UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS JARDINS DE ANITA

CIÊNCIA DE DADOS

ERICK COUTINHO

RELATÓRIO: REDES NEURAIS

ITAPAJÉ, CE

2023

INTRODUÇÃO

O dataset em questão é referente ao diagnóstico de câncer de mama originado pela Universidade de Wisconsin. Ele representa um problema de classificação na área de saúde e medicina, onde o objetivo é prever se uma massa mamária é benigna (B) ou maligna (M) com base em características extraídas de imagens digitalizadas de aspirados com agulha fina (PAAF) das massas.

Descrição do Problema:

O problema em questão é a classificação de massas mamárias como benignas ou malignas com base em características extraídas de imagens. Esta é uma tarefa crítica na detecção precoce de câncer de mama, permitindo intervenções médicas oportunas para melhorar as taxas de sobrevivência e tratamento.

Descobertas:

A normalização das características antes da aplicação dos modelos contribuiu significativamente para o desempenho, especialmente em modelos sensíveis à escala.

A avaliação de métricas como acurácia, Mean Squared Error e matriz de confusão proporcionou uma compreensão detalhada do desempenho e da precisão do modelo em relação aos diagnósticos reais.

Essas descobertas têm implicações significativas na aplicação de métodos de aprendizado de máquina no diagnóstico de câncer de mama, destacando a importância da escolha do modelo e da preparação adequada dos dados para obter resultados confiáveis e clinicamente relevantes.

METODOLOGIA

O conjunto de dados utilizado neste estudo é proveniente do Banco de Dados de Diagnóstico de Câncer de Mama de Wisconsin. Ele consiste em 569 instâncias, cada uma com 30 atributos. Esses atributos representam características extraídas de núcleos celulares em imagens digitalizadas de aspirados com agulha fina (PAAF) de massas mamárias. Os atributos descrevem diversas propriedades dos núcleos celulares, como raio, textura, suavidade, compacidade, concavidade e outros.

Para preparar os dados para a aplicação de algoritmos de aprendizado de máquina, algumas etapas de pré-processamento foram realizadas:

- Normalização:

Dado que os algoritmos de aprendizado de máquina podem ser sensíveis à escala das características, foi aplicada a normalização padrão (z-score) aos dados, usando o método: “StandardScaler”

- Conversão de Rótulos:

Os rótulos originais ('M' para maligno e 'B' para benigno) foram convertidos para valores binários (1 para maligno e 0 para benigno) para fins de classificação.

RESULTADOS

Após a análise do conjunto de dados e a aplicação de modelos de aprendizado de máquina, as descobertas principais incluem:

A aplicação de modelos Random Forest, k-NN e MLP resultou em acurácias notáveis, indicando que esses modelos são promissores para a classificação de câncer de mama.

Desempenho DO KNN

- Acurácia: 0.96
- Mean Squared Error: 0.0351
- Precision: 1.00
- Recall: 0.91
- F1 Score: 0.95

Matriz de Confusão:

68 (Verdadeiros Positivos)	0 (Falso positivo)
4 (Falso negativo)	42 (Verdadeiros Negativos)

Desempenho do Random Forest:

- Acurácia (Random Forest): 0.96
- Mean Squared Error (Random Forest): 0.0351
- Precision: 0.98
- Recall: 0.93
- F1 Score: 0.96

Matriz de Confusão (Random Forest):

67 (Verdadeiros Positivos)	1 (Falso positivo)
3 (Falso negativo)	43 (Verdadeiros Negativos)

Desempenho Rede Neural MLP:

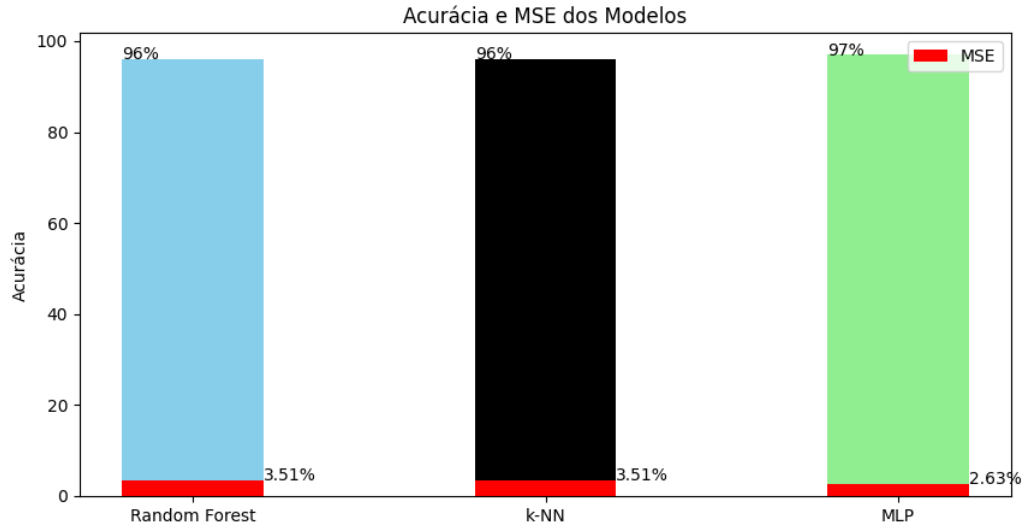
- Acurácia (Random Forest): 0.97
- Mean Squared Error (Random Forest): 0.0263
- Precision: 1.00
- Recall: 0.93
- F1 Score: 0.97

