SCC0277 - Competições de Ciências de Dados
Primeiro Projeto Prático

Neste projeto vamos nos aprofundar no desenvolvimento do problema **House Prices - Advanced Regression Techniques** do Kaggle. 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.

Questões

Embora as questões a seguir estejam numeradas - caso ache necessário - você pode desenvolver os código em outra ordem. Só se lembre de deixar clara a resposta de cada uma das questões.

Questão 1 (valor 2,5 pontos)

a) Proponha 3 novas variáveis explicativas que façam sentido do ponto de vista prático do problema;
b) Faça uma análise ilustrando a (possível) qualidade das variáveis propostas.

Questão 2 (valor 2,5 pontos)

a) Aprimore o pipeline desenvolvido em aula para que seja possível comparar mais de uma técnica de modelagem (por exemplo, KNN e Árvore de decisão) usando um mesmo grid-search;
b) Faça uso de uma técnica de modelagem ainda não falada na disciplina. Essa técnica não precisa ser uma modelo de regressão, ela pode ser uma técnica de agrupamento para ajudar na criação de novas variáveis. **Considere como técnicas já vistas: KNN, árvores e ensembles.**

Questão 3 (valor 2,5 pontos)

a) Comparar vários modelos candidatos e escolher o melhor;
b) Gerar os resultados do melhor modelo na base de teste e enviar para o Kaggle.

Questão 4 (valor 2,5 pontos)

a) Identificar um possível problema com a métrica de erro usada (RMSE). Justificar com explicações, exemplos e análises;
b) Apontar uma nova métrica que não tenha esse problema. Justificar com explicações, exemplos e análises.