Seleção de Atributos para Predição de Tempo de Manutenção em Vagões Ferroviários: Uma Análise com Métodos de Filtro

Josemar Coelho Felix[1], Rodrigo César Pedrosa Silva[1], Andrea Gomes Campos Bianchi[1]

[1] Programa de Pós-Graduação em Ciência da Computação - Universidade Federal de Ouro Preto – (UFOP)

Abstract. *This study investigated filter methods for selecting relevant attributes in predicting railcar maintenance time, using a database from the company MRS Logística and its experts. The results highlighted the sensitivity of the algorithms in attribute selection, with the Groupfs method proving efficient in predictive accuracy, despite having a higher computational cost. However, all methods faced challenges in distinguishing between pertinent and irrelevant attributes. As next steps, the integration of multiple company databases is planned to incorporate data on human and material costs, aiming to further optimize the railway maintenance process.*

Resumo. *Este estudo investigou métodos de filtro para selecionar atributos relevantes na predição do tempo de manutenção de vagões ferroviários, utilizando um banco de dados da empresa MRS Logística e a opinião de seus especialistas. Os resultados destacaram a sensibilidade dos algoritmos na escolha de atributos, com o método Groupfs mostrando-se eficiente em precisão preditiva, embora tenha um custo computacional maior. No entanto, todos os métodos enfrentaram desafios na distinção entre atributos pertinentes e irrelevantes. Como próximos passos, planeja-se integrar múltiplos bancos de dados da empresa para incorporar dados sobre custos humanos e materiais, visando otimizar ainda mais o processo de manutenção ferroviária.*

1. Introdução

A gestão eficiente da manutenção ferroviária é crucial para a operação segura e eficaz das ferrovias, e a seleção adequada de variáveis para prever o tempo de manutenção de vagões desempenha um papel fundamental nesse cenário, influenciando diretamente a disponibilidade dos ativos, os custos operacionais e a satisfação dos clientes (Felix et al., 2022). Este estudo propõe-se a investigar a escolha dos melhores métodos de filtro para a seleção de variáveis relevantes na construção de modelos que preveem a manutenção de vagões. Nesse contexto, serão considerados o custo computacional, a eficácia na escolha dos atributos e os erros dos modelos construídos com o método de aprendizado de máquina denominado árvore de decisão, para determinar o método de filtro mais adequado.

Atualmente, existem lacunas no conhecimento sobre quais métodos são mais eficazes para lidar com conjuntos de dados complexos como os da gestão ferroviária, o que motiva esta pesquisa (Silva et al., 2023). Além disso, modelos simplificados

[1] O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de NíıvelSuperior - Brasil (CAPES) e do Laboratório de Computação de Sistemas Inteligentes (CSI-Lab)

utilizando esses métodos apresentam desempenho semelhante ou superior ao modelo original, tornando-os mais eficientes e interpretáveis, sem prejuízo ou até com melhora da precisão preditiva (Bommert et al., 2020).

A relevância do tema reside na necessidade de otimizar processos de manutenção, reduzir custos e melhorar a eficiência operacional das empresas ferroviárias (Felix & Silva, 2024). Um método de filtro para seleção de atributos é uma técnica utilizada em aprendizado de máquina e análise de dados que visa identificar e selecionar as variáveis mais relevantes em um conjunto de dados, como no caso da manutenção de vagões. O método de filtro avalia individualmente as características de cada variável com base em diferentes métricas, independentemente do modelo preditivo a ser utilizado (Hou et al., 2020). Diferentemente de outros métodos, como os métodos wrapper e embedded, que consideram a interação entre as variáveis e o modelo durante a seleção, os métodos de filtro são rápidos e escaláveis (Alyasiri et al., 2022). Paralelamente, já existem pesquisas focadas em identificar os parâmetros mais relevantes para aumentar a eficiência dos processos de tratamento de resíduos orgânicos, buscando comparar e classificar os melhores atributos no descarte de resíduos (Rao & Baral, 2011).

Os objetivos desta pesquisa são explorar métodos de seleção de atributos através de uma análise quantitativa e exploratória do banco de dados da empresa ferroviária MRS Logística, comparar a escolha de atributos feita por métodos de filtro com a seleção realizada por especialistas da empresa, avaliar o desempenho de modelos preditivos utilizando árvores de decisão e analisar o custo computacional associado aos métodos de seleção de atributos. Pretende-se com este estudo contribuir metodologicamente ao validar a seleção por filtros específicos em contextos industriais complexos como o ferroviário, além de oferecer uma abordagem sistemática para a escolha e avaliação de atributos que pode ser aplicada em outras realidades. Em termos práticos, os resultados deste estudo podem beneficiar gestores ferroviários ao proporcionar insights sobre quais atributos influenciam a produtividade e ao auxiliar a estruturação de pesquisas aplicadas na área computacional.

