Ponta do Iceberg: primeiros passos na Ciência de dados

Plano de Trabalho para Solicitação de Bolsa de Iniciação Científica PIBIC-Nota 10

Edital PIBi 03/2021 - PIBIC NOTA 10

Aluno: Daniel Brito dos Santos

Orientadora: Annabell Del Real Tamariz

Curso: Bacharelado em Ciência da Computação

Novembro de 2020

Campos dos Goytacazes – RJ

1. JUSTIFICATIVA

Sites acompanham cada clique de cada usuário, celulares registram sua localização, hábitos de uso, padrões de sono e movimento. “Pulseiras fit” medem frequência cardíaca, oxigenação sanguínea e permeabilidade da pele. Assistentes residenciais processam vozes para extrair comandos, além de aumentar o banco de comportamentos e rotinas dos usuários. Assim, tudo o que acessamos, compramos, lemos e ouvimos gera dados. Analogamente, governos, indústrias, laboratórios e a própria internet representam uma gigantesca teia de informações entrelaçadas. Enterrado em algum lugar dessa vastidão estão respostas para incontáveis perguntas, muitas ainda sequer imaginadas. (GRUS, 2015)

Nesse sentido surge a Ciência de Dados (DS) que pode ser definida como o campo de conhecimento fundamentalmente interdisciplinar; responsável por aplicar, organizar e expandir o conjunto de ferramentas, técnicas e conceitos necessários no processo de transformação de dados brutos em informações relevantes. Entretanto, considerando a amplitude de sua “caixa de ferramentas”, a necessidade de repertório para aplicá-las, a diversidade dos problemas abordados e a velocidade com que todo o campo se desenvolve, a formação de cientista de dados é um desafio inerente à própria natureza desse campo.

Nesse sentido, propomos percorrer os primeiros passos na ciência de dados, através de um projeto representativo, a fim de aprender, praticar e mapear as técnicas, ferramentas, fundamentos e referências necessárias para cada etapa, e dessa forma, desenvolver o ferramental básico de um cientista de dados. Bem como contextualizar e disponibilizar todo o código produzido além do referencial teórico utilizado em um caderno interativo online. Dessa forma, ao final deste ano teremos um caminho percorrido a ser sugerido para se abordar a ciência de dados, o que entendemos estar de acordo com uma aprendizagem eficaz, autodidata e exploradora.

2. OBJETIVOS

1. Ganhar fluência nas ferramentas que fundamentam a prática da Ciência de Dados (DS): SQL, Jupyter, Flask e principalmente Python com suas bibliotecas da área: Pandas, Numpy, Scikit Learn, Matplotlib, Seaborn.

2. Estudar os métodos empregados atualmente na prática da Ciência de Dados. Como por exemplo técnicas de: Feature Engineering, Data Mining e Machine Learning.

3. Integrar informações da literatura visando compreender e explicitar as bases computacionais e matemáticas de cada método empregado.

4. Consolidar aprendizagem por meio da execução do projeto “Titanic” da plataforma Kaggle, de acordo com o pipeline de um projeto DS, a fim de compreender cada uma de suas etapas, desde a definição da pergunta a ser respondida até o deployment do modelo.

5. Documentar em um caderno interativo online todo esse processo de modo a mapear as referências utilizadas e correlaciona-las à sua aplicação prática.

3. METODOLOGIA

Visando direcionar os estudos e abranger as principais ferramentas básicas do arsenal de um cientista de dados, propomos executar o projeto “Titanic” disponível na plataforma Kaggle[1]. Essa plataforma é um importante recurso de aprendizagem de DS por apresentar minicursos, fóruns, e principalmente sediar competições abertas de Ciência de Dados[2] nas quais cada usuário pode submeter a sua solução que fica acessível em um ranking, o que promove a troca de ideias e a dinâmica de aprendizado. A principal vantagem desse projeto é justamente sua popularidade, pois a vasta produção ao seu respeito traz uma importante diversidade de abordagens, discussões e exemplos, o que vai permitir o contato com um grande vocabulário de técnicas, métodos e fundamentos teóricos, promovendo, dessa forma, o desenvolvimento de um repertório próprio.

