Projeto prático em equipe

Construção e análise de conjunto de dados

1. Descrição geral:

A equipe deverá construir um conjunto de dados e analisar esse conjunto de dados, levantando hipóteses, levantando regras de associação e indicando anomalias.

O conjunto de dados poderá ser construído por meio de:

- mesclagem de três conjuntos de dados vindos da plataforma de conjuntos de dados Kaggle;
- dados de uso cotidiano de uma empresa que estejam armazenados em múltiplos arquivos ou em um banco de dados sql;
- Web Scraping.

A equipe deve ter no mínimo 4 e no máximo 5 integrantes. Este projeto só pode ser entregue em equipe. Todos os membros da equipe serão entrevistados individualmente e separadamente sobre os resultados apresentados pela equipe.

A equipe apresentará seus resultados com um relatório detalhado explicando e justificando suas decisões.

Esse projeto vale 10 pontos.

2. Requisitos do projeto

1. O projeto deve ser desenvolvido na plataforma Google Colab. **O não atendimento a esse requisitos implica em nota 0.**
2. Caso a equipe escolha construir o conjunto de dados a partir de dados do Kaggle ou via Web Scrapping, ela deverá escolher entre um dos temas abaixo para a construção do conjunto de dados:
    a. Filmes e afins.
    b. Emprego.
    c. E-commerce.

d. Esporte.
e. Imóveis.
f. Outro tema a ser combinado com os professores.

3. Se os dados vierem do Kaggle, a equipe deverá escolher e mesclar ao menos 3 conjuntos de dados diferentes dessa plataforma. Os conjuntos mesclados devem ser do mesmo tema. Um novo atributo deve permitir identificar de que conjunto de dados original uma instância veio.
4. O conjunto de dados construído deve ter ao menos 8 atributos diferentes.
5. O conjunto de dados construído deve ter pelo menos 2000 amostras.
6. A equipe deverá apresentar o conjunto de dados construído em um arquivo CSV atualizado a até pelo menos 7 dias antes da entrega.
7. A equipe deverá apresentar o arquivo Notebook Colab que constrói o conjunto de dados. Caso tenha sido construído a partir de arquivos de uma empresa, a equipe deverá apresentar também esses arquivos.
8. Os nomes de variáveis, funções, métodos, classes e qualquer outra estrutura definida pela equipe no código devem ser autodescritivas, não permitindo dúvidas quanto a sua finalidade.
9. Todas as etapas de análise e desenvolvimento devem ser descritas usando estruturas de texto do Notebook Colab. As descrições devem atender à norma culta da língua, além de ser concisas, claras e corretas.
10. Todos os membros do grupo deverão estar identificados nos arquivos enviados entregues, sob pena do membro ter sua nota zerada por não ter sua participação confirmada pela equipe. A constituição da equipe tem que estar no início do arquivo.
11. Seguir as seguintes práticas recomendadas para visualização de dados, conforme livro “Practical Data Science with Python”:
    a. Evitar lixo gráfico.
    b. Usar as cores com sensatez
    c. Usar os gráficos corretos conforme o tipo de análise de dados. Por exemplo:
        i. Gráficos de barras - para gráficos categóricos
        ii. Histogramas – para distribuição de valores contínuos
        iii. Gráficos de linha - para séries temporais
        iv. Gráficos de dispersão - para relações entre duas variáveis contínuas
        v. Heatmaps – para relações entre duas variáveis contínuas e correlações
    d. Rotular claramente os eixos e conjuntos de dados e usar um único tamanho de fonte com uma fonte sem serifa
    e. Adaptar as visualizações ao público

12. A equipe deverá entregar um Notebook Colab com a análise do conjunto de dados, apresentando:
    a. Descrição dos procedimentos de coleta de dados e significado dos atributos
    b. Discussão de limitações da coleta de dados
    c. Hipóteses sobre como os diversos atributos devem se comportar no mundo real (ex. o valor de vendas por produto deve apresentar distribuição enviesada para valores menores)
    d. Tratamento de dados ausentes e duplicados.
    e. Análise numérica de estatísticas descritivas
        i. Média, moda, mediana, desvio padrão, intervalo interquartil etc.
    f. Visualização e análise de estatísticas descritivas
        i. Gráficos de histogramas, Box Plots, Scatter Plots, Violin Plots etc.
    g. Indicações de anomalias.
    h. Análise de tendências (quando aplicável)
    i. Análise de Correlações

3. Entregas e prazos:

- Definição das equipes: até 10/11/2022
- P1 – Entrega de resultados preliminares – 20%
    - Apresentação: 10/11/2022
    - Deve apresentar o conjunto de dados construído.
    - A equipe deve apresentará o projeto para o professor atendendo aos requisitos associados aos seguintes itens:
        - Descrição dos procedimentos de coleta de dados e significado dos atributos
        - Discussão de limitações da coleta de dados
        - Hipóteses sobre como os diversos atributos devem ser comportar no mundo real (ex. o valor de vendas por produto deve apresentar distribuição enviesada para valores menores)
        - Tratamento de dados ausentes e duplicados.
        - Análise numérica de estatísticas descritivas
- P2 – Entrega final – 80%
    - Entrega final: 23/11/2022
        - Deve atender a todos os requisitos.

        - Deve ser enviado:
            - Arquivo CSV com o conjunto de dados atualizado até pelo menos 7 dias antes da entrega.
            - Arquivo Notebook Colab que constrói o conjunto de dados.
            - Arquivo Notebook Colab com a análise do conjunto de dados, levantando hipóteses, levantando regras de associação e indicando anomalias.
        - Os arquivos Notebooks devem ter explicações explicitas para as decisões tomadas no desenvolvimento do código.
    - Apresentação: 24/11/2022 a 01/12/2022
        - A data exata de cada equipe será definida pelos professores. Todas equipes devem estar preparadas para apresentar no dia 24/11/2022.
        - A equipe apresentará o projeto para o professor, demonstrando o atendimento aos requisitos e justificando as decisões tomadas.
        - O professor poderá fazer perguntas e solicitar alterações do código a cada membro da equipe após a apresentação, sendo esse um dos critérios a ser utilizado pelo professor para pontuar a equipe. Essa entrevista individual define a nota individual até o máximo da nota da equipe. Caso o aluno não tenha bom desempenho, a nota individual poderá ser a nota da equipe com multa de 50%. O aluno que não comparecer terá nota 0.
        - Todos os membros do projeto devem ter conhecimento total sobre todos os aspectos dele, não importando que parte específica ficou responsável individualmente.

4. Observações Gerais

Entregas deverão ser feitas pelo Ulife a não ser que seja informado outro destino.

Durante as aulas, no momento das práticas, as equipes podem e devem fazer consultas ao professor sobre o projeto.