PROGRAMA INSTITUCIONAL DE BOLSAS DE INICIAÇÃO CIENTIFÍCA E TECNOLÓGICA

UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE DARCY RIBEIRO
Centro CCT
Labotatório LCMAT

Plano de Trabalho para Renovação de Bolsa de Iniciação Científica

Bolsista: Daniel Brito dos Santos
Matricula: 00119110393
Orientadora: Prof. Dra. Annabell Del Real Tamariz
Curso: Bacharelado em Ciência da Computação

Titulo do Projeto: Project-driven Data Science: Aprendendo e Mapeando
Título do Plano de Trabalho: Aplicação Exploratória de *Data Science* em um Projeto Real de Dados.
Fonte financiadora: PIBIC/UENF

1 Justificativa

Ciência de dados pode ser definida como o conjunto de técnicas, conceitos e ferramentas utilizados para extrair informação relevante de dados do mundo real. Logo percebemos o seu enorme potencial de solucionar problemas e inovar, fato comprovado pela crescente demanda por profissionais capazes de operacionalizar essa ciência nos mais diversos contextos [Res]. Entretanto, pela sua própria natureza multidisciplinar, dinâmica e aplicada, a formação de um cientista de dados é um dos seus principais desafios.

Dessa forma, o principal objetivo dessa pesquisa como um todo é cartografar o atual campo da ciência de dados, evidenciando os principais conceitos e ferramentas introdutórias. Neste segundo ano de trabalho, nosso objetivo é consolidar e aprofundar os aprendizados do ano anterior (2021). Assim, aliado aos fundamentos construídos anteriormente, identificamos importantes oportunidades de aprofundamento em diversas áreas e conceitos da ciência de dados. Desse modo, selecionamos os seguintes tópicos para desenvolvermos nesse próximo ano: Banco de Dados, Modelos Preditivos e Visualização de dados.

Assim, no segundo ano na bolsa de Iniciação Científica (IC), propomos mapear e aplicar cada uma dessas áreas por meio de um projeto real de dados. Tal projeto deve ter parte interessada e dados a serem estruturados, armazenados e analisados. Desse modo, teremos uma aprendizagem baseada na resolução de problemas, direcionada pelas necessidades do projeto, fundamentos construídos no ano anterior (2021) e material relevante de cada novo tópico estudado.

2 Objetivos

1. Identificar um problema real e relevante no qual possamos definir e estruturar um Projeto de Dados para aborda-lo. De modo que possamos aplicar os conceitos estudados no primeiro ano da bolsa de IC, bem como direcionar os aprofundamentos deste segundo ano.
2. Executar de acordo com as diretrizes estudadas no ano anterior o projeto definido.
3. Estudar e aplicar no projeto os principais conceitos e ferramentas das seguintes áreas:
   (a) Banco de dados, no contexto da Obtenção de Dados especialmente na estruturação e consulta a bancos de dados com a linugagem SQL.
   (b) Modelos Preditivos, como sub-área do Aprendizado de Máquina.
   (c) Visualização de Dados, como uma ciência própria da comunicação visual em todas as suas dimensões.

3 Metodologia

Seguindo a ideia de aprendizagem baseada em projeto [KB06], nesse segundo ano de bolsa de IC iremos buscar um projeto real, aliado aos materiais de referência relevantes para direcionar o estudo de cada tópico planejado. Dessa forma, a cada etapa do trabalho estabeleceremos perguntas e buscaremos suas respostas em um processo cíclico inspirado nas metodologias ágeis [BBVB+01] tendo em vista gerar valor à cada etapa.

3.1 Bancos de Dados

No ano anterior, identificamos a importância da linguagem SQL no arsenal de um cientista de dados, bem como elencamos recursos para nos aprofundarmos na linguagem:

- Os livros [DeB22, SKS19] para referência da linguagem SQL e conceitos mais amplos de Bancos de Dados.
- Os sites SQLbolt[1], Hacker Rank[2] e PostgreSQL Exercises[3] para prática deliberada de resolução de problemas com SQL.

