IEPG SUMMIT'25
Pensando o futuro com inteligência artificial e consciência social

SELEÇÃO ÓTIMA DE CARACTERÍSTICAS PARA ALGORITMOS DE APRENDIZADO DE MÁQUINA COM O F-WILKS

Matheus Costa Pereira – matheusc_pereira@hotmail.com
Anderson Paulo de Paiva – andersonppaiva@unifei.edu.br

Palavras-chave: Lambda de Wilks, Seleção de Características, Engenharia de Atributos, Planejamento de Experimentos, Aprendizado de Máquina.

1. INTRODUÇÃO

A relevância da *Feature Selection* (FS) e da *Feature Importance* (FI) tem crescido na literatura científica, principalmente no contexto de *Machine Learning* (ML). Embora relacionadas, a FS é aplicada antes do treinamento, reduzindo as variáveis utilizadas no treinamento, enquanto a FI é utilizada após o modelo para avaliar a contribuição de cada variável nas previsões (Mauro *et al*., 2021; Venkatesh e Anuradha, 2019; Wang *et al*., 2022; Anysz *et al*., 2020; Saarela e Jauhiainen, 2021).

O interesse em FS e FI relaciona-se ao desafio da alta dimensionalidade, que afeta tanto a interpretação quanto o custo computacional. A FS busca selecionar subconjuntos de variáveis mais relevantes, ao passo que a FI avalia a influência de cada variável no desempenho do modelo. Contudo, muitos estudos aplicam essas técnicas somente após o treinamento, o que exige que os algoritmos processem inicialmente todas as variáveis disponíveis. Embora útil para avaliação, essa prática pode ser computacionalmente ineficiente, sobretudo em modelos complexos e com grandes volumes de dados (Ioannidou *et al*., 2022; Kaissis *et al*., 2019; Mohan, Thirumalai e Srivastava, 2019; Shi *et al*., 2019; Velusamy e Ramasamy, 2021; Verma e Chandra, 2023).

Diante desse contexto, este estudo propõe uma metodologia de FS baseada em métricas estatísticas multivariadas, em especial o Lambda de Wilks (Λ*) e a estatística F (*F-Value*). O propósito é aplicar FS antes do treinamento, reduzindo o espaço de características e otimizando o uso computacional. Para isso, empregam-se conceitos de *Design of Experiments* (DOE), com modelos de superfícies de resposta posteriormente otimizados por métodos de otimização multiobjetivo.

Adicionalmente, a abordagem integra a FS à *Principal Component Factor Analysis* (PCFA), que combina a *Principal Component Analysis* (PCA) e a *Factor Analysis* (FA), transformando as variáveis originais em escores fatoriais rotacionados. Esse processo visa reduzir a dimensionalidade e aumentar a precisão dos modelos.

Com base nesse contexto, os objetivos desta pesquisa são:

- Propor uma metodologia de FS aplicada previamente, sem a necessidade de executar todo o processo de ML, ou seja, identificar as variáveis mais relevantes antes da modelagem;
- Avaliar o uso DOE na análise de espaços de características reduzidos;
- Comparar estratégias de pré-processamento, considerando variáveis originais, PCA e PCFA.

2. DESENVOLVIMENTO

2.1 Referencial teórico

A FS e FI são amplamente aplicadas para reduzir dimensionalidade e aprimorar o desempenho de algoritmos de ML. Mohan *et al*. (2019) aplicaram FS baseada na entropia de árvores de decisão para melhorar a previsão de doenças cardiovasculares. Verma e Chandra (2023) utilizaram FI em modelos como *Decision Tree*, *LightGBM* e *XGBoost*, combinando-a ao *Mutual Information Gain*, idenficando variáveis relevantes na detecção de ataques em Internet das Coisas.

Carletti *et al*. (2023) exploraram o método *Isolation Forest Depth-based Feature Importance* para selecionar variáveis em tarefas de anomalias. Já Ali *et al*. (2020), empregaram o *Information Gain* para eliminar variáveis ruidosas em sistemas de monitoramento de saúde voltados à detecção de doenças.

Estudos comparativos de Bommert *et al*. (2020) e Reddy *et al*. (2020) evidenciam limitações de custo computacional e contrastam métodos supervisionados, como *Linear Discriminant Analysis* e não supervisionados, como PCA. Singh e Singh (2020) analisaram técnicas de filtragem baseadas em distância entre classes, bem como métodos de envoltório que avaliam o desempenho do classificador.

