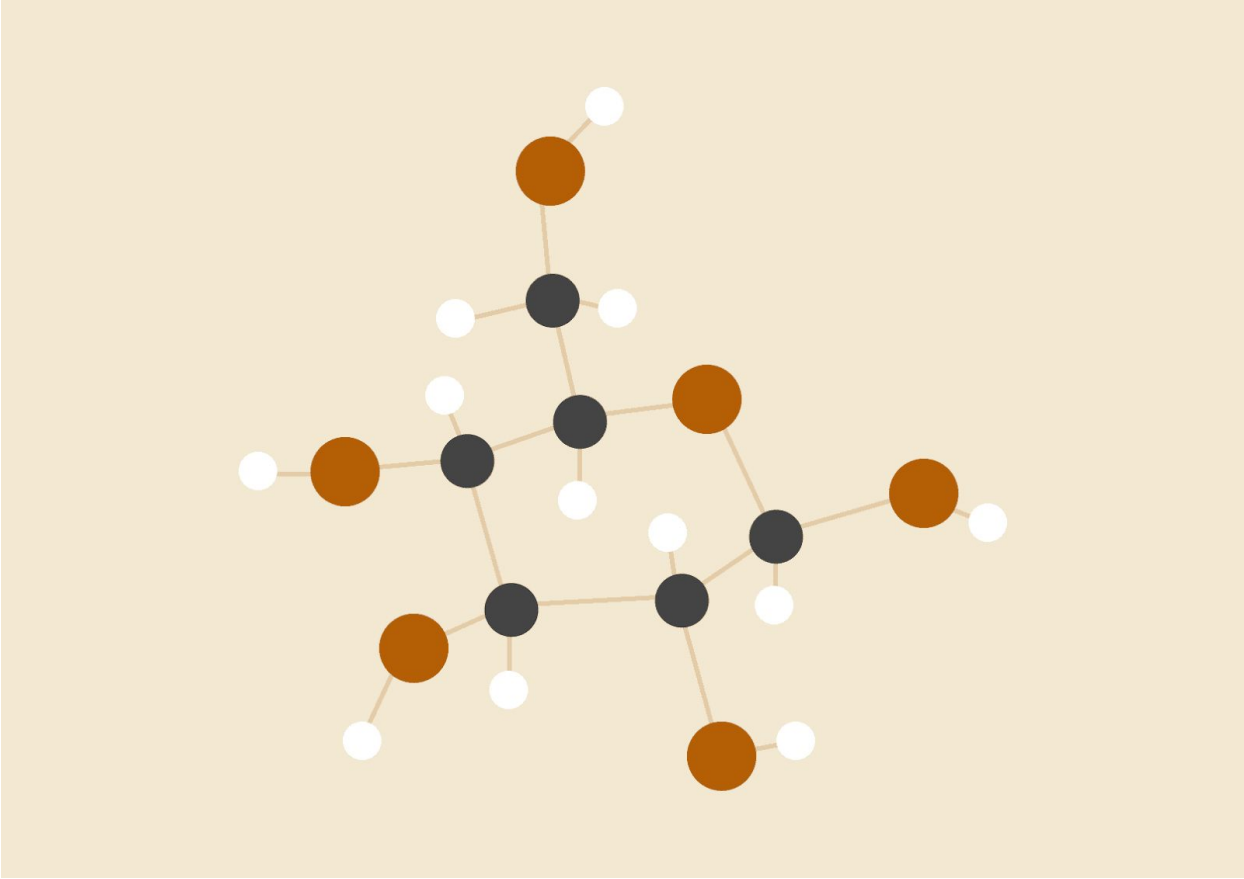
IMPLEMENTAÇÃO DE UM CLASSIFICADOR (KNN) E UM AGRUPADOR (K MEANS) PARA AS BASE DE DADOS: IRIS E BOSTON



João Kevin Gomes Rodrigues

Luiz Felipe Duarte Fiuza Pereira

16.05.2019

Introdução à inteligência Artificial

INTRODUÇÃO

Foram escolhidos duas bases de dados, a base IRIS e a base BOSTON. O objetivo será aplicar os algoritmos KNN (K-Nearest Neighbors) e o K-Means clustering e.observar os resultados.

