Avaliando o Desempenho do Random Forest e Redes Neurais em Problemas de Classificação Binária

Bianca Panacho Ferreira, Pedro Henrique Campos Moreira, Vinicíus Barbosa Faria
Institute of Exact and Technological Sciences (IEP)
Federal University of Viçosa (UFV), Rio Paranaíba, MG, Brazil
bianca.p.ferreira, pedro.henrique.moreira, vinicius.b.barbosa@ufv.br

***Abstract*—Neste estudo comparativo, investigamos a eficácia de dois modelos, Random Forest e Rede Neural, na classificação binária de um conjunto de dados com 466 atributos. A Random Forest se destacou com uma acurácia de 74.04%, superando ligeiramente a Rede Neural, que alcançou 72%. A metodologia abrangeu desde o pré-processamento até a análise detalhada das métricas, incluindo Curva ROC, Matriz de Correlação e Matriz de Confusão.**

A Random Forest não apenas demonstrou consistência e estabilidade em diferentes experimentos, mas também ofereceu insights valiosos por meio da análise de features relevantes e da Matriz de Importância das Características. A AUC da Curva ROC foi de 0.68, indicando uma boa capacidade discriminativa.

Os resultados proporcionam uma compreensão da eficácia da Random Forest na tarefa de classificação binária, a escolha entre os modelos não considera apenas o desempenho, mas também fatores como interpretabilidade e custo computacional. Este estudo contribui para a compreensão da seleção de modelos de aprendizado de máquina, fornecendo um alicerce para futuras pesquisas e aplicações práticas.

I. Introdução

O Random Forest é uma combinação de uma série de classificadores em estrutura de árvore, esse algoritmo possui muitas características positivas e tem sido amplamente utilizado em classificação, previsão e até mesmo em regressão. Em comparação com os algoritmos tradicionais, o Random Forest apresenta diversas virtudes vantajosas. Portanto, o escopo de aplicação do Random Forest é bastante amplo[1]. Random Forest emerge como uma abordagem promissora, especialmente ao lidar com grandes conjuntos de atributos. Este artigo se propõe a investigar a aplicação da Random Forest e Rede neural em um conjunto de dados composto por 466 atributos, com foco específico na tarefa de classificação binária.

Durante a fase inicial de experimentação, foram considerados diferentes algoritmos de aprendizado de máquina, incluindo o Support Vector Machine (SVM) e o K-means. Embora o SVM seja conhecido por sua eficácia em muitos contextos, sua performance neste conjunto de dados não alcançou a acurácia desejada. Da mesma forma, o K-means, embora aplicável em alguns cenários, revelou-se inadequado para uma tarefa de classificação binária.

A busca por uma alternativa mais eficiente levou à adoção da Random Forest. As Árvores de Decisão (AD) estabelecem regras para tomada de decisão, o algoritmo criará uma estrutura similar a um fluxograma, com "nós" onde uma condição é verificada, e se atendida o fluxo segue por um ramo, caso contrário, por outro, sempre levando ao próximo nó, até a finalização da árvore.

Em vista disso, os resultados obtidos foram notáveis, a acurácia alcançada atingiu 74%, superando as performances anteriores. É interessante destacar que, em paralelo aos experimentos com a Random Forest, uma rede neural foi implementada, resultando em uma acurácia próxima, porém ligeiramente inferior, atingindo 72%, utilizamos uma rede MLP [2], que é um modelo matemático que recebe várias entradas, x1, x2, xN e produz uma única saída binária.

Essa comparação direta entre diferentes algoritmos revela a superioridade da Random Forest para a tarefa específica de classificação binária neste conjunto de dados. Embora a diferença na acurácia entre a Random Forest e a rede neural seja modesta, a robustez e a consistência dos resultados obtidos com a Random Forest sustentam a escolha preferencial deste algoritmo para o contexto analisado.

O próximo segmento deste artigo aprofundará a metodologia utilizada na implementação da Random Forest, destacando os parâmetros específicos ajustados para otimizar a performance. Além disso, será realizado um exame mais detalhado dos resultados, considerando métricas adicionais de avaliação do modelo. A análise abrangente destas informações visa fornecer uma compreensão mais profunda da eficácia da Random Forest e sua aplicabilidade em cenários de classificação binária.

