A Métrica CFIS: Uma Nova Perspectiva na Análise da Importância das Features em Modelos de Machine Learning

VINICIUS GODOY MARQUES

Resumo

A crescente demanda por modelos de aprendizado de máquina que ofereçam respostas rápidas e precisas impulsiona a criação de novas técnicas para aprimorar a performance e interpretabilidade desses modelos. Neste contexto, o presente trabalho propõe a métrica CFIS (*Combined Feature Importance Score*), uma nova métrica e abordagem inovadora para avaliar a importância das *features* em problemas de aprendizado de máquina. O CFIS combina diferentes métodos, como a importância de permutação, coeficientes de modelos de regressão e correlação com a variável alvo, a fim de fornecer uma visão mais abrangente da relevância das *features*. Ao integrar essas métricas, o CFIS busca superar as limitações de abordagens individuais, oferecendo uma análise mais robusta e detalhada sobre como cada *feature* contribui para a performance do modelo. A aplicação do CFIS pode beneficiar áreas diversas, permitindo que modelos de aprendizado de máquina sejam mais transparentes e eficazes em suas previsões.

Palavras-chave*:* CFIS; Importância das *Features*; Aprendizado de Máquina; Regressão Logística; Importância de Permutação.

The CFIS Metric: A New Perspective on Feature Importance Analysis in Machine Learning Models

Abstract

The growing demand for machine learning models that deliver quick and accurate responses drives the development of new techniques to enhance model performance and interpretability. In this context, this paper proposes the CFIS (Combined Feature Importance Score), an innovative metric and approach to evaluate feature importance in machine learning problems. CFIS combines different methods, such as Permutation Importance, regression model coefficients, and correlation with the target variable, to provide a comprehensive view of feature relevance. By integrating these metrics, CFIS aims to overcome the limitations of individual approaches, offering a more robust and detailed analysis of how each feature contributes to model performance. The application of CFIS can benefit various fields, enabling machine learning models to be more transparent and effective in their predictions.

Keywords: *CFIS; Feature Importance; Machine Learning; Logistic Regression; Permutation Importance.*

1 INTRODUÇÃO

A inteligência artificial, especialmente a subárea de *Machine Learning*, tem avançado significativamente, permitindo que sistemas computacionais aprendam padrões a partir de grandes volumes de dados (Russell; Norvig, 2016). Esse aprendizado é fundamental para gerar previsões e decisões que refletem com precisão a realidade, desde que os dados sejam devidamente tratados e as variáveis corretas sejam identificadas (Pudjihartono *et al*., 2022). No entanto, um dos desafios mais críticos no desenvolvimento de modelos de aprendizado de máquina é a compreensão da importância das *features* no processo de tomada de decisão do modelo (Jie *et al*., 2018).

A importância das *features* é uma área de estudo vital, pois compreender o impacto de cada variável no desempenho do modelo pode melhorar a interpretabilidade, a confiabilidade e a eficácia dos modelos (Kumar; Chong, 2018). Métodos tradicionais, como a importância de permutação, coeficientes de regressão e correlações, oferecem *insights* sobre a relevância das *features*, mas cada abordagem possui limitações intrínsecas. Por exemplo, a importância de