SOBRE AS BASES DE DADOS

A base da IRIS consiste em uma tabela, contendo 150 linhas e 4 colunas. Cada linha representa uma flor de Iris, e cada coluna representa uma medida, da altura e largura, para a sépala e para a pétala.

Tendo estas medidas, é possível classificar a flor em 3 tipos: versicolor, setosa e virginica.

A base de dados Boston, é uma base para a classificação de imóveis.

Ela contém 506 imóveis, avaliados em 13 critérios: taxa de criminalidade per capita, proporção de terrenos residenciais destinados a lotes, proporção de hectares comerciais não varejistas, Variável dummy de Charles River (1 se o setor delimita rio; 0 caso contrário), concentração de óxidos nítricos, número médio de quartos por habitação, proporção de unidades ocupadas pelo proprietário construídas antes de 1940, distâncias ponderadas pelo 'DIS' para cinco centros de emprego em Boston, índice de acessibilidade a rodovias radiais, taxa de imposto sobre propriedades, proporção aluno-professor por cidade, proporção de negros, valor mediano de residências ocupadas pelo proprietário.

A classificação é feita em valores que variam entre 5 e 50. Por questões práticas, classificamos estes valores em: 0, para os valores menores que 18.8, 1 para valores entre 18.8 e 23.7, e 2 para valores maiores que 23.7. Assim, 0 seria uma área de baixo custo, 1 para custo mediano e 2 para alto custo.

ALGORITMOS UTILIZADOS

K-Nearest Neighbors (KNN): É um método de aprendizagem supervisionada que consiste em comparar o quão próximo o dado é de outros k dados classificados e assim, classificar este dado.

K-Means: É um algoritmo de classificação não supervisionada que consiste em definir k centróides aleatórios , calcular a média das distâncias, recalcular a posição dos centróides até que estas não se alterem mais. Atingindo o ponto de convergência.

TAXA DE ACERTO (Média em 100 vezes)

Base de dados	KNN	K-Means
Iris	83.88%	78.66%
Boston	82.66%	-

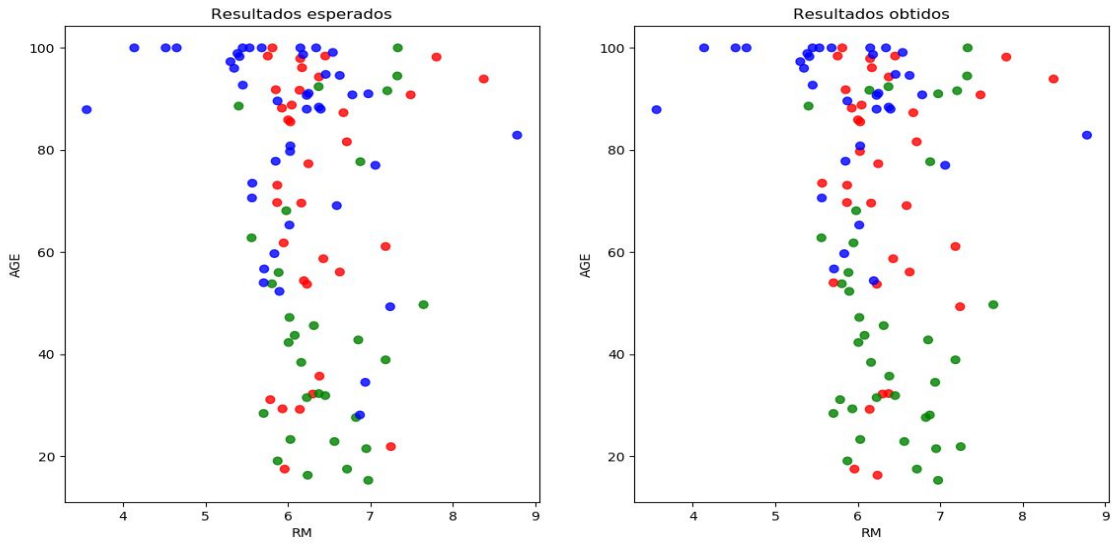
Resultados

Aplicando o algoritmo KNN na base de dados Boston

Como a base de dados Boston contém 13 atributos, e poderia gerar muitos gráficos, por questões práticas, escolhemos apenas 2 atributos: 'AGE' e 'RM' , que são número médio de quartos por habitação e proporção de unidades ocupadas pelo proprietário construídas antes de 1940.