Matriz de Confusão (MLP):

68 (Verdadeiros Positivos)	0 (Falso positivo)

3 (Falso negativo)	43 (Verdadeiros Negativos)

Análise gráfica dos resultados:

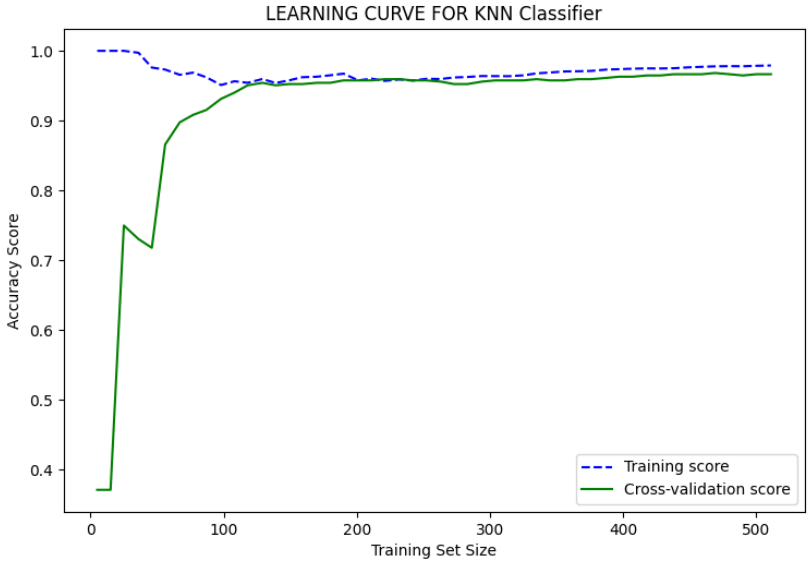


Os resultados destacam a utilidade desses modelos como ferramentas de apoio à decisão em diagnósticos médicos. A alta acurácia e as baixas taxas de erro indicam que esses modelos têm potencial para serem incorporados a práticas clínicas para aprimorar a eficiência na detecção de câncer de mama.

- **Verificando Overfitting**

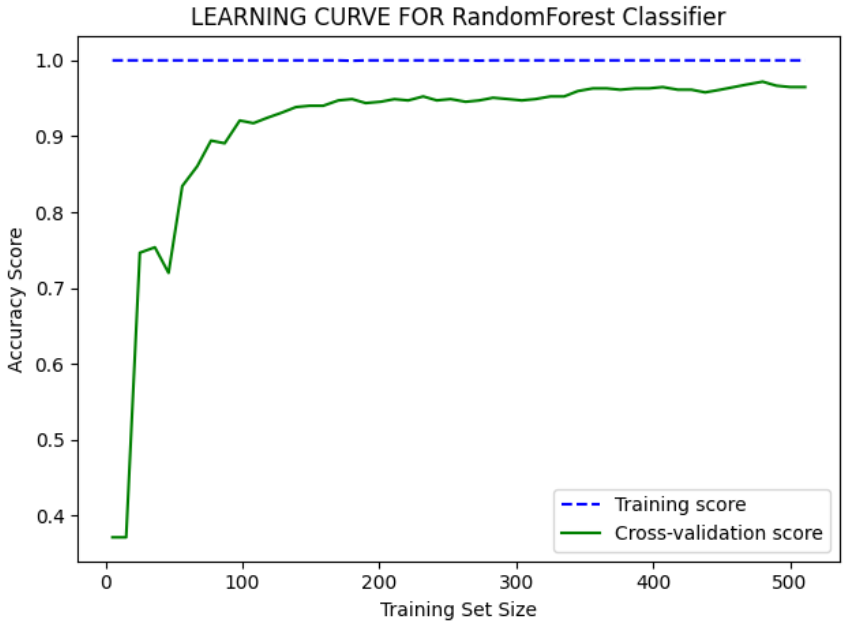
Overfitting ocorre quando um modelo de aprendizado de máquina se ajusta muito bem aos dados de treinamento, capturando até mesmo o ruído e os detalhes irrelevantes, resultando em um desempenho inferior ao generalizar para novos dados não vistos. Isso pode levar a uma falta de capacidade do modelo de se adaptar a padrões mais amplos, comprometendo sua eficácia em situações do mundo real.

Curva de aprendizagem do KNN:



Observando o gráfico, percebemos que as curvas de validação e treino estão próximas, sugerindo um bom equilíbrio entre variância e viés. Isso significa que o modelo não está superajustando (overfitting) nem subajustando (underfitting) os dados de treinamento. A estabilidade das curvas aumenta a confiança de que o modelo terá um bom desempenho em novos dados, uma vez que não há grandes flutuações durante o treinamento.