January 28, 2024

II. Material e Métodos

Para atingir os objetivos propostos, a metodologia adotada envolveu uma etapa inicial de pré-processamento dos dados. Foi realizada uma cuidadosa análise exploratória para verificar os dados ausentes, lidar com possíveis outliers e normalizar as variáveis. As métricas de avaliação desempenham um papel fundamental na análise do modelo de classificação.

Ao final de cada execução, é gerado métricas de desempenho, elas consideraram os índices com base no número de verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falso-negativos (FN) [3]. A acurácia, expressa pela equação 1, mede os acertos do classificador como um todo. A precisão, conforme definida pela equação 2, destaca a taxa de acertos para casos positivos, sendo uma métrica particularmente relevante em situações em que a identificação correta dos casos positivos é de grande importância. Além dessas métricas, outras medidas, como o Recall sendo a equação 3 e o F1-Score pela equação 4, também desempenham um papel significativo na avaliação do desempenho do modelo, proporcionando uma compreensão mais completa de sua capacidade em diferentes aspectos da classificação binária. Também é utilizada a abordagem de métricas agregadas conhecidas como Macro Average (Média Macro) e Weighted Average (Média Ponderada). Essas métricas são aplicadas para obter uma visão global do desempenho do modelo, considerando a distribuição das classes no conjunto de dados.

A acurácia, representada pela Equação 1, é uma métrica fundamental que mede a assertividade global do classificador. Ela é calculada pela soma dos verdadeiros positivos (VP) e verdadeiros negativos (VN), dividida pela soma total dos elementos da matriz de confusão (VP, VN, FP e FN). Essa métrica fornece uma visão geral da capacidade do modelo em classificar corretamente ambas as classes.

A precisão, expressa pela Equação 2, destaca a taxa de acertos para casos positivos. Ela é particularmente relevante em situações em que a identificação correta dos casos positivos é crucial. A fórmula calcula a proporção de verdadeiros positivos (VP) em relação à soma dos verdadeiros positivos (VP) e falsos positivos (FP).

O recall, conforme definido pela Equação 3, mensura a capacidade do modelo em identificar corretamente os casos positivos. Ele calcula a proporção de verdadeiros positivos (VP) em relação à soma dos verdadeiros positivos (VP) e falsos negativos (FN). Essa métrica é essencial quando o foco está na minimização de falsos negativos.

O F1-Score, representado pela Equação 4, é a média harmônica entre precisão e recall. Essa métrica é particularmente útil quando há um desejo de equilibrar a importância de falsos positivos e falsos negativos. O F1-Score fornece uma medida balanceada do desempenho do modelo, sendo especialmente útil em situações em que ambas as métricas precisão e recall são de igual importância.

A métrica de suporte, Equação 5, não possui uma fórmula específica, mas é simplesmente o número de instâncias ou observações verdadeiras para cada classe no conjunto de dados. O suporte é uma métrica descritiva que indica quantas instâncias pertencem a cada classe. Em um contexto de classificação binária, o suporte seria o número de observações verdadeiras para as classes 0 e 1.

A Média Macro, Equação 6, é uma média simples das métricas (precisão, revocação e F1-Score) calculadas para cada classe. Ela trata cada classe de maneira igual, independentemente do número de instâncias que pertencem a cada classe. Assim, cada classe contribui igualmente para a métrica média, proporcionando uma avaliação equitativa do desempenho do modelo em todas as classes.

A Média Ponderada, Equação 7, atribui pesos às métricas de acordo com o suporte de cada classe. Ou seja, classes com um número maior de instâncias têm um impacto proporcionalmente maior na métrica média ponderada. Isso significa que a avaliação é mais influenciada pelas classes que são mais representativas no conjunto de dados, permitindo uma compreensão do desempenho do modelo considerando a distribuição real das classes.