Hannousse e Yahiouche (2021) combinaram filtros de classificação ao algoritmo Boruta, aumentando a precisão na detecção de *phishing*. Em *Deep Neural Networks*, Khaki e Wang (2019) aplicaram FS após o treinamento, reduzindo a dimensionalidade sem comprometer a acurácia. Zhang *et al*. (2021) propuseram o método *Conditional Weight Joint Relevance*, comparando-o com outras técnicas e validando em diversos conjuntos de dados.

No geral, os estudos confirmam a importância da redução de variáveis, mas expõem limitações das abordagens aplicadas apenas após o treinamento, o que justifica a proposta deste trabalho.

2.2 Metodologia

A proposta deste trabalho é estruturada em cinco etapas principais. A primeira consiste na criação de um planejamento fatorial fracionário que mapeia o espaço completo das características sem explorar todas as iterações possíveis. Na segunda, aplica-se a PCFA para reduzir a dimensionalidade, preservar a maior variância explicada e identificar estruturas latentes, originando um novo subconjunto de *feature*s.

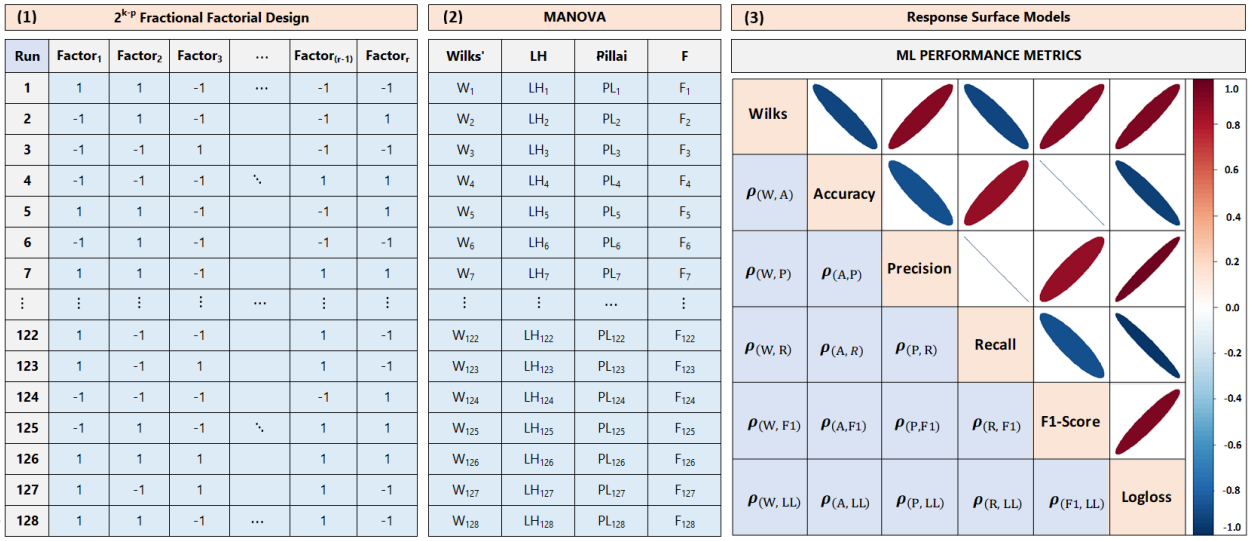
Na terceira etapa, realiza-se a *Multivariate Analysis of* Variance (MANOVA) para cada combinação, com o cálculo e armazenamento de Λ^* e F. A quarta etapa envolve a análise estatística do DOE, modelando Λ^* e de F em função das principais características e de suas interações. Com base nos *Likelihood Ratio Tests* e na lógica de formação de *clusters*, busca-se o subconjunto que minimize Λ^* e maximize F. Por fim, a quinta etapa aplica uma rotina de otimização não linear para encontrar esse subconjunto ótimo.

Após a metodologia, são executados algoritmos de ML com balanceamento de dados, validação cruzada e ajuste de hiperparâmetros, avaliando métricas de desempenho para identificar a melhor configuração.

O método proposto permite identificar combinações de variáveis que reduzem falsos positivos e falsos negativos, ao mesmo tempo em que maximizam verdadeiros positivos, sendo os resultados comparados a métricas tradicionais de FI.

