Instituto Federal de Educação, Ciência e Tecnologia

Câmpus Campinas
INTL6 – Inteligência Artificial

Professor: Samuel Martins (samuel.martins@ifsp.edu.br)

Trabalho Individual 1

Prazo: 21/12/2020 às 23:59

1) Especificação

Neste primeiro trabalho individual, cada aluna(o) deverá realizar uma **análise exploratória** de um conjunto de dados *(dataset)* escolhido. Cada a(o) aluna(o) deverá descrever e motivar o problema a ser explorado (por que o problema é relevante?), detalhar as características da base de dados, elencar um conjunto de perguntas/hipóteses sobre o problema e suas respectivas respostas a serem confirmadas/achadas pela análise do dataset. As respostas podem envolver a apresentação de medidas descritivas, visualizações, etc.

Cada a(o) aluna(o) deverá preparar um Jupyter Notebook com todos os trechos requeridos (explicações em texto, códigos e visualização).

2) Critérios de Avaliação

Seguem os critérios a serem avaliados. Cada critério tem um conjunto de pontos que servirão como um guia para seu desenvolvimento. Outros pontos não mencionados aqui também podem ser considerados.

- Descrição e motivação do problema: **[1.5 pontos]**
  - Qual o problema que você trabalhará?
  - Qual sua motivação?
  - Por que ele é relevante?
  - Quais são os pontos em aberto a serem explorados/analisados?
- Descrição da base de dados escolhida: **[1.5 pontos]**
  - Onde você a coletou?
  - O que significa cada linha?
  - Quais são os principais atributos (colunas) e seus tipos?
  - Quão complexa é o dataset?
- Preparação da base de dados: **[1 ponto]**
  - Você teve dificuldades para preparar o dataset para uso (p. ex., nome ruins para as colunas, arquivos dos datasets não possuíam boa organização, etc)?
  - Você aplicou algum pré-processamento nos dados?
    - Você removeu outliers/ruídos no dataset? Se sim, como e quais?
    - Alguns exemplos (linhas) possuíam dados faltantes para alguns atributos (colunas)? Se sim, como você resolveu este problema?
  - Caso esta etapa não tenha sido aplicada, sua pontuação será usada na análise exploratória;
- Análise exploratória: **[5 pontos]**

    - Elaboração das perguntas/hipóteses para a análise exploratória;
        - Mínimo de 8 perguntas/hipóteses;
        - O que você quer explorar sobre o problema usando tal dataset?
        - Quais são perguntantes relevantes para o problema?
    - Respostas às perguntas/hipóteses:
        - Algumas possibilidades:
            - Utilização de filtragem das tabelas;
            - Medidas descritivas;
            - Gráficos (https://datavizcatalogue.com)
    - Conclusões
- Relatório (Jupyter Notebook): **[1 ponto]**
    - Organização do relatório;
    - Clareza na apresentação dos textos e códigos;
    - Qualidade do código;
- Uso de algum modelo/algoritmo de Machine Learning para a predição de alguma etapa/ponto do problema **[1 ponto extra]**;

Você poderá enviar um jupyter notebook (.ipynb) ou o link do repositório online com o código (ex., github) ou o link do Google Colab ou um .zip com todos os arquivos necessários. No caso dos links para repositórios ou plataformas online, serão considerados apenas aqueles com atualização até o prazo de entrega desta atividade.

A submissão deverá ser realizada na atividade correspondente no Moodle, até o dia **21/12/20 às 23:59.**

3) Datasets e Sugestão de Temas

Datasets:

- Kaggle: https://www.kaggle.com/data
- UCI: https://archive.ics.uci.edu/ml/datasets.php
- Dados Abertos Brasileiros: https://dados.gov.br/dataset

Sugestão de temas:

- Dados relacionados a doenças (COVID-19, câncer, hemofilia, ...)
- Estudo sobre os gastos públicos em:
    - Campanhas eleitorais;
    - Saúde;
    - Educação;
    - etc