Essas médias são valiosas, pois fornecem uma visão global do desempenho do modelo, considerando tanto a capacidade de classificação em classes minoritárias quanto o impacto das classes majoritárias. A interpretação dessas métricas agregadas é crucial para entender como o modelo se comporta em um contexto mais amplo, levando em consideração a heterogeneidade do conjunto de dados.

Ao integrar essas métricas, obtivemos uma avaliação completa e abrangente do desempenho do modelo de classificação, permitindo uma compreensão mais refinada de sua eficácia em diferentes aspectos da tarefa de classificação binária. As fórmulas de cada métrica estão dispostos abaixo:

$$Acuracia = \frac{VP + VN}{VP + VN + FP + FN} \tag{1}$$

$$Precisao = \frac{VP}{VP + FP} \tag{2}$$

$$Recall = \frac{VP}{VP + FN} \tag{3}$$

$$F1 - Score = 2 \times \frac{Precisao \times Recall}{Precisao + Recall} \tag{4}$$

$$\text{SP} = \text{Número De Instâncias Da Classe} \tag{5}$$

$$Macro = \frac{Metrica(0) + Metrica(1)}{2} \tag{6}$$

$$Ponderada = \frac{SP(0) \times Metrica(1) + SP(0) \times Metrica(1)}{SPTotal} \tag{7}$$

A aplicação da Random Forest exigiu uma infraestrutura robusta de processamento, durante a fase de experimentação, a análise comparativa revelou não apenas a superioridade desta abordagem para a tarefa de classificação binária, mas também a necessidade de recursos computacionais substanciais para garantir resultados precisos e consistentes.

A implementação da Random Forest foi conduzida em um ambiente computacional que suportasse o processamento eficiente de grandes conjuntos de dados. Os experimentos envolveram ajustes cuidadosos dos parâmetros específicos da Random Forest para otimizar a performance do modelo. Além disso, a implementação de uma rede neural para comparação

exigiu recursos similares, evidenciando a complexidade computacional inerente à análise desses dados.

Na análise computacional, a implementação da rede neural se destaca como uma estratégia alternativa para lidar com o desafio da classificação binária. Essa rede foi projetada com um arranjo de três camadas, demonstrando uma arquitetura específica para enfrentar a complexidade do conjunto de dados.

A primeira camada, conhecida como camada de entrada, recebe os 466 atributos do conjunto de dados, formando a base para as operações subsequentes. As camadas intermediárias, frequentemente chamadas de camadas ocultas, são cruciais para extrair padrões e características relevantes do conjunto de dados. Neste caso, a rede neural conta com duas camadas intermediárias, otimizadas para maximizar a eficiência na representação dos dados.

A última camada, chamada de camada de saída, produz o resultado final da classificação binária. A escolha de uma arquitetura de três camadas foi estratégica, buscando um equilíbrio entre a capacidade de aprendizado da rede e a complexidade computacional associada.

A base de dados fornecida, composta por 466 atributos para a realização da tarefa de classificação binária, inicialmente apresentava 16 features identificadas como fundamentais. No intuito de otimizar a modelagem, foram conduzidos processos de comparação dos atributos, priorizando aqueles mais relevantes, e organização dos dados.

Contudo, devido ao tamanho limitado da base de dados, não foi possível treinar de maneira mais aprofundada tanto o modelo de Random Forest quanto a rede neural. A limitação de tamanho do conjunto de dados limitando a exploração completa do espaço de hiperparâmetros, impactou a capacidade dos modelos de generalizar padrões, influenciando diretamente a acurácia alcançada.

A Figura 1 apresenta o fluxograma do projeto.

