

ECOM08A – Inteligência Artificial - Prof. João Paulo R. R. Leite

Leia atentamente as instruções abaixo:

- **Grupos:** Este trabalho deve ser feito **individualmente ou em dupla**.
 - **Coloque os nomes dos integrantes no início dos dois notebooks.**
- **Prazo para entrega:** até o dia **01/12/2024**, às 23:59h.
- **Nota:** **60% da N2**.
- **Formato de arquivo:** **PDF**. Devem ser entregues dois arquivos **em formato PDF** (regressão e classificação). Não serão aceitos outros formatos. Os arquivos devem conter os códigos, textos explicativos, gráficos e **resultados obtidos**.

O que deve ser feito?

Escolha dois conjuntos de dados de fontes públicas (como o Kaggle, UCI *Machine Learning Repository*, Physionet ou qualquer repositório de dados): um para regressão e outro para classificação. Escolha dados sobre assuntos do interesse dos integrantes do grupo.

A seguir, faça dois notebooks (Jupyter):

- um para classificação: use **regressão logística, árvore de decisão, SVM e MLP**.
- outro para regressão: use **regressão linear, árvore de decisão, SVM e MLP**.

Siga o passo a passo básico ensinado em sala de aula:

1. Carregue os dados.
2. Limpe e pré-processe os dados. Normalize se necessário, remova outliers e dados faltantes, filtre se achar importante.
3. Realize uma análise exploratória dos dados, encontrando relações interessantes entre os dados de entrada e saída, e mostrando visualmente no notebook.
4. Divida os dados em conjuntos de treinamento e teste.
5. Treine e teste cada modelo, avaliando a seguir, sua capacidade de generalização a partir de métricas adequadas para cada tipo de problema (MAE, MAPE, RMSE, MSE e R^2 para regressão; Acurácia, Precisão, Recall, F1-score e matriz de confusão para classificação).
6. Compare (textualmente e/ou visualmente) os resultados obtidos, e faça uma análise sobre os (possíveis) motivos que levaram determinado modelo a ser melhor do que outros no teste. Discuta a complexidade e a capacidade de generalização de cada modelo, além de suas vantagens e desvantagens para o conjunto de dados escolhido.

Requisitos:

- Cada célula de código ou gráfico deve ter uma célula correspondente de texto que explique o que se passa na célula seguinte/anterior. Analise o que está sendo feito!
- É necessário que cada notebook contenha no mínimo dois gráficos (um para análise dos dados e outro para análise de resultados).
- Desempenho mínimo: O melhor modelo avaliado deve ter um R-quadrado de no mínimo 0.75 na regressão e acurácia de no mínimo 75% na classificação. Os demais modelos podem ter desempenhos inferiores. Caso não consiga para a base de dados escolhida, é necessário buscar outra base!

Bom trabalho!