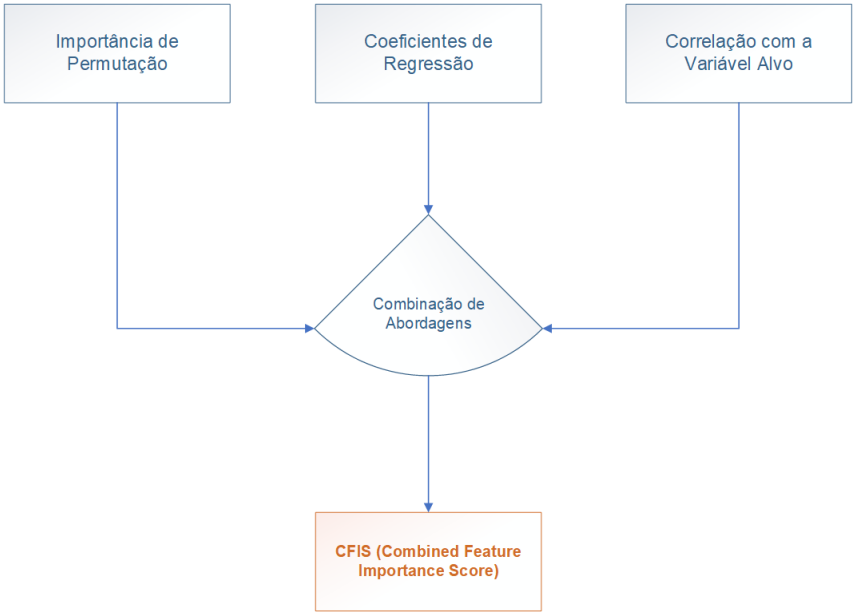
permutação pode ser sensível ao ruído nos dados, enquanto os coeficientes de regressão são específicos a modelos lineares e podem não capturar relações complexas (Brownlee, 2020).

Com o objetivo de superar essas limitações, este trabalho propõe o CFIS (*Combined Feature Importance Score*), uma métrica inovadora que combina múltiplas abordagens para fornecer uma avaliação mais robusta e completa da importância das *features*. O CFIS integra a importância de permutação, os coeficientes de modelos de regressão e a correlação com a variável alvo, criando uma métrica unificada que reflete diferentes aspectos da relevância das variáveis. Essa abordagem combinada permite capturar tanto relações lineares quanto não lineares, além de considerar a variabilidade nas relações entre as *features* e a variável alvo.

A avaliação da importância das *features* é uma área de estudo vital, pois compreender o impacto de cada variável no desempenho do modelo pode melhorar não apenas a interpretabilidade, mas também a eficiência computacional (Hall, 1999). *Features* que não possuem correlação significativa com a variável alvo ou que apresentam baixa pontuação no CFIS podem ser eliminadas, reduzindo a complexidade do modelo e o consumo de recursos computacionais. A remoção de *features* irrelevantes não só simplifica o modelo, mas também diminui o tempo de treinamento e a necessidade de armazenamento de dados, sem comprometer a precisão das previsões (pudjihartono *et al*., 2022).

A Figura 1 ilustra o conceito do CFIS, mostrando como diferentes abordagens são integradas para gerar uma pontuação combinada. Cada método contribui com uma perspectiva única, e o CFIS resulta na média ponderada dessas contribuições, oferecendo uma visão mais abrangente sobre a importância de cada *feature*.

Figura 1 – Diagrama de integração das abordagens de importância de *feature* no CFIS.



Fonte: Elaborado pelo autor (2024).

Este trabalho adota uma abordagem quantitativa e experimental para desenvolver e validar o CFIS, o desenvolvimento dessa métrica tem grande potencial para aprimorar a análise de modelos em diversas áreas, como a medicina, onde a identificação de variáveis críticas pode impactar diretamente o diagnóstico e o tratamento de doenças, e o setor financeiro, onde a compreensão das principais fatores determinantes pode melhorar a previsão de riscos. Além disso, ao oferecer uma métrica de importância mais completa, o CFIS contribui para o desenvolvimento de modelos de aprendizado de máquina mais transparentes, permitindo que decisões automatizadas sejam justificadas e mais bem compreendidas pelos envolvidos.

2 DESENVOLVIMENTO

2.1 Resumo da Etapa de Identificação de Limitações das Técnicas de Análise de Features

Esta etapa visa apresentar as principais limitações das técnicas de análise de *features* de forma individual que foram selecionadas para a criação do CFIS.

2.1.1 Limitações na Importância de Permutação

Esta técnica avalia o impacto de cada *feature* na performance do modelo ao permutar seus valores e medir a degradação do desempenho. Embora ofereça uma visão prática sobre a contribuição das *features*, é sensível a ruídos e variações nos dados. Além disso, pode ser computacionalmente intensiva, especialmente com um grande número de *features* e repetições (BROWNLEE, 2020).

