Smarkio – Teste Prático Ciência de Dados

Para realização do teste serão utilizados os dados do arquivo “teste_smarkio.xls”. Os dados estão disponíveis em duas abas, sendo a primeira para a realização das questões de 1 a 4 e a segunda para a questão de número 5.

Considerações:

1. O teste precisa estar disponível no Github em repositório público;
2. Os resultados também devem ser exportados em formato PDF;
3. Todas as instruções de como devemos proceder para rodar o código devem estar no ReadMe;
4. Arquivos utilizados e/ou gerados também devem estar junto do repositório.
5. Use bibliotecas e ferramentas open source.

Os datasets possuem as seguintes colunas:

1. Primeira aba - Análise_ML:
    a. pred_class - A classe que foi identificada pelo modelo;
    b. probabilidade - A probabilidade da classe que o modelo identificou;
    c. status - status da classificação de acordo com um especialista (approved);
    d. true_class - A classe verdadeira (se nula, assumir o pred_class);

    Obs: Se pred_class é igual a true_class, temos que o modelo acertou.

2. Segunda aba - NLP:
    a. letra - trecho de música;
    b. artista - cantora referente a letra.

Dessa forma, realize as seguintes atividades:

1. Análise exploratória dos dados utilizando estatística descritiva e inferencial, considerando uma, duas e/ou mais variáveis;
2. Calcule o desempenho do modelo de classificação utilizando pelo menos **três** métricas;
3. Crie um classificador que tenha como output se os dados com status igual a **revision** estão corretos ou não (Sugestão : Técnica de cross-validation K-fold);
4. Compare **três** métricas de avaliação aplicadas ao modelo e descreva sobre a diferença;
5. Crie um classificador, a partir da segunda aba - NLP do arquivo de dados, que permita identificar qual trecho de música corresponde às respectivas artistas listadas (Sugestão: Naive Bayes Classifier).

****As sugestões são apenas para direcionamento, podendo o candidato optar ou não em utilizá-las****