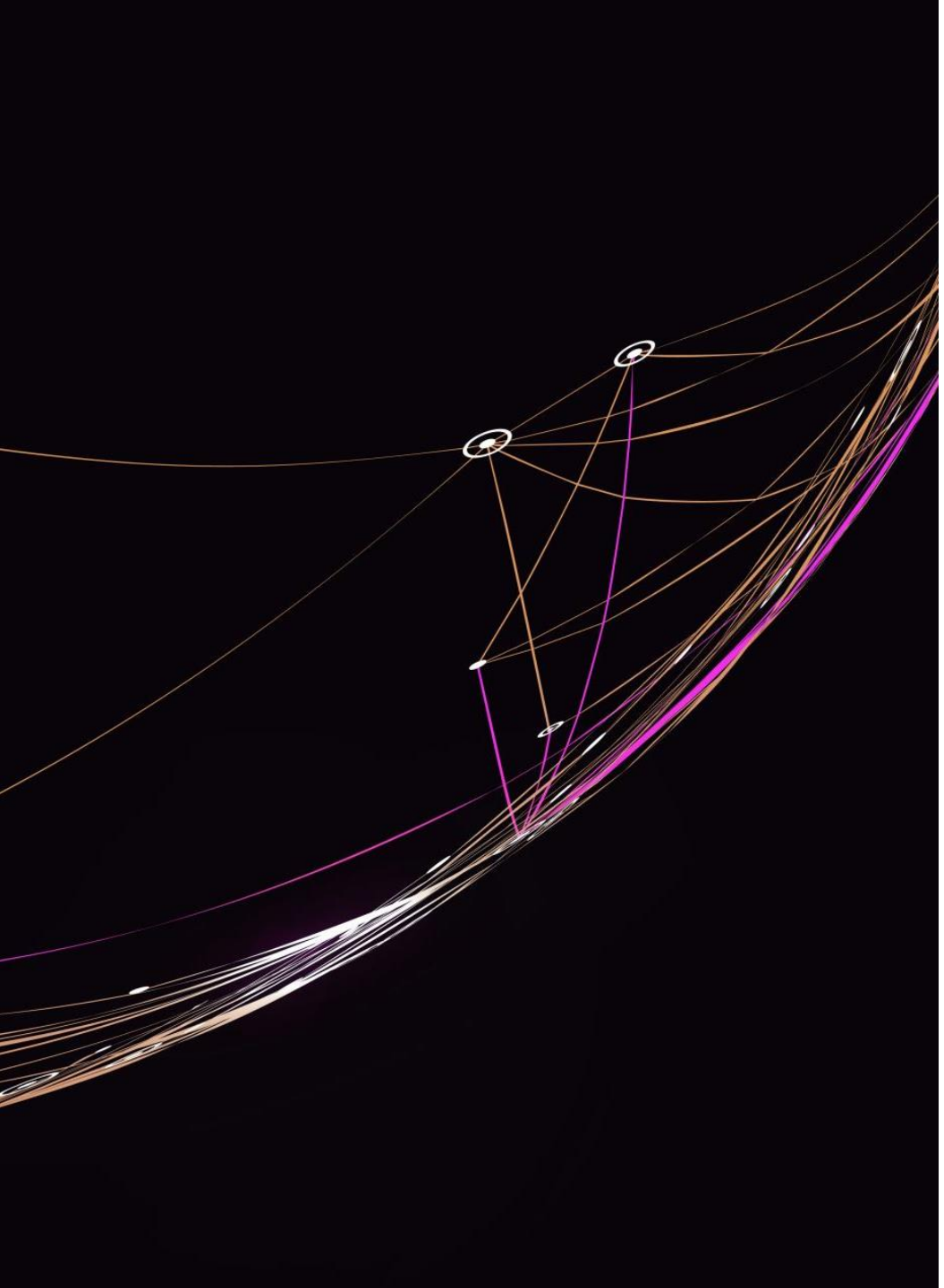


ADMINISTRAÇÃO

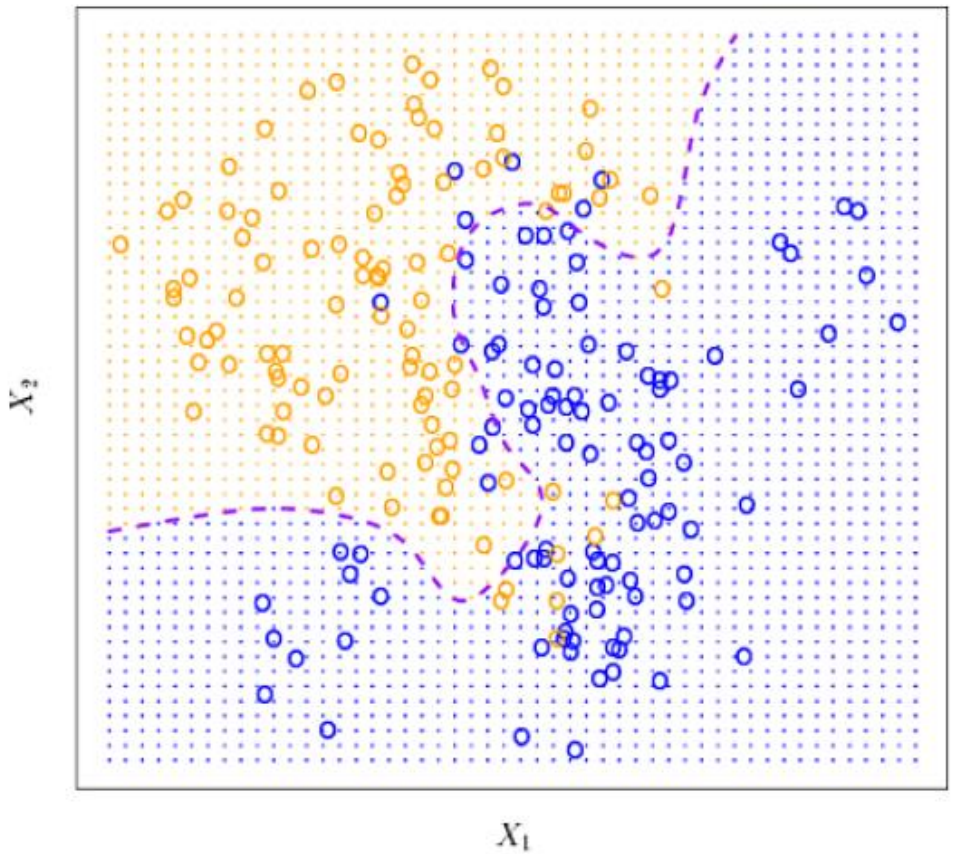
IBM0112 DATA MINING

K-Nearest Neighbors (KNN)

Cassius Figueiredo

Fronteira de decisão

A fronteira de decisão é formada dos valores de *x* onde existe indeterminação sobre classes, i.e. a probabilidade de selecionar a classe laranja é a mesma de selecionar a classe azul:



KNN – K-Vizinhos mais Próximos

- Depende de uma medida de distância entre pontos. A mais utilizada é a distância euclidiana.