2.1.2 Limitações nos Coeficientes de Regressão

Em modelos lineares, como a Regressão Logística, os coeficientes das *features* fornecem uma medida direta da importância, refletindo a força e a direção das relações com a variável alvo. No entanto, essa abordagem assume que as relações são lineares, o que pode não ser o caso para todas as variáveis. Interações complexas entre *features* podem não ser capturadas adequadamente (Russell; Norvig, 2016)..

2.1.3 Limitações na Correlação com a Variável Alvo

A correlação mede a força da relação linear entre cada *feature* e a variável alvo. Enquanto é útil para identificar relações lineares, não captura interações não lineares ou dependências complexas entre as *features*. A correlação isolada pode fornecer uma visão limitada da importância das variáveis (Russell; Norvig, 2016).

2.2 Etapa de Identificação das Complementaridades das Técnicas

Embora cada abordagem ofereça uma perspectiva valiosa sobre a importância das *features*, suas limitações individuais são superadas quando combinadas. A Importância de Permutação, com sua análise prática da contribuição das *features*, pode ser sensível a ruídos, o que pode ser mitigado ao integrar outras abordagens. Os Coeficientes de Regressão fornecem uma visão direta das relações lineares entre *features* e a variável alvo, mas não capturam interações não lineares. A Correlação, por sua vez, é útil para identificar relações lineares, mas não reflete dependências mais complexas (Russell; Norvig, 2016).

Ao combinar essas técnicas, o CFIS (*Combined Feature Importance Score*) aproveita a robustez prática da Importância de Permutação, a clareza dos Coeficientes de Regressão e a análise estatística da Correlação. Essa combinação permite uma avaliação mais completa da importância das *features*, incorporando tanto relações lineares quanto não lineares, e fornecendo uma medida consolidada que é menos suscetível às limitações de qualquer abordagem isolada. Assim, o CFIS oferece uma análise mais detalhada e confiável da relevância das variáveis no modelo, preenchendo as lacunas deixadas pelas análises individuais e fornecendo uma visão holística da importância das *features*.

2.3 Resumo da Etapa da Criação do CFIS

Nesta seção será abordado, de forma detalhada, o processo de desenvolvimento do *Combined Feature Importance Score* (CFIS). Serão descritos os procedimentos e cálculos necessários para a criação dessa métrica, começando pela coleta e normalização das importâncias das *features*, passando pelo cálculo dos coeficientes de regressão e das correlações com a variável alvo, até a integração dessas três abordagens para gerar uma pontuação combinada. A apresentação será feita em uma sequência lógica e estruturada, permitindo uma compreensão clara do desenvolvimento do CFIS e de como ele se diferencia ao proporcionar uma análise mais robusta da importância das *features*.

2.3.1 Cálculo da Importância das Features Usando Permutation Importance

Esta etapa calcula a importância das *features* com base no *Permutation Importance*. O método *permutation_importance* avalia o impacto de cada *feature* na performance do modelo ao permutar seus valores e medir a degradação da performance. O parâmetro *n_repeats* define o número de repetições para garantir a robustez dos resultados, e *random_state* assegura a reprodutibilidade dos resultados.

Figura 2 – Cálculo da impotância das *features* através do *Permutation Importance*.

```
# Calcular a importância das features usando Permutation Importance
result = permutation_importance(logreg, X, y, n_repeats=10, random_state=42)
```

Fonte: Elaborado pelo autor (2024).

2.3.1.1 Coleta das Importâncias das Features

Nesta etapa a partir do resultado da *Permutation Importance*, esta linha de comando extrai a média das importâncias das *features*. O importances_mean fornece uma média das importâncias calculadas em várias permutações, oferecendo uma medida consolidada da contribuição de cada *feature* para o modelo.

Figura 3 – Trecho do código que extrai a média das importâncias das *features*.

```
# Coletar as importâncias das features
importances_permutation = result.importances_mean
```

Fonte: Elaborado pelo autor (2024).

