



MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PIAUÍ
CAMPUS SENADOR HELVÍDIO NUNES DE BARROS – CSHNB
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO
DISCIPLINA DE TÓPICOS EM COMPUTAÇÃO

Atividade Prática X

1. A base de dados Statlog (German Credit Data) é um conjunto utilizado em problemas de classificação binária, cujo objetivo é prever se um solicitante deve ser classificado como bom ou mau pagador. Essa base contém **1.000 instâncias** descritas por **20 atributos** numéricos e categóricos, relacionados a informações financeiras, socioeconômicas e comportamentais dos clientes. O *dataset* possui desbalanceamento moderado entre classes, presença de outliers, ruído e variáveis categóricas complexas. Nesse contexto, desenvolva um comparativo consistente de classificação utilizando todo o conhecimento adquirido na disciplina. Para realizar a tarefa, utilize os algoritmos *Random Forest*, *KNN*, *SVM* e *MLP*, seguindo os passos descritos.

Passos Necessários:

2. Dados:
 - Faça o download do arquivo da base de dados (<https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data>) e carregue os dados em Python.
3. Pré-processamento:
 - Corrija os problemas de *outliers*, dados faltantes e atenua os efeitos negativos das variáveis complexas;
 - Utilize alguma técnica que permite criar dados sintéticos (por exemplo, SMOTE!) e corrigir o problema de classes desbalanceadas.
4. Cenários de Classificação:
 - Divida os experimentos em dois cenários: com o uso de *PCA* (3 componentes), dados normais.
5. Sistemática Experimental
 - Os experimentos devem utilizar a validação cruzada ($k=5$) com a divisão de 80% para treino e 20% para teste.
 - Defina um espaço de hiperparâmetros para cada algoritmo classificador. Em seguida, associe as estratégias de otimização: *Grid Search*, *Random Search*, Bayesiana e Genética.
 - Calcule as métricas de acurácia, precisão, *recall*, F1, AUC ROC e *kappa*. Além disso, relate a média e desvio-padrão no CV, desempenho final no conjunto de teste com os melhores parâmetros obtidos e a técnica que obteve o melhor desempenho (descrever os principais parâmetros dos melhores resultados em cada cenário).
 - Analisar desempenho e custo em cada cenário com os otimizadores. Houve casos em que o algoritmo conseguiu um melhor resultado somente utilizando um otimizador específico?



MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PIAUÍ
CAMPUS SENADOR HELVÍDIO NUNES DE BARROS – CSHNB
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO
DISCIPLINA DE TÓPICOS EM COMPUTAÇÃO

- Elabore insights consistentes sobre os resultados, comparações e particularidades dos experimentos.

6. Relatório Final:

- Desenvolva um relatório (utilizando o template da SBC) contendo:
 - a) Introdução, desenvolvimento e conclusão.
 - b) Descrição detalhada das etapas e métodos aplicados.
- Submeta, via SIGAA, o relatório e o código-fonte com todo o processo descrito até o dia **02/11**.
- Elabore uma apresentação de, no máximo, 10 minutos. Os alunos deverão apresentar o relatório em sala no dia **04/11**.