

FIAP

Aprendizado Supervisionado - Classificação

Objetivos da aula:

- Apresentar e utilizar o classificador k-nearest neighbours (kNN)
- Apresentar a técnica de separação de dados (treino e teste)
- Avaliar Aprendizagem do modelo

- Vamos dar continuidade ao nosso estudo de aprendizagem de máquina, já vimos:
 - Tudo começa, conhecendo os dados disponíveis.
 - Como carregar um data frame
 - Como visualizar os dados em gráficos (histograma, box plot, violin plot)
 - Fizemos uma breve introdução sobre análise exploratória buscando correlacionar os dados para gerar informações.
 - Hoje, vamos seguir nossa jornada e finalizar nosso estudo aplicando a técnica de KNN.

k-Nearest Neighbors

- O KNN(K vizinhos mais próximos) é considerado um dos algoritmos mais simples dentro da categoria de aprendizagem supervisionada sendo muito utilizado para problemas de classificação, porém também pode ser utilizado em problemas de regressão.

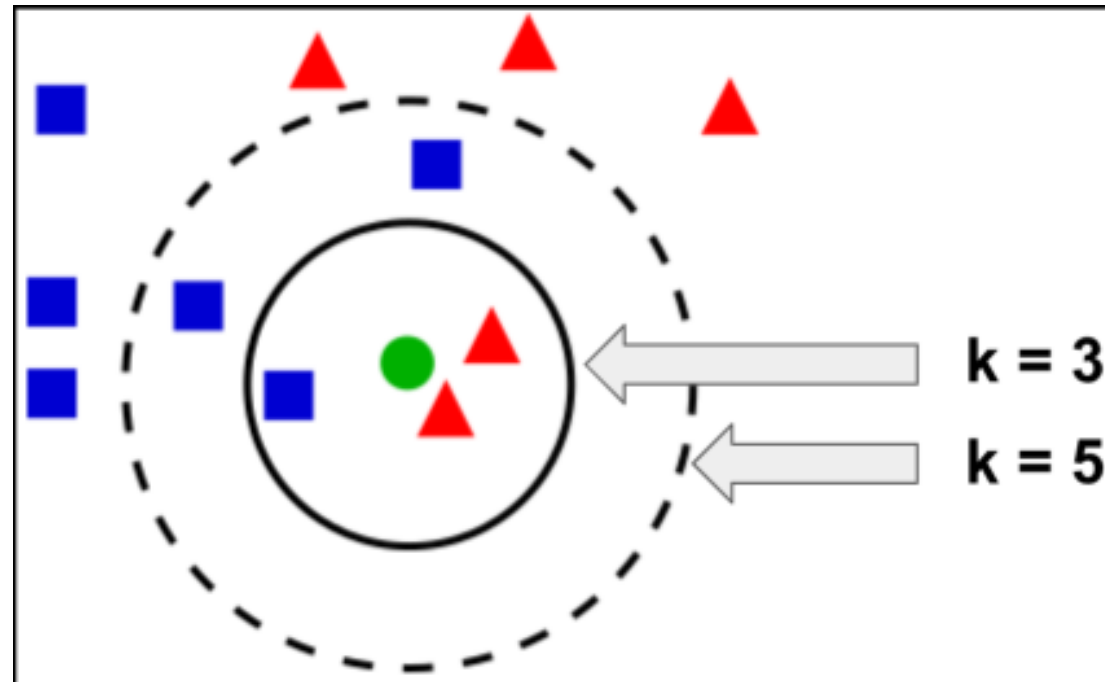
- **Problemas de classificação** = Vale lembrar que em problemas de classificação não estamos interessados em valores exatos, queremos apenas saber se um dado pertence ou não a uma dada classe.

Uma intuição sobre o método

- Para realizar a classificação o KNN calcula a distância objeto desconhecido (target) para todos os outros elementos, encontra os mais K vizinhos mais próximos faz uma contagem dos rótulos e considera que o objeto desconhecido pertence ao rótulo de maior contagem.