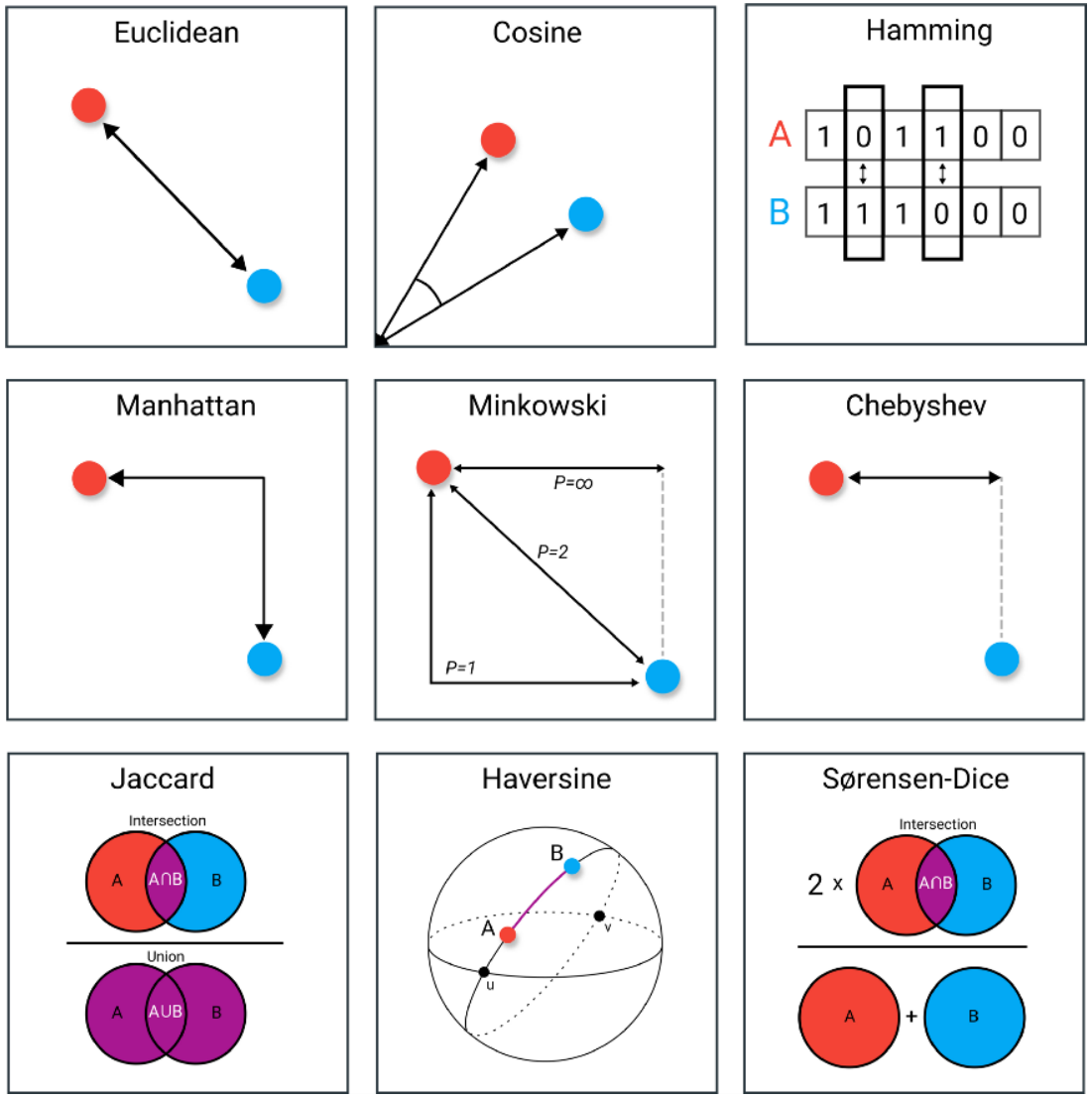
Distância entre duas instâncias $\mathbf{p}_i$ e $\mathbf{p}_j$ definida como:

$$d = \sqrt{\sum_{k=1}^{n} (p_{ik} - p_{jk})^2}$$

$\mathbf{p}_{ik}$ e $\mathbf{p}_{jk}$ para $k = 1, \ldots, \mathbf{n}$ são os $\mathbf{n}$ atributos que descrevem as instâncias $\mathbf{p}_i$ e $\mathbf{p}_j$, respectivamente