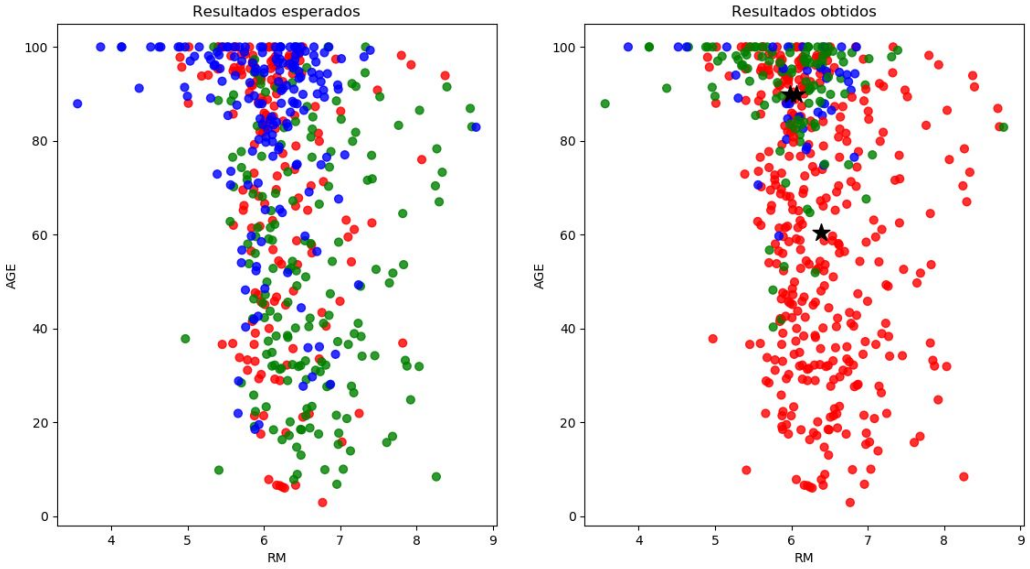
É observado uma concentração de pontos no eixo x, entre 5 e 8. Já no eixo Y, os valores variam entre 0 e um teto de 100, com grande variação nas classificações.

Pontos vermelhos representam custo baixo, pontos verdes, representam custo médio e pontos azuis custo alto.



Aplicando o algoritmo KMeans na base de dados Boston

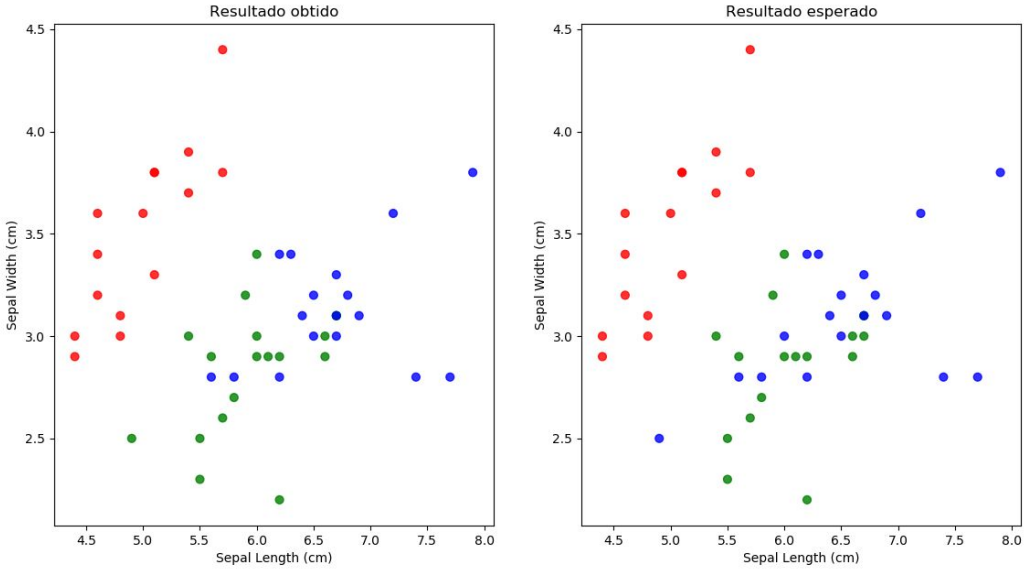
Utilizando as mesmas colunas de antes, temos estes gráficos. Neste, foram usados todos os indivíduos, criando assim uma densidade maior. Os centróides ficaram muito próximos, causando uma taxa de erro considerável.



