COMPUTATIONAL INTELLIGENCE SOCIETY UnB

Universidade de Brasília Faculdade de Tecnologia, Departamento de Engenharia Elétrica

Treinamento CIS - 1º Período (Regressão e Classificação)

1. **Conteúdos do Período**
   a. Conceitos básicos de Data Science (DS);
   b. Python para DS;
   c. Manipulação de base de dados;
   d. Visualização de dados;
   e. Pandas, numpy, seaborn e matplotlib;
   f. Normalização de parâmetros;
   g. Tipos de dados;
   h. Pré-processamento de dados;
   i. Redução de dimensionalidade;
   j. PCA;
   k. Datasets de Treinamento x Validação;
   l. Regressão Linear;
   m. Regressão Logística.
   n. Algoritmos de Classificação

2. **Conteúdos Essenciais**
   a. Tipos de Dados - Artigo com os tipos de dados em modelos estatísticos;
   b. PCA StatQuest - Vídeo explicando detalhadamente PCA;
   c. Data Imputation - Lidando com dados faltantes;
   d. Data Encoding - Como tratar variáveis categóricas;
   e. Padronização e Normalização - Alterando a escala dos dados;
   f. Treino e Teste - Como avaliar o seu modelo;
   g. Linear Regression StatQuest - Playlist sobre regressão linear;
   h. Logistic Regression StatQuest - Playlist sobre regressão logística.
   i. Métricas de avaliação - Como avaliar a performance do modelo em um problema de regressão.

3. **Conteúdos Complementares**
   a. Fundamentos de Python para Análise de Dados - Curso DataScience Academy, Módulos recomendados - 5, 8 e 9;
   b. Cursos do Kaggle - Cursos recomendados para o primeiro período - Python, Data Visualization, Pandas e Data Cleaning;
   c. Playlist - Introdução à Data Science, Visualização de Dados e Machine Learning - Playlist brasileira que explica dos conceitos básicos até as primeiras implementações dos conceitos citados;
   d. Python DataScience HandBook - Livro sobre Numpy, Pandas, Matplotlib, relativamente curto e direto;
   e. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow - Livro completo: Para o primeiro período recomendam-se os capítulos 1, 2, 8 e 9;
   f. Data Analysis with Dr Mike Pound - Playlist com diversos conceitos de análise de dados avançada;
   g. Machine Learning Sentdex - Vídeos iniciais da playlist sobre Machine Learning;

h. Tipos de modelos de regressão;
i. CRISP-DM - Metodologia para projetos de dados.

4. **Desafio: *Wine Quality***

O Vinho verde é um produto único da região de Minho, do Noroeste de Portugal. Médio em álcool, este vinho é particularmente apreciado devido ao seu frescor, especialmente no verão. Com base nisso, foram coletadas as seguintes informações sobre o vinho:

1 - acidez fixa (fixed acidity)
2 - acidez volátil (volatile acidity)
3 - acidez cítrica (citric acidity)
4 - açúcar residual (residual sugar)
5 - concentração de cloretos (chlorides)
6 - concentração de dióxido sulfúrico livre(free sulfur dioxide)
7 - concentração total de dióxido sulfúrico (total sulfur dioxide)
8 - densidade (density)
9 - pH (pH)
10 - concentração de sulfatos (sulphates)
11 - concentração alcoólica (alcohol)

Output (based na avaliação média de especialistas):
12 - qualidade (quality) (score entre 0 e 10)

Há dois datasets, relacionados às variantes vermelha e branca do “vinho verde”, de Portugal. Escolha um dos datasets contendo informações de uma das variantes (vermelha ou branca) e crie um modelo de regressão para prever a qualidade do vinho. A entrega é individual e deverá ser colocada no seu GitHub pessoal.