Centro Universitário do Estado do Pará
Curso de Ciência de Computação
Disciplina: Modelagem Estatística
Professor: Pedro Girotto

PROJETO 2° BIMESTRE

INTRODUÇÃO

O projeto tem o objetivo de colocar em práticas as habilidades:(i) conduzir análise exploratória de dados com limpeza, tratamento de ausências e investigação das relações entre variáveis dependentes e independentes; (ii) implementar e comparar algoritmos de IA para regressão (linear, múltipla e polinomial) e classificação (Naive Bayes e Regressão Logística); (iii) avaliar desempenho com métricas apropriadas e (iv) otimizar modelos (validação cruzada e tuning) reportando ganhos e limitações.

Ferramentas obrigatórias (Python): pandas, seaborn, statsmodels, sklearn e pycaret.

METODOLOGIA

1. Procurar um dataset: escolher um conjunto de dados público e documentar a fonte e licença, descrevendo a variável-alvo (regressão ou classificação) e hipóteses de negócio;

2. EDA e preparação: inspecionar esquema, tipos e estatísticas descritivas, tratar valores ausentes e inconsistências, lidar com outliers e visualizar distribuições e relações (histogramas, boxplots, pairplots, heatmap de correlação);

3. Modelagem: implementar regressão linear simples, múltipla e polinomial (usando statsmodels para interpretação e sklearn/pycaret para pipelines), implementar Naive Bayes e Regressão Logística, e realizar a divisão treino/validação/teste com definição de baseline;

4. Avaliação do desempenho: para regressão, reportar MAE, RMSE e R^2 e conduzir diagnóstico de resíduos (normalidade, homocedasticidade, multicolinearidade/VIF); para classificação, reportar accuracy, precision, recall, F1, AUC-ROC e matriz de confusão;

5. Otimização: aplicar validação cruzada, usar pycaret para comparação e tune_model e, no sklearn, Grid/Random Search, registrando parâmetros, resultados e principais trade-offs;

6. Relatório: estruturar com Markdown e código as seções de introdução e objetivos, descrição dos dados e licença, EDA e testes, modelagem, avaliação, otimização, conclusões e próximos passos, além de referências.

ENTREGAS

Repositório no GitHub:

- Código, requirements.txt (ou environment.yml), dados (ou script de download) e README contendo descrição do projeto, instruções de instalação/execução e organização do repositório.
- Licença e citação da fonte de dados.

Relatório: Feito no próprio jupyter notebook, dataset escolhido, EDA e insights, modelos implementados e resultados iniciais, processo de otimização e novos resultados (comparativos com tabelas e gráficos). Inclua discussões sobre limitações e possíveis vieses.

AVALIAÇÃO

(20%) EDA & Insights — 0,6 pt

- **Insuficiente**: limpeza incompleta; gráficos irrelevantes; não trata ausências.
- **Regular**: tratamento básico; poucas visualizações; testes estatísticos ausentes ou mal aplicados.
- **Bom**: EDA consistente com testes (t/ANOVA/Qui-Quadrado, correlação) corretos; insights úteis.
- **Excelente:** EDA profunda, bem visualizada, com justificativas (CLT/suposições) e insights acionáveis.

(40%) Implementação de IAs & Resultados Iniciais — 1,2 pt

- **Insuficiente**: modelos faltantes ou incorretos; sem baseline.
- **Regular**: implementa parte dos modelos requeridos; avaliação limitada.
- **Bom**: todos os modelos requeridos implementados com métricas adequadas e comparação inicial.
- **Excelente**: pipelines reprodutíveis, boa interpretação (coeficientes/p-valores em statsmodels) e análise de erros.

(30%) Otimização & Novos Resultados — 0,9 pt

- **Insuficiente**: sem tuning nem validação cruzada.
- **Regular**: tuning superficial; ganhos pouco claros.
- **Bom**: tuning sistemático (pycaret/sklearn) com CV e melhoria comprovada.
- **Excelente**: busca bem planejada (hiperparâmetros, VIF/regularização), comparação robusta e discussão de trade-offs.

(10%) Qualidade do Notebook — 0,3 pt

- **Insuficiente**: desorganizado; sem Markdown ou referências
- **Regular**: organização parcial; pouca explicação.
- **Bom**: narrativa clara, seções completas, gráficos legíveis e reprodutibilidade básica.
- **Excelente**: comunicação excelente, código limpo, reprodutibilidade total e referências adequadas.

Observações: respeitar a ética no uso de dados; versionar o trabalho; documentar decisões. Recomendado incluir testes automatizados simples (ex.: verificação de schema) e seed aleatória para reprodutibilidade.