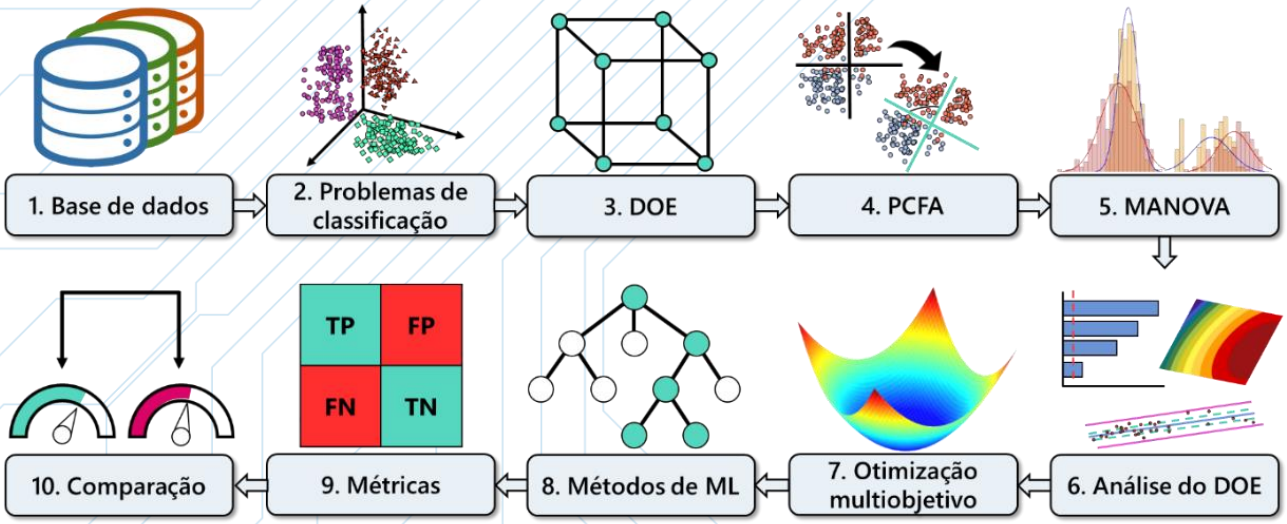
As Figura 1Figura 3 e a Figura 2 ilustram a relação entre as métricas provenientes da MANOVA e as métricas de desempenho ML, bem como o fluxo metodológico detalhado. Os *datasets* empregados são tabulares, com foco em problemas de classificação.

Figura 1 - Relação entre F, Λ* e as métricas de desempenho de ML



Fonte: Elaborada pelo autor (2024).

Figura 2 - Esquema detalhado da metodologia passo a passo



Fonte: Elaborada pelo autor (2024).

A metodologia proposta busca estabelecer um procedimento de FS para definir previamente o subconjunto ideal de atributos, reduzindo a dimensionalidade e preservando a variabilidade dos dados. Assim, é possível identificar as *features* mais relevantes sem executar todo o ciclo de treinamento, validação e teste dos modelos de ML, torando o processo mais eficiente e menos custoso do ponto de vista computacional.

3. RESULTADOS E DISCUSSÃO

Para avaliar a abrangência e a eficácia, a metodologia proposta foi testada em 10 *datasets* listados na Tabela 1. O método F-Wilks prior foi aplicado em domínios diversos, incluindo física, engenharia, química, biologia, saúde, medicina, clima e meio ambiente, contemplando diferentes quantidades de atributos originais, classes e instâncias.

A Tabela 2 apresenta a correlação entre Λ* e cinco métricas de desempenho, considerando o melhor algoritmo de ML em cada conjunto de dados. Observa-se forte correlação de Λ* com os algoritmos de maior desempenho, independentemente do número de atributos, classes ou instâncias.

Tabela 1 - Informações gerais sobre os conjuntos de dados utilizados

Dataset	Área de estudo	*Features*	Classes	Instâncias
Wine	Physics and Chemistry	13	3	177
Dry Bean	Biology	16	7	13.612
Chemical Composition	Physics and Chemistry	17	2	87
Connectionist Bench	Physics and Chemistry	60	2	208
Algerian Forest Fire	Biology	10	2	122
Forest Type	Biology	27	4	523
Vehicle Silhouettes	Other	18	4	846
QSAR Biodegradation	Other	32	2	1.055
Diabetic Retinopathy	Health and Medicine	16	2	1.151
Image Segmentation	Other	16	7	2.100
Landsat Satellite	Climate and Environment	36	6	6.435

Fonte: Elaborada pelo autor (2024).

