UNIVERSIDADE FEDERAL DO CEARÁ

Campus de Sobral

Curso de Engenharia da Computação

Tópicos Especiais em Computação 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 são 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.

Datasets a ser analisado:

A) Dados de quantidade de pratos servidos ao longo dos anos no Restaurante Universitário da UFC - Campus de Sobral para análise de predição (séries temporais e/ou regressão). Disponível no SIGAA.

B) Dados de clientes de telefonia com possibilidade de desistência - dataset Telco Customer Churn - para classificação. Disponível no SIGAA.

Entrega relacionadas ao trabalho:

1) Notebooks com a descrição do processso OSEMN para cada tipo de problema;
2) Apresentar DataApps em uma apresentação de até 15 (quinze) minutos em sala de aula;
3) Prazo de Entrega: 13/07/2023, via SIGAA. Apresentação em sala: **12/07/2023**.

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, séries temporais e classificação nos dados para avaliar qual o melhor modelo de análise dos mesmos;
3) Etapa Interpretação - apresentar em um DataApp com uma visualização dos dados trabalhados em acordo com as conclusões obtidas a partir das hipóteses lançadas inicialmente.

Formação das Equipes

Sugestões de formações de equipes (até 9 membros, com liberdade de mudança das formações das mesmas):

Equipe 1: Abraão, Ailton, Akyla, Alanna, Alex, Ananda, André Luis, André Veras, Antonia Thamires

Equipe 2: Antonio Assis, Antonio Eraldo, Clara, Clézio, Daniel, Fca Janielly, Fco Anderson, Fco Emerson

Equipe 3: Fco Evandro, Galatas, Gideao, Guilherme, Igor, Inácio, Ismael, Israel, João Gabriel

Equipe 4: Jonatas, José Caio, José Darlyson, José Matheus, Klayver, Maxela, Pedro, Raquel, Roberto

Equipe 5: Robson, Salmo, Samuel, Samyle, Tailson, Thiago, Vitor, Vitoria, Yara