2.3.1.2 Normalização das Importâncias das Features

Nesta etapa, as importâncias das *features* são normalizadas para garantir que a soma das importâncias seja igual a 1. Isso é feito dividindo cada importância pela soma total das importâncias. A normalização permite uma comparação mais fácil entre as *features*.

Figura 4 – Trecho do código que normaliza as importâncias das *features*.

```
# Normalizar as importâncias das features para que a soma seja igual a 1
normalized_importances_permutation = importances_permutation / np.sum(importances_permutation)
```

Fonte: Elaborado pelo autor (2024).

2.3.2 Coleta dos Coeficientes do Modelo

Esta etapa coleta os coeficientes do modelo de Regressão Logística ajustado. Os coeficientes indicam a força e a direção da relação entre cada *feature* (x) e a variável alvo no modelo (y).

Figura 5 – Trecho do código que coleta os coeficientes do modelo.

```
# Coletar os coeficientes do modelo após o ajuste
coefficients = logreg.coef_[0]
```

Fonte: Elaborado pelo autor (2024).

2.3.2.1 Normalização dos Coeficientes do Modelo

Os coeficientes são normalizados dividindo cada coeficiente pela norma dos coeficientes. Esta normalização é feita para garantir que a soma dos coeficientes seja comparável, facilitando a análise de sua importância relativa.

Figura 6 – Trecho do código que normaliza os coeficientes do modelo.

```
# Normalizar os coeficientes
normalized_coefficients = coefficients / np.linalg.norm(coefficients)
```

Fonte: Elaborado pelo autor (2024).

2.3.3 Cálculo da Correlação entre Features e a Variável Alvo

A correlação entre cada *feature* e a variável alvo é calculada usando a função corr. A aplicação da função lambda permite calcular a correlação para cada *feature* individualmente. Esses valores fornecem uma medida da relação linear entre cada *feature* e a variável alvo.

Figura 7 – Trecho do código que normaliza os coeficientes do modelo.

```
# Calcular a correlação entre as features e a variável alvo
correlation_scores = X.apply(lambda feature: feature.corr(y))
```

Fonte: Elaborado pelo autor (2024).

2.3.4 Resumo da Etapa de Cálculo do CFIS

O cálculo do CFIS (*Combined Feature Importance Score*) será detalhado a seguir, explicando a fórmula de forma teórica utilizada para combinar as métricas de importância de permutação, coeficientes de regressão e correlação. Esta fórmula visa fornecer uma análise abrangente da importância das *features*, integrando diferentes perspectivas em uma única métrica. Posteriormente será apresentado os comandos de cálculo do CFIS e sua normalização.

2.3.4.1 Fórmula do CFIS

A Pontuação Combinada de Importância do Recurso (CFIS: *Combined Feature Importance Score*) é definida na eq.(1):

$$CFIS = \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{PI_i}{\sum_{j=1}^{N} PI_j}\right) + \left(\frac{CM_i}{\|CM\|}\right) + \left(\frac{Corr_i}{\|Corr\|}\right)\right] \quad (1)$$

Onde:

PI_i: Representa a importância de permutação individual da i-ésima feature.
CM_i: Corresponde ao coeficiente do modelo para a i-ésima feature.
$Corr_i$: Refere-se à correlação da i-ésima feature com a variável alvo.
N: É o número total de features.
$\sum_{j=1}^{N} PI_j$: Soma das Importâncias de Permutação de todas as features.
$\|CM\|$: Norma do vetor de coeficientes do modelo.
$\|Corr\|$:Norma do vetor de correlações

2.3.4.2 Cálculo do CFIS

O CFIS é calculado como a média das métricas de importância, coeficientes e correlações. A combinação das três métricas permite uma análise mais robusta da importância das *features*, considerando diferentes aspectos das suas contribuições.

