Projeto da disciplina de Programação para Ciência de Dados

A atividade deverá ser realizada em grupos de até no máximo 5 pessoas. O projeto consiste em realizar um projeto que passe por quatro etapas da ciência de dados, sendo elas:

- Coleta de Dados (3 pontos);
- Tratamento de Dados (3 pontos);
- Aplicação de um algoritmo de aprendizado de máquina (2 pontos);
- Visualização ou interpretação do resultado (2 pontos).

Coleta de Dados

O grupo deverá buscar uma base de dados em um repositório na internet, gerando um arquivo CSV a partir do uso de um web scrapping ou simplesmente realizando uma transformação simples. Alguns exemplos de repositórios:

- https://archive.ics.uci.edu/ml/index.php
- http://dados.gov.br/
- http://dados.prefeitura.sp.gov.br
- https://www.kaggle.com/datasets
- https://www.v7labs.com/blog/best-free-datasets-for-machine-learning

Tratamento de Dados

Os dados deverão ser mantidos na biblioteca Pandas e, se aplicável, utilizar a biblioteca Numpy para alguma transformação. O grupo deverá utilizar técnicas de tratamento de dados, como remoção de duplicatas, inserção ou exclusão de valores ausentes, transformação de escala de atributos, entre outras, de acordo com a necessidade do algoritmo escolhido ou do algoritmo selecionado.

Aplicação de um algoritmo de Aprendizado de Máquina

Dependendo do tipo de exploração desejada ou questão a ser respondida usando os dados, o grupo deverá escolher um algoritmo classificador, de regressão ou de agrupamento.

Alguns exemplos de algoritmos de classificação e regressão são:

- Naïve Bayes; (probabilístico)
- C4.5; (Arvore de decisão)
- Regressão Linear Multipla;
- Regressão logística;

- Support Vector Machines;
- K-Vizinhos Mais Próximos; (Lazy Learning)
- Multilayer Perceptron (Rede Neural Multicamada)

Alguns exemplos de algoritmos de agrupamento são:

- K-Means (função de custo);
- DBscan (densidade);
- AGNES (hierárquico);

Visualização ou interpretação do resultado

A partir do resultado obtido nos dados, o grupo deverá realizar uma análise desses resultados, podendo utilizar visualizações com bibliotecas, como a matplotlib, ou de forma textual. Se o algoritmo é de agrupamento, é ideal aplicar alguma técnica de índice de validação, como o Silhouette. Se o algoritmo for de classificação ou regressão é possível utilizar métricas como acurácia ou estatística como o F1-score; além disso, para avaliar um algoritmo de classificação é desejável utilizar o método de validação cruzada.

Entrega final: 15/12