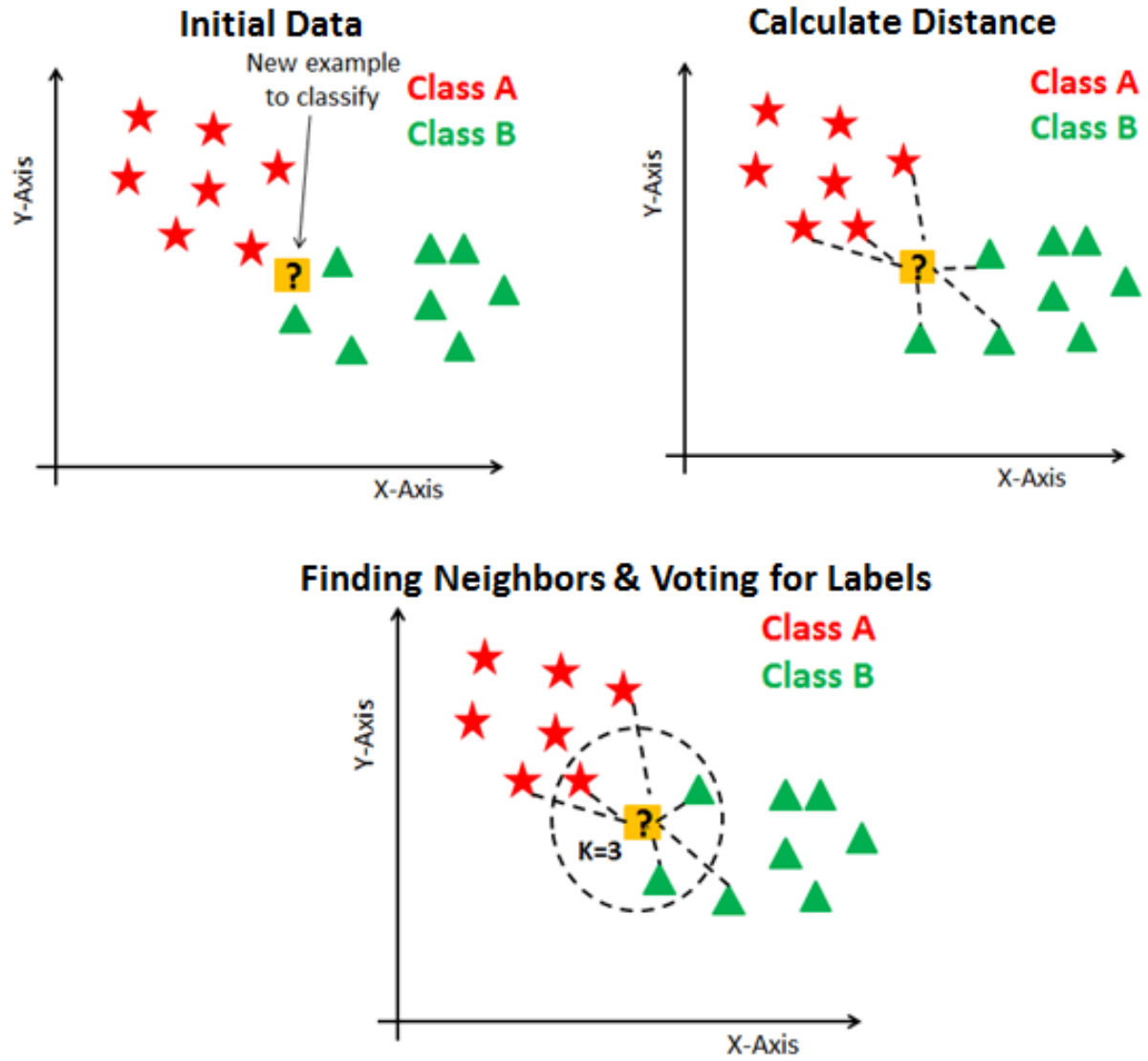
- O número de vizinhos K funciona com uma taxa de regularização. Valores pequenos podem levar ao overfitting em quanto