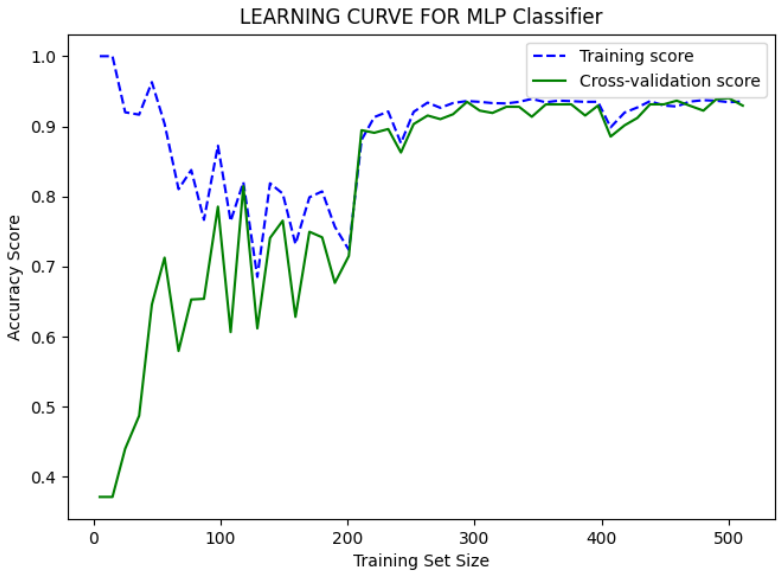
Curva de aprendizagem do Random Forest:



No modelo Random Forest, um score de treinamento de 100% pode ser comum e não necessariamente indicativo de overfitting. Isso ocorre porque cada árvore na floresta pode se ajustar perfeitamente aos dados de treinamento, e a diversidade entre as árvores na floresta ajuda a controlar o overfitting.

O importante a ser observado na curva de aprendizado da Random Forest é como a pontuação de validação cruzada se comporta. A pontuação da validação cruzada sobe à medida que que o tamanho do conjunto aumenta, isso indica um bom sinal, o modelo está generalizando bem para dados não vistos.

Curva de aprendizagem da MLP:



Como vemos no gráfico, no início, com um conjunto de treinamento muito pequeno, o modelo é mais sensível a variações nos dados devido à sua capacidade de memorizar padrões pequenos. Isso resultou em oscilações nos scores de treinamento e validação cruzada.

À medida que o conjunto de treinamento aumenta, o modelo tem mais exemplos para aprender e a variabilidade diminui. Estabilizando e melhorando a consistência do desempenho, indicado pelo aumento contínuo nos escores de treinamento e validação cruzada após um certo ponto.

CONCLUSÕES

Este trabalho explorou a aplicação dos modelos de aprendizado de máquina: Random Forest, k-NN e uma Rede Neural MLP, no diagnóstico de câncer de mama com base em características extraídas de imagens PAAF. Os resultados obtidos foram promissores, demonstrando a eficácia desses modelos na classificação precisa de massas mamárias como benignas ou malignas.

Principais Descobertas:

- A acurácia elevada e os baixos erros de previsão indicam que os modelos são robustos e têm potencial para serem integrados em ambientes clínicos.
- A Rede Neural MLP se destacou, atingindo uma acurácia de 97%, evidenciando sua eficácia na tarefa de diagnóstico de câncer de mama.
- Considerações éticas, transparência nas decisões dos modelos e a normalização prévia dos dados são fundamentais para garantir a confiabilidade dos resultados.

Trabalhos Futuros:

- Validação Externa:

Expandir a validação dos modelos em conjuntos de dados externos para garantir a generalização e a robustez em diferentes contextos clínicos.

- Interpretabilidade do Modelo:

Investigar técnicas para melhorar a interpretabilidade dos modelos, especialmente em cenários médicos onde a transparência é crucial.

- Aprimoramento da Coleta de Dados:

Explorar a possibilidade de incluir dados adicionais e mais variáveis para aprimorar a representatividade do conjunto de dados.

- Estudos Clínicos:

Conduzir estudos clínicos para avaliar como esses modelos podem ser integrados à prática médica, considerando a colaboração com profissionais de saúde.

- Otimização de Hiperparâmetros:

Realizar uma otimização mais aprofundada de hiperparâmetros para melhorar ainda mais o desempenho dos modelos.

Considerações Finais:

Este trabalho oferece uma visão promissora sobre a aplicação de modelos de aprendizado de máquina no diagnóstico de câncer de mama. No entanto, é crucial reconhecer que esses modelos devem ser vistos como ferramentas de apoio à decisão e não substitutos da expertise médica. A colaboração contínua entre a comunidade de aprendizado de máquina e profissionais de saúde é essencial para garantir que essas tecnologias sejam implementadas de maneira ética, transparente e responsável. O caminho para avanços significativos na detecção precoce e tratamento do câncer de mama envolve uma abordagem interdisciplinar e um compromisso contínuo com aprimoramento e validação.