Uma Abordagem Analítica do Fluxo de Trabalho de um Cientista de Dados

Alexei Alves de Souza

Agosto de 2020

1 Introdução

Ciência de Dados talvez não seja a "ciência" que vem a mente primeiro quando pensamos nesse termo, porém a cada dia certamente é a que está mais presente no nosso cotidiano.

Quem nunca abriu um site que pediu para "aceitar cookies"? Ou ficou curioso em relação a uma previsão em um jornal sobre determinado assunto?

Determinar perfis pessoais de um determinado usuário, prever resultados, encontrar padrões são apenas algumas das atividades abordadas na ciência de dados.

Porém dados não são como um horóscopo com textos fáceis de ler e que tentam dizer sobre você ou sobre o seu dia pelos atros. Eles são conjuntos de valores armazenados em bancos de dados muitas vezes gigantescos em que é necessário uma análise e uma visão científica para extrair algum resultado ou interpretação de algo tão bruto.

Esse trabalho é feito pelo Cientista de Dados o qual deverá encontrar a melhor e mais adequada maneira de tratar esses dados definindo o "processo de ciência de dados".

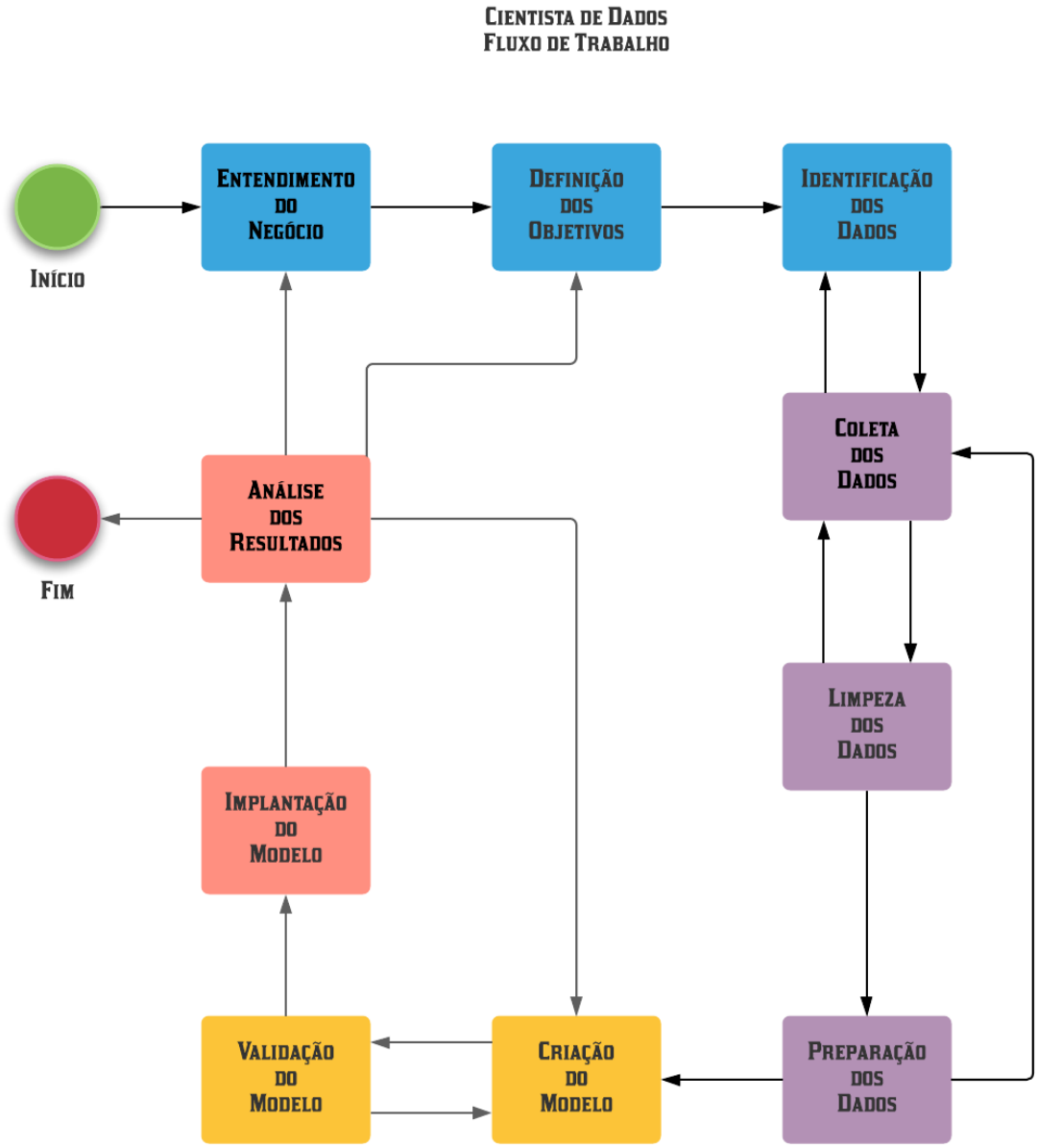
Existem de fato muitas abordagens sobre essse processo porém em geral é tomada como base a metodologia do Crisp-DM(Cross Industry Standard Process for Data Mining) projetada originalmente para a mineração de dados. Os processos desta metodologia sao mostrados a seguir:

