Ponta do Iceberg: primeiros passos na Ciência de dados

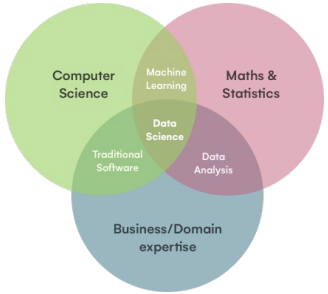
Daniel Brito dos Santos (danielbrito@uenf.br); Annabell Del Real Tamariz

dados da Instituição do apresentador

Universidade Estadual do Norte Fluminense Darcy Ribeiro

O que é

Ciência de Dados é a disciplina responsável pela sistematização das técnicas, conceitos e ferramentas usadas na resolução de um problema por meio de dados, desde a concepção de uma pergunta respondível, até a apresentação de uma possível resposta.



Área do conhecimento na interseção entre estatística, computação e domínio do contexto do problema.

Problema → Dados brutos → Informação relevante e aplicável

Objetivo e metodologia

O objetivo desse projeto é mapear o ferramental básico de um cientista de dados por meio do estudo bibliográfico e da execução de um projeto representativo.

Para tanto, estamos seguindo a metodologia CRISP-DM (Cross Industry Standard Process for Data Mining), abordando cada uma das etapas na busca de seu respectivo arcabolso prático e teórico. O que se dará principalmente por meio do estudo de livros-texto, prática deliberada e recursos disponíveis na plataforma Kaggle.

Site do grupo de pesquisa, se houver

CRISP-DM

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment

Resultados

Construção de vocabulário em:

- **Fundamentos gerais**:
  - mapa/glossário dos principais conceitos;
  - familiarização com Python e Pandas;
  - conceitos de SQL.
- **Business Understanding:**
  Conceitos para estruturar uma solução baseada em dados para determinado problema.
- **Data Understanding:**
  Abordagens e cuidados com os dados em si, como por exemplo na identificação das variáveis de interesse, tratamento de duplicatas e valores faltantes.

Discussão e Conclusão

Nesses primeiros meses de trabalho, o enfoque foi na construção tanto do vocabulário fundamental quanto das duas primeiras etapas da metodologia, fundamentais para a confiabilidade dos resultados e assertividade do projeto. Em seguida produziremos um repositório que será disponibilizado no Github documentando todo o desenvolvimento, e iniciaremos os trabalhos com os dados do projeto em si, onde estudaremos a preparação de dados, modelagem, avaliação do modelo e deployment.

Assim, entendemos que essa construção prévia do arcabouço teórico comum a grande maioria dos projetos de dados para em seguida catalisar a aprendizagem em contextos de maior complexidade e amplitude, tem se mostrado um caminho promissor para os objetivos do projeto.