INTELIGÊNCIA ARTIFICIAL

Projeto

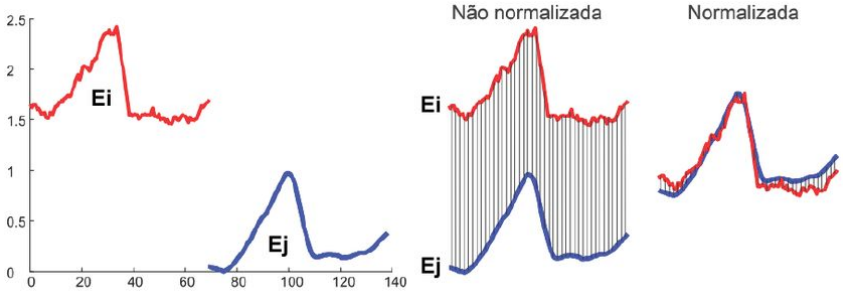
Marcos de Souza
mso2@cesar.school

C.E.S.A.R school

Resolvendo problemas com Machine Learning

1. Coleta de dados
2. Preparação dos dados
3. Escolha do modelo
   a. Classificação binária/multiclasse (predição)
   b. Regressão (previsão)
   c. Clustering
4. Treinamento
5. Avaliação
6. Aprimoramento dos parâmetros
7. Deploy

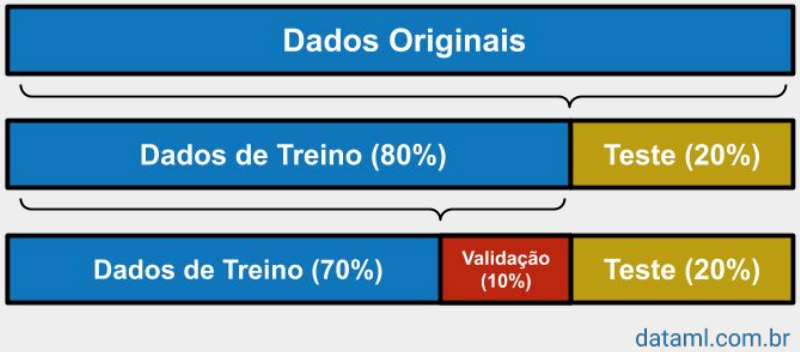
Preparação de Dados



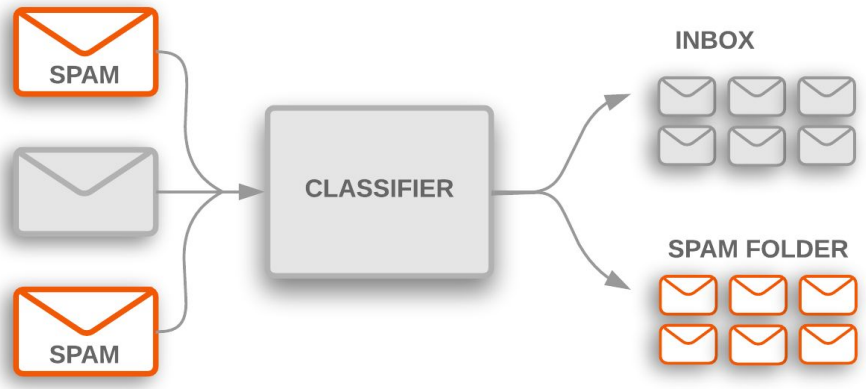
Limpeza, normalização e filtros a partir do que faz sentido no problema estudado