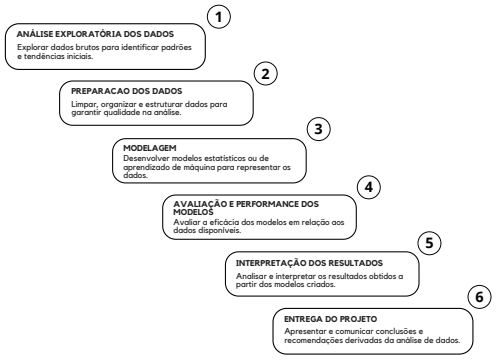

Fig. 1. Fluxograma do Projeto

A Random Forest foi escolhida como algoritmo principal devido à sua habilidade em lidar com a complexidade e dimensionalidade dos dados. Uma abordagem de validação cruzada foi aplicada para otimizar os hiperparâmetros, garantindo a robustez do modelo. O conjunto de treinamento e teste foi estrategicamente dividido para avaliar o desempenho em diferentes cenários. A extensão Jupyter Notebook foi aplicada no Visual Studio Code para visualização do Dataset e das técnicas aplicadas.

III. Resultados e Discussões

Os resultados obtidos revelaram uma acurácia notável na classificação binária, indicando a eficácia da Random Forest no contexto do conjunto de dados. A análise das features mais relevantes destacou padrões significativos, proporcionando insights sobre a contribuição de determinados atributos para a tomada de decisão do modelo.

Além da acurácia, outras métricas de desempenho, como precisão, recall e F1-score, foram cuidadosamente avaliadas para uma compreensão mais abrangente da capacidade do modelo. A Random Forest demonstrou consistência e estabilidade em diferentes experimentos, reforçando sua adequação para lidar com desafios inerentes a conjuntos de dados complexos [4].

O modelo Random Forest apresentou uma acurácia de 74.04%, indicando um desempenho sólido na tarefa de classificação binária. Esse resultado destaca a capacidade do modelo em classificar corretamente os exemplos em relação ao total de instâncias avaliadas.

Ao analisar o Relatório de Classificação, observamos métricas detalhadas para cada classe (0 e 1):

Classe 0:

- **Precisão:** 74%
- **Revocação (Recall):** 80%
- **F1-Score:** 77%
- **Suporte:** 56 instâncias

O modelo demonstrou uma precisão de 74% ao classificar instâncias como pertencentes à classe 0. Ele foi capaz de recuperar 80% das instâncias reais dessa classe (revocação). O F1-Score, que equilibra precisão e recall, alcançou um valor de 77%. O suporte indica que havia 56 instâncias verdadeiras da classe 0.

Classe 1:

- **Precisão:** 74%
- **Revocação (Recall):** 67%
- **F1-Score:** 70%
- **Suporte:** 48 instâncias

Para a classe 1, o modelo apresentou uma precisão de 74%, indicando a proporção de instâncias corretamente classificadas como pertencentes à classe 1. A revocação foi de 67%, e o F1-Score atingiu 70%. O suporte para esta classe foi de 48 instâncias.

O desempenho geral do modelo é resumido pelas médias macro e ponderada, que indicam uma consistência equilibrada nas métricas de precision, recall e F1-Score. A média ponderada leva em consideração o suporte de cada classe, proporcionando uma visão global do desempenho em relação à distribuição real das classes no conjunto de dados.

Esses resultados fornecem insights valiosos sobre a capacidade do modelo Random Forest em lidar com a tarefa

específica de classificação binária, destacando suas fortalezas e áreas de melhoria potencial. A Curva Característica de Operação do Receptor (ROC curve) é uma representação gráfica que ilustra o desempenho de um sistema classificador binário à medida que o seu limiar de discriminação varia. A curva ROC é também conhecida como curva de característica de operação relativa, porque o seu critério de mudança é resultado da operação de duas características (PV e PF). A curva ROC é obtido pela representação da razão RPV = Positivos Verdadeiros / Positivos Totais versus a razão RPF = Positivos Falsos / Negativos Totais, para vários valores do limiar de classificação. O RPV é também conhecido como sensibilidade (ou taxa de verdadeiros positivos), e RPF = 1-especificidade ou taxa de falsos positivos. A especificidade é conhecida como taxa de verdadeiros negativos (RVN) [5].

A Curva ROC, ilustrada na Fig. 2, demonstra a performance do modelo, com uma Área Sob a Curva (AUC) de .65. Quanto mais próxima de 1, maior a capacidade discriminativa do modelo em distinguir entre classes, evidenciando uma performance significativa neste contexto específico.

