UNIVERSIDADE ESTADUAL DE LONDRINA
Especialização em Machine Learning e Big Data
Fundamentos Matemáticos e Computacionais de Machine Learning

Profa. Dra. Juliana Felix
jufelix16@uel.br

Scikit-Learn e Datasets

Instruções: Você pode utilizar o ambiente de desenvolvimento de sua preferência. As respostas devem ser feitas em arquivo do tipo .ipynb ou .py.

Se optar por fazer todas as questões em um único arquivo, organize-o de forma adequada para permitir a identificação de cada questão.

Se optar por fazer as respostas para cada questão em arquivos separados, submeta um arquivo .zip contendo todos os arquivos de solução, identificados por questão.

Em relação às questões em que há perguntas para serem respondidas/comentadas, você pode comentá-las no próprio código ou enviar um documento pdf com seus comentários.

Data de entrega: Até 23h59 do dia 26/05/2023

1. Faça um código em Python que leia o arquivo price-house.txt e realize o split em conjuntos de treino/teste com as seguintes proporções:

a) 90/10;
b) 80/20;
c) 75/25;
d) 70/30;
e) 60/40;

Plote um gráfico de scatter plot colorindo o conjunto de treinamento e o de teste em cada divisão. Você observou algum padrão nas figuras geradas?

2. Escreva um script em Python usando o Scikit-Learn para treinar e estimar com a regressão linear sobre a base de dados Diabetes Dataset.

- A base está disponível no Scikit-Learn e pode ser carregada conforme instruções em:

Profa. Dra. Juliana Felix
jufelix16@uel.br

- https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_diabetes.html#sklearn.datasets.load_diabetes

A base contém 11 colunas, em que as 10 primeiras são features monitoradas, e a 11a coluna mostra o progresso da doença em um ano.

- Maiores detalhes sobre a base de dados podem ser obtidos em:
  - https://scikit-learn.org/stable/datasets/toy_dataset.html#diabetes-dataset

a. Os dados estão normalizados? Precisam ser normalizados? Precisam ser de-normalizados? Aplique as técnicas que achar necessário e explique o que foi feito com as suas próprias palavras.
b. Faça um gráfico de dispersão, em que no eixo horizontal constam os valores reais, e no eixo vertical os valores estimados.

3. Escreva um script em Python usando o Scikit-Learn para treinar e classificar com a regressão logística sobre a base de dados Wine Dataset.

- A base está disponível no Scikit-Learn e pode ser carregada conforme instruções em:
  - https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html#sklearn.datasets.load_wine

A base contém 13 colunas, em que as 12 primeiras são features preditivas, e a 13a coluna contém a definição da classe.

- Maiores detalhes sobre a base de dados podem ser obtidos em:
  - https://scikit-learn.org/stable/datasets/toy_dataset.html#wine-dataset

Os dados estão normalizados? Precisam ser normalizados? Precisam ser de-normalizados? Aplique as técnicas que achar necessário e explique o que foi feito com as suas próprias palavras.