SCC0275 - Introdução à Ciência de Dados
Terceiro Projeto Prático

Esse projeto tem como objetivo fazer um estudo de um dataset sobre bike sharing e tentar criar um modelo que ajude a prever o número total de bicicletas que serão alugadas a cada hora. O dataset se encontra no link:

- https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

Utilize a biblioteca scikit-learn para criar os modelos pedidos. Você pode usar outras bibliotecas para fazer as demais análises =)

Todas as respostas devem ser justificadas com base em:

1. **Código Python mostrando a(s) análise(s) e/ou o(s) modelo(s) feitos;**
2. **O resultado da(s) análise(s) e/ou do(s) modelo(s) e**
3. **Uma explicação textual (pode ser breve) da conclusão obtida.**

Em caso de plágio (mesmo que parcial) o trabalho de todos os alunos envolvidos receberá nota ZERO.

Desorganização excessiva do código resultará em redução da nota do projeto.
Exemplos:

- **Códigos que devem ser rodados de forma não sequencial;**
- **Projeto entregue em vários arquivos sem um README;**
- **...**

Bom projeto,
Tiago.

Questão 1 (valor 2.5 pontos)

a) Baixe os dados e carregue-os no Python.
b) Entende o que cada coluna significa. Faça um histograma mostrando a distribuição da variável resposta (total de aluguéis) e um gráfico mostrando a relação entre a temperatura do dia e a variável resposta.

Dica: use a base de dados sobre aluguéis a cada hora (arquivo `hour.csv`). Este arquivo deve ser usado em todas as questões do projeto!!!

Questão 2 (valor 2.5 pontos)

a) Diga quais variáveis (colunas) são explicativas (fazem parte do X).
b) Diga quais variáveis são futuras, mas não são a variável resposta (total de aluguéis naquela hora).
c) Diga quais colunas são metadados.

Questão 3 (valor 2.5 pontos)

a) Divida a base em treino e teste. Mantenha os dados de 2011 no treinamento e use os de 2021 para teste.
b) Use o OneHotEncoder do sklearn para transformar todas as variáveis qualitativas em dummies. Lembre-se de analisar todas as variáveis para determinar corretamente quais são qualitativas.

Questão 4 (valor 2.5 pontos)

Treine os seguintes modelos de rede neural na base de treino e compute os seus MAE (mean absolute error) na base de teste:

1. Modelo sem nenhuma cada intermediária;
2. Modelo com uma camada intermediária com 10 neurônios;
3. Modelo com duas camadas intermediárias com 10 neurônios cada.

Use random_state=42 para todos os experimentos. Os demais hiperparâmetros devem ser deixados com o valor padrão do sklearn.