2. Metodologia

Este estudo quantitativo exploratório foi conduzido utilizando um estudo de caso na empresa ferroviária MRS Logística. A pesquisa analisa a influência de diversos atributos no tempo necessário para realizar tarefas de manutenção, utilizando técnicas de seleção de atributos e modelagem preditiva. A empresa MRS Logística forneceu um banco de dados abrangente contendo 140 mil registros de tarefas de manutenção de vagões. Os atributos estudados incluíram: Descrição da Operação, Dia da Semana, Feriado, Tempo de Empresa, Recursos Humanos, Exigência de Manutenção, Descrição de Materiais, Lote do Ativo e Tipo de Manutenção. Outras variáveis registradas pelo sistema de informação da empresa, como ordem de serviço, data e condição do item, foram consideradas não representativas para influenciar a variável resposta.

Para a seleção de atributos, foram utilizados métodos de filtro, baseados na revisão da literatura (Pilnenskiy & Smetannikov, 2020). Esses algoritmos foram usados para identificar quais atributos tinham maior relevância na previsão do tempo de manutenção. A escolha dos atributos selecionados pelos algoritmos foi comparada com a escolha de atributos de três especialistas da empresa que forneceram os dados e, para garantir a robustez dos modelos, foi utilizada a técnica de cross-validation com um k-fold de 5. Essa seção descreve a metodologia utilizada para avaliação do desempenho dos modelos. O desempenho foi comparado com base em duas métricas principais

sugeridas pela literatura (Penha et al., 2016): Mean Absolute Error (MAE) e Mean Squared Error (MSE). A escolha dessas métricas se deve à sua capacidade de avaliar, respectivamente, a magnitude absoluta dos erros e a penalização de erros maiores de forma quadrática, o que permite uma análise abrangente da acurácia preditiva e robustez dos modelos. A estratégia utilizada para identificar melhorias na tomada de decisão na gestão da manutenção ferroviária é possível de visualizar na Figura 1.

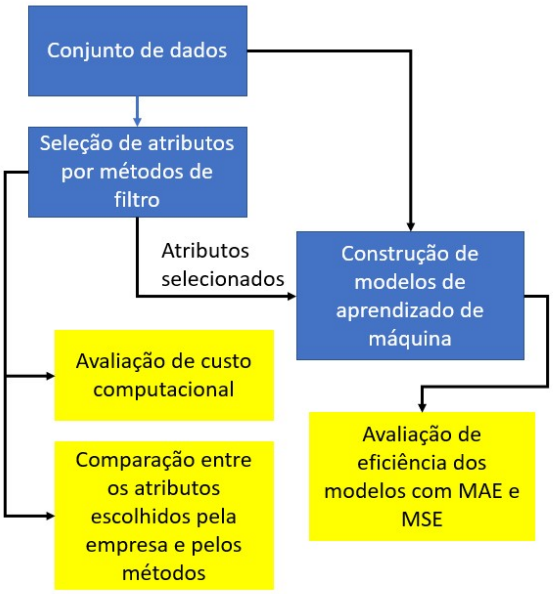

Figure 1. Adaptação do método experimental de PARMEZAN et al. (2017).

3. Resultados

A Tabela 1 apresenta os métodos de filtro utilizados e as porcentagens de escolhas de atributos baseadas na base de dados, comparando-as com a escolha de atributos feita pelos especialistas da empresa. A tabela fornece uma visão detalhada sobre a eficácia e a precisão dos diferentes métodos de filtro na seleção de atributos relevantes para a previsão do tempo de manutenção de itens de vagões ferroviários. Observa-se que metade dos algoritmos de filtro selecionaram todos os atributos do banco de dados como importantes para a previsão do tempo de manutenção. Esse resultado aponta para modelos excessivamente complexos, com maior custo computacional e risco de overfitting. A Tabela 1 mostra os algoritmos "Information Gain Score", "DIRS Score" e "JMI Score" destacaram-se por selecionar todos os atributos do banco de dados, incluindo aqueles que os especialistas da empresa consideraram irrelevantes para o processo. Esse comportamento pode ser explicado pelo fato de esses algoritmos não diferenciarem adequadamente entre atributos realmente informativos e aqueles que são apenas ruídos. Esse resultado sugere uma limitação intrínseca dos algoritmos de filtro em distinguir completamente os atributos pertinentes dos irrelevantes.