Tabela 2 - Correlação entre Λ* e métricas de desempenho em ML

Dataset	Algoritmo	*Log Loss*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Wine	LR	0,9604	-0,9569	-0,9564	-0,9569	-0,9582
Dry Bean	KNN	0,9138	-0,8863	-0,8826	-0,8863	-0,8847
Chemical Composition	RF	0,9461	-0,9584	-0,9550	-0,9584	-0,9580
Connectionist Bench	LDA	0,9678	-0,8601	-0,8492	-0,8601	-0,8521
Algerian Forest Fire	SVM	0,9501	-0,9134	-0,8949	-0,9134	-0,9117
Forest Type	DT	0,9544	-0,9544	-0,9550	-0,9544	-0,9546
Vehicle Silhouettes	LR	0,9692	-0,9104	-0,8983	-0,9104	-0,9086
QSAR Biodegradation	LDA	0,9931	-0,9378	-0,8660	-0,9378	-0,9529
Diabetic Retinopathy	LDA	0,9915	-0,9380	-0,8418	-0,9380	-0,8544
Image Segmentation	GB	0,9026	-0,8774	-0,8786	-0,8774	-0,8781
Landsat Satellite	XGB	0,9658	-0,9673	-0,9625	-0,9673	-0,9638

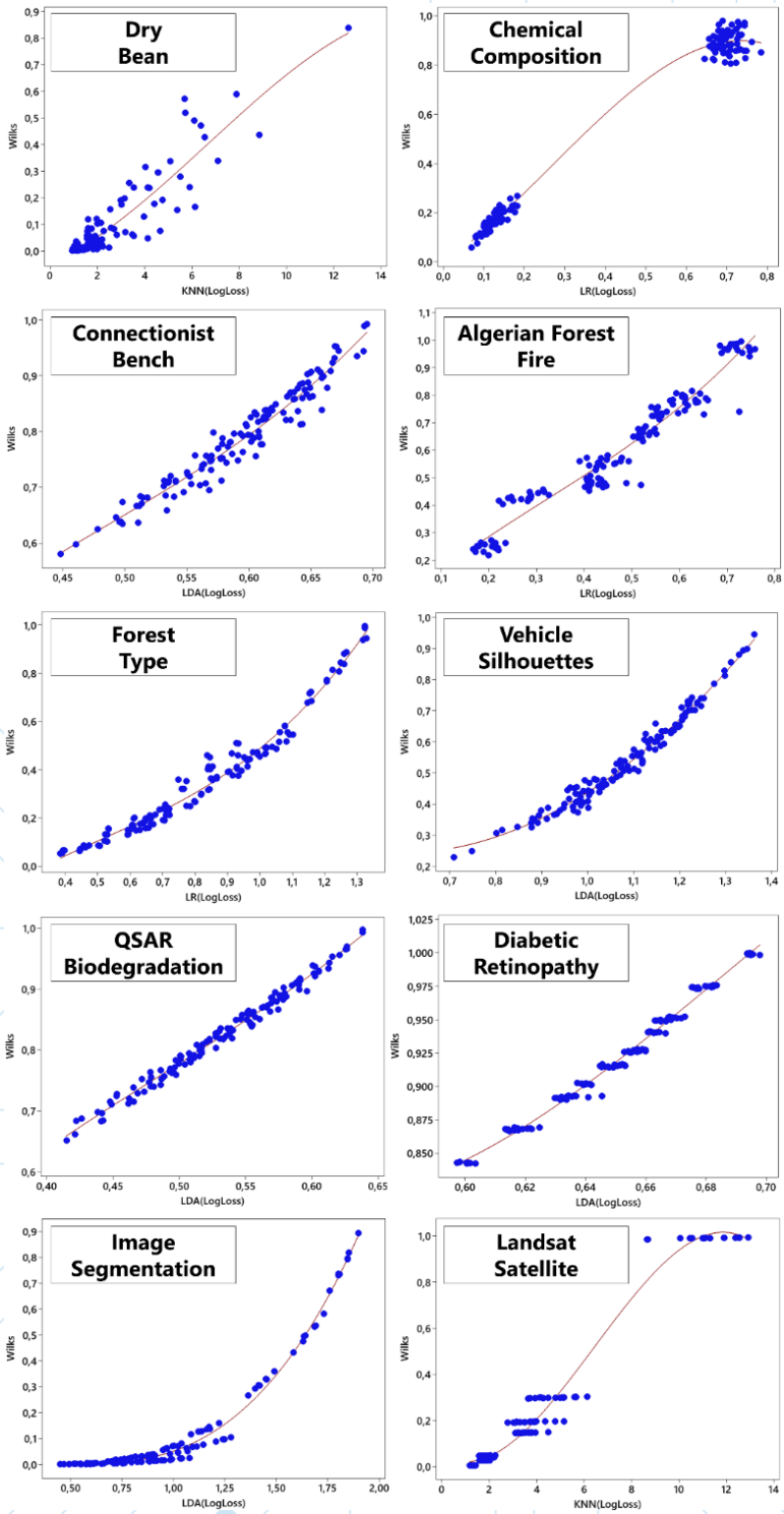
Fonte: Elaborada pelo autor (2024).