valores grandes podem levar ao underfitting.

Distâncias em Ciência de Dados



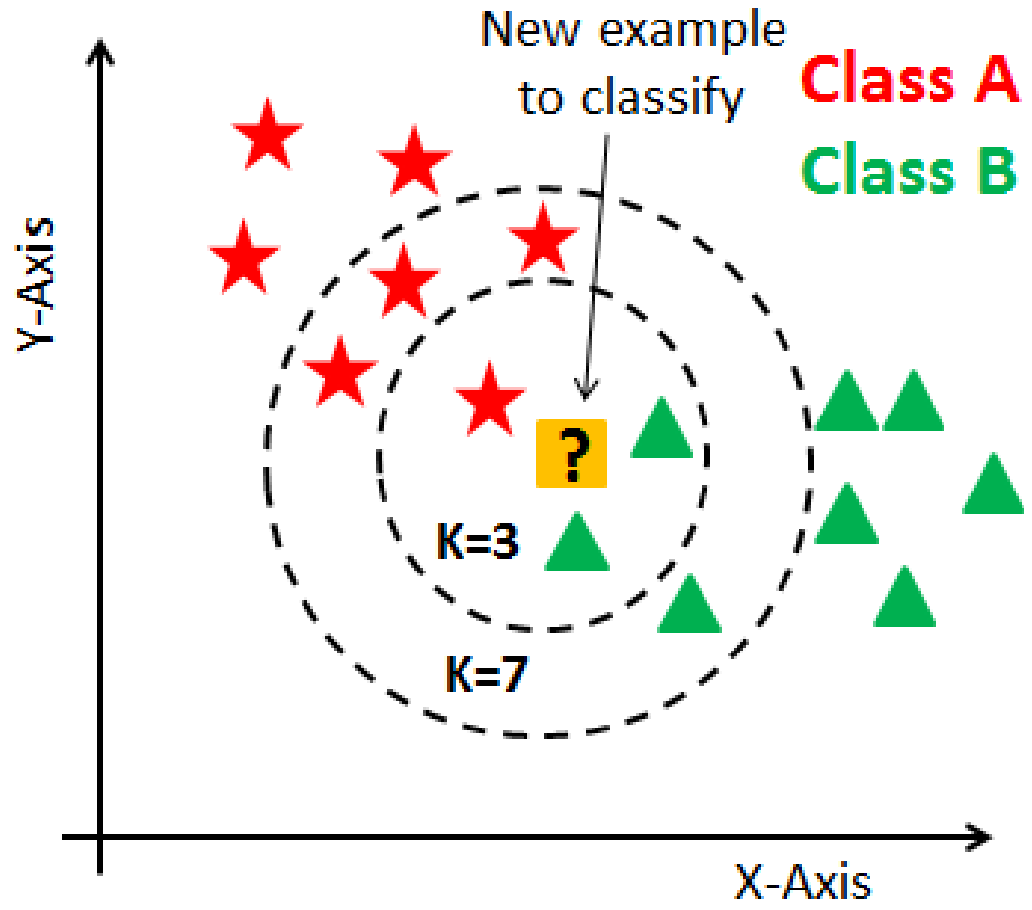
Fonte: 9 Distance Measures in Data Science

