



Treinamento CIS - 1º Período (Regressão)

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.

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. Tarefas

Para esse período, foi criada uma [competição no kaggle](#):

- a. Fazer uma análise no dataset utilizando as ferramentas aprendidas no período;
- b. Demonstrar como pré-processar e representar os diferentes tipos de dados;
- c. Demonstrar a utilização de técnicas como one-hot encoding, redução de dimensionalidade e PCA;
 - i. Como a redução da dimensionalidade afetou no desempenho do modelo?
- d. Aplicar regressão nos dados - dividir o dataset em treinamento e validação e aplicar, pelo menos, 3 algoritmos distintos de regressão para prever o preço do imóvel. Testar e visualizar os resultados, comparando o desempenho dos modelos treinados;
 - i. Qual modelo performou melhor?
- e. Avaliar o desempenho do(s) modelo(s) treinado(s). A métrica da competição será o erro quadrático médio (MSE), porém, recomenda-se a avaliação do modelo por outras métricas adicionais, como o R^2 e o erro médio absoluto (MAE).

A entrega é individual e deverá ser colocada no seu GitHub pessoal.