INTRODUÇÃO À CIÊNCIA DE DADOS

Introdução às Bibliotecas Scikit-Learn, MatPlotLib e Seaborn

BIBLIOTECA SCIKIT-LEARN

- Scikit-learn é uma biblioteca de aprendizado de máquina (machine learning) de código aberto que oferece suporte ao aprendizado supervisionado e não supervisionado.
- Ela também fornece várias ferramentas para ajuste de modelo, pré-processamento de dados, seleção e avaliação de modelo e muitos outros utilitários.
- Construída sobre NumPy, SciPy e Matplotlib.
- Veremos Machine Learning mais adiante nessa disciplina, e a biblioteca base para todas nossas implementações será a Scikit-learn, iremos aprofundar os estudos com ela!

BIBLIOTECA SCIKIT-LEARN

- Ela já oferece alguns datasets que podem ser utilizados para realizar testes e aprender a utilizá-la.
    - Boston House-Prices Dataset
    - Breast Cancer Dataset
    - Diabetes Dataset
    - Digits Dataset
    - Iris Dataset

BIBLIOTECA SCIKIT-LEARN

- Vamos observar o site da ferramenta e ver como ele pode ser útil para o aprofundamento nos estudos.
- https://scikit-learn.org/stable/index.html
- Demonstração Site

VISUALIZAÇÃO

- Um dos primeiros artifícios de comunicação da humanidade.
- "Uma imagem vale mais que mil palavras".
- A ideia principal de uma visualização de dados por meio de imagens, gráficos ou cartografia é simplificar conteúdos de forma a permitir a compreensão de uma ideia geral do todo.
- Visualização envolve a arte de facilitar a percepção do todo!

VISUALIZAÇÃO

- A visualização é um recurso que facilita o processo de compreensão e de tomada de decisão em praticamente todas as áreas do conhecimento.
- Sabia que é possível processar uma imagem de 250 megapixels e tomar uma decisão a partir de uma imagem nunca anteriormente vista em milésimos de segundos?

VISUALIZAÇÃO

- A visualização é um dos pontos mais importantes de um processo que envolve DS. Se todos os procedimentos são bem feitos mas a entrega não atende às necessidades, tudo que foi feito antes perde o sentido.
- É muito importante entregar bem uma análise de dados para quem vai consumi-la.
- Importante saber explorar todos os recursos possíveis no processo de visualização, de forma a demonstrar com clareza o resultado da análise, impedindo que sejam tomadas decisões equivocadas por má interpretação da visualização.

VISUALIZAÇÃO

- Dominar o processo de visualização é uma arte, há muitas estratégias para fazer isso!
- Entender para que serve cada tipo de gráfico, e a melhor forma de visualizar os dados, é parte de um diferencial muito grande dos profissionais de DS.
  - Já conhece o site do Data Viz Project? Esse ambiente nos ajuda a conhecer um pouco mais sobre que tipo de gráfico usar para cada situação.
  - https://datavizproject.com/
- Abordaremos mais adiante as questões que podem envolver dados enviesados, falta de ética e visualizações que sugerem algo que não representa a análise.

BIBLIOTECAS MATPLOTLIB E SEABORN

- As bibliotecas MatPlotLib e Seaborn, do Python, fazem parte de um conjunto de bibliotecas open source para visualização de dados. Para geração de gráficos.
- A biblioteca Seaborn é baseada na MatPlotLib. Há também outras bibliotecas que também são baseadas na MatPlotLib.
- Há um conjunto muito grande de ferramentas além destas.
- E por que usar uma ou outra??
  - Há vários motivos para isso: praticidade, visual mais bonito, diversidade de possibilidades (detalhes), configurações automáticas, entre outros.

BIBLIOTECAS MATPLOTLIB E SEABORN

- A biblioteca MatPlotLib por ser a base para grande parte de outras bibliotecas, é uma das mais completas, entretanto muitas das configurações precisam ser feitas manualmente, com linhas de programação.
- A biblioteca Seaborn já se apresenta com um conjunto de parâmetros que se autoconfiguram, para se adaptar ao seu conjunto de dados e muitas das informações não precisam ser explicitadas, elas já estão configuradas. Irá perceber também que há uma pitada de bom gosto na arte final dos gráficos da biblioteca Seaborn.
- Vamos olhar uma demonstração dessas bibliotecas!

FINALIZANDO

- Visualização é muito importante para qualquer projeto de DS.
- Saber como apresentar os dados demanda conseguir compreender objetivos e pessoas, algo que vai além de programação.
- Conhecer as bibliotecas de visualização permite que seja possível implementar com programação algo que está desenhado na ideia de um profissional de DS, algo que vai impactar e facilitar o uso dos dados pelo usuário que vai consumir os dados.
- Visualização é uma arte!

INTRODUÇÃO À CIÊNCIA DE DADOS

Introdução às Bibliotecas Scikit-Learn, MatPlotLib e Seaborn