Etapa de Treinamento

Separação de dados



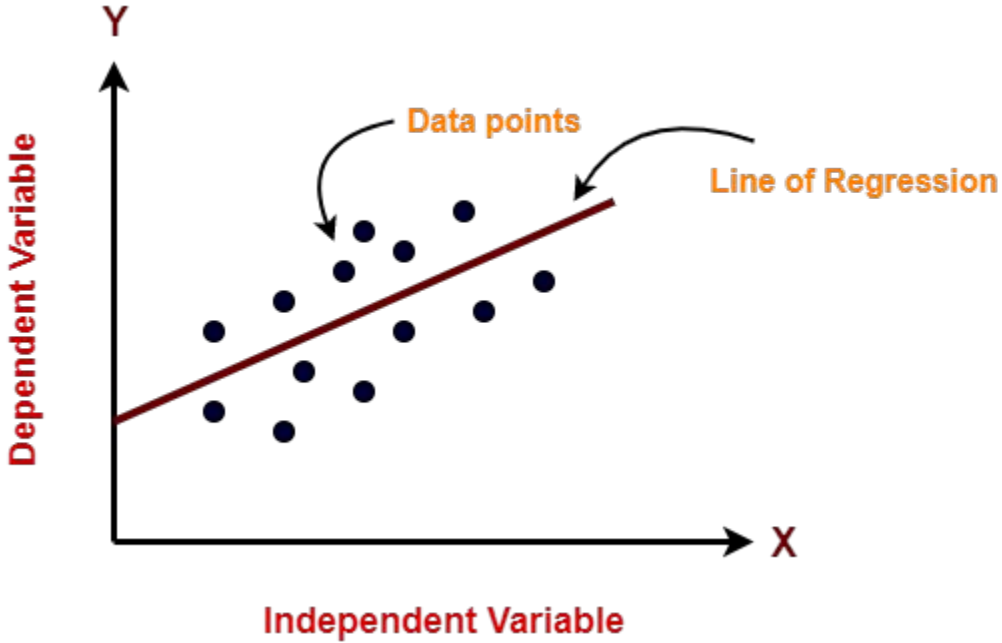
Problemas a serem estudados



Classificação

Classificação

1. Carregue a base de dados "***Non verbal tourists data***"
   a. A base está disponivel em https://archive.ics.uci.edu/ml/machine-learning-databases/00620/non-verbal%20tourist%20data.csv
   b. Considere como variável de saída a coluna ***"Tipo/classe do cliente"***
2. Valide e trate os dados disponíveis na base:
   a. Visualize a distribuição dos dados;
   b. Calcule as estatísticas necessárias da sua base;
   c. Trate os valores que não foram informados;
   d. Realize operações para facilitar o treinamento do seu classificador;
   e. Remova colunas que não agregam valor na sua base de dados
3. Divida a base em um conjunto de treinamento (75%) e um de teste (25%);
4. Determine uma ou mais métricas para avaliar a performance do seu classificador;
5. Treine um KNN para fazer a classificação da base de dados;
6. Determine o melhor valor para os vizinhos. Demonstre graficamente porque esse é o melhor valor;
7. Apresente o matriz de confusão da sua base de dados.



