SCC0277 - Competições de Ciências de Dados
Terceiro Projeto Prático

Neste projeto vamos nos aprofundar no desenvolvimento do problema Digit Recognizer 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) Faça o plot de uma imagem de cada uma das classes do problema.
b) Faça um histograma mostrando a distribuição das classes do problema. Discuta este histograma.

Questão 2 (valor 2,5 pontos)

a) Detecte variáveis que não são úteis para o modelo e remova elas da base de dados. Justifique a remoção de tais variáveis. Obs.: as variáveis não podem ser nenhum pouco úteis.
b) Usando o critério do item anterior é possível inferir que algumas variáveis devem ter pouca importância para os modelos que iremos construir?

Questão 3 (valor 2,5 pontos)

a) Comparar vários modelos candidatos e escolher o melhor de acordo com a acurácia. Usar pelo menos uma rede neural.
b) Gerar os resultados do melhor modelo na base de teste e enviar para o Kaggle. Colocar alguma evidência do envio no projeto entregue.

Questão 4 (valor 2,5 pontos)

a) Calcular a acurácia e a matriz de confusão do melhor modelo. Para fazer este item siga os seguintes passos:
    - Divida a base de treinamento em 2 partes (70/30%);
    - Treine o modelo na primeira parte e calcule os resultados na segunda;
    - Use os hiperparâmetros do melhor modelo da questão 3.
b) O modelo tem um resultado parecido para todas as classes, ou há uma grande diferença em seu comportamento (quantidade acertos e erros) para diferentes classes?