Capacitação IA e IoT
Softex, MCTI
Prof.: Thiago Moura

Projeto Final – Ciclo 2

O objetivo do projeto é consolidar o conhecimento adquirido durante todo o ciclo 2. Aplicar os modelos de aprendizado de máquina aprendidos, configurando seus parâmetros, comparando os resultados obtidos na predição de bases de classificação ou regressão. Validar o resultado dos modelos através da validação cruzada, modelos de ensemble, seleção de *features* e gerenciar todo o ciclo de vida dos modelos através do MLflow.

1) Obter um *dataset* de classificação ou de regressão, diferentes dos já apresentados durante a capacitação (ciclo I e II). Os *datasets* podem ser obtidos dos seguintes locais:

- *Datasets* do SkLearn: https://scikit-learn.org/stable/datasets/toy_dataset.html
- *Datasets* do Seaborn: https://github.com/mwaskom/seaborn-data
- *Datasets* do UCI: https://archive.ics.uci.edu/datasets
- *Datasets* do Kaggle: https://www.kaggle.com/datasets
- Qualquer fonte de dados da internet, empresa que trabalha etc.

ATENÇÃO: O ideal seria o aluno trazer algum problema real com o qual tem contato na faculdade, empresa, ou algum problema que acha interessante analisar.

2) Analisar o *dataset* usando os métodos do Pandas. Comparar as *features* e verificar se há valores nulos, correspondência entre *features*, converter os *labels* para números e *features* categóricas usar *One-Hot Encoder*. Normalizar as *features* na escala 0..1.

3) Rodar experimentos e guardar as execuções no MLflow. Criar uma bateria de experimentos chamada: "exp_projeto_ciclo_2"

- 3.1) Escolher 3 algoritmos de aprendizado de máquina diferentes com 3 variações de parâmetros para cada um deles. Um total de 9 variações;
- 3.2) Testar também o Bagging e RandomForest com parâmetros padrão ou alguma parametrização customizada (caso queira);
- 3.3) Executar experimentos com o GradientBoosting, XGBoost e LightGBM, com parâmetros padrão ou alguma parametrização customizada (caso queira);
- ~~3.4) Testar todos os métodos de seleção dinâmica (OLA, LCA, KNORA-U, KNORA-E e MCB) com os modelos gerados com o Bagging e também com os modelos gerados com o RandomForest;~~ **(métodos de seleção dinâmica serão um *plus* para quem fizer e não será mais obrigatório!)**
- 3.5) Para todos os experimentos o cálculo das métricas (Acurácia, Precision, Recall, Specificity, AUC (***classificação***) ou MSE, RMSE,

MAPE (***regressão***)) deve ser a média dos 10 *folds* da validação cruzada;

- 3.6) Utilizar 1 método de seleção de *features* para todos os experimentos (**usar só um método de seleção de features**);
- 3.7) Guardar todas as execuções no Tracking do MLflow;
- 3.8) Guardar todos os parâmetros e métricas dos modelos usando o ~~`autolog()`~~ no MLflow (**não usar o autolog(), guardar tudo explicitamente**);
- 3.9) Registrar os 3 (três) modelos que tiveram o melhor desempenho (baseado na acurácia);
- 3.10) Escrever um código Cliente para carregar os modelos registrados e exibir as informações desses modelos com a sua descrição.

DATA DE ENTREGA: 06/06/2024

HORA: até 20:00h

Observação: PROJETO INDIVIDUAL!