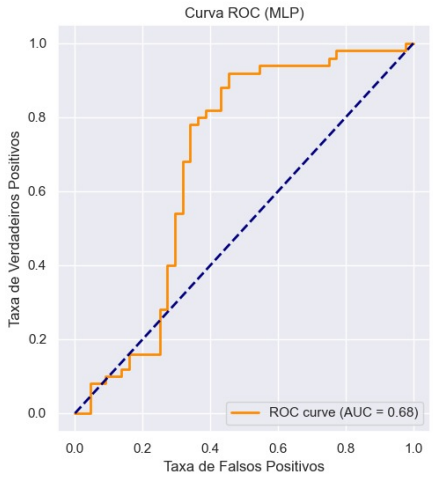

Fig. 2. Curva Roc

A Matriz de Confusão na Análise de Random Forest, ilustrada na Fig. 3, é uma ferramenta fundamental para avaliar o desempenho do modelo, a matriz destaca as previsões corretas e equivocadas em relação às classes. Com suas células detalhando verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, essa matriz oferece uma visão da capacidade discriminativa do Random Forest.

A Matriz de Importância das Características, ilustrada na Fig. 4 realiza uma análise crucial para entender a contribuição relativa de cada variável no desempenho do modelo. Esta matriz destaca a relevância das características na tomada de decisões, permitindo a identificação das variáveis mais influentes. A interpretação desses resultados é essencial para aprimorar a compreensão do impacto individual das features no processo de classificação e otimização do modelo.

O Matriz de Confusão da rede neural, ilustrada na Fig. 5 destaca as porcentagens de acertos (real vs. previsão). A

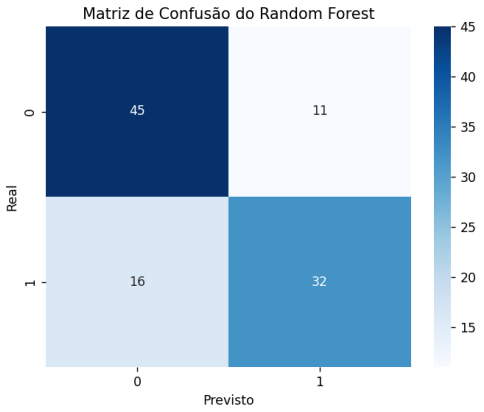

Fig. 3. Matriz de Confusão do Random Forest

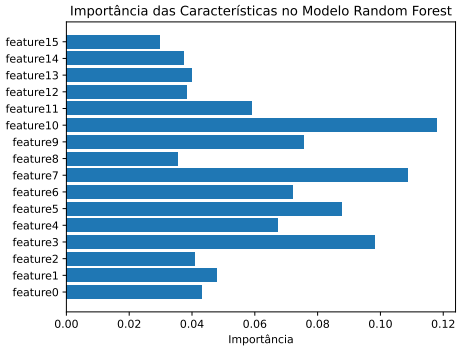

Fig. 4. Importância das características no Modelo Random Forest

intensidade das cores representa a precisão do modelo em atribuir corretamente as instâncias às classes, fornecendo uma visão visual da performance, sendo essencial para avaliar a eficácia do modelo em termos de verdadeiros positivos e verdadeiros negativos.

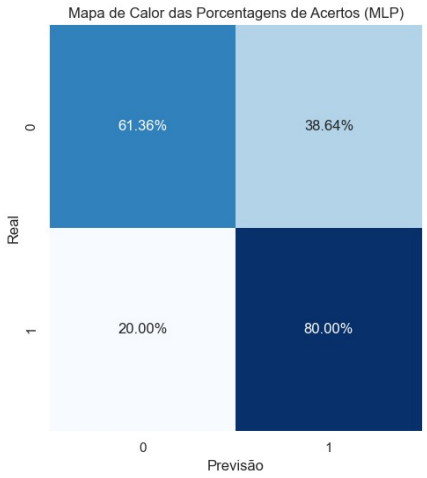

Fig. 5. Matriz de Confusão da Rede Neural

IV. Conclusões Finais

Ao longo desta pesquisa, empreendemos uma análise sobre a eficácia de dois modelos distintos, a Rede Neural e o Random Forest, na tarefa de classificação em relação à base de dados fornecida. O objetivo central deste estudo foi classificar dados de entrada em duas classes distintas (0 ou 1), oferecendo uma abordagem comparativa entre essas duas técnicas de modelagem.