KNN – Como funciona?



Fonte: Datacamp

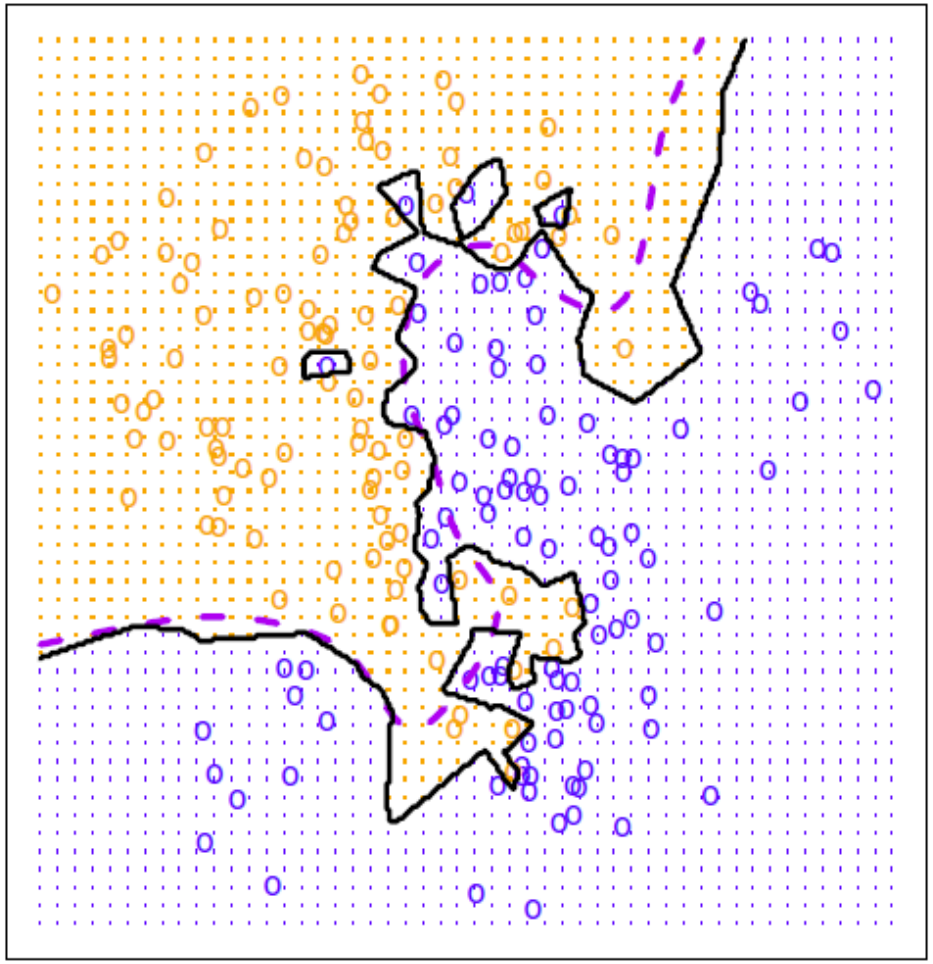
Escolha do parâmetro K



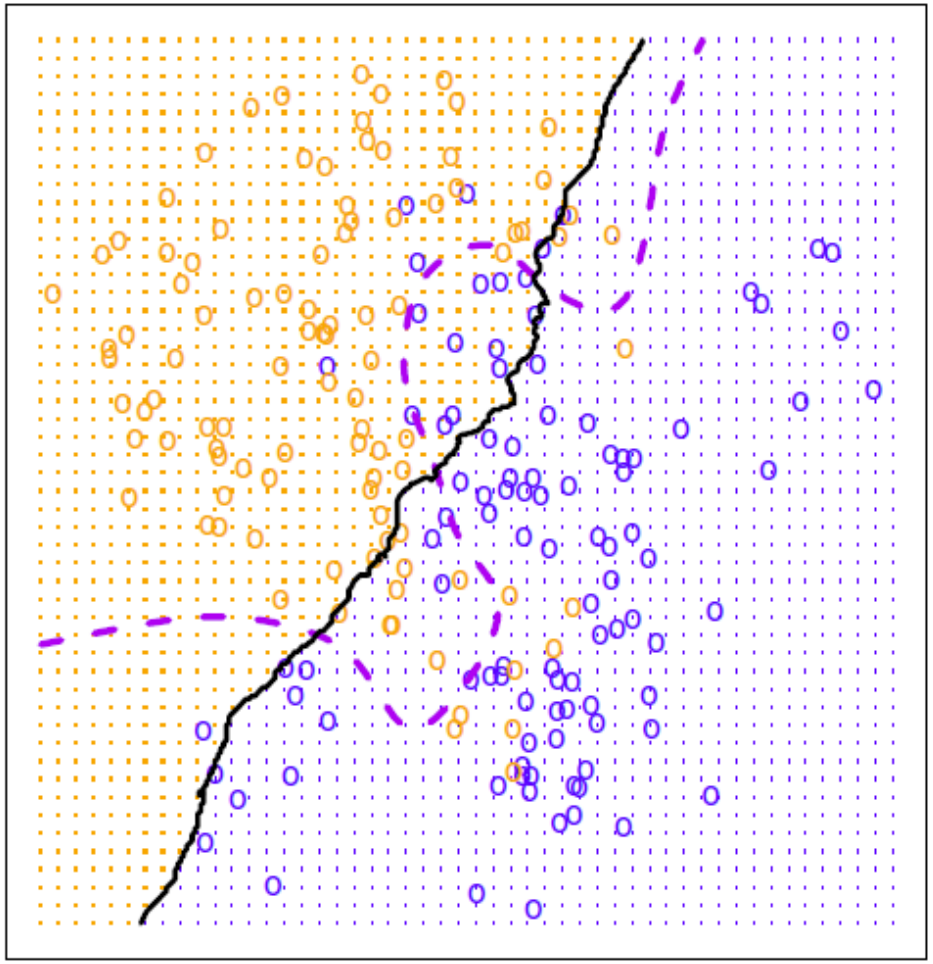
Fonte: Datacamp

Escolha do parâmetro K

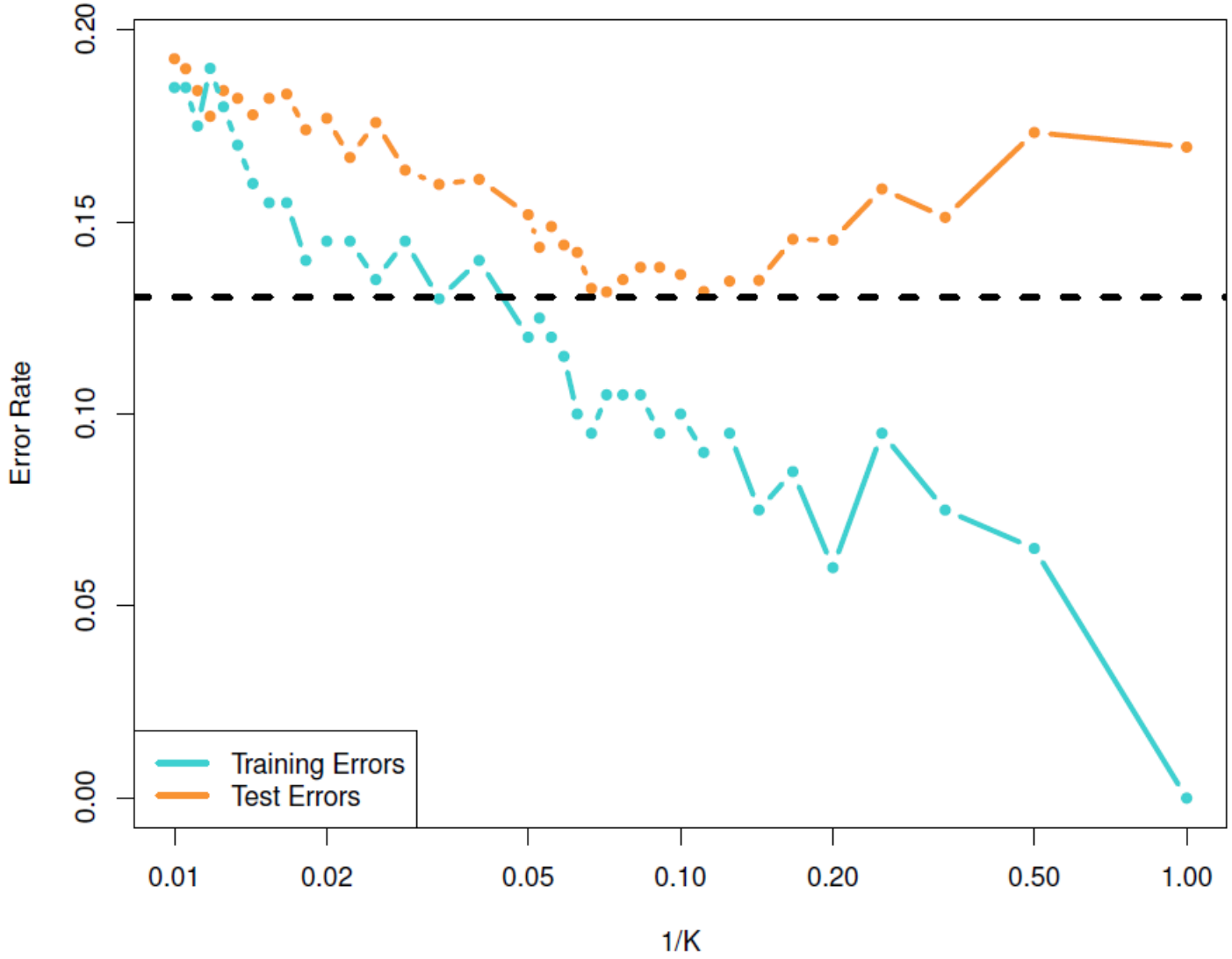
KNN: K=1



KNN: K=100



Escolha do parâmetro K



Escolha do parâmetro K

- A escolha tradicional é por K ímpar caso tenhamos um número de classes par. Isso nos dá uma configuração melhor para a escolha da classe dominante pelo modelo.
- Podemos também testar o modelos para vários valores de K e comparar a performance, baseada em uma métrica apropriada.

Vantagens

- Intuitivo e simples: fácil de entender e implementar.
- Sem premissas: é um algoritmo não-paramétrico.
- Resposta rápida: o algoritmo responde rápido às alterações trazidas por novos dados.
- Multi-classe: funciona muito bem em problemas multi-classe.
- Classificação e regressão: pode ser usado nos dois tipos de problemas.
- Hiperparâmetro único: pode-se demorar um pouco para chegar ao K ideal, porém é o único hiper parâmetro necessário, o que torna o investimentop de tempo viável.
- Diversos critérios de distância: diversas funções de distância podem ser utilizadas (Euclideana, Manhattan, etc.)

Desvantagens

- Lenta convergência.
- Maldição da dimensionalidade: funciona bem para poucas variáveis mas sofre para convergir com muitas.
- Scaling: para distâncias Euclideanas ou Manhattan, requer que os dados estejam na mesma escala.
- Ruim para dados desbalanceados.
- Bastante sensível a outliers pois usa apenas a distância escolhida como critério de decisão.
- Não trata dados faltantes de forma própria, necessitando de tratamento prévio.