UNIVERSIDADE FEDERAL DO CEARÁ

Campus de Sobral

Curso de Engenharia da Computação

Tópicos Especiais em Automação e Controle I

Processo de Ciência de Dados

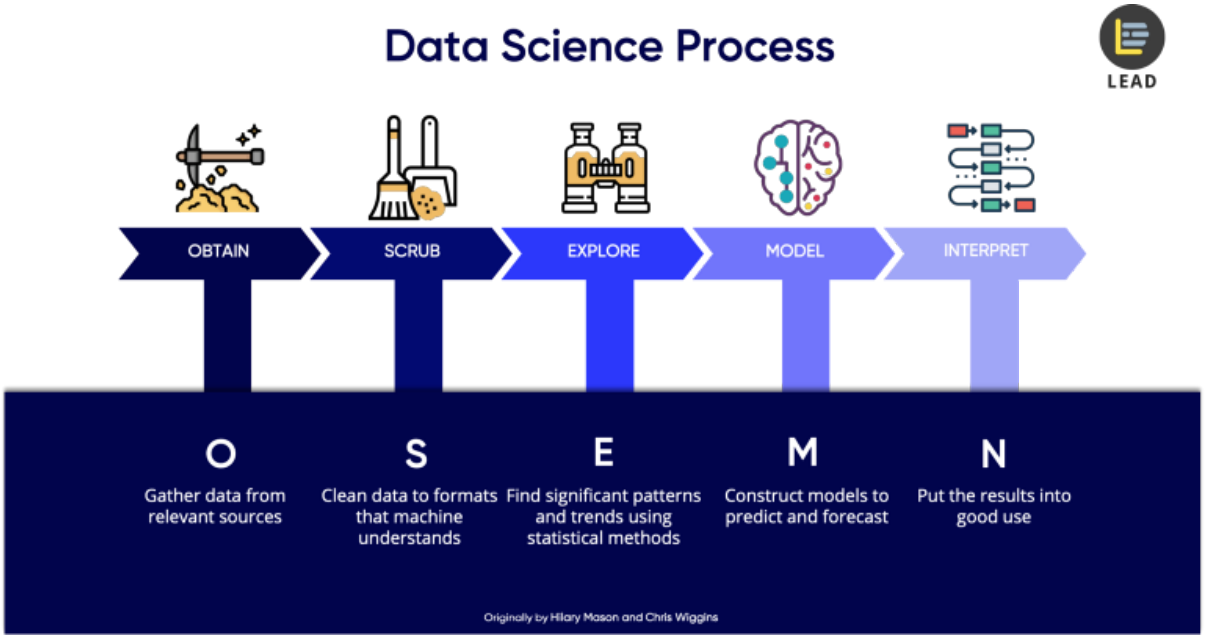
Uma metodologia para soluções de problemas ligados a ciência de dados pode ser definida a partir da aplicação do processo OSEMN. Este mesmo é definido por um conjunto de etapas recomendadas para desenvolvimento da solução em 5 (cinco) momentos bem específicos. A primeira etapa envolve obter os dados (***O**btain*). Os dados podem ser coletados praticamente de qualquer lugar, como redes sociais, exames médicos, sensores, APIs, datasets públicos e privados, etc. A maioria das bases coletadas apresentam falhas, como dados faltantes, por exemplo. Para realizar o tratamento desses dados é aplicada a segunda etapa do processo OSEMN, definido por limpeza (***S**crub*), que atuará na remoção ou substituição dos dados desnecessários. Na terceira etapa, relacionada à exploração (***E**xplore*), a propriedade dos dados é verificada. Em uma base de dados há diferentes tipos de dados, como numérico, categóricos, datas, etc. Para cada um desses dados faz-se necessário realizar um tratamento diferente, seja para extração de novos dados ou para conversão. O quarto passo associa-se à modelagem (**M**odel), em que os algoritmos de aprendizado de máquina serão utilizados para realizar classificação ou regressão sobre os dados. Este passo é completamente dependente da etapa anterior, o que reforça que uma boa análise exploratória dos dados influi diretamente nas predições do modelo. Após o uso do modelo e assim alcançar o resultado de suas predições, faz-se necessário interpretar os dados alcançados. Esta é a última etapa, que se trata da interpretação (*i**N**terpret*). Este passo se mostra relevante para dar significado ao que o modelo apresentou como saída, o que aquela predição representa e como ela pode ser aplicada. Esse tipo de inferência pode ser apresentada de forma gráfica, permitindo um melhor entendimento por parte do público-alvo da solução.

Dataset a ser analisado: Dados abertos sobre Acidentes da Polícia Rodoviária Federal
Disponível em: https://portal.prf.gov.br/dados-abertos-acidentes (oficial) ou https://bit.ly/2FvcONR (Google Drive para quando a base oficial estiver indisponível).

Entregas relacionadas ao processo OSEMN:

1) Etapa Obter - não se aplica porque dataset já foi disponibilizado
2) Etapa Limpeza e Exploração - apresentar notebook no Google Colaboratory (Colab): **28/09/2020**
3) Etapa Modelagem - apresentar notebook no Google Colab: **05/10/2020**
4) Etapa Interpretação - apresentar DataApp no Google Meet: **13/10/2020**

Figura - Etapas do Processo OSEMN. Fonte: https://bit.ly/2RD216s



O que se busca em cada etapa?

1) Etapa Limpeza e Exploração - avaliar se ainda há demanda de limpeza e realizar uma exploração estatística para observar que hipóteses podem ser lançadas sobre estes dados;
2) Etapa Modelagem - desenvolver algumas técnicas de regressão de dados para avaliar qual o melhor modelo de análise dos mesmos;
3) Etapa Interpretação - apresentar em um DataApp uma visualização dos dados trabalhados com as conclusões obtidas a partir das hipóteses lançadas inicialmente.

Formação das Equipes
(divulgação e acompanhamento no servidor **API Publico** no **Discord**):

Grupos de até 6 (seis) membros que devem ser informados ao professor, pelo e-mail: ialis@sobral.ufc.br. A primeira equipe que me notificar será a Equipe 01 e assumirá os canais de texto e voz #equipe-1, e assim por diante. A apresentação final deve durar até 10 (dez) minutos.

Os envios das entregas podem ser feitos no SIGAA ou Google Classroom. As equipes que terminarem antes, podem agendar apresentação antecipadamente pelo mesmo e-mail do professor. Por questão do momento de pandemia e potenciais dificuldades de acesso à internet, receberei trabalhos e apresentações das equipes (que me notifiquem sobre esse problema) até o dia 31/10/2020.