Executaremos no ecossistema de Python o projeto “Titanic” de acordo com a sequência de cinco passos que constituem um “Pipeline” de Data Science (DS) (OZDEMIR, 2005), listados a seguir:

1. Definição da pergunta;
2. Obtenção dos dados;
3. Data Analysis (limpeza e exploração);
4. Machine Learning (modelagem);
5. Comunicação dos resultados (Data Visualization e deployment).

Assim, simultaneamente à própria resolução exploratória de cada etapa, estudaremos seu conteúdo teórico e prático da seguinte forma:

0. Bases Gerais:

Utilizaremos diversas mídias como fonte de conteúdo em função da própria natureza multidisciplinar, dinâmica, e recente da Ciência de Dados (DS), portanto, além dos livros e artigos que permanecem fundamentais, há um contingente considerável de conteúdos em blogs, vídeos, plataformas e cursos online. Elencamos a seguir algumas dessas referências que utilizaremos:

[1] https://www.kaggle.com/c/titanic
[2] https://www.kaggle.com/competitions

Os livros “Data Science From Scratch”, “Python Data Science Handbook”, “Principles of Data Science” e “Doing Data Science” serão as bases gerais deste projeto, lidos de acordo com a necessidade e o desenvolvimento do mesmo. Visto que a maneira estruturada na qual eles apresentam seu conteúdo permite abordar a DS com o rigor e a abrangência que buscamos, ainda que nesse plano de trabalho o enfoque seja introdutório.

Complementando os livros, e considerando a velocidade com que a área se desenvolve, temos os blogs Towards Data Science[3], Analytics Vidhya[4], KDnuggets[5], e os canais no youtube sentdex[6], RichardOnData[7], Ken Jee[8], Mario Filho - Machine Learning[9].

Bugs e dúvidas pontuais em relação a parte computacional serão consultadas no StackOverflow[10], fóruns do Kaggle, blogs e vídeos de referência bem como na documentação oficial de cada ferramenta.

Utilizaremos a ferramenta Jupyter Notebook[11] como IDE para construir esse projeto em células executáveis e interativas, que serão disponibilizadas ao final do projeto.

1. Definição da pergunta

Na maior parte dos projetos de DS, é papel do Cientista de Dados formular, dado o problema em questão, qual é a pergunta passível de resposta através de dados, e em seguida elencar o necessário para respondê-la.

Sendo o “Titanic” um projeto Kaggle e talvez o mais clássico da DS, ele já apresenta um objetivo pré estabelecido, nesse caso, determinar “Quais foram as pessoas com maiores chances de sobreviver ao naufrágio do Titanic”. Em outras palavras, determinar quais características mais influenciaram na sobrevivência dos tripulantes.

Para responder esta pergunta devemos obter os dados dos passageiros, nos familiarizarmos com os mesmos e analisá-los, para em seguida criar um modelo preditivo de classificação binária, visto que queremos classificar cada passageiro como sobrevivente ou vítima, e finalmente criarmos visualizações que apresentem de forma clara e elucidativa a resposta à pergunta inicial.

Nesse tema, estudaremos os materiais de referência na busca de mapear as possíveis técnicas para formulação de perguntas respondíveis por meio de dados.

---

[3] https://towardsdatascience.com/
[4] https://www.analyticsvidhya.com
[5] https://www.kdnuggets.com
[6] https://www.youtube.com/user/sentdex
[7] https://www.youtube.com/c/RichardOnData
[8] https://www.youtube.com/c/KenJee1
[9] https://www.youtube.com/c/MarioFilhoML
[10] https://stackoverflow.com
[11] https://jupyter.org/

2. Obtenção dos dados

A partir da definição do problema concluímos a necessidade de obtermos os dados a respeito de cada passageiro, nesse caso, através do próprio sistema do Kaggle. Portanto estudaremos as ferramentas e código necessário para importação dos dados que passam a ser denominados “Dataset”.

