Relatório técnico

Tech Challenge Fase 1

Tema: Predição de Câncer de Mama.

Aluno: Charles Andrews William Kulkauski – RM368344.

1. **Introdução**

- Esse relatório tem como objetivo analisar o desenvolvimento do modelo de aprendizado de máquina capaz de prever o diagnóstico de câncer de mama, com base em variáveis numéricas derivadas de imagens de tumores.
- Os modelos utilizados foram Suport Vector Machine, Random Forest e K-Nearest Neighbors. Para preparação dos dados foi utilizado o StandardScaler para padronizar as variáveis numéricas e Seaborn e Matplotlib para os gráficos.

2. **Dataset**

O dataset utilizado foi Breast Câncer Dataset, que contém informações obtidas a partir de imagem de células mamárias representando diferentes características geométricas e estatísticas do tumor.

Target: Representada pela coluna/feature Diagnosis (M = Maligno, B = Benigno).

Tratamento: Leitura utilizando pandas, remoção de colunas vazias, conversão de variáveis categóricas.

3. **Análise Exploratória**

Nesta etapa observamos a distribuição das variáveis e identificamos possíveis correlações com o diagnóstico.

- Foi encontrada a necessidade de padronização dos dados antes da modelagem, para não afetar desempenho do modelo.
- O Heatmap de correlação mostra a alta correlação entre váriaveis como radius_mean area_mean, perimeter_mean.
- Os gráficos gerados possuem diferentes escalas e distribuições assimétricas.

4. **Preparação dos dados**

- Utilizado a padronização.
- Separação no train_test_split com 70% treino e 30% teste.
- Variáveis altamente correlacionadas foram analisadas para evitar sobreajuste.

5. **Modelagem**
   - Foram escolhidos modelos supervisionados de classificação com foco no F1-Score, na robustez contra overfitting, facilidade de implementação, minimização dos erros de classificação, sendo eles **KNN**, **SVM** e **Random Forest**.
   - Para otimizar parâmetros do modelo, foram utilizadas a validação cruzada e GridSearchCV,

6. **Avaliação do Modelo e Resultados**
   - As métricas para validar o modelo foram **Matrizes de confusão, Acurácia, F1-score, Recall e Precision**

   Para esta análise considerei principalmente a média do F1-Score como delimitador de qual melhor modelo a ser utilizado, pois, o F1-Score trás a média entre Precisão e Recall que é útil para este Dataset visto que ele tem dados bastante desbalanceados. No gera o modelo se mostrou inferior apenas na precisão média em comparação com os outros modelos, porém tendo o F1-Score, Recall e Acurácia mais altos.