Já a Figura 3 mostra a relação entre Λ* e *Log Loss* dos algoritmos que apresentaram melhor ajuste em cada estudo. Os subgráficos evidenciam que, conforme o *Log Loss* diminui (melhor desempenho preditivo), Λ* também tendem a reduzir, indicando associação consistente entre os indicadores.

Por sua vez, Figura 4 amplia a análise ao correlacionar Λ* com métricas baseadas de classificação (*Accuracy*, *Precision*, *Recall* e *F1-Score*). Os pontos coloridos representam variações por algoritmo e o *dataset*. Nota-se padrão de correlação inversa: valores mais baixos de Λ* associam-se a desempenhos mais elevado, reforçando o potencial de descritores multivariados como indicadores preliminares de qualidade do conjunto de variáveis.

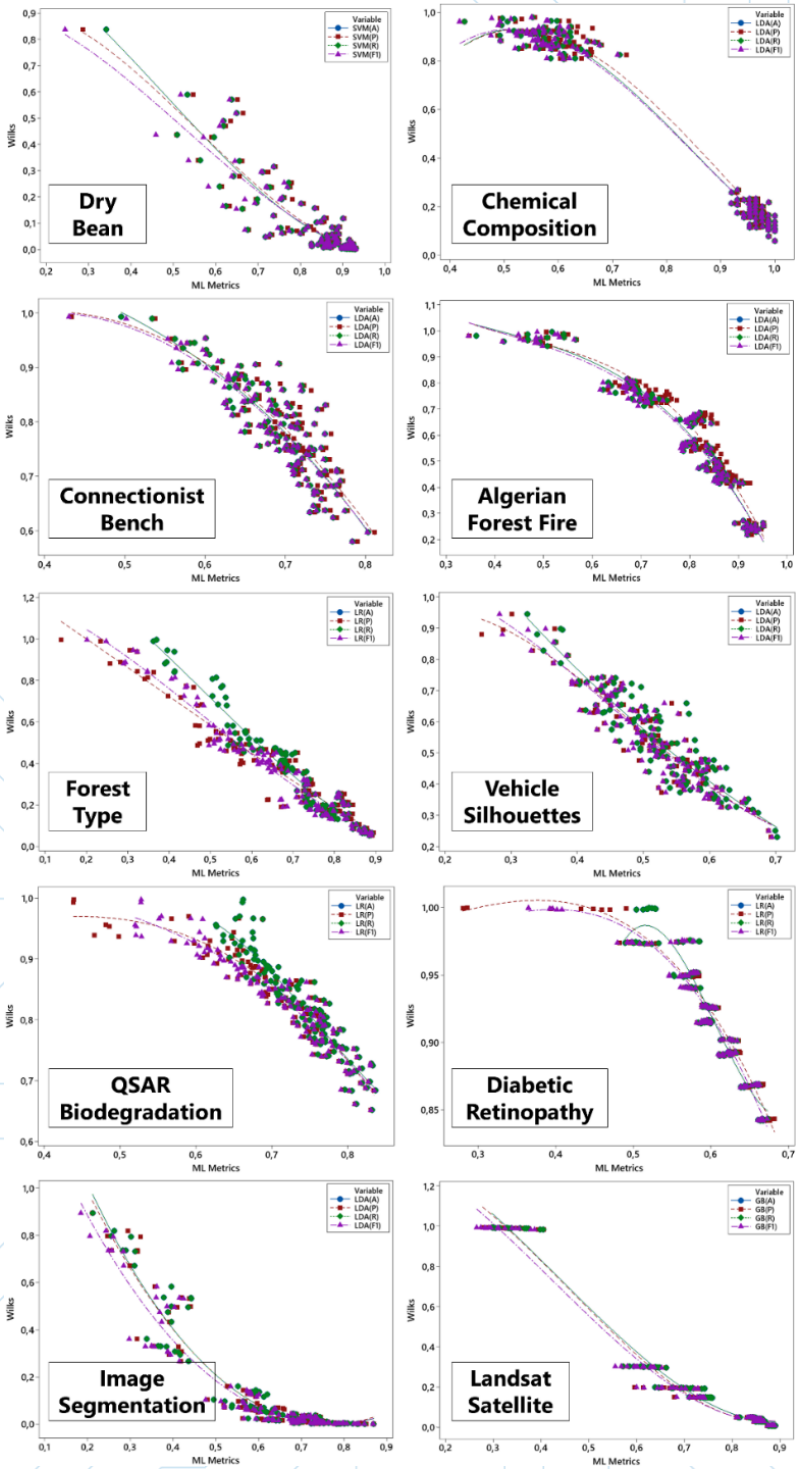
Embora requeiram cautela quanto à generalização, os resultados indicam que estatísticas multivariadas, como o Λ*, podem atuar como etapa preliminar na triagem em processos de FS, inclusive em cenários mais complexos. Os coeficientes de correlação entre Λ* e as métricas de desempenho em diferentes algoritmos de ML mostraram-se elevados e estatisticamente significativos, fortalecendo a robustez da evidência.