Figura 8 – Trecho do código que calcula o CFIS.

```
# Calcular o CFIS como a média das métricas de importância, coeficientes e correlações
combined_score = (normalized_importances_permutation + normalized_coefficients + correlation_scores) / 3
```

Fonte: Elaborado pelo autor (2024).

2.3.4.3 Normalização das Pontuações CFIS

As pontuações do CFIS são normalizadas para que a soma seja igual a 100. Esta normalização converte os valores do CFIS em porcentagens, facilitando a interpretação e comparação das importâncias das *features*.

Figura 9 – Trecho do código que normaliza o cálculo do CFIS.

```
# Normalizar as pontuações CFIS para que a soma seja igual a 100
cfis_percentage = (combined_score / np.sum(combined_score)) * 100
```

Fonte: Elaborado pelo autor (2024).

2.3.5 Resumo da Etapa de Apresentação dos Resultados do CFIS

Nesta etapa, é detalhado o processo de organização, formatação e exibição dos resultados obtidos a partir do cálculo da métrica CFIS (*Combined Feature Importance Score*). Após o cálculo das pontuações de importância, coeficientes e correlações das *features*, as informações são compiladas em um DataFrame, onde são normalizadas e apresentadas de maneira clara e ordenada. A ordenação dos resultados pelo CFIS, a aplicação de formatações visuais e a exibição final das *features* classificadas buscam facilitar a interpretação dos resultados e a identificação das *features* mais relevantes para o modelo.

2.3.5.1 Criação do DataFrame com as Pontuações de CFIS das *Features*

Nesta etapa, um DataFrame é criado para organizar e apresentar as pontuações das *features*. O DataFrame inclui as importâncias normalizadas de permutação, coeficientes, correlações, pontuações CFIS e suas porcentagens. As *features* são indexadas pelos nomes das colunas conforme podemos visualizar na Figura 10.

Figura 10 – Trecho do código que cria um dataframe com as pontuações de CFIS das *features*.

```
# Criar um DataFrame com as features e suas pontuações de Permutation Importance, coeficientes, correlações,
# pontuações combinadas e porcentagens de importância CFIS
feature_scores = pd.DataFrame({
    'Permutation Importance': normalized_importances_permutation,
    'Coefficient': coefficients,
    'Correlation': correlation_scores,
    'CFIS': combined_score,
    'CFIS (%)': cfis_percentage
}, index=X.columns)
```

Fonte: Elaborado pelo autor (2024).

2.3.5.2 Ordenação dos Resultados pelo CFIS

Os resultados foram ordenados de acordo com o CFIS em ordem decrescente. Isso permite identificar rapidamente as *features* mais importantes com base na métrica CFIS conforme podemos visualizar na Figura 11.

Figura 11 – Trecho do código que ordena os resultados pelo CFIS.

```
# Ordenar os resultados de acordo com o CFIS em ordem decrescente
sorted_features = feature_scores.sort_values(by='CFIS', ascending=False)
```

Fonte: Elaborado pelo autor (2024).

2.3.5.3 Formatação da Coluna CFIS

Nesta etapa, a função highlight_cfis é definida para formatar a coluna CFIS no DataFrame. *Features* com valores positivos são destacadas com um fundo verde claro, enquanto valores negativos são destacados com um fundo salmão. Isso facilita a visualização das importâncias conforme podemos visualizar na Figura 12.

Figura 12 – Trecho do código que formata a coluna CFIS.

```
# Formatar a coluna CFIS para destacar com uma cor diferente
def highlight_cfis(value):
    if value > 0:
        return 'background-color: lightgreen'
    elif value < 0:
        return 'background-color: salmon'
    else:
        return ''
```

Fonte: Elaborado pelo autor (2024).

2.3.5.4 Aplicação da Formatação e Ajustes na Exibição de Porcentagem

Nesta etapa, a formatação é aplicada ao DataFrame para destacar a coluna CFIS com cores e ajustar a formatação das porcentagens. O objetivo final é melhorar a legibilidade dos resultados. Podemos visualizar o trecho do código para esta etapa na figura 13.