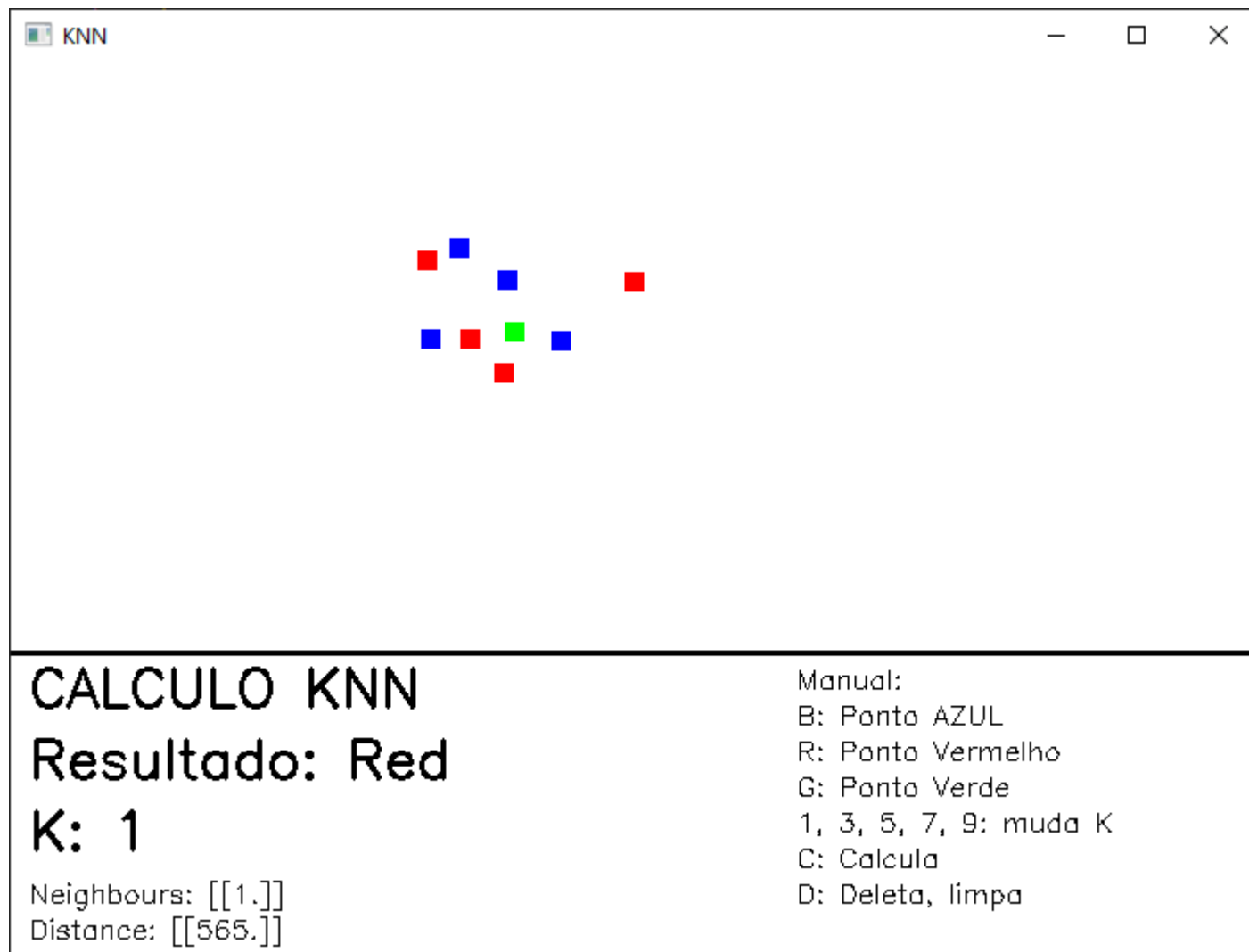
Exemplo

- A imagem abaixo exemplifica o funcionamento, mas se ficou um pouco complicado de entender, rode o script python ***iknn.py*** e faça algumas simulações para compreender.