Figura 3 - Relação entre *Log Loss*, Λ* e métricas de ML em múltiplos conjuntos de dados



Fonte: Elaborada pelo autor (2024).

Figura 4 - Relação entre *Accuracy*, *Precision*, *Recall* e *F1-Score* e o Λ* no contexto de ML em múltiplos conjuntos de dados



Fonte: Elaborada pelo autor (2024).

4. CONSIDERAÇÕES FINAIS

Este artigo apresentou uma metodologia de FS baseada em descritores estatísticos multivariados, com destaque para Λ* e F obtidos pela MANOVA. Os resultados mostraram que essas são capazes de identificar as variáveis mais relevantes de um conjunto de dados original para alimentar algoritmos de ML.

Para reduzir a dimensionalidade, foram utilizadas variáveis latentes extraídas via PCFA. A relação entre os descritores e o desempenho dos algoritmos foi investigada por meio de planejamento fatorial fracionário, o que reduziu a quantidade de iterações necessárias. A estratégia de DOE possibilitou a construção de modelos de superfície de resposta para os descritores e métricas de desempenho, posteriormente submetidos a um problema de otimização multiobjetivo, resultando na determinação do subconjunto ótimo.

Foram realizadas 32 execuções de confirmação, nas quais o subconjunto ótimo de variáveis foi aplicado a oito algoritmos distintos de ML. Os testes de hipótese indicaram desempenho consistente em todos os modelos, sem indícios de *overfitting*. Esse resultado é expressivo, considerando o número de *datasets* e variáveis analisadas, e reforça a capacidade de generalização da metodologia proposta, evidenciando seu potencial como ferramenta de seleção de *features* em cenários diversos e complexos.

AGRADECIMENTOS

Os autores agradecem à FAPEMIG, CAPES e CNPq pelo apoio financeiro, e ao NOMATI-UNIFEI pela infraestrutura laboratorial, materiais e conhecimento técnico.

REFERÊNCIAS

ALI, F. et al. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. **Information Fusion**, v. 63, p. 208–222, 2020.

ANYSZ, H. et al. Feature Importance of Stabilised Rammed Earth Components Affecting the Compressive Strength Calculated with Explainable Artificial Intelligence Tools. **Materials**, v. 13, n. 10, p. 2317, 2020.

BOMMERT, A. et al. Benchmark for filter methods for feature selection in high-dimensional classification data. **Computational Statistics & Data Analysis**, v. 143, p. 106839, 2020.

CARLETTI, M.; TERZI, M.; SUSTO, G. A. Interpretable Anomaly Detection with DIFFI: Depth-based feature importance of Isolation Forest. **Engineering Applications of Artificial Intelligence**, v. 119, p. 105730, 2023.