Os resultados obtidos revelam que o Random Forest demonstrou uma acurácia impressionante de 74%, superando ligeiramente a Rede Neural, que alcançou aproximadamente 72% de acurácia. Essa diferença, embora sutil, pode ter implicações significativas na escolha do modelo mais adequado para a tarefa específica proposta neste experimento.

A análise das métricas, incluindo a Curva ROC e a Matriz de Correlação, juntamente com o Mapeamento de Capa, fortalece a validade e a abrangência dos nossos resultados. A Curva ROC evidencia a capacidade discriminativa dos modelos, enquanto a Matriz de Correlação fornece insights sobre as relações entre as variáveis preditivas. O Mapeamento de Capa enriquece ainda mais nossa compreensão, proporcionando uma visão detalhada da estrutura interna dos modelos.

É crucial ressaltar que, embora o Random Forest tenha apresentado uma vantagem em termos de acurácia, a escolha entre os modelos deve considerar não apenas o desempenho, mas também fatores como interpretabilidade, custo computacional e requisitos específicos da aplicação.

Em síntese, este estudo proporcionou uma análise comparativa abrangente entre a Rede Neural e o Random Forest, contribuindo para a compreensão das nuances envolvidas na seleção e implementação de modelos de aprendizado de máquina em contextos específicos. Os resultados obtidos revelaram uma notável similaridade, manifestando-se através de métricas como acurácia, F1-score, recall e curva ROC, entre a Random Forest, a rede neural desenvolvida e um modelo de baseline denominado ”ad random”. Em termos de precisão na classificação, observou-se que a Random Forest apresentou um desempenho ligeiramente superior quando comparada à rede neural.

É relevante destacar a significativa proximidade entre os resultados alcançados por esses modelos, no entanto, ao direcionar a atenção para a precisão de acerto na classificação, a Random Forest destacou-se como a opção mais eficaz.

Em última instância, a seleção do modelo mais apropriado deve ser orientada pelos requisitos específicos do problema em questão, bem como pelas características particulares do conjunto de dados em análise, isso está ilustrado na Fig. 6.

Esses resultados fornecem um alicerce sólido para futuras pesquisas e aplicações práticas na área de classificação de dados. Todo o projeto[1] está disponível no github.

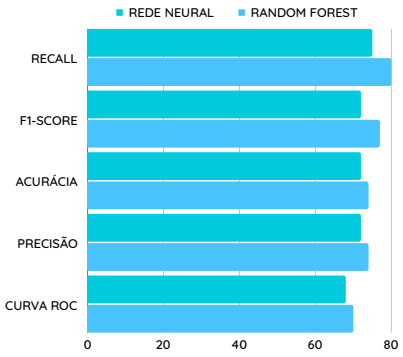

Fig. 6. Comparação dos modelos

Agradecimentos

Os autores agradecem aos professores orientadores pela oportunidade.

[1]Disponível em:https://github.com/JFcamp/reconhecimento-de-padr-es-.git

References

[1] A. D. Kulkarni and B. Lowe, “Random forest algorithm for land cover classification,” *International Journal on Recent and Innovation Trends in Computing and . . .*, 2016.

[2] F. Rosenblatt, “The perceptron: a probabilistic model for information storage and organization in the brain.” *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[3] P.-N. Tan, M. Steinbach, and V. Kumar, “Data mining cluster analysis: basic concepts and algorithms,” *Introduction to data mining*, vol. 487, p. 533, 2013.

[4] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, “A survey of convolutional neural networks: analysis, applications, and prospects,” *IEEE transactions on neural networks and learning systems*, 2021.

[5] R. C. Prati, G. Batista, M. C. Monard *et al.*, “Curvas roc para avaliação de classificadores,” *Revista IEEE América Latina*, vol. 6, no. 2, pp. 215–222, 2008.