Figura 13 – Trecho do código que aplica a formatação e ajustes na exibição das porcentagens de CFIS .

```
# Aplicar a formatação de cores à coluna CFIS e ajustar a formatação da porcentagem
sorted_features_styled = sorted_features.style.applymap(highlight_cfis, subset=['CFIS']).format({'CFIS (%)': '{:.2f}%'}).set_table_styles([{
    'selector': 'th',
    'props': [('text-align', 'left')]
}])

return sorted_features_styled
```

Fonte: Elaborado pelo autor (2024).

2.3.5.5 Chamada da Função para Gerar o DataFrame Ordenado

Após a definição das funções que calcula e organiza as pontuações das *features* com base na métrica CFIS, é necessário exibir os resultados obtidos. No trecho de código na figura 14 a função combined_*feature*_selection(df) é chamada com o DataFrame df como argumento, que contém as *features* do modelo. Essa chamada executa todas as etapas previamente descritas, desde o cálculo das métricas de importância até a formatação dos resultados, gerando como saída um DataFrame com as *features* ordenadas de acordo com a métrica CFIS. Esse DataFrame, armazenado na variável sorted_*features*, permite a visualização e análise direta das *features* mais relevantes.

Figura 14 – Trecho do código que chama a função para gerar o DataFrame ordenado.

```
# Chamar a função com o dataframe df
sorted_features = combined_feature_selection(df)
```

Fonte: Elaborado pelo autor (2024).

2.3.5.6 Exibição das *Features* Classificadas com Valores de CFIS

Nesta etapa, finalmente, as *features* são exibidas com base na métrica CFIS. A impressão do DataFrame permite a visualização das *features* classificadas conforme a importância combinada, na Figura 15 podemos visualizar o trecho do código para a exibição das *features* classificadas com valores de CFIS.

Figura 15 – Trecho do código que exibe as *features* classificadas com valores de CFIS.

```
# Exibir as features classificadas de acordo com a métrica CFIS
print("Features classificadas de acordo com a métrica CFIS(Combined Feature Importance Score):")
sorted_features
```

Fonte: Elaborado pelo autor (2024).

Na Figura 16, é possível visualizar a saída do código que exibe as *features* com os valores calculados através da métrica CFIS. Para este exemplo, foi utilizado o dataset breast_cancer.csv, que contém as seguintes *features*: *Bare Nuclei, Clump Thickness, Bland Chromatin, Uniformity of Cell Shape, Marginal Adhesion, Mitoses, Normal Nucleoli, Uniformity of Cell Size e Single Epithelial Cell Size*. A métrica CFIS foi aplicada para avaliar a importância dessas *features*, combinando informações de importância de permutação, coeficientes de regressão e correlações, resultando em uma análise mais robusta da relevância de cada variável no modelo.

Figura 16 – Resultado da métrica CFIS utilizando como exemplo o dataset breast_cancer.csv.

	Permutation Importance	Coefficient	Correlation	CFIS	CFIS (%)
Bare Nuclei	0.409471	0.381048	0.822696	0.532035	15.76%
Clump Thickness	0.334262	0.525734	0.714790	0.517059	15.32%
Bland Chromatin	0.147632	0.433038	0.758228	0.439818	13.03%
Uniformity of Cell Shape	0.030641	0.311252	0.821891	0.383269	11.35%
Marginal Adhesion	0.011142	0.320957	0.706294	0.341327	10.11%
Mitoses	0.086351	0.482733	0.423448	0.323618	9.59%
Normal Nucleoli	-0.008357	0.211017	0.718677	0.303954	9.00%
Uniformity of Cell Size	-0.005571	0.011726	0.820801	0.275476	8.16%
Single Epithelial Cell Size	-0.005571	0.097666	0.690958	0.259556	7.69%

Fonte: Elaborado pelo autor (2024).

