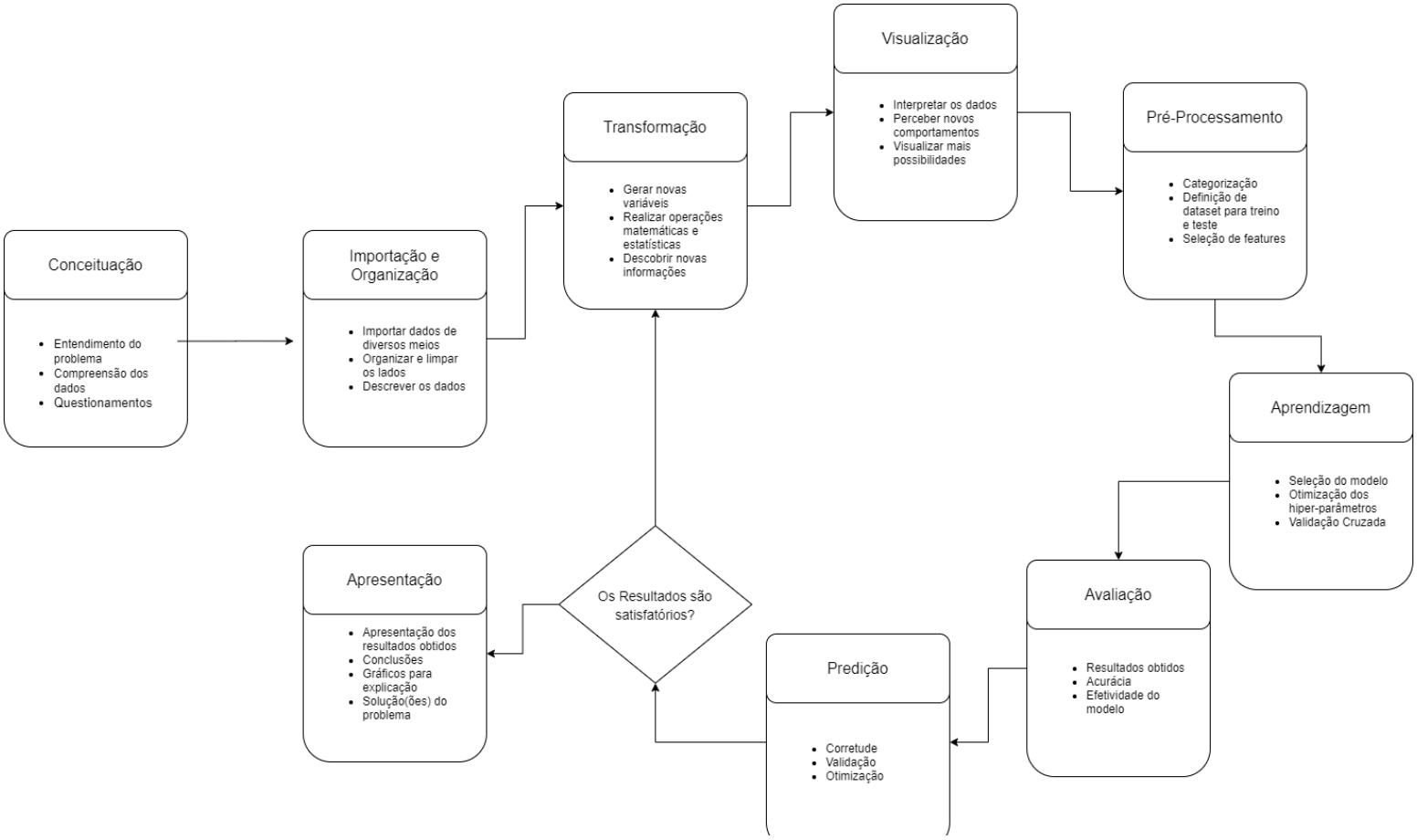
Trabalho 2

Fluxograma



O fluxograma acima define os principais passos de um cientista de dados. Iniciamos o processo no passo de:

Conceituação: “Primeiramente, um cientista de dados deve estudar e compreender o problema proposto, através de pesquisa na área que apresenta os dados. Dessa forma, ele é capaz de entender os reais obstáculos e vislumbrar possíveis objetivos para serem aplicados na análise do dado.”

Logo após:

Importação e Organização: “Logo após a compreensão do assunto e do problema, um cientista de dados deve extrair o conjunto de dados necessário para a análise. Esses dados podem vir de banco de dados, planilhas e diversas outras fontes de informação. Além disso, é necessário organização e descrição dos dados extraídos, tornando o trabalho de analisá-los mais claro. Para isso, o cientista deve documentar os dados extraídos, deve organizar e limpar os dados de qualquer tipo de sujeira que tenha sido originada a partir da extração.”