Regressão Linear

Regressão Linear

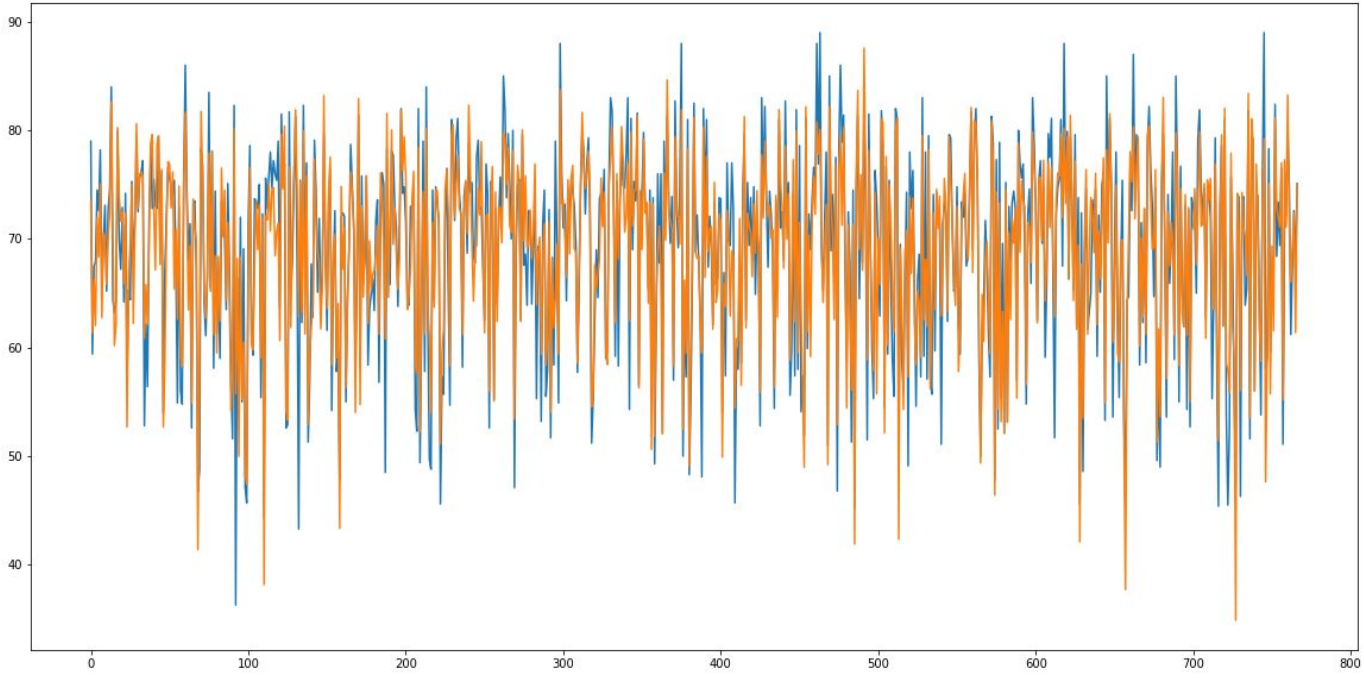
1. Carregue o dataset **Life Expectancy Data**
   a. para esse projeto, vamos considerar que a variável de saída é a "**Life expectancy**"
   b. Disponível em: https://drive.google.com/uc?export=download&id=1dHH13FgfqADSRgkNvkrf2Qn8o_Ph-_sn
2. Faça uma exploração dos dados do dataset, procurando verificar:
   a. quais são as features, observando os tipos delas e se precisam de algum pré-processamento;
   b. as informações estatísticas básicas das colunas do dataset;
   c. se há dados faltantes e decida o que fazer: preencher com algum valor default, descartar as linhas/colunas;
   d. a matriz de correlação das entradas com a saída.
   e. com base nas correlações das features com a saída, você acredita que esse dataset oferece condições de predizer a variável de saída?
3. Divida o dataset em conjunto de treinamento (70%) e de teste (30%);
4. Aplique a padronização, de forma separada, nos conjuntos de treinamento e de teste;
5. Use a regressão linear para predizer a expectativa de vida;
6. Crie gráficos de linha para visualizar a performance do modelo.
7. Avalie o modelo quantitativamente utilizando as métricas aprendidas e confirme sua impressão visual;

Regressão Linear

Resultados esperados

não precisa ser necessariamente igual, pois o resultado depende de algumas decisões tomadas durante o caminho que podem ser também consideradas corretas