Além dessa importação direta, estudaremos os outros métodos de coleta de dados como o Web Scraping[12] nos materiais de referência (livros e blogs) e faremos o mini-curso de SQL disponível no Kaggle[13], a língua franca dos bancos de dados estruturados.

3. Data Analysis (limpeza e exploração)

Nesta etapa aplicaremos diversas técnicas de limpeza e análise de dados a fim de compreender a natureza das variáveis e identificar possíveis relações entre as mesmas, de modo a melhor estruturar o modelo que desenvolveremos a seguir. Para tanto, estudaremos Data Mining[14], Data Wrangling[15] [16] [17] e as poderosas bibliotecas de Python: Pandas[18] [19], que é a referência absoluta para manipulação de tabelas, e NumPy[20], que traz uma robusta gama de implementações matemáticas, principalmente aplicações em álgebra linear.

Nesse sentido, além dos capítulos referentes ao tema nos livros de base, utilizaremos de referência computacional o livro “Python for Data Analysis” (MCKINNEY, 2017), e considerando a base estatística dessas análises, o curso de estatística e probabilidade da Khan Academy[21], maior plataforma de aprendizado online, e principalmente o livro “Practical Statistics for Data Scientists” (BRUCE; GEDECK, 2020).

4. Machine Learning (modelagem)

Nesta fase, estudaremos o que é machine learning, bem como os seus principais algoritmos a fim de compreender e mapear os fundamentos, vantagens, limitações de cada um e quais métricas podemos utilizar para compará-los. Também estudaremos a biblioteca SciKit Learn[22], que inclui implementações em estado da arte dos principais algoritmos de Machine Learning (ML). Dessa forma,

---

[12] Data Science Skills: Web scraping using python | by Kerry Parker
[13] https://www.kaggle.com/learn/intro-to-sql
[14] 5 Data Mining Techniques Every Data Scientist Should Know | by Sara A. Metwalli
[15] https://www.elderresearch.com/blog/what-is-data-wrangling-and-why-does-it-take-so-long/#:~:text=Data%20wrangling%20is%20the%20process,20%25%20for%20exploration%20and%20modeling.
[16] Top Data Wrangling Skills Required for Data Scientists | by ODSC - Open Data Science
[17] https://www.kaggle.com/learn/data-cleaning
[18] https://pandas.pydata.org/
[19] https://www.kaggle.com/learn/pandas
[20] https://numpy.org/
[21] https://www.khanacademy.org/math/statistics-probability/
[22] https://scikit-learn.org/stable/index.html

selecionaremos um dos algoritmos estudados para implementar nosso modelo para classificar os sobreviventes do “Titanic”.

Como referência temos principalmente os capítulos de ML nos livros base, o curso de Machine Learning ministrado por Andrew Ng em Stanford disponibilizado no Coursera[23], a documentação oficial da SciKit Learn[24], e os minicursos do Kaggle[25].

5. Data Visualization

Métodos de visualização são fundamentais durante todo o processo de análise, não obstante é nesse passo que devemos estruturar a resposta final de modo a expressá-la com clareza para as partes interessadas, o que geralmente é feito por meio de apresentações visuais como gráficos.

Para tanto, por meio dos blogs, livros e minicurso Kaggle[26] estudaremos as bibliotecas gráficas de Python: Seaborn[27] e Matplotlib[28]. Além disso, pesquisaremos sobre o tema mais amplo de comunicação científica por artifícios gráficos tendo como principal referência os livros do professor Alberto Cairo: “The Functional Art” e “How Charts Lie”.

