IMPLEMENTAÇÃO DE INTELIGÊNCIA ARTIFICIAL UTILIZANDO FLORESTA DE ÁRVORES RANDÔMICAS NO DIAGNÓSTICO CÂNCER DE MAMA

Marcelo Pontes Rodrigues[1]; Marcelo de Oliveira Rego[2]
[1] Programa de Pós-Graduação em Tecnologias Aplicadas a Animais de Interesse Regional–UFPI;
[2]Programa de Doutorado em Cência da Computação Associação UFMA/UFPI/DCCMAPI/CCET
[1]marcelo.rodrigues@ufpi.edu.br; [2]oliveira@ufdpar.edu.br;

Resumo

A pesquisa sugeriu uma abordagem para o diagnóstico do câncer de mama, utilizando características nucleares que foram extraídas de imagens digitalizadas adquiridas por meio de aspiração com agulha fina (FNA, do inglês, *Fine Needle Aspirations*) e armazenada no *dataset Breast Cancer Wisconsin (Diagnostic)*. Os atributos são divididos em 10 características principais, com cada uma medida de três formas (média, erro padrão e pior valor). Com isso, resulta em 30 variáveis ao todo. Um dos principais métodos empregados foi a elaboração de árvores de decisão por meio do algoritmo de aprendizado de máquina baseado em um conjunto de classificadores (*ensemble*) denominado Random Forest, o que culminou em um modelo eficaz com precisão atingindo 97%, recall 95%, f1-score 96% e acurácia 97%.

Palavras-chave: câncer de mama, marchine learning, árvores de decisão e Random Forest.

INTRODUÇÃO

O propósito da pesquisa foi criar um modelo para o diagnóstico do câncer de mama, utilizando **características nucleares obtidas a partir de imagens digitalizadas** por meio de **aspiração por agulha fina** (FNA, do inglês, *Fine Needle Aspirations*). O desenvolvimento desse conjunto de dados tinha como objetivo oferecer uma base de dados confiável e padronizada, destinada a apoiar investigações em **aprendizado de máquina e análises preditivas** na medicina. As informações foram coletadas no ***Clinical Sciences Center*** da Universidade de ***Wisconsin***. As coletas ocorreram durante o início da década de 1990, sendo a primeira publicação relevante sobre os dados apresentada em 1993. O desenvolvimento e refinamento do dataset continuaram até aproximadamente 1995, quando ele foi disponibilizado no***UCI Machine Learning Repository***(UCI Machine Learning Repository 1988), onde continua sendo amplamente utilizado para pesquisas em aprendizado de máquina e análise médica.

O algoritmo escolhido para este estudo foi o Random Forest(Breiman 2001), um método de aprendizado de máquina baseado em um conjunto de classificadores (*ensemble*) que combina múltiplas árvores de decisão. Essa abordagem aumenta a precisão e reduz o risco de *overfitting*, proporcionando um modelo robusto. O Random Forest é amplamente utilizado tanto em problemas de classificação quanto de regressão, devido à sua versatilidade e eficácia.

METODOLOGIA

O estudo foi conduzido de forma abrangente, revisitando cada etapa várias vezes para avaliar, ajustar e validar as hipóteses propostas. Em cada iteração, o tamanho do conjunto de testes foi mantido constante em 30% para efeitos de comparação, e análises detalhadas foram realizadas para decidir sobre a aceitação ou rejeição das hipóteses. Inicialmente, realizamos uma rodada de testes com os dados "crus," cujas métricas estão apresentadas na Tabela 1.

Classe	**Precisão**	**Recall**	**F1-Score**	**Suporte**
B (Benigno)	0,96	0,99	0,98	108
M (Maligno)	0,98	0,94	0,96	63
Acurácia				0,9708
Macro Avg	0,97	0,96	0,97	171
Weighted Avg	0,97	0,97	0,97	171

Tabela 1: Relatório de classificação: dados "crus"e hiperparâmetros padrão, exceto random_state

	Predito Benigno	Predito Maligno
Real Benigno	107	1
Real Maligno	4	59

Figura 1: Matriz de confusão: dados ”crus”e hiperparâmetros padrão, exceto random_state

Após uma análise aprofundada dos dados em conjunto dos resultados da matriz de confusão Figura 1, formulamos duas hipóteses e justificativas principais:
(i) Redução de dimensionalidade;
(ii) Ajuste de parâmetros.