Item 6: plot com todos os pontos

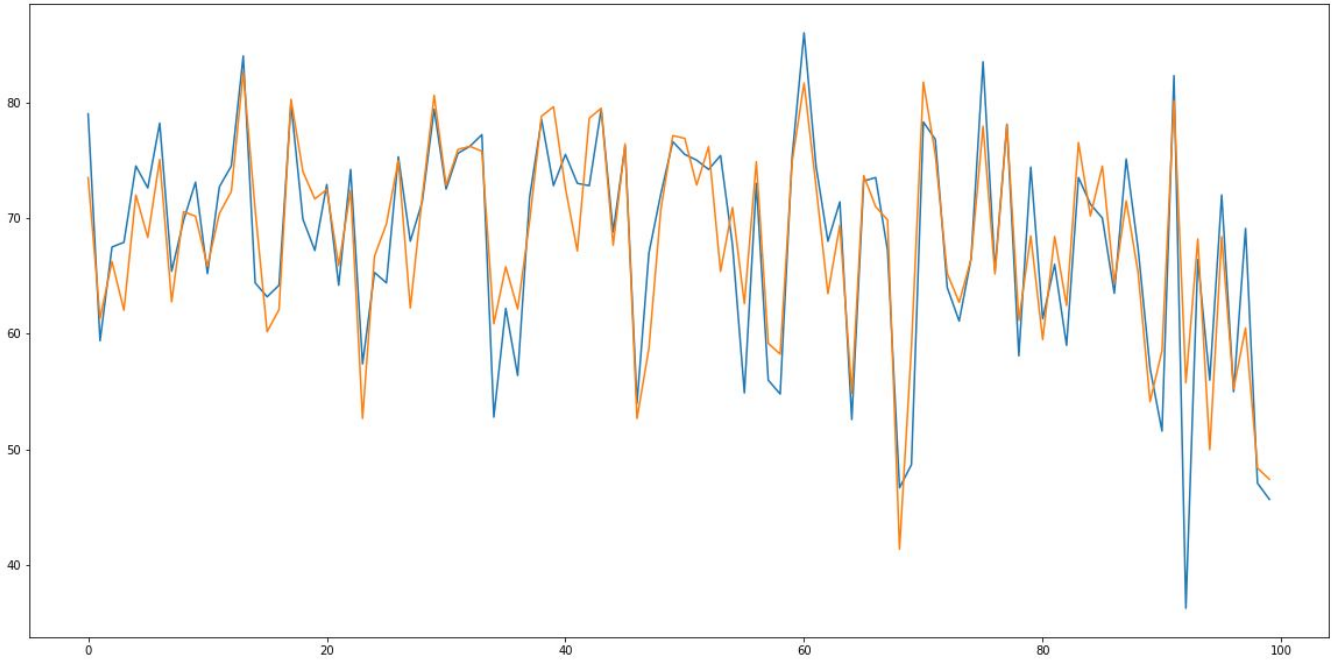


Regressão Linear

Resultados esperados

não precisa ser necessariamente igual, pois o resultado depende de algumas decisões tomadas durante o caminho que podem ser também consideradas corretas