6. Deployment

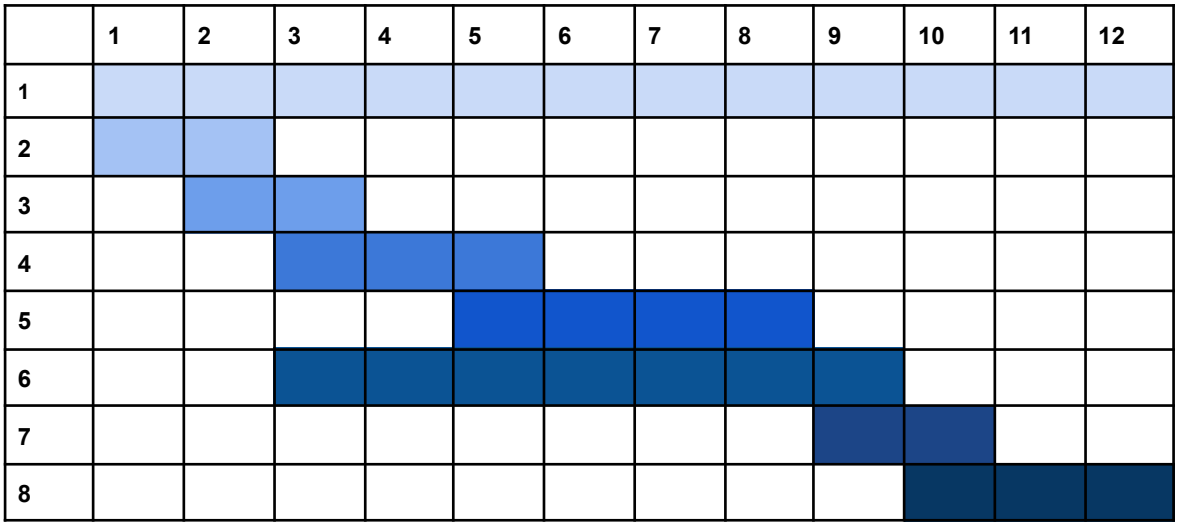
Finalmente, estudaremos por meio de blogs o framework web Flask para encapsular nosso modelo preditivo e disponibilizá-lo online.[29] Tendo em vista que o processo de deployment consiste em transferir o modelo criado no ambiente de desenvolvimento para um ambiente de produção, de modo a torná-lo acessível para as outras pessoas.

Em suma, o possível modelo é inicialmente pensado junto ao escopo do problema, treinado com dados que importamos, limpamos e analisamos de modo a melhor compreender as possíveis relações e características, que utilizamos como parâmetros na implementação de um modelo dentre vários possíveis. A partir de então, criamos visualizações e apresentamos os resultados de nossa análise, encapsulamos e transferimos nosso modelo para o ambiente de produção, onde ele enfim será disponibilizado ao público, fechando o ciclo de vida básico de um projeto de dados.

---

[23] https://www.coursera.org/learn/machine-learning
[24] https://scikit-learn.org/stable/user_guide.html
[25] https://www.kaggle.com/learn/intro-to-machine-learning
[26] https://www.kaggle.com/learn/data-visualization
[27] https://seaborn.pydata.org/
[28] https://matplotlib.org/
[29] https://faculty.ai/blog/creating-data-science-apis-with-flask/

5. ETAPAS



1. Bases gerais
2. Definição de problema
3. Obtenção de dados
4. Data Analysis
5. Machine Learning
6. Data Visualization
7. Deployment
8. Elaboração do Relatório Final

6. REFERÊNCIAS BIBLIOGRÁFICAS

BRUCE, Peter; BRUCE, Andre; GEDECK, Peter. **Practical Statistics for Data Scientists**. [S. l.]: O'Reilly, 2020.

CAIRO, Alberto. **The Functional Art.** [S. l.]: New Riders, 2013.

GRUS, Joel. **Data Science from Scratch**. [S. l.]: O'Reilly, 2015.

MCKINNEY, Wes. **Python for Data Analysis**. Data Wrangling with Pandas, NumPy, and IPython. [S. l.]: O'Reilly, 2017.

OZDEMIR, Sinan. **Principles of Data Science**. [S. l.]: Packt, 2016.

SCHUTT, Rachel; O'NEIL, Cathy. **Doing Data Science**: Straight Talk From the Frontlines. [S. l.]: O'Reilly, 2014.

VANDERPLAS, Jake. **Python Data Science Handbook**: Essential Tools for Working with Data. [S. l.]: O'Reilly, 2017.