UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS SOBRAL

Curso de Engenharia da Computação

Tópicos Especiais em Computação I

Processo de Ciência de Dados

Uma metodologia para soluções de problemas ligados à 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 serem analisados:

1) Boletim de Acidente de Trânsito, disponibilizado pelo portal de Dados Abertos da Polícia Rodoviária Federal (PRF) para análise de predição (classificação e/ou regressão). Pode ser assumido o primeiro conjunto disponibilizado "Documento CSV de Acidentes 2025 (Agrupados por ocorrência)" ou do mesmo tipo de conjunto de algum ano anterior. Disponível em: https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf .
2) Expectativa de Vida - Life Expectancy (WHO). Disponibilizado na plataforma Kaggle, que aborda sobre expectativa de vida em vários países entre 2000 e 2015. Ele inclui variáveis como taxas de vacinação, mortalidade, escolaridade, PIB e outros fatores

de saúde e sociais. A proposta é analisar como essas variáveis influenciam a expectativa de vida e aplicar modelos de regressão para prever esse valor. Disponível em: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?resource=download;

Entrega relacionadas ao trabalho:

1) Notebook(s) com a descrição do processo OSEMN para os problemas analisados;
2) Apresentar Data Apps (um para regressão e outro para classificação) em uma apresentação de até 10 (quinze) minutos em sala de aula;
3) Prazo de Entrega: 02/08/2025, via SIGAA. Apresentação em sala: **30/07/2025**.

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 e classificação nos dados para avaliar qual o melhor modelo de análise dos mesmos. A equipe deve explorar o dataset da PRF para classificação, e escolher entre qual dos dois mencionados pode desenvolver um modelo de regressão;
3) Etapa Interpretação - apresentar os Data Apps 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 8 membros, com liberdade de mudança das formações das mesmas). As descrições dos membros das equipes devem ser realizadas neste link: https://bit.ly/grupoTopicos25-1.