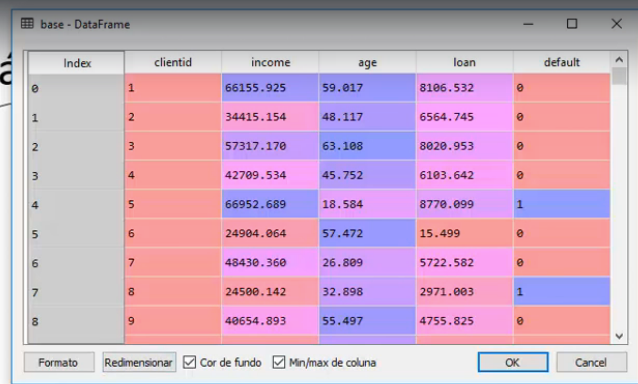
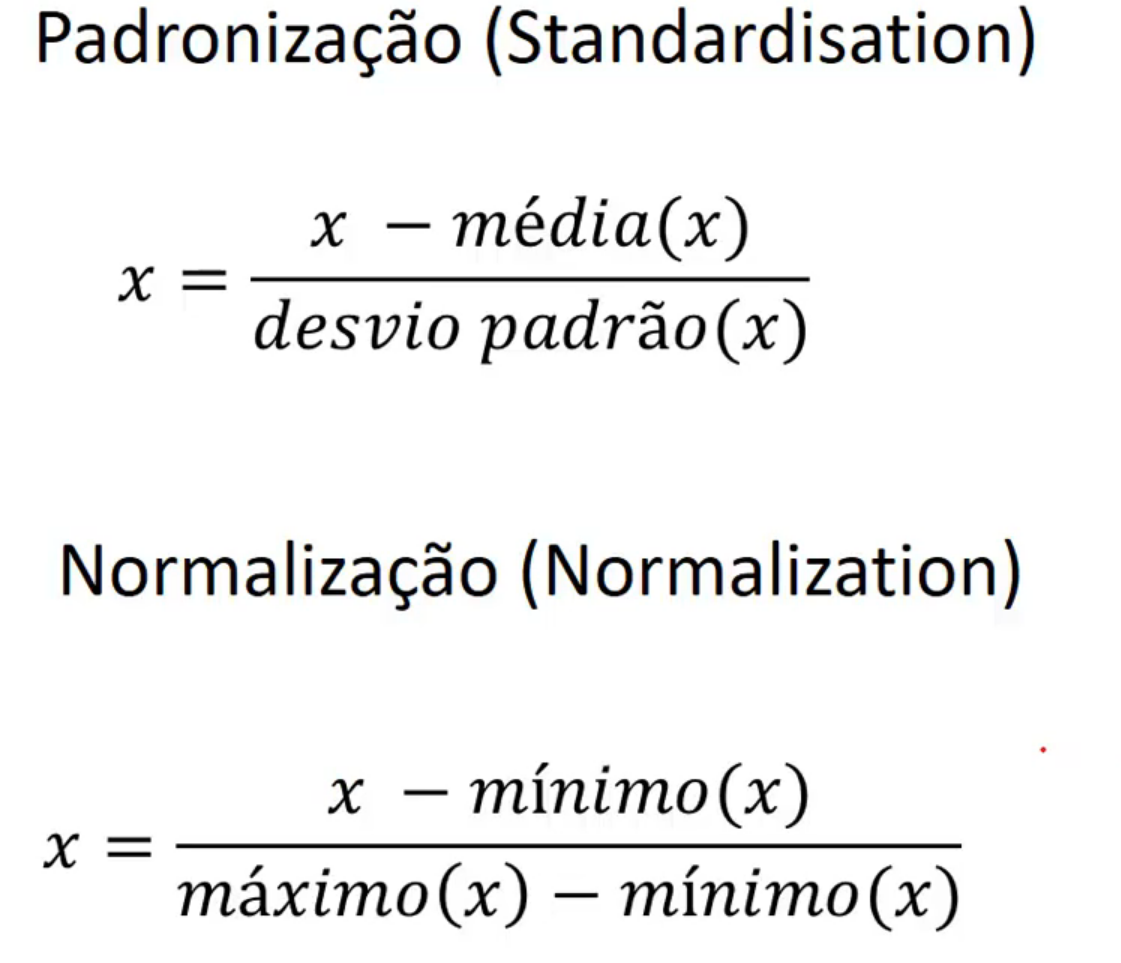
anotações - módulo 3



* Classificou clientid sendo variável do tipo nominal, porque é único.
* Income sendo contínua
* age sendo continua pq tem a idade quebrada, no caso 59.017 anos, por exemplo. Se fosse 59 apenas seria discreta
* loan sendo contínua
* defaut sendo discreta, pq são números inteiros.

escalonamento de atributos (deixa atributos na mesma escala)

* diferença de escala é a diferença entre valores, tipo coluna X possui valores 10 e 1000, já coluna Y possui valores 10 e 20. Logo valores da coluna Y possui uma escala menor.
* Quando se usa algoritmo de ML que faz uso da distância euclidiana (principalmente o KNN), é necessário fazer o escalonamento para não gerar resultados inconsistentes, pois considera mais importante os atributos que tivesse maior distância, mas na verdade o que tem são escalonamentos diferentes.