Item 6: plot com apenas os 100 primeiros pontos

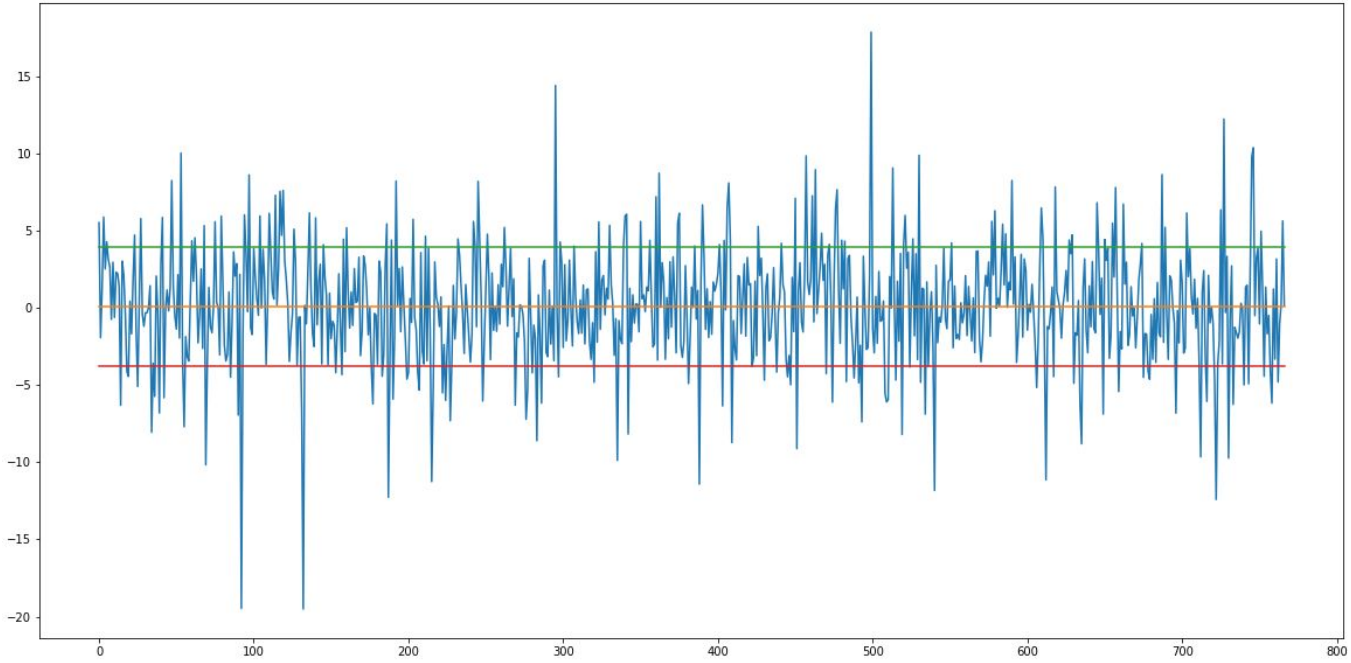


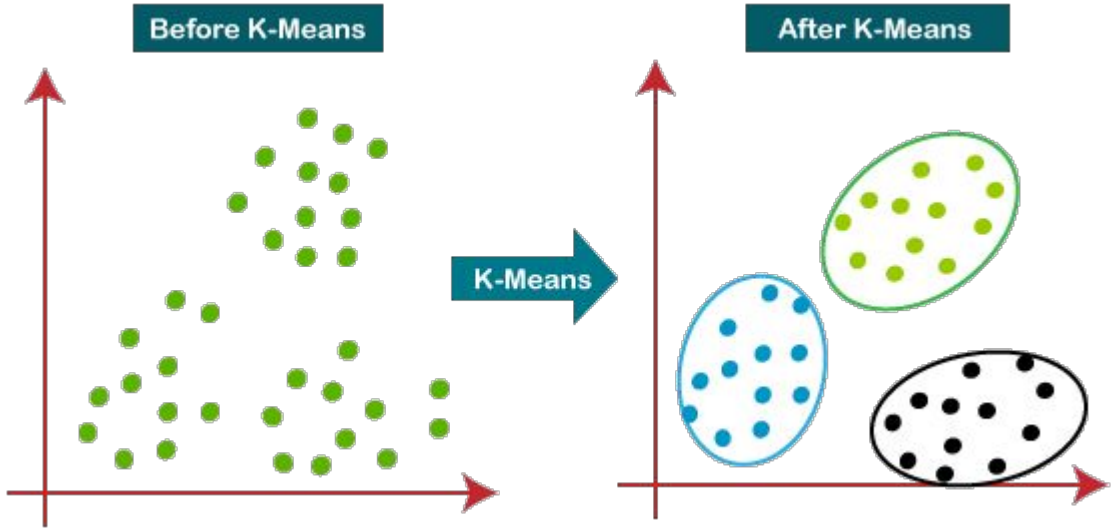
Regressão Linear

Resultados esperados

não precisa ser necessariamente igual, pois o resultado depende de algumas decisões tomadas durante o caminho que podem ser também consideradas corretas

Item 6: plot do erro de todos os pontos (**desconsidere** as linhas verde, vermelha e laranja horizontais na imagem)





Agrupamento / Clustering

Agrupamento / Clustering

1. Carregue o conjunto de dados Iris
   - Nesse projeto iremos considerar a variável de saída: `Iris-setosa`
   - Disponível em: https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
2. Aplique a padronização nos dados
3. Plote um gráfico de dispersão para identificar visualmente o número de grupos. Realize esse plot para cada par de atributos, de modo a obter a melhor visualização.
4. Selecione as duas melhores features de acordo com as visualizações do passo anterior
5. Aplique o método elbow:
   - Realize uma busca a partir de 2 grupos até 10
   - Utilize a inertia para avaliar a variação;
   - Plote os valores da inertia em cada K
6. Execute o k-means com o K escolhido de forma automática da etapa anterior
7. Imprima os índices da silhueta e NMI e plote o gráfico de dispersão com a saída do k-means
   - Para calcular o NMI será necessário um tratamento para converter a variável de saída do tipo textual em numérico

Links de Suporte

https://machinelearningmastery.com/machine-learning-in-python-step-by-step/

- Scikit
  - Classificação
  - Regressão Linear
  - Clusterização
    - KMeans
  - Avaliação de Modelos (Métricas)
    - Validação cruzada
- Tutorial
  - Pandas
  - Numpy
  - MatplotLib 1
  - MatplotLib 2
  - Seaborn

C . E . S . A . R

Pessoas impulsionando inovação.

Inovação impulsionando negócios.

NOSSO CONTATO

mso@cesar.org.br

mso2@cesar.school