[1] sqlbolt.com/
[2] www.hackerrank.com/domains/sql
[3] pgexercises.com

Nesse ano, iremos utilizar tais recursos para estudar a linguagem SQL bem como buscaremos expandir a pesquisa para conceitos mais amplos, visando construir conhecimento sobre a criação e consulta de bancos de dados. Nesse sentido, iremos utilizar os exercícios disponibilizado nos sites mencionados para direcionar o aprendizado SQL na prática. Também utilizaremos os livros [DeB22, SKS19] para estudo e consulta dos conceitos de banco de dados ao mesmo tempo em que criaremos um sistema de banco de dados para armazenar e organizar os dados do Projeto escolhido.

3.2 Modelos Preditivos

Conforme vimos anteriormente, modelos preditivos são o principal componente do Aprendizado de Máquina, através deles conseguimos construir programas capazes de encontrar padrões e relações entre variáveis e criar uma abstração matemática, esta que pode ser utilizada para compreender e principalmente predizer uma variável alvo a partir de inputs. Vimos também a complexidade dessa área mas que paradoxalmente podemos aplicar algum dos modelos mais conhecidos com grande facilidade por meio da biblioteca Scikit Learn em Python. Dessa forma, localizamos os cinco modelos mais representativos e os aplicamos no projeto Titanic desenvolvido no primeiro ano de bolsa de IC (2021). Nesse ano, propomos nos aprofundarmos nos mecanismos e casos de uso de cada um desses modelos.

- *Logistic Regression*
- *Deciosion Tree Classifier*
- *Random Forest Classifier*
- *KNN*
- *Suppor Vector Classifier*

Para tanto estudaremos os livros [Van16, Aur17], buscando documentar cada um dos modelos e suas aplicações. Além de aplicá-los na prática, conforme necessidade do projeto.

3.3 Visualização de Dados

A visualização de dados diz respeito ao conjunto de conceitos e ferramentas utilizados na busca de uma comunicação visual efetiva. Nesse sentido utilizaremos o livro [Kna15] como nossa maior referência. Também percorreremos as referências encontradas no artigo [CFGT20] que mapeou a prática da ciência de dados buscando evidenciar o papel da visualização em cada processo. Dessa forma, além de estruturar uma visão holística e fundamentada desse campo, executaremos o Projeto selecionado visando sempre aplicar o ferramental na prática, bem como identificar as necessidades de visualização do projeto para buscar respaldo na literatura.

4 Etapas

As etapas foram divididas de modo a realizarmos duas etapas simultaneamente a cada mês, cada objetivo foi dividido de modo a otimizar o tempo e as necessidades do projeto.

	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º
A - Definição do Projeto de Dados	■	■										
B - Execução do Projeto de Dados			■	■	■		■		■		■	
C - Banco de Dados		■		■				■				
D - Modelos Preditivos					■		■			■		
E - Visualização de Dados			■			■			■		■	
F - Elaboração do Relatório Final						■		■		■		■

Tabela 1: Etapas do plano de trabalho

Referências

[Aur17] Géron Aurélien. Hands-on machine learning with scikit-learn & tensorflow. *Geron Aurelien*, 2017.

[BBVB+01] Kent Beck, Mike Beedle, Arie Van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, et al. Manifesto for agile software development. 2001.

[CFGT20] Anamaria Crisan, Brittany Fiore-Gartland, and Melanie Tory. Passing the data baton: A retrospective analysis on data science work and workers. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1860–1870, 2020.

[DeB22] Anthony DeBarros. *Practical SQL: A Beginner's Guide to Storytelling with Data*. No Starch Press, 2022.

[KB06] Joseph S Krajcik and Phyllis C Blumenfeld. *Project-based learning*. na, 2006.

[Kna15] Cole Nussbaumer Knaflic. *Storytelling with data: A data visualization guide for business professionals*. John Wiley & Sons, 2015.

[Res] Grand View Research. Data science platform market size, 2020-2027.

[SKS19] Abraham Silberschatz, Henry Korth, and S Sudarshan. *Database System Concepts*. McGraw-Hill, New York, NY, 7 edition, April 2019.

[Van16] Jake VanderPlas. *Python data science handbook: Essential tools for working with data*. "O'Reilly Media, Inc.", 2016.