Na **redução de dimensionalidade**, foram excluídas variáveis que demonstraram baixo impacto na predição. Para isso, calculou-se um erro base, e, em seguida, implementamos um processo em que os valores de cada atributo foram embaralhados individualmente, enquanto todos os outros atributos foram mantidos constantes. Em seguida, o modelo foi treinado e testado novamente, conforme ilustrado na **Figura 2**. Estipulamos que as variáveis fossem removidas caso apresentassem um valor de importância ≤ 0.0000, resultando em um total de 21 variáveis excluídas, sendo elas: `smoothness_mean`, `concave points_mean`, `symmetry_mean`, `texture_se`, `concavity_se`, `fractal_dimension_se`, `area_worst`, `fractal_dimension_worst`, `area_mean`, `fractal_dimension_mean`, `perimeter_se`, `concave points_se`, `compactness_worst`, `symmetry_worst`, `radius_se`, `area_se`, `smoothness_se`, `symmetry_se`, `smoothness_worst`, `concave points_worst` e `perimeter_worst`. De modo que o conjunto de dados passou a incluir apenas 9 variáveis a saber: `radius_mean`, `texture_mean`, `perimeter_mean`, `compactness_mean`, `concavity_mean`, `compactness_se`, `radius_worst`, `texture_worst` e `concavity_worst`.

```python
import numpy as np
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from sklearn.utils import shuffle

# Carregar o conjunto de dados breast-cancer-wisconsin.csv
# Substitua 'diagnosis' pelo nome real da coluna alvo, se for diferente
df = pd.read_csv(file_path)
df.drop('Unnamed: 32', axis=1, inplace=True)

# Assumindo que a coluna 'diagnosis' é o alvo e as demais são características
X = df.drop(columns=['diagnosis', 'id'], axis=1)

# Converter 'M' para 1 e 'B' para 0
y = df['diagnosis'].map({'M': 1, 'B': 0})

# Treinar uma floresta aleatória no conjunto de dados
model = RandomForestClassifier(n_estimators=100, random_state=42, oob_score=True)
model.fit(X, y)

# Calcular a acurácia OOB original
baseline_accuracy = accuracy_score(y, model.oob_decision_function_.argmax(axis=1))

# Função para calcular a importância das variáveis via permutação
importances = {}
```

```python
for coluna in X.columns:
    # Copiar o conjunto de dados para fazer a permutação
    X_permuted = X.copy()

    # Embaralhar os valores da coluna atual
    X_permuted[coluna] = shuffle(X[coluna].values, random_state=42)

    # Treinar o modelo novamente com a coluna permutada e calcular a nova acurácia OOB
    model_permuted = RandomForestClassifier(n_estimators=100, random_state=42, oob_score=True)
    model_permuted.fit(X_permuted, y)
    permuted_accuracy = accuracy_score(y, model_permuted.oob_decision_function_.argmax(axis=1))

    # A importância da variável é a diferença na acurácia
    importance = baseline_accuracy - permuted_accuracy
    importances[coluna] = importance


importances = dict(sorted(importances.items(), key=lambda item: item[1], reverse=True))
# Exibir a importância das variáveis
for coluna, importance in importances.items():
    print(f"Importância da variável {coluna}: {importance:.4f}")

# Gerar lista de chaves onde o valor associado é maior que 0.0000
chaves_filtradas = [chave for chave, valor in importances.items() if valor <= 0.0000]

print(chaves_filtradas)
```

Figura 2: Algoritmo cálculo Importância

A segunda hipótese consistiu em **ajustar os parâmetros** do modelo utilizando a biblioteca `sklearn` em Python. Definimos o tamanho da floresta com `n_estimators` igual a 100, `max_samples_leaf` igual a $\log_2$ e o número mínimo de amostras nas folhas como `min_samples_leaf` igual a 9. Esses valores foram selecionados após a execução de um script que testou valores de `n_estimators` de 100 a 1000, com incrementos de 50, `max_samples_leaf` variando entre $\log_2$, $\sqrt{.}$, ”None”e valores de `min_samples_leaf` de 2 a 10.