A Tabela 1 faz uma comparação entre os atributos selecionados pelos especialistas e os aceitos pelo método de filtro, gerando uma porcentagem que informa o grau de concordância entre as escolhas. Os métodos "VDM", "MRMR", "FWE" e "FPR" consideraram que nenhum dos atributos era relevante para a variável resposta no problema estudado, o que sugere uma possível inadequação desses métodos para este conjunto de dados específico (Tabela 1). A Figura 2 apresenta a avaliação do custo computacional associado à escolha de cada atributo pelos diferentes métodos de filtro.

Houve diferenças consideráveis no tempo necessário para a seleção de atributos, com o gráfico destacando o tempo de processamento para cada método. No entanto, como a avaliação foi feita em segundos, a diferença de tempo de processamento nessa aplicação específica foi irrelevante.

Tabela 1. A porcentagem de alinhamento de escolhas de atributos entre o banco de dados original e as escolhas dos especialistas.

Métodos	% de escolha das variáveis do banco	% de escolha das variáveis do Especialista	Métodos	% de escolha das variáveis do banco	% de escolha das variáveis do Especialista
stepwiselm	81,82%	100,00%	DIRS Score	100,00%	100,00%
stepwiseglm	81,82%	100,00%	Information Gain Score	100,00%	100,00%
groupfs	45,45%	33,33%	Gini Index	100,00%	66,67%
graphfs	45,45%	100,00%	Fit Criterion Score	100,00%	66,67%
LS_L21 Score	45,45%	33,33%	Fehner Score	100,00%	66,67%
LL_L21 Score	45,45%	100,00%	fdr	100,00%	66,67%
Fisher Score	45,45%	33,33%	F-ratio	100,00%	66,67%
MIM	27,27%	33,33%	Chi2 Score	100,00%	66,67%
misfs	18,18%	33,33%	ANOVA Score	100,00%	66,67%
jmi	9,09%	33,33%	MIFS Score	100,00%	33,33%
icap	9,09%	33,33%	JMI Score	100,00%	100,00%
CMIM Score	100,00%	66,67%	ICAP Score	100,00%	66,67%
Backward selection	100,00%	33,33%	VDM; MRMR; FEW; FPR;	0%	0%

Os métodos de filtro "FWE", "Groupfs" e "FDR" destacaram-se por apresentarem os melhores tempos computacionais para processamento. No entanto, ao avaliarmos os resultados na Tabela 1, observamos que o método "FDR" teve um desempenho superior na escolha de atributos, selecionando as variáveis mais relevantes para a previsão do tempo de manutenção. Em contraste, foi possível verificar que alguns métodos consideraram que nenhum dos atributos era relevante para a predição, o que indica uma possível limitação ou inadequação desses métodos para este conjunto de dados específico. O método "Groupfs" selecionou um número menor de variáveis que impactavam na variável resposta e escolheu apenas uma pequena porcentagem (33%) de atributos considerados irrelevantes pelos especialistas, demonstrando uma eficiência equilibrada entre a escolha de atributos relevantes e a minimização de atributos irrelevantes.

A discussão entre as Figuras 2, 3 e 4 revela aspectos cruciais sobre a avaliação do desempenho do modelo de predição para o tempo de manutenção de itens ferroviários. A Figura 2 ilustra o custo computacional associado à seleção de cada atributo pelos métodos de filtro utilizados. Observou-se uma variação no tempo de processamento entre os diferentes métodos, o que destaca a necessidade de considerar

não apenas a precisão na escolha dos atributos, mas também o impacto do aumento significativo no volume de dados para a construção dos modelos.

