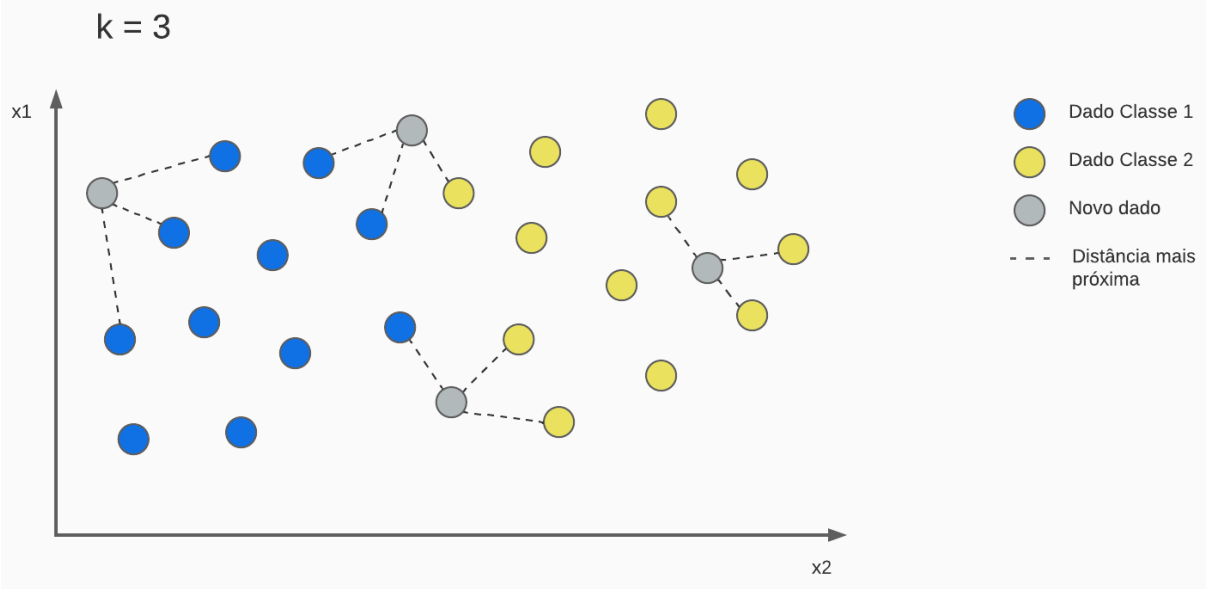
KNN

K-nearest neighbors (KNN) é um algoritmo de aprendizado supervisionado de classificação. Ele funciona comparando uma nova amostra de dados aos k exemplos mais próximos já classificados no conjunto de treinamento. A classe mais comum entre os k vizinhos mais próximos é então atribuída à nova amostra.

Para entender como o KNN funciona, considere um conjunto de dados com dois recursos (como x_1 e x_2 coordenadas em um plano cartesiano) e uma classe associada a cada ponto de dados. Quando uma nova amostra é dada, o algoritmo KNN calcula a distância entre essa nova amostra e cada ponto de dados no conjunto de treinamento. Ele então seleciona os k pontos de dados com as menores distâncias (os k vizinhos mais próximos) e atribui a classe mais comum entre esses k pontos para a nova amostra.



As vantagens e desvantagens de estabelecer um valor alto ou baixo para o parâmetro k no algoritmo KNN são:

Vantagens de estabelecer um k alto:

- Menor sensibilidade a ruído: Quanto maior o valor de k, menor é a influência dos pontos de dados isolados (ou ruído) no conjunto de treinamento. Isso pode resultar em uma previsão mais precisa.
- Menor overfitting: Quanto maior o valor de k, menor é a chance de overfitting, já que é preciso uma maior quantidade de vizinhos para uma amostra ser classificada.

Desvantagens de estabelecer um k alto:

- Menor flexibilidade: Quanto maior o valor de k, menor é a flexibilidade para adaptar-se a formas complexas de dados, já que precisa de mais amostras para se ajustar ao modelo.
- Menor capacidade de lidar com a classificação de pontos de dados raros: Quanto maior o valor de k, menor é a chance de um ponto de dado raro ser classificado corretamente, já que é preciso mais amostras para se ajustar ao modelo.

Vantagens de estabelecer um k baixo:

- Maior flexibilidade: Quanto menor o valor de k, maior é a flexibilidade para adaptar-se a formas complexas de dados, já que precisa de menos amostras para se ajustar ao modelo.
- Maior capacidade de lidar com a classificação de pontos de dados raros: Quanto menor o valor de k, maior é a chance de um ponto de dado raro ser classificado corretamente, já que é preciso menos amostras para se ajustar ao modelo.

Desvantagens de estabelecer um k baixo:

- Maior sensibilidade a ruído: Quanto menor o valor de k, maior é a influência dos pontos de dados isolados (ou ruído) no conjunto de treinamento. Isso pode resultar em uma previsão menos precisa.
- Maior overfitting: Quanto menor o valor de k, maior é a chance de overfitting, já que é preciso uma menor quantidade de vizinhos para uma amostra ser classificada.

Em geral, é uma boa ideia experimentar diferentes valores de k e ver como eles afetam a precisão do modelo em seus dados de teste. Um bom ponto de partida é um valor médio como 5 ou 10, mas pode ser necessário ajustar esse valor dependendo do seu conjunto de dados.

Por fim, é importante realizar a normalização dos dados antes de treinar uma KNN, pois ela ajuda a equilibrar as escalas das características, o que é crucial para o funcionamento correto do algoritmo. Isso ocorre porque o algoritmo KNN utiliza a distância euclidiana para medir a similaridade entre as amostras, e essa medida é sensível às escalas das características. Se as escalas das características forem muito diferentes entre si, a medida de distância pode ser dominada por uma ou poucas características, o que pode prejudicar a performance do algoritmo.

É importante notar que é necessário normalizar os dados tanto para o conjunto de treinamento quanto para o conjunto de teste, para evitar vieses na avaliação do modelo.