Relatório Técnico - Solução de Machine Learning

1. Enunciado do Problema

O objetivo deste projeto é desenvolver um modelo preditivo capaz de classificar tumores de mama como **Malignos** ou **Benignos** com base em características extraídas de imagens digitalizadas de aspirados por agulha fina (FNA) de massas mamárias. O conjunto de dados utilizado é o **Breast Cancer Wisconsin (Diagnostic) Dataset**.

Este é um problema clássico de classificação binária supervisionada, onde a precisão e a sensibilidade (recall) são cruciais para o diagnóstico médico.

2. Explicação da Estratégia

Para abordar este problema, adotamos a seguinte estratégia:

1. **Pré-processamento de Dados**:
    - Separação dos dados em conjuntos de treinamento e teste (80% treino, 20% teste) para simular o desempenho em dados não vistos.
    - Padronização das variáveis (StandardScaler) para garantir que todas as características (como raio, textura, perímetro) tenham a mesma escala, o que é fundamental para modelos lineares e baseados em distância.
2. **Pipeline de Machine Learning**:
    - Construção de um `Pipeline` que integra o pré-processamento e o modelo, evitando vazamento de dados (data leakage) durante a validação cruzada.
3. **Otimização de Hiperparâmetros**:
    - Utilização de `GridSearchCV` para testar sistematicamente diferentes combinações de hiperparâmetros.

3. Justificativa das Ferramentas e Modelos Escolhidos

- **Python e Scikit-Learn**: Escolhidos por serem o padrão da indústria para ciência de dados, oferecendo ferramentas robustas e eficientes.
- **Modelo: Regressão Logística**:
    - Foi escolhida por ser um modelo linear simples, interpretável e altamente eficaz para problemas de classificação binária.
    - Permite a aplicação direta de técnicas de regularização (L1 e L2) para controle de complexidade.
- **Métricas de Avaliação**: Acurácia, Precision, Recall e F1-Score foram escolhidas para fornecer uma visão completa do desempenho, especialmente importante em contextos médicos onde falsos negativos são perigosos.

4. Validação Cruzada e Regularização

Validação Cruzada

Para garantir que o modelo generalize bem e não sofra de *overfitting* (ajuste excessivo aos dados de treino), utilizamos a **Validação Cruzada K-Fold Estratificada** (com k=5). Isso divide os dados em 5 partes, treinando e validando o modelo 5 vezes em subconjuntos diferentes, garantindo que a proporção de classes seja mantida em cada dobra.

Regularização

A regularização é aplicada para penalizar modelos excessivamente complexos e evitar overfitting. No `GridSearchCV`, exploramos:

- **Tipos de Penalidade**:
    - `l1` (Lasso): Pode zerar coeficientes de características menos importantes (seleção de features).
    - `l2` (Ridge): Reduz a magnitude dos coeficientes, distribuindo o peso entre as características correlacionadas.
- **Parâmetro C**: O inverso da força de regularização. Testamos valores como `[0.01, 0.1, 1, 10, 100]` para encontrar o equilíbrio ideal entre viés e variância.