**Desafio: Previsão de Inadimplência de Crédito com Python (AT1 – N2)**

**1. Objetivos**

* Exercitar técnicas de leitura e limpeza de dados (Pandas)
* Desenvolver análise exploratória (EDA) e visualizações (Matplotlib/Seaborn)
* Aplicar algoritmos de classificação (scikit-learn)
* Avaliar performance dos modelos (matriz de confusão, AUC, recall, precision)
* Criar uma apresentação interativa (notebook bem documentado ou app simples em Streamlit)

**2. Contextualização**

Você é analista de dados de uma fintech e recebeu um conjunto de dados de clientes que solicitaram empréstimo. A planilha **credit\_data.csv** contém variáveis como renda, idade, saldo bancário, histórico de crédito e informação se houve default (inadimplência). Seu desafio é construir um modelo capaz de prever quais clientes têm maior risco de não pagar o empréstimo.

**3. Requisitos Técnicos**

1. **Ambiente**
   * Python 3.8+
   * Colab ou Spider ou Jupyter Notebook
   * Bibliotecas mínimas: pandas, numpy, matplotlib (ou seaborn), scikit-learn, opcionalmente streamlit
2. **Carga de dados**
   * Leitura do CSV em DataFrame
   * Identificação e tratamento de valores faltantes/inconsistentes
3. **Análise Exploratória (EDA)**
   * Estatísticas descritivas
   * Gráficos de distribuição (histogramas, boxplots)
   * Análise de correlação
4. **Pré-processamento**
   * Tratamento de outliers (se necessário)
   * Codificação de variáveis categóricas
   * Escalonamento (StandardScaler ou MinMaxScaler)
5. **Modelagem**
   * Divisão em treino e teste (por exemplo, 70%/30%)
   * Treinamento de ao menos dois algoritmos de classificação (ex.: Regressão Logística, Random Forest, XGBoost)
   * Ajuste de hiperparâmetros (GridSearchCV ou RandomizedSearchCV)
6. **Avaliação**
   * Cálculo de métricas: acurácia, precisão, recall, F1-score e AUC-ROC
   * Exibição de matriz de confusão
7. **Entrega Interativa**
   * **Opção A**: Jupyter Notebook bem organizado, com seções narrativas (“markdown”) e resultados claros
   * **Opção B**: Pequena aplicação em **Streamlit** que permita ao usuário ajustar parâmetros do modelo e visualizar métricas em tempo real

**4. Etapas Sugeridas**

1. **Planejamento** (1–2 dias)
   * Revisão da documentação das bibliotecas
   * Exploração inicial da estrutura do CSV
2. **Implementação do EDA e pré-processamento** (2–3 dias)
3. **Desenvolvimento e comparação de modelos** (3–4 dias)
4. **Avaliação e refinamento** (1–2 dias)
5. **Documentação e apresentação** (1–2 dias)

**5. Entregáveis**

* Código-fonte (Notebook .ipynb ou pasta com scripts + requirements.txt)
* Relatório/resumo (PDF ou markdown) contendo:
  + Principais insights do EDA
  + Descrição dos modelos e métricas obtidas
  + Conclusões e recomendações de negócio
* (Se optar por Streamlit) Link ou instruções para rodar a aplicação

Você pode acessar bases de dados gratuitas de “crédito e inadimplência” em diversas fontes. Aqui vão três opções bastante utilizadas:

1. **Kaggle – Default of Credit Card Clients**  
   – é exatamente um CSV com variáveis financeiras e a coluna “default” (inadimplência).  
   – Acesse: <https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset>
2. **UCI Machine Learning Repository – Default of Credit Card Clients**  
   – mesma base do Kaggle, direto no repositório UCI.  
   – Acesse: <https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>
3. **UCI Machine Learning Repository – German Credit Data**  
   – uma outra base clássica para riscos de crédito, com 20 variáveis e alvo binário.  
   – Acesse: <https://archive.ics.uci.edu/ml/datasets/statlog+%28german+credit+data%29>