Script

FIAP



Let`s go

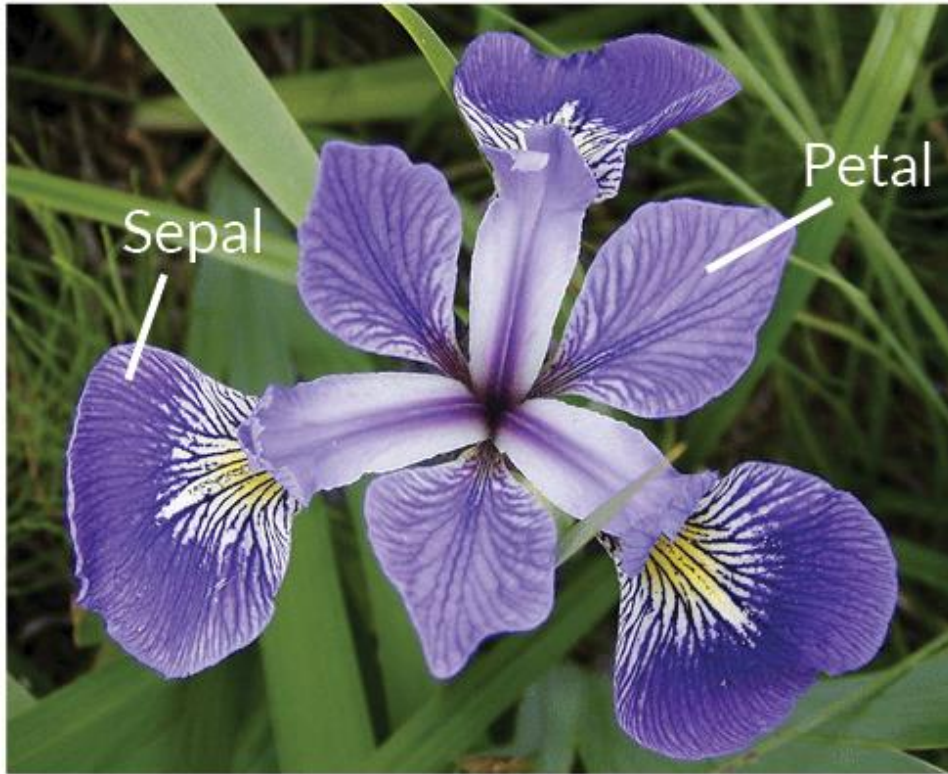
Vamos juntos realizar nosso primeiro projeto, do começo ao fim, de aprendizagem de máquina.

Definição do problema

A primeira coisa que precisamos fazer é a definição do problema. Neste primeiro caso vamos trabalhar com o mesmo dataset da última aula, dataset iris. Vamos desenvolver um sistema de machine learning capaz de classificar sua espécie com base nos dimensionais da pétala.

Observação

- São 150 exemplares de flor de íris, pertencentes a três espécies diferentes: setosa, versicolor e virginica, sendo 50 amostras de cada espécie. Os atributos de largura e comprimento de sépala e largura e comprimento de pétala de cada flor foram medidos manualmente.



Iris Versicolor



Iris Setosa



Iris Virginica

Dataframe

Out[10]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Atributos

Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
 - Iris Setosa
 - Iris Versicolour
 - Iris Virginica

classificador-knn.ipynb



classificador-knn.ipynb

Desafio 4

- Lembra o dataset 'breast-cancer', faça um modelo de predição que informa se o câncer é maligno ou não.



**Copyright © 2023 Prof. Arnaldo Jr/Yan
Coelho**

Todos direitos reservados. Reprodução ou divulgação total ou parcial deste documento é expressamente proibido sem o consentimento formal, por escrito, do Professor (autor).