O cientista então começa fazendo uma análise generalista e básica sobre os dados para extrair possíveis informações que não sejam imediatamente óbvias:

Transformação: “Aqui são realizadas análises iniciais sobre os dados, novas variáveis são geradas a partir dos dados anteriores e são aplicadas diversas operações

básicas de matemática e estatística para tentar extrair alguns conceitos básicos sobre o dado, por exemplo: médias, medianas e proporções. Além disso, dependendo do objetivo o dado pode ser transformado e transposto para atender melhor a um determinado objetivo do cientista."

Visualização: "Nessa fase o cientista de dados utiliza-se de diversas modelagens e gráficos para visualizar informações gerais sobre o conjunto de dados. Através desta visualização é que podem-se gerar os primeiros conhecimentos que irão levar à solução do problema. Uma boa visualização dos dados pode revelar comportamentos inesperados e até mesmo definir novos objetivos para o cientista."

E então, com uma visão mais clara do que os dados realmente representam e possíveis objetivos e correlações obtidos, o cientista de dados parte para uma análise aprofundada utilizando modelos de aprendizado e categorizando o dado:

Pré-processamento: "Nesse passo alguns objetivos estão definidos então é necessário começar a analisá-los de forma aprofundada. Para isso geralmente são usados modelos de aprendizado e para alimentar esses modelos é necessária uma categorização e separação dos dados. Os que serão úteis para o aprendizado, o conjunto que será utilizado no treinamento do modelo, o conjunto que será utilizado como teste e as categorias que o modelo vai aprender."

Aprendizagem: "Aqui o modelo é escolhido, seja uma rede neural ou aprendizado de máquina, o algoritmo para aprendizado até o tipo de neurônio utilizado na rede neural. Através de alguns testes, são realizados ajustes nos hiperparâmetros do modelo para melhor comportar o dado e são utilizadas técnicas de cross-validation para observar a generalidade e corretude do modelo aplicado."

Assim, o cientista pode avaliar a escolha de modelo e correlações para observar se o problema pode ou não ser resolvido, se foi realmente resolvido ou se algum erro ocorreu em passos anteriores:

Avaliação: "Checar com os próprios dados se a categorização que o modelo aprendeu está correta. Assim, o cientista pode calcular a acurácia do seu modelo, sabendo assim a efetividade do método aplicado e podendo pensar se existe como melhorar esse modelo ou se ele já é satisfatório para ser testado no próximo passo.

Predição: "Nesse passo o cientista testa seu modelo com diversos dados, podendo ou não fazerem parte do seu conjunto inicial. O ideal é que a predição do modelo seja testada com outros conjuntos para identificar problemas de underfit ou overfit de dados e o cientista saber para onde retornar para tornar o modelo mais correto."

Avaliados os resultados da predição o cientista decide se eles são suficientes para resolver o problema ou se são insuficientes, no último caso, o cientista pode refazer todo o processo desde a transformação dos dados para testar novos métodos e chegar a diferentes conclusões. Caso os resultados resolvam o problema e a predição seja satisfatória:

Apresentação: "Na última fase o resultado de todo o trabalho é mostrado, o que foi obtido e o que não foi, os modelos que geraram resultados positivos, as categorias e features que foram extraídas e utilizadas, etc. Essa fase é o produto de todo cientista de dados, por isso deve ser organizada, limpa e clara. Tendo em vista que, sem uma explicação clara e coesa todo o trabalho não irá fazer sentido nos olhos do cliente que requisitou o serviço. Então gráficos são gerados, explicados e debatidos com base nos resultados obtidos, todos sendo apresentados para o cliente."

Conhecimentos de um cientista de dados: Como observamos, o cientista de dados deve está sempre se adaptando para solucionar problemas em diversas áreas, trazendo soluções coerentes e corretas. Além disso, deve possuir conhecimento matemático e estatístico para abordar o dado de forma generalista e saber quais relações buscar dentro do dado. Ademais, deve possuir conhecimento em técnicas de inteligência artificial para implementar algoritmos de predição e assim gerar modelos que podem categorizar o dado estudado. Por fim, deve ter conhecimento sobre visualização computacional para tornar seu próprio trabalho mais fácil, observar coisas talvez inesperadas no dado e poder apresentar os resultados obtidos com clareza.

Bibliografia:

- https://medium.com/data-hackers/como-se-tornar-um-cientista-de-dados-bdda45047be1
- https://www.abgconsultoria.com.br/blog/afinal-o-que-e-data-science/
- https://creately.com/diagram/example/ixh78j671/Data%20Science%20Project%20Work%20flow
- https://www.researchgate.net/figure/Flow-of-development-for-FDQ-KDT_fig4_283430974
- http://www.ipsr.ku.edu/naddi/about.shtml