2.4 Materiais e Métodos

Nesta etapa, serão detalhados os materiais e tecnologias utilizados para o desenvolvimento da métrica CFIS, assim como as bibliotecas, ambiente de programação, linguagem utilizada e o dataset selecionado como exemplo para as análises. A linguagem de programação escolhida para a implementação da métrica CFIS foi o Python, devido à sua ampla gama de bibliotecas dedicadas à ciência de dados e aprendizado de máquina, além de ser amplamente acessível e de fácil utilização.

- **Biblioteca Pandas:** Utilizada para a manipulação e tratamento dos dados, permitindo a criação de DataFrames e a execução eficiente de operações de análise de dados.
- **Biblioteca NumPy:** Empregada para a execução de cálculos numéricos, fornecendo funções matemáticas essenciais para a normalização das métricas e o cálculo do CFIS.
- **Google Colaboratory:** Ambiente de desenvolvimento acessado via navegador, utilizado para a elaboração do código em Python. Este ambiente já integra todas as bibliotecas necessárias, como Pandas e NumPy, e oferece poder de processamento em servidores remotos do Google, dispensando a necessidade de recursos locais de hardware.
- **Dataset:** O conjunto de dados utilizado é o breast_cancer.csv, contendo informações sobre 699 observações de pacientes com diagnóstico positivo (maligno) ou negativo (benigno) para câncer de mama. O dataset é ideal para a análise de regressão logística, devido à sua estrutura e às variáveis independentes que permitem a classificação da variável dependente.

Esses materiais e métodos foram cruciais para a criação e validação da métrica CFIS, proporcionando uma análise robusta e detalhada da importância das *features* no contexto de um problema real.

2.5 Resultados e Discussões

A métrica CFIS (*Combined Feature Importance Score*) foi aplicada ao dataset de câncer de mama para avaliar a importância das características na classificação da doença. Os resultados indicam que a *feature* "*Bare Nuclei*" apresenta a maior importância no modelo, com um CFIS

de 0.532035, representando 15.76% da importância total. Em seguida, "*Clump Thickness*" e "*Bland Chromatin*" destacam-se como as características mais relevantes, com CFISs de 0.517059 e 0.439818, respectivamente. Esses achados corroboram a literatura existente, que também aponta "*Bare Nuclei*" e "*Clump Thickness*" como variáveis cruciais para a classificação de câncer de mama (Wolberg et al., 1995; UCI Machine Learning Repository, 2024). A métrica CFIS demonstra que as características com maiores pontuações têm uma contribuição significativa para a previsão, validando sua eficácia em identificar as variáveis mais impactantes no modelo. Portanto, a métrica CFIS revela uma visão integrada da importância das características, oferecendo um método robusto para a seleção de *features* e corroborando os resultados encontrados na literatura.

3 CONSIDERAÇÕES FINAIS

Este estudo apresentou a métrica CFIS (*Combined Feature Importance Score*) como uma abordagem eficaz para avaliar a importância das características em modelos de aprendizado de máquina. A aplicação da métrica ao dataset de câncer de mama destacou "*Bare Nuclei*", "*Clump Thickness*" e "*Bland Chromatin*" como as características mais significativas, alinhando-se com pesquisas anteriores (Wolberg et al., 1995; UCI Machine Learning Repository, 2024). Os resultados evidenciam a importância da CFIS na identificação das variáveis mais relevantes, mostrando que uma análise integrada pode aprimorar a precisão dos modelos preditivos. A métrica permite uma avaliação detalhada e consolidada da contribuição das características, facilitando a construção de modelos mais eficientes e interpretáveis. Conclui-se que a CFIS é uma ferramenta valiosa para a análise de *features*, destacando a necessidade de um tratamento cuidadoso das características para alcançar resultados confiáveis e eficazes em problemas de aprendizado de máquina. A qualidade dos dados e a seleção apropriada das variáveis são essenciais para otimizar o desempenho dos modelos.

