[2021.1] Ciência de Dados

Projeto de Ciência de Dados de Ponta a Ponta

Trabalho Final

VISÃO GERAL

Ao longo da presente disciplina, você aprendeu a utilizar diversas ferramentas e algoritmos comuns no dia-a-dia de um cientista de dados. Agora, você deverá colocar em prática os conhecimentos aprendidos em um projeto de Ciência de Dados inédito. Seu projeto deve contemplar todos os passos que um projeto real deve ter, como obtenção de dados, análise de informações e treinamento de modelos.

OBJETIVOS

1. Desenvolver um projeto inédito de Ciência de Dados o mais próximo possível de um projeto real, a fim de demonstrar os conhecimentos adquiridos ao longo da disciplina.
2. Comparar os algoritmos escolhidos com conjuntos de dados reais utilizando métricas de avaliação vistas ou não na disciplina.

ESPECIFICAÇÕES

Neste trabalho, você deverá escolher um conjunto de dados para desenvolver sua solução.

Cada projeto pode ser desenvolvido por uma equipe de, no máximo, 5 componentes. Vale lembrar que equipes maiores têm correções mais rigorosas.

Os membros da equipe devem preencher a seguinte planilha com o nome completo de cada membro, matrícula, nível (graduação ou pós) e tema do trabalho até o dia 18 de Julho.

Link para a planilha: Equipes - Trabalho Final.

Ao longo do desenvolvimento do trabalho, você deve entregar alguns checkpoints:

1. **22 de Julho** - The Machien Learning Canvas (v0.4),
    a. Entregar cópia da apresentação no Google Slides preenchido com a sua proposta
2. **29 de Julho** - Análise exploratória dos dados

    a. Entregar um notebook no Google Colab
3. **18 de Agosto** - Entrega Final
    a. Entregar um notebook no Google Colab
4. **19 a 31 de Agosto** - Apresentações dos Trabalhos
5. **02 de Setembro** - Entrega da versão final corrigida

OBS: Scripts / Notebooks de pré-processamento, se forem muito extensos, podem ser entregues separados (no mesmo deadline, claro). Nesse caso, é necessário prover as instruções para execução do trabalho.

Seu projeto final deve contemplar os seguintes pontos:

1. Motivação do projeto,
2. Obtenção de dados,
3. Análise exploratória dos dados,
4. Transformações e limpeza dos dados,
5. Escolha dos modelos a serem utilizados, bem como métricas de avaliação,
6. Análise dos resultados.

Por fim, o projeto será apresentado em aula.

AVALIAÇÃO

Como não sabemos o dataset que será escolhido, decidimos não definir um número mínimo de análises a serem feitas. Porém, utilizaremos os seguintes critérios:

- **Organização do notebook** - seu relatório será todo feito usando o Jupyter Notebook (o qual será o artefato de envio do trabalho). Logo, ele precisa está bem organizado, combinando código python com markdown;
- **Organização dos experimentos** - não basta treinar os modelos e apresentar uma métrica qualquer. Procure métricas que possam ser mais adequadas para seu problema. Além disso, utilize recursos para deixar sua experimentação mais consistente, como *grid search*, *cross validation*, etc.
- **Reprodutividade do relatório** - seu notebook será executado no momento da avaliação. Garanta que as células estejam na ordem correta e que qualquer instrução adicional para a reprodutibilidade esteja devidamente apresentada no notebook;
- **Tamanho da equipe** - trabalhos feitos por equipes possuem avaliações mais rígidas, sendo a dificuldade proporcional (não necessariamente de forma linear) ao número de membros.

LINKS IMPORTANTES

1. Kaggle
2. Dados do twitter a partir da API
3. Portal Brasileiro de Dados Abertos
4. UCI Machine Learning Repository
5. Google Dataset Search
6. Dados Abertos do Governo Americano
7. Google Acadêmico
8. Sidra IBGE
9. KD Nuggets
10. IBM Data Asset eXchange