Foram realizadas várias rodadas de testes para avaliar as hipóteses propostas, ajustando parâmetros como a proporção do conjunto de teste (variando entre 10% e 40%, com incrementos de 5%) e os critérios de avaliação, incluindo entropia e índice de Gini. Cada ajuste teve como objetivo otimizar as métricas de desempenho e identificar a configuração mais eficaz para o modelo.

O estudo foi implementado na linguagem python e está disponível no repositório do GitHub (Pontes e Rego 2024)

RESULTADOS E DISCUSSÕES

Os resultados obtidos após a aplicação das diferentes hipóteses e ajustes de parâmetros demonstraram variações significativas nas métricas de desempenho. Percebemos que com 30% dos dados sendo destinados para o teste obtivemos as melhores métricas portanto fizemos todos os testes subsequentes já com essa proporção. Inicialmente com os dados "crus" as métricas de precisão, recall, f1-score na classe Maligna foram de 98%, 94%, 96% , respectivamente e acurácia de 97%. Esses resultados serviram como base de comparação para as etapas seguintes.

Após implementar a hipótese de redução de dimensionalidade conjuntamente com a escolha dos hiperparâmetros, tanto as métricas como a matriz de confusão apresentaram uma melhora, com podemos verificar na Tabela 2 e Figura 3

Classe	**Precisão**	**Recall**	**F1-Score**	**Suporte**
B (Benigno)	0,9727	0,9907	0,9817	108
M (Maligno)	0,9836	0,9524	0,9677	63
Acurácia				0,9766
Macro Avg	0,9782	0,9716	0,9747	171
Weighted Avg	0,9767	0,9766	0,9765	171

Tabela 2: Relatório de classificação: dados após remoção de variáveis e ajuste dos hiperparâmetros

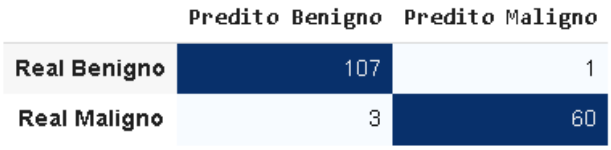

Figura 3: Matriz de confusão: dados e hiperparâmetros ajustados

com precisão atingindo 98,36%, recall 95,24%, f1-score 96,77% e acurácia 97,66%. Esse aumento reforça a justificativa de exclusão das variáveis, otimizando o modelo sem perda de informações relevantes.

CONCLUSÃO

O estudo confirmou a eficácia de abordagens específicas para otimizar a precisão e robustez do modelo de classificação. A hipótese de redução de dimensionalidade, aplicada por meio da exclusão de variáveis que demonstraram pouca **Importância**, mostrou-se particularmente vantajosa.Esses resultados indicam que a simplificação do conjunto de variáveis pode ser uma estratégia eficaz para evitar redundâncias e melhorar o desempenho do modelo.

Esses insights fornecem uma base sólida para estudos futuros, especialmente para aprimorar modelos preditivos em cenários onde a correlação entre variáveis é significativa.

Referências

BREIMAN, L. Random forests. *Machine Learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

PONTES, M.; REGO, M. *Repositório de Inteligência Artificial utilizando Árvore de Decisão no Diagnóstico Câncer de Mama*. 2024. Acessado em: 7 de novembro de 2024. Disponível em: <https://github.com/infopontes/inteligencia_artificial/>.

UCI Machine Learning Repository. *Breast Cancer Wisconsin (Original) Data Set*. 1988. Acessado em: 30-out-2024. Disponível em: <https://archive.ics.uci.edu/dataset/14/breast+cancer>.