Desenvolvimento de Software Multiplataforma
Aprendizagem de Máquina – 2024-1
2ª Atividade em grupo

1. Especificação

Nesta atividade, cada grupo, **de no máximo 3 integrantes**, deverá aplicar os conceitos de **regressão** ou **clusterização** em algum dataset, a escolha do grupo.

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 sucinta do problema e da base de dados: **[0.5 ponto]**
    - Qual o problema a ser resolvido?
    - O que significa cada instância do dataset?
    - Quais são os principais atributos e seus tipos?
- Análise Exploratória de Dados: **[1.0 pontos]**
    - Como cada variável se distribui?
    - Correlação de variáveis;
    - Gráficos que gerem insights para o tratamento de dados e/ou treinamento dos modelos:
        - P. ex: detecção de ruídos via scatter plot;
    - Discussão dos principais achados da análise exploratória de dados;
- Limpeza e preparação da base de dados: **[1.5 pontos]**
    - Remoção de duplicidade e/ou outliers;
    - Preenchimento de dados faltantes;
    - Feature scaling;
    - Class imbalance; etc
    - Discussão sucinta sobre a razão de cada etapa de limpeza e pré-processamento considerada;
- Treinamento e Validação de modelos: [6 pontos]
    - Comparar ao menos 3 algoritmos de classificação diferentes;
    - Cross-Validation;
    - Métricas consideradas para o problema;
    - Discussão dos resultados;
        - Há overfitting ou underfitting? Etc.
    - Fine-tunning
    - Avaliação no conjunto de teste:
        - Avaliar os melhores modelos no conjunto de teste;
        - **Discussão dos resultados.**
    - Trabalhos Futuros:
    - Discussão sobre estratégias/ideias/sugestões para a melhoria dos modelos;
- Relatório (Notebook): [1 ponto]
    - Organização do relatório;
    - Clareza na apresentação dos textos e códigos;
    - Qualidade do código;
- Atividades opcionais: [até 1 ponto extra]
    - Uso de técnicas não vistas em sala;

- Abordagem de negócios:
- Motivação e descrição mais detalhada sobre o problema, com enfoque na resolução de problemas de negócio;
- Definição de um baseline;
- Comparação dos resultados com o baseline;
- Conversão dos resultados (medidas técnicas) em medidas/performance de negócio:
    - P. ex, o que os 10% a mais de acurácia de seu modelo, frente ao baseline, impactaram no negócio da empresa?

2. Entregáveis

Cada grupo deverá preparar um **único jupyter notebook** com os códigos feitos para a resolução dos problemas, bem como comentários e discussões sobre os mesmos.

3. Submissão (prazo final: 18/06/24)

- A submissão da atividade será feita em tarefa específica no Teams da disciplina.
- O grupo poderá enviar um jupyter notebook (.ipynb) ou o link do repositório online com o código (ex., Google Colab, GitHub, Kaggle).
- ◦ No caso dos links para repositórios ou plataformas online, serão considerados apenas aqueles com atualização até o prazo de entrega desta atividade.
- Apenas **um membro do grupo** deverá submeter a atividade.