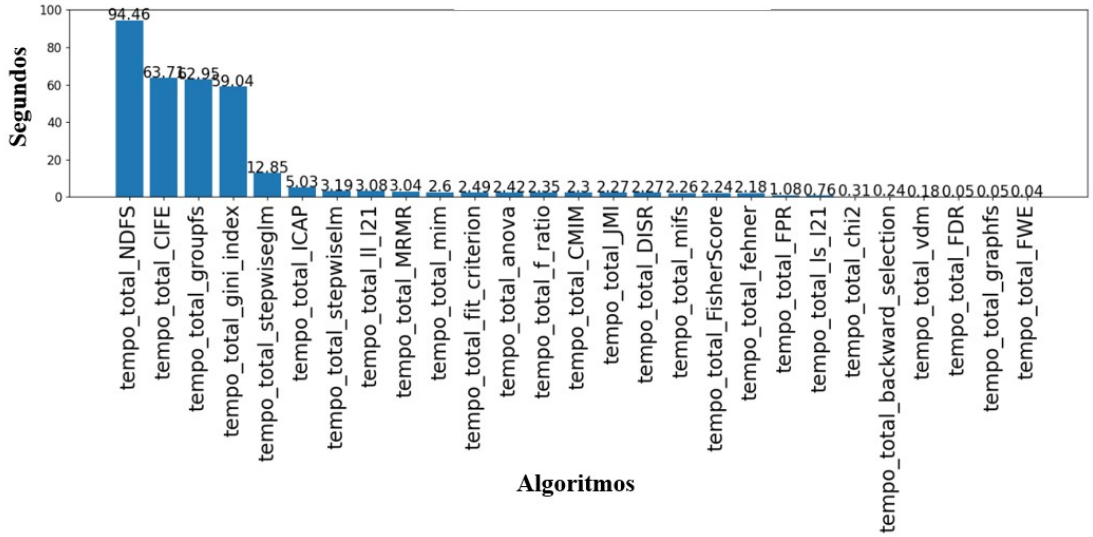

Figure 2. Avaliação do custo computacional da escolha de cada atributo

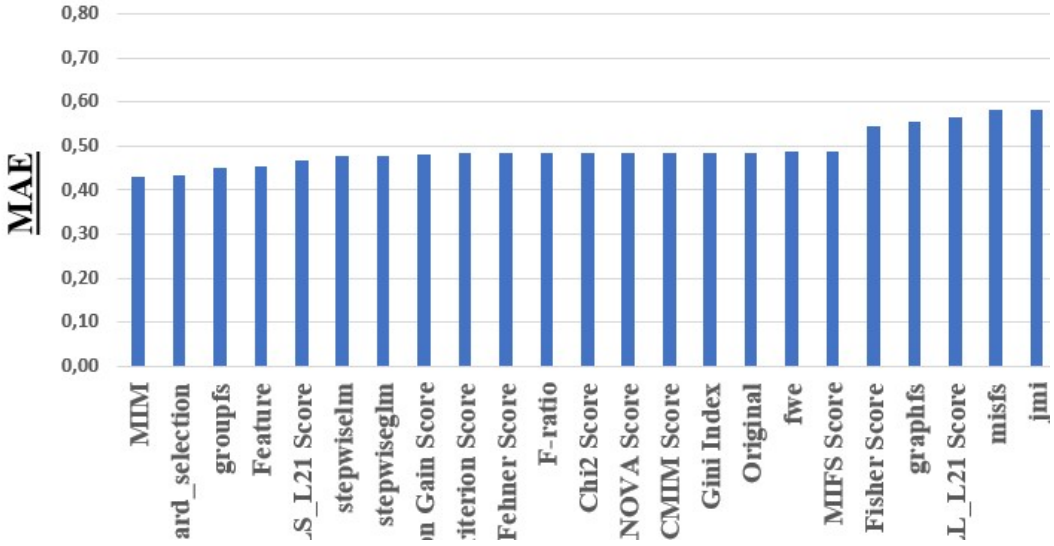

Figure 3. Avaliação do MAE após a construção do modelo.

Por outro lado, a Figura 3 apresenta a avaliação do MAE após a construção do modelo utilizando o algoritmo de árvore de decisão. Entre os métodos de filtro utilizados, MIM, backward_selection e Groupfs destacaram-se por alcançar o melhor MAE. Os atributos selecionados pelos especialistas, que têm um impacto direto no tempo de manutenção, mostraram-se associados a um MAE mais alto e a um MSE mais baixo, refletindo sua importância na precisão das previsões, como podemos observar na Figura 4. Um MAE alto, junto a um MSE baixo, indica que, embora alguns erros individuais sejam consideráveis em termos absolutos, a dispersão geral dos erros quadráticos é menor, sugerindo uma boa capacidade do modelo em ajustar-se aos dados de maneira consistente.

Analisando os resultados de forma abrangente, o método de filtro Groupfs emergiu como a escolha mais consistente e destacada em todas as análises realizadas. Este método não apenas proporcionou eficiência computacional na seleção de atributos,

mas também contribuiu significativamente para a precisão geral do modelo, representando uma abordagem robusta e eficaz para o problema de predição do tempo de manutenção ferroviária.