CHANTAR, H. et al. Feature selection using binary grey wolf optimizer with elite-based crossover for Arabic text classification. **Neural Computing and Applications**, v. 32, n. 16, p. 12201–12220, 2020.

CHEN, K.; ZHOU, F.-Y.; YUAN, X.-F. Hybrid particle swarm optimization with spiral-shaped mechanism for feature selection. **Expert Systems with Applications**, v. 128, p. 140–156, 2019.

EFFROSYNIDIS, D.; ARAMPATZIS, A. An evaluation of feature selection methods for environmental data. **Ecological Informatics**, v. 61, p. 101224, 2021.

HANNOUSSE, A.; YAHIOUCHE, S. Towards benchmark datasets for machine learning based website phishing detection: An experimental study. **Engineering Applications of Artificial Intelligence**, v. 104, p. 104347, 2021.

IOANNIDOU, M. et al. Assessing the Added Value of Sentinel-1 PolSAR Data for Crop Classification. **Remote Sensing**, v. 14, n. 22, p. 5739, 2022.

KAISSIS, G. et al. A machine learning model for the prediction of survival and tumor subtype in pancreatic ductal adenocarcinoma from preoperative diffusion-weighted imaging. **European Radiology Experimental**, v. 3, n. 1, p. 41, 2019.

MAFARJA, M. et al. Binary grasshopper optimisation algorithm approaches for feature selection problems. **Expert Systems with Applications**, v. 117, p. 267–286, 2019.

MAURO, M. DI et al. Supervised feature selection techniques in network intrusion detection: A critical review. **Engineering Applications of Artificial Intelligence**, v. 101, p. 104216, 2021.

MOHAN, S.; THIRUMALAI, C.; SRIVASTAVA, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. **IEEE Access**, v. 7, p. 81542–81554, 2019.

OMUYA, E. O.; OKEYO, G. O.; KIMWELE, M. W. Feature Selection for Classification using Principal Component Analysis and Information Gain. **Expert Systems with Applications**, v. 174, p. 114765, 2021.

POLAT, H.; POLAT, O.; CETIN, A. Detecting DDoS Attacks in Software-Defined Networks Through Feature Selection Methods and Machine Learning Models**. Sustainability**, v. 12, n. 3, p. 1035, 2020.

REDDY, G. T. et al. Analysis of Dimensionality Reduction Techniques on Big Data. **IEEE Access**, v. 8, p. 54776–54788, 2020.

SAARELA, M.; JAUHIAINEN, S. Comparison of feature importance measures as explanations for classification models. **Applied Sciences**, v. 3, n. 2, p. 272, 2021.

SHI, X. et al. A feature learning approach based on XGBoost for driving assessment and risk prediction. **Accident Analysis & Prevention**, v. 129, p. 170–179, 2019.

SINGH, D.; SINGH, B. Investigating the impact of data normalization on classification performance. **Applied Soft Computing**, v. 97, p. 105524, 2020.

TOO, J.; ABDULLAH, A. R.; MOHD SAAD, N. A New Quadratic Binary Harris Hawk Optimization for Feature Selection. Electronics, v. 8, n. 10, p. 1130, 2019.

TSAI, C.-F.; CHEN, Y.-C. The optimal combination of feature selection and data discretization: An empirical study. **Information Sciences**, v. 505, p. 282–293, 2019.

VELUSAMY, D.; RAMASAMY, K. Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset. **Computer Methods and Programs in Biomedicine**, v. 198, p. 105770, 2021.

VENKATESH, B.; ANURADHA, J. A Review of Feature Selection and Its Methods. Cybernetics and Information Technologies, v. 19, n. 1, p. 3–26, 2019.

VERMA, R.; CHANDRA, S. RepuTE: A soft voting ensemble learning framework for reputation-based attack detection in fog-IoT milieu. **Engineering Applications of Artificial Intelligence**, v. 118, p. 105670, 2023.

WANG, J. et al. An adaptively balanced grey wolf optimization algorithm for feature selection on high-dimensional classification. **Engineering Applications of Artificial Intelligence**, v. 114, p. 105088, 2022.

WANG, X. et al. Aircraft taxi time prediction: Feature importance and their implications. **Transportation Research Part C: Emerging Technologies**, v. 124, p. 102892, 2021.

ZHANG, P. et al. A conditional-weight joint relevance metric for feature relevancy term. **Engineering Applications of Artificial Intelligence**, v. 106, p. 104481, 2021.