REFERÊNCIAS

BROWNLEE, Jason. **Data Preparation for Machine Learning: Data Cleaning, *Feature* Selection, and Data Transforms in Python**. Machine Learning Mastery, 2020. 398 p. Disponível em: https://books.google.com.br/books?hl=pt-BR&lr=&id=uAPuDwAAQBAJ&oi=fnd&pg=PP1&dq=machine+learning+*feature*+selection&ots=Cm2IuhdMxY&sig=MJ_tcogEDloexTh-y9-juSgz_bk&redir_esc=y#v=onepage&q=machine%20learning%20*feature*%20selection&f=false. Acesso em: 19 ago. 2024.

GODOY MARQUES, Vinicius. **CFIS: *Combined Feature Importance Score* - Repositório**. GitHub, 2024. Disponível em: https://github.com/vinigodoy1/cfis. Acesso em: 7 ago. 2024.

GOOGLE. **Google Colaboratory**. Disponível em: https://colab.research.google.com/. Acesso em: 20 ago. 2024.

GOPIKA, N.; MEENA KOWSHALAYA, A. **Correlation Based *Feature* Selection Algorithm for Machine Learning**. In: 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2019, Washington, DC. Anais... Piscataway, NJ: IEEE, 2019. p. 169-178. Disponível em: https://ieeexplore.ieee.org/abstract/document/8723980. Acesso em: 19 ago. 2024.

HALL, Mark A. **Correlation-based *feature* selection for machine learning**. 1999. Tese (Doutorado em Ciência da Computação) – University of Waikato, Hamilton, 1999. Disponível em: https://researchcommons.waikato.ac.nz/items/12a40834-bf51-4d87-89ef-d00c2740f0d8. Acesso em: 17 ago. 2024.

KHOURDIFI, Youness; BAHAJ, Mohamed. ***Feature* Selection with Fast Correlation-Based Filter for Breast Cancer Prediction and Classification Using Machine Learning Algorithms**. In: 2018 International Conference on Smart Electrical and Electronic Devices (ISAECT), 2018, Rabat, Morocco. Anais... Piscataway, NJ: IEEE, 2018. Disponível em: https://ieeexplore.ieee.org/abstract/document/8618688. Acesso em: 21 ago. 2024.

KUMAR, Sunil; CHONG, Ilyoung. **Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States**. International Journal of Environmental Research and Public Health, v. 15, n. 12, p. 2907, 2018. Disponível em: https://www.mdpi.com/1660-4601/15/12/2907. Acesso em: 18 ago. 2024.

NUMPY. **NumPy documentation**. NumPy, 2022. Disponível em: https://numpy.org/doc/. Acesso em: 8 ago. 2024.

PANDAS. **Pandas documentation**. Pandas, 2022. Disponível em: https://pandas.pydata.org/docs/. Acesso em: 8 ago. 2024.

PUDJIHARTONO, Nicholas; FADASON, Tayaza; KEMPA-LIEHR, Andreas W.; O'SULLIVAN, Justin Martin. A review of *feature* selection methods for machine learning-based disease risk prediction**. Frontiers in Bioinformatics**, v. 2, p. 927312, jun. 2022. Disponível em: https://www.frontiersin.org/articles/10.3389/fbinf.2022.927312/full. Licença CC BY 4.0. Acesso em: 13 ago. 2024.

RUSSELL, Stuart; NORVIG, Peter. **Inteligência Artificial: uma abordagem moderna**. 3. ed. São Paulo: Pearson, 2016. 1152 p.
JIE, Cai; JIAWEI, Luo; SHULIN, Wang; SHENG, Yang. ***Feature* selection in machine learning: A new perspective**. Neurocomputing, v. 300, p. 70-79, 2018. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0925231218302911. Acesso em: 17 ago. 2024.

UCI MACHINE LEARNING. **Breast Cancer Wisconsin (Diagnostic) Data Set**. Kaggle, 2016. Disponível em: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data. Acesso em: 13 ago. 2024.

WOLBERG, W. H.; STINGL, J. C.; HALL, S. A. **Wisconsin Breast Cancer Dataset**. In: UCI Machine Learning Repository. University of California, Irvine, 1995. Disponível em: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic). Acesso em: 21 ago. 2024.