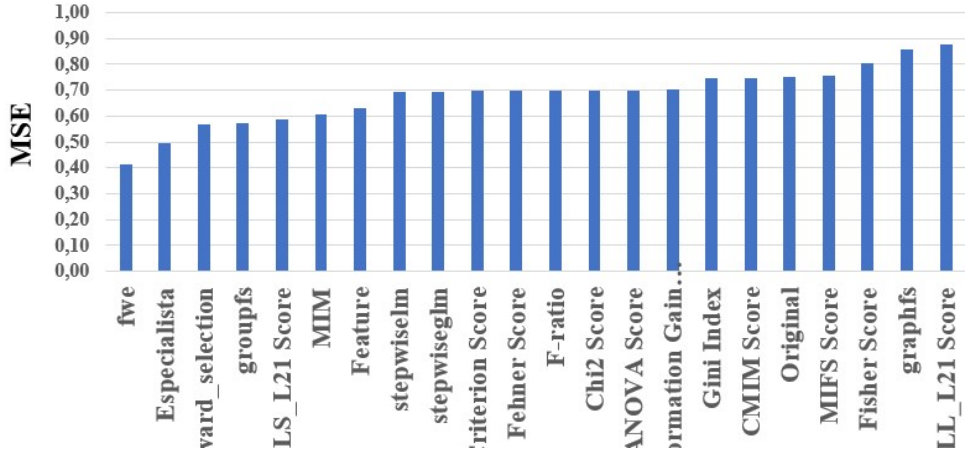

Figure 4. Avaliação do MSE após a construção do modelo.

4. Conclusão

Este estudo explorou métodos de seleção de atributos aplicados a um banco de dados de manutenção ferroviária, comparando suas eficácias e custos computacionais. Os resultados indicaram que é possível um equilíbrio significativo entre a eficiência computacional e a precisão dos modelos preditivos. A comparação com a seleção de atributos realizada por especialistas revelou que, embora as escolhas humanas sejam valiosas, os algoritmos de filtro podem identificar padrões e relações menos evidentes que contribuem para a melhoria da previsão do tempo de manutenção. A análise detalhada do MAE e do MSE colocou o algoritmo Groupfs como o mais consistente e eficaz. Os resultados obtidos podem orientar gestores ferroviários na tomada de decisões. Em suma, a aplicação de métodos de filtro na seleção de atributos mostrou-se uma estratégia eficaz para melhorar a tomada de decisão no setor ferroviário.

Futuros estudos podem expandir esta análise para outros contextos, como a operação de trens, com o objetivo de selecionar os atributos que integram produção e movimentação do material rodante, bem como as condições e necessidades de manutenção ferroviária. Pretende-se também expandir esta pesquisa, integrando o planejamento orçamentário da manutenção, de modo a orientar especialistas na alocação de custos humanos e materiais.

5. Referências

Alyasiri, O. M., Cheah, Y. N., Abasi, A. K., & Al-Janabi, O. M. (2022). Wrapper and hybrid feature selection methods using metaheuristic algorithms for English text classification: A systematic review. IEEE Access, 10, 39833-39852.

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics & Data Analysis, 143, 106839.

Hou, C. K. J., & Behdinan, K. (2022). Dimensionality reduction in surrogate modeling: A review of combined methods. Data Science and Engineering, 7(4), 402-427.

Felix, J. C., Oliveira, V. M., & Silva, R. (2022, November). A Machine Learning with an Inlier/Outlier Separation Approach for the Prediction of Wagon Maintenance Times. In Anais do X Symposium on Knowledge Discovery, Mining and Learning (pp. 9-16). SBC.

Rao, P. V., & Baral, S. S. (2011). Attribute based specification, comparison and selection of feed stock for anaerobic digestion using MADM approach. Journal of Hazardous Materials, 186(2-3), 2009-2016

Silva, L. R., & Nascimento, D. C. (2023). Avaliando o Processo de Seleção de Características na Tarefa de Junção de Similaridade. In Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados (pp. 348-353). SBC.

Parmezan, A. R. S., Lee, H. D., & Wu, F. C. (2017). Metalearning for choosing feature selection algorithms in data mining: Proposal of a new framework. Expert Systems with Applications, 75, 1-24.

Penha, G., Cardoso, T. N., da Silva, A. P. C., & Moro, M. M. (2016). Análise de métodos de Inferência Ecológica em dados de redes sociais. In Anais do XXXI Simpósio Brasileiro de Bancos de Dados (pp. 109-114). SBC.

Pilnenskiy, N., & Smetannikov, I. (2020). Feature selection algorithms as one of the python data analytical tools. Future Internet, 12(3), 54.