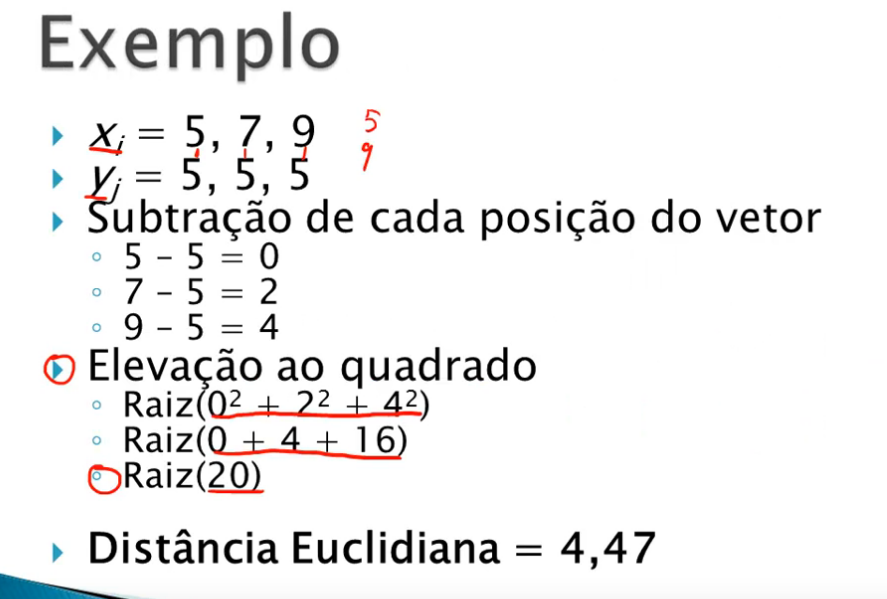
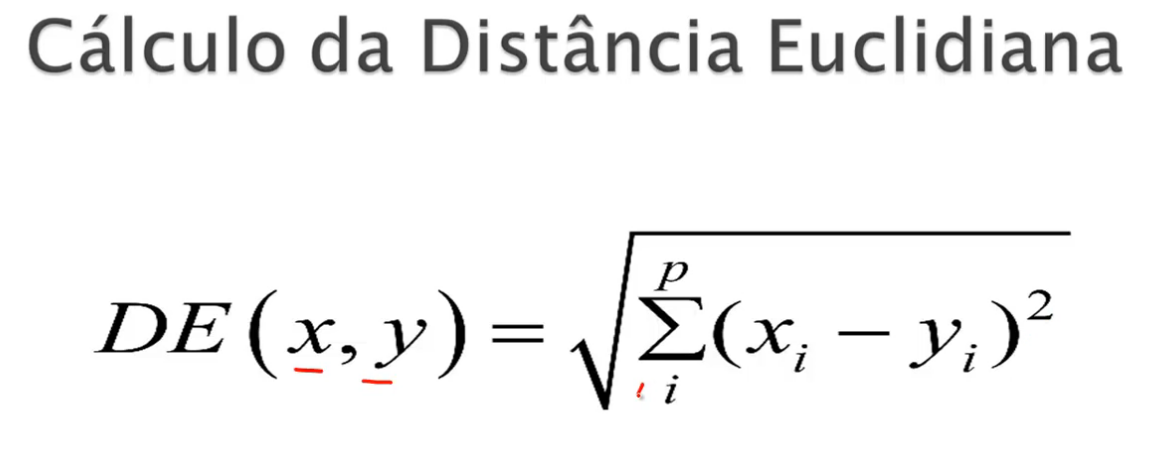


As duas fórmulas acima representa forma de fazer o escalonamento, deixando os atributos na mesma escala. Existe biblioteca no Python que faz. O mais indicado é o padronização, pois ele trata outliers.

Quanto maior valor da distância euclidiana (formúla para calcular e exemplo imagens abaixo), maior distância (semelhância) entre dois pontos.

Um exemplo do uso desse tipo de algoritmo é sistema de recomendações.

vídeo com explicação: <https://www.youtube.com/watch?v=KBh4vnsCDc4>



Para medir precisão (acurácia) dos resultados do preditor é necessário ter uma base de teste, logo para predição precisa de uma grande base de treinamento e uma para testes.

As bases de dados de treinamento e testes precisam ser diferentes, ou seja, ter registros diferentes.