Resumo

O objetivo do modelo selecionado, que é um modelo Random Forest de classificação, considerando a estrutura do conjunto de dados Breast Cancer Wisconsin, é solucionar a tarefa de diagnóstico de câncer de mama. O conjunto de dados contém informações clínicas e características de biópsias de tumores de mama, juntamente com o diagnóstico (maligno ou benigno).

O modelo pretende solucionar o problema de classificar corretamente os tumores de mama em duas categorias principais: maligno e benigno. Esse é um problema crucial na área da saúde, pois a identificação precisa do câncer de mama é fundamental para o tratamento e prognóstico dos pacientes.

O modelo Random Forest foi escolhido devido à sua capacidade de lidar com conjuntos de dados complexos e realizar classificações precisas. Ele pode aproveitar as características do conjunto de dados para fazer previsões confiáveis quanto à natureza do tumor, ajudando os profissionais de saúde a tomar decisões informadas.

Em resumo, o objetivo do modelo é fornecer um sistema de apoio ao diagnóstico que auxilie médicos e especialistas na identificação precisa do câncer de mama, contribuindo para uma detecção mais precoce e eficaz da doença. Isso pode ter um impacto significativo na qualidade de vida dos pacientes e nas taxas de sobrevivência.

Análise e Descrição das Métricas

No contexto do diagnóstico de câncer de mama, a ênfase está em minimizar os falsos negativos, uma vez que classificar erroneamente um câncer maligno como benigno pode ter graves consequências. Portanto, o recall é uma métrica crítica nesse cenário.

Em resumo, no diagnóstico de câncer de mama, as métricas de precisão, recall e F1-Score são particularmente importantes, uma vez que o foco principal é a detecção de casos malignos. Além disso, a análise da matriz de confusão permite uma visão detalhada do desempenho do modelo em termos de verdadeiros e falsos positivos e negativos. Essas métricas são essenciais para avaliar o quão bem o modelo está realizando a tarefa de classificação de tumores de mama.

Feature engineering

1. Carregamento do Conjunto de Dados:

O conjunto de dados Breast Cancer Wisconsin foi carregado a partir da biblioteca sklearn.datasets. Ele contém informações clínicas e características de biópsias de tumores de mama.

2. Criação de um DataFrame:

Os dados foram transformados em um DataFrame do Pandas, tornando-os mais acessíveis para análise e modelagem.

3. Verificação de Dados Faltantes:

Foi realizada uma verificação para garantir que não houvesse dados faltantes no conjunto de dados. Nenhum dado faltante foi encontrado.

4. Criação de Variáveis:

Uma nova variável, 'mean\_radius\_texture', foi criada no conjunto de dados. Essa nova variável é a soma das variáveis 'mean radius' e 'mean texture'. A criação de novas variáveis é uma técnica de feature engineering que pode fornecer informações adicionais aos modelos.

5. Exclusão de Variáveis:

Duas variáveis, 'worst smoothness' e 'worst symmetry', foram excluídas do conjunto de dados. Isso foi feito com o objetivo de simplificar o conjunto de dados, removendo variáveis que podem não ser relevantes para a tarefa de classificação.

6. Ajuste de Variável:

A variável 'mean area' foi ajustada elevando-a ao quadrado. Essa etapa de ajuste pode destacar relações específicas entre características ou ampliar o impacto de uma variável no modelo.

7. Normalização ou Padronização:

Os dados foram normalizados utilizando o StandardScaler da biblioteca sklearn.preprocessing. A normalização é importante para garantir que todas as variáveis tenham a mesma escala, evitando que algumas variáveis dominem outras nos modelos de machine learning.