Ponta do Iceberg: Primeiros passos na Ciência de dados

Daniel Brito dos Santos, Annabell Del Real Tamariz

A Ciência de Dados pode ser definida como a área do conhecimento na interseção entre estatística, computação e conhecimento de causa, que se ocupa de aplicar, organizar e expandir as técnicas, ferramentas, e conceitos necessários no processo de transformar dados brutos em informação relevante e aplicável. Entretanto, o repertório necessário, à diversidade dos problemas abordados, e à velocidade na qual o campo se desenvolve, tornam a formação profissional qualificada um desafio inerente à Data Science. Nesse sentido, o presente projeto propõe mapear os métodos, ferramentas e fundamentos do ferramental básico de um cientista de dados por meio do estudo bibliográfico e da execução de um projeto representativo. Para tanto, seguiremos a metodologia CRISP-DM (Cross Industry Standard Process for Data Mining), que divide um projeto de dados em 6 etapas: compreensão de negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e deployment. De modo que abordaremos cada uma buscando seu respectivo conjunto de conceitos e boas práticas, o que se dará principalmente por meio do estudo de livros-texto, e dos recursos disponíveis no Kaggle, a maior plataforma de Data Science online com minicursos, fóruns, diversos datasets, e principalmente as competições abertas onde cada usuário pode submeter a sua solução ao problema proposto, que fica acessível em um ranking, de modo a promover a troca de ideias e a dinâmica de aprendizado promovendo o desenvolvimento de um repertório próprio. Desses projetos escolhemos o Titanic, no qual desenvolveremos um modelo preditivo de classificação binária da sobrevivência de cada tripulante do famoso navio a partir das informações disponíveis como idade e Porto de embarque. Nesses primeiros meses de trabalho o enfoque foi na construção dos fundamentos gerais e das duas primeiras etapas mencionadas. O que se traduz em um glossário e mapa de conceitos principais, familiarização com o slack de Python na figura da biblioteca Pandas, conceitos de SQL, e conjuntos de métodos para abordar a compreensão do problema e dos dados. Estamos na transição para a etapa de preparação dos dados onde de fato trabalharemos no dataset do projeto Titanic e consumirmos o conteúdo das competições Kaggle. Entendemos ter sido importante essa construção prévia em ambiente controlado do arcabouço teórico comum a grande maioria dos projetos de dados, para em seguida catalisar a aprendizagem em contextos de maior complexidade e amplitude. Assim, a compreensão dos fundamentos, sistematização e prática, partindo de um recorte encadeado e exposição deliberada ao “mundo real” tem se mostrado um caminho promissor para os objetivos do projeto.

Instituição do Programa de IC, IT ou PG: UENF
Fomento da bolsa (quando aplicável):