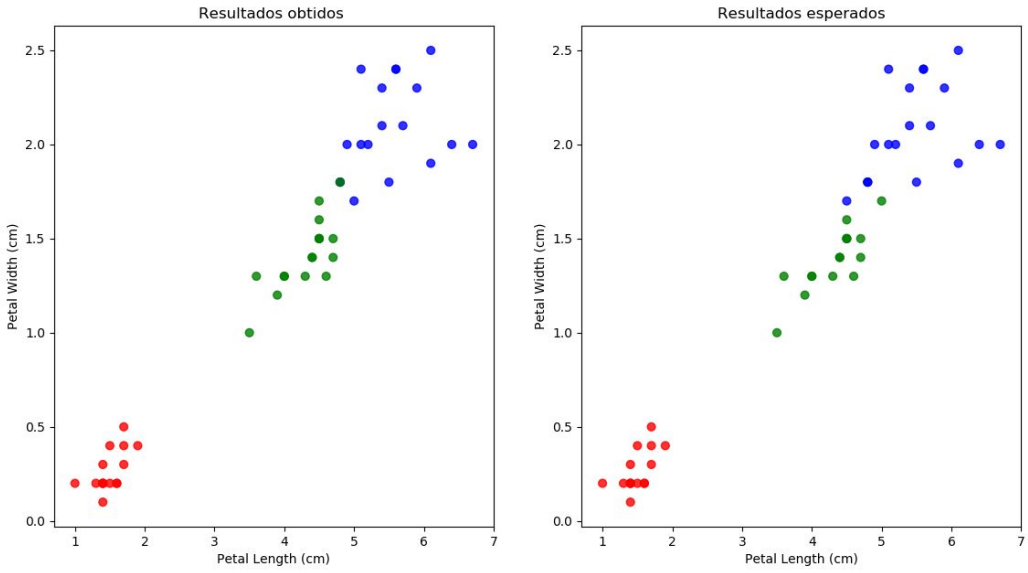
Aplicando o algoritmo KNN na base de dados IRIS

Foram gerados dois gráficos, que representam as medidas de largura e comprimento da sépala e da pétala. Abaixo temos a relação da altura e comprimento das sépalas das flores. Os pontos vermelhos representam as flores da espécie ‘setosa’, azuis para as flores da espécie ‘virginica’ e verde para ‘versicolor’.

Percebemos que existe uma separação até considerável dos dados. As medidas de largura são maiores que 3cm e o comprimento das sépalas de setosa são menores que 5.5cm. As sépalas de setosa são as com menores comprimentos e maiores alturas. Enquanto as sépalas de virginica são as de maior comprimento, e contêm largura mediana. As sépalas de versicolor são as menos largas, mas com comprimento mediano.



