

Relatório técnico

Tech Challenge Fase 1

Tema: Predição de Câncer de Mama.

Aluno: Charles Andrews William Kulkauskis – 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.
- No processo geral foram utilizadas bibliotecas para limpeza e conversão de features do Dataset. Os modelos utilizados para treinamento foram Support Vector Machine, Random Forest e K-Nearest Neighbors. Para a padronização dos dados foi utilizado o StandardScaler, 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).

3. Análise Exploratória

Nesta etapa observamos a distribuição das variáveis, suas características no Dataset e identificamos possíveis correlações com o diagnóstico.

- Observados colunas vazias e tipo da variável na coluna target como String.
- Foi encontrado no modelo valores muito esparsos nas features.
- Os gráficos gerados possuem diferentes escalas e distribuições assimétricas.
- O Heatmap de correlação mostra a alta correlação nas variáveis como radius_mean, area_mean, perimeter_mean.

4. Preparação dos dados

- Tratamento na target alteração do tipo da variável de string para integer através do mapeamento.

- Remoção de uma coluna vazia.
- Utilizada padronização nos dados esparsos através do StandardScaler.
- 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 principal na métrica F1-Score (explicação posteriormente), 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 exibidas para o resultado do modelo foram **Matrizes de confusão, Acurácia, F1-score, Recall e Precision**.
- Os modelos apresentam pouca diferença entre os resultados e estão bem treinados de acordo com os resultados do script.
- O algoritmo que se sobressai na análise é o Random Forest, que é a escolha principal de utilização do modelo no treinamento de precisão na análise de resultados de câncer de mama.

Para esta análise considere principalmente a média do F1-Score como delimitador final na escolha, que de acordo com os resultados é o melhor modelo a ser utilizado, pois o resultado do F1-Score trás a média entre Precisão e Recall que é útil para este Dataset visto que ele tem classes desbalanceadas entre. No geral 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.