- Compreensão de Negócios
- Entendimento de Dados
- Preparação de Dados
- Modelagem
- Avaliação
- Desenvolvimento

  A seguir irei mostrar um fluxograma mais abragente destes processos e seus desdobramentos baseados nos fluxos de trabalhos de grandes empresas como a IBM e a Microsoft.

2 Fluxograma - Cientista de Dados

2.1



2.2 Business Understanding

- **Entendimento do Negócio:**

  Nessa estapa será definido o problema que será tratado, analisando sua viabilidade e quais serão os recursos fornecidos .

- **Definição dos Objetivos:**

  Definição de quais são as metas que se deseja alcançar com o projeto, prazos a serem alcançados e qual o resultado final esperado.

- **Identificação dos Dados:**

  Serão analisados os dados disponíveis e estabelecidas as fontes de onde serão obtidos esses dados, considerando quais tipos de dados serão possivelmente necessários.

2.3 Data Engineering

- **Coleta de Dados:**

  Nesta fase os dados necessários serão coletados, armazenados e organizados , podendo ser preciso a busca por novas fontes caso encontradas lacunas nos dados voltando assim para o estágio anterior.

- **Limpeza dos Dados:**

  Serão corrigidas falhas nos dados como dados inválidos, tratadas as necessidades específicas para leitura de um determinado dataset e realizada a padronização dos dados.

- **Preparação dos Dados:**

  Com os datasets já obtidos, é necessário preparar e definir quais e como os dados serão utilizados, sendo realizadas combinações, filtragem por valores específicos, gerando assim os datasets específicos com todas as informações necessárias para construção do modelo.

  É importante ressaltar que nessa abordagem o tipo de modelagem somente será decidido nesta fase de acordo com o resultado final dos dados obtidos.

2.4 Modeling

- **Criação do Modelo**

  Serão desenvolvidos modelos preditivos ou descritivos com base na abordagem anteriormente descrita, utilizando-se de conceitos de aprendizado de máquina e redes neurais.

- **Validação do Modelo**

  É analisado se o modelo é aborda o problema e atende aos objetivos definidos de forma completa, sendo submetido a cenários de teste específicos com o objetivo de buscar falhas e apresentando seus resultados em forma de relatórios inteligíveis com gráficos e tabelas.

2.5 Deployment

- **Implamntação do Modelo**

  O modelo será submetido a um cenário de teste ou ambiente de produção ,tendo seus resultados analisados pelo patrocinadores podendo ter várias fases de teste, as quais a primeira em geral é apenas para analisar o desempenho.

- **Ánalise dos Resultados**

  É realizado o Feedback dos resultados baseados em como ele foi efetivo durante determinado tempo e se cumpre com o esperado. Nesta fase é também discutido possíveis ajustes, melhorias ou mesmo projetos complementares.

3 Conclusão

Enfim, nesta leitura foram abordadas as etapas que um cientista de dados realiza durante um projeto genérico, trazendo em contraste com outras abordagens uma definição de caráter mais individual em que o fluxo é definido em paralelo com a interpretação dos resultados obtidos em antítese por exemplo com a abordagem da IBM em que o tipo de modelagem que será utilizada já é discutido durante a análise inicial.

References

[1] https://www.bbvadata.com/accelerating-data-science-workflows/

[2] https://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science

[3] https://developer.ibm.com/articles/ba-intro-data-science-1/

[4] https://azure.github.io/learnAnalytics-UsingAzureMachineLearningforAIWorkloads/lab01-tdsp_and_aml/0_README.html