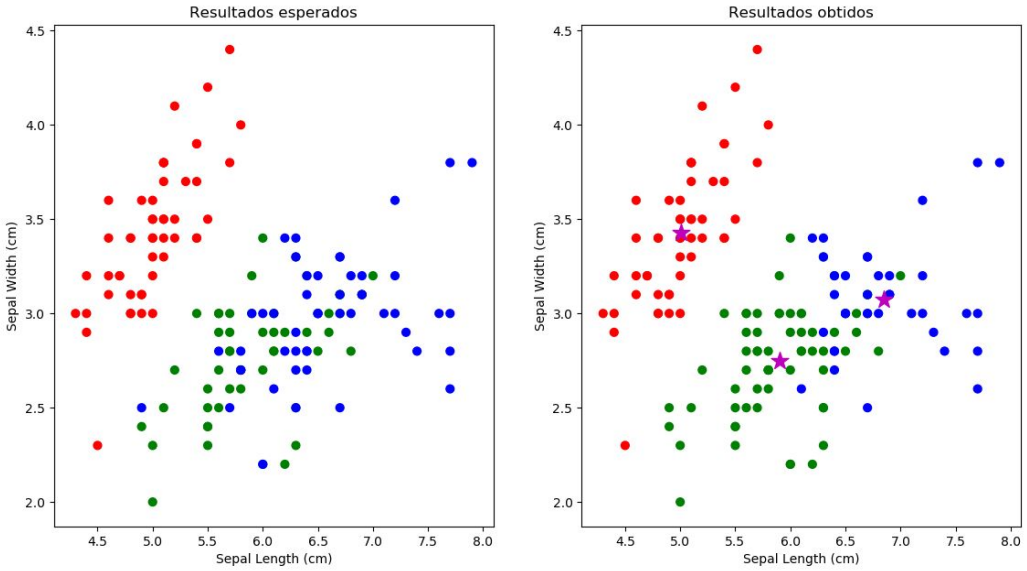
Analisando as medidas das pétalas das flores, podemos notar uma acentuada diferença entre as espécies e pontos mais agrupados, sendo as pétalas de setosa as menores, tanto em comprimento, quanto em largura. As pétalas de versicolor ocupam a posição mediana e as da espécie virginica são as maiores.



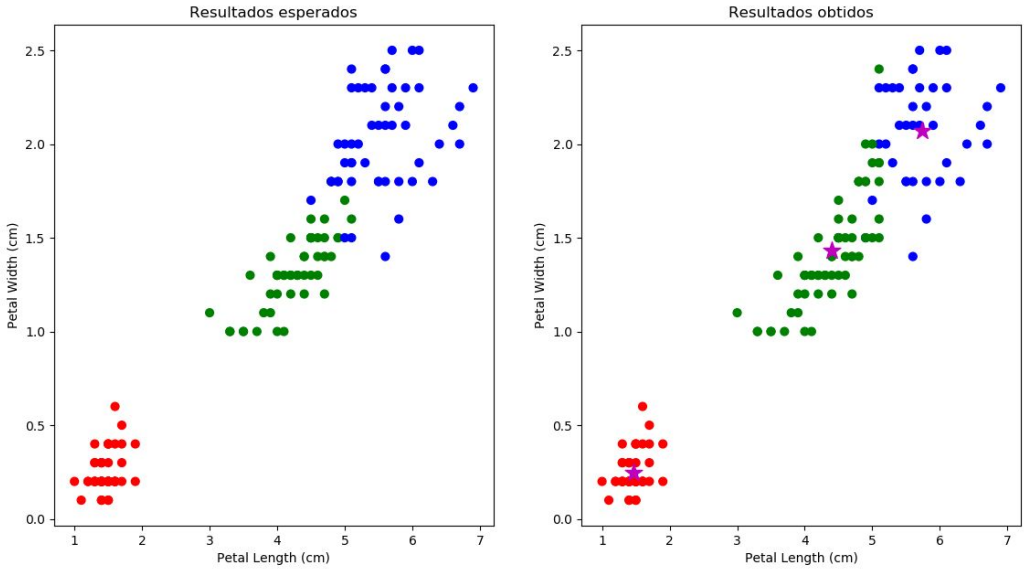
Aplicando o algoritmo K-Means a base de dados IRIS

Podemos ver marcações em roxo dos centróides. Nota-se que neste gráficos das sépalas, os resultados obtidos por meio do K-Means são mais imprecisos, já que se trata de um

agrupamento, e, no centróide mais próximo ao centro, as medidas das sépalas de virginica e versicolor se aproximam muito.



Nas medidas das pétalas, os resultados já se mostram bem mais definidos, e os centróides, visualmente em um ótimo posicionamento.



REFERÊNCIAS

http://people.csail.mit.edu/dsontag/courses/ml13/slides/lecture14.pdf

https://medium.com/@belen.sanchez27/predicting-iris-flower-species-with-k-means-clustering-in-python-f6e46806aaee

https://minerandodados.com.br/index.php/2018/02/02/algoritmo-k-means-python-passo-passo/

http://constantgeeks.com/playing-with-iris-data-kmeans-clustering-in-python/

https://scikit-learn.org/stable/modules/clustering.html#k-means

http://minerandodados.com.br/index.php/2017/12/12/entenda-o-algoritmo-k-means/