

Atividade em Grupo

1. Especificação

Nesta atividade, cada grupo (máx. 3 integrantes) deverá aplicar os conceitos de regressão para o projeto de estudos abaixo: São Paulo Real State.

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: **[1 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: **[2 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: **[2 pontos]**
 - Exemplos:
 - ◆ Remoção de duplicidade e/ou outliers;
 - ◆ Preenchimento de dados faltantes;
 - ◆ Feature scaling;
 - ◆ etc
 - Discussão sucinta sobre a razão de cada etapa de limpeza e pré-processamento considerada;
- Treinamento e Validação de modelos de regressão: **[4 pontos]**
 - Comparar ao menos 3 modelos diferentes;
 - ◆ Podem ser 3 modelos do mesmo algoritmo mas com hiperparâmetros diferentes;
 - ◆ Podem ser algoritmos diferentes (mais desejável);
 - Cross-Validation;
 - ◆ Métricas consideradas para o problema;
 - ◆ Discussão dos resultados;
 - ◆ Há overfitting ou underfitting?
 - 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]**
 - Utilização de **Pipelines** do Scikit-Learn
 - Fine-tuning
 - ◆ Aplicar alguma técnica para buscar os melhores hiperparâmetros para os modelos mais promissores da etapa anterior;
 - Uso de *ensemble methods*;
 - 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 dupla deverá preparar um **único jupyter notebook** com os códigos feitos e executados para a resolução dos problemas, bem como comentários e discussões sobre os mesmos.

Observações:

- **Não** há necessidade da criação de vários notebooks simulando os *Sprints* como feitos em sala. O grupo pode até seguir tal estratégia, mas para fins de organização interna.
- Apenas um **único** notebook final, com os principais achados, deverá ser entregue;

3. Submissão (prazo: 05/10/23)

- A submissão da atividade será feita em tarefa específica no Moodle 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.