PREDIÇÃO DA EFICIÊNCIA DE COMBUSTÍVEL DE VEÍCULOS UTILIZANDO *MACHINE LEARNING*E MODELAGEM ESTATÍSTICA: UMA ANÁLISE COMPARATIVA ENTRE RANDOM FOREST E REGRESSÃO LINEAR MÚLTIPLA

Eduarda da Silva Brito[1], Romário de Sousa Almeida[1,2], Maria Helena Ramos de Macedo[1], Arthur Victor Ramos Vidal Negreiros[1], Elizeu Silva de Medeiros[1], José Wanderson Ferreira de Freitas[1]

[1]Universidade Estadual da Paraíba – UEPB, Campina Grande, Brasil
(eduarda.brito@aluno.uepb.edu.br)
[2]Universidade Federal de Lavras - UFLA, Lavras, Brasil

Resumo: Este estudo compara o desempenho preditivo de modelos estatísticos e de aprendizado de máquina na estimativa da eficiência de combustível de veículos, utilizando o conjunto de dados Auto MPG. Foi adotada a regressão linear múltipla como modelo de referência e o algoritmo Random Forest para capturar relações não lineares e interações entre variáveis. O pré-processamento envolveu a imputação de valores ausentes, codificação de variáveis categóricas e divisão em conjuntos de treinamento (70%) e teste (30%). Os modelos foram ajustados utilizando validação cruzada k-fold repetida (10 partições e 3 repetições). O desempenho preditivo foi avaliado por meio de métricas estatísticas. Os resultados demonstraram que o Random Forest apresentou maior precisão preditiva e melhor capacidade de generalização em relação ao modelo linear. Adicionalmente, análises de superfície de resposta evidenciaram interações relevantes entre peso e potência do motor, destacando a importância de abordagens não lineares na modelagem do consumo de combustível. Os achados reforçam o potencial de técnicas de aprendizado de máquina para aplicações em eficiência energética e sistemas de transporte.

Palavras-chave: Análise de regressão; Aprendizado de máquina; Modelagem do consumo de energia; inteligência artificial.

INTRODUÇÃO

A crescente demanda por eficiência energética, aliada à necessidade de mitigação das emissões de gases poluentes, tem impulsionado o desenvolvimento de tecnologias voltadas à otimização do consumo de combustível em veículos automotores (IEA, 2023).

A avaliação da eficiência energética é estratégica tanto para o setor industrial quanto para a formulação de políticas ambientais, uma vez que o consumo de combustível está diretamente associado a impactos econômicos e ambientais relevantes no setor de transportes (YOO *et al.*, 2025).

Nesse contexto, a aplicação de técnicas avançadas de modelagem e *machine learning* tem se destacado como uma abordagem promissora para a análise do desempenho energético, permitindo avanços significativos na predição da eficiência de combustível (ANANTHA*et al.*, 2025).

Com o avanço das tecnologias digitais e da crescente disponibilidade de dados, métodos baseados em *machine learning* têm sido amplamente empregados na modelagem e predição de variáveis relacionadas ao desempenho energético. Taisabordagens permitem identificar padrões complexos e capturar relações não lineares entre variáveis, superando limitações frequentemente observadas em métodos estatísticos tradicionais (MAHMOOD *et al*., 2024).

Nesse sentido, estudos recentes como o de Nguyen *et al.* (2025) destacam a integração de técnicas avançadas de modelagem em sistemas energéticos, enquanto Arévalo-Correra *et al*. (2024) evidencia a aplicação de algoritmos de aprendizado de máquina na análise de sistemas energéticos em veículos, reforçando a relevância dessas abordagens na área.

Apesar desses avanços, modelos estatísticos clássicos, como a regressão linear múltipla, permanecem amplamente utilizados devido à sua simplicidade, interpretabilidade e baixo custo

computacional. Esses modelos possibilitam aavaliação direta da influência das variáveis explicativas sobre a variável resposta, sendo frequentemente empregados como referência em estudos comparativos (PEINADO *et al*., 2025).

Nesse sentido, abordagens tradicionais baseadas em modelagem estatística e simulação continuam relevantes na análise do desempenho energético de veículos (SONG*et al.*,2022). Entretanto, evidências recentes indicam que modelos de *machine learning* apresentam desempenho superior na modelagem de relações complexas e não lineares (WANG*et al.*,2024), destacando limitações inerentes aos modelos lineares em cenários de alta complexidade.

Entre os algoritmos de *machine learning*, o Random Forest destaca-se pela elevada capacidade preditiva e robustez frente a dados ruidosos e não lineares. Esse método baseia-se na combinação de múltiplas árvores de decisão, permitindo modelar interações complexas entre variáveis e reduzir a variância das previsões (MUTALE *et al*., 2024).

Estudos recentes demonstram que técnicas de *machine learning* e *deep learning* apresentam alto desempenho na predição do consumo de combustível e das emissões veiculares, especialmente em cenários com dados reais e alta variabilidade (ÇEKEN; GEMCI, 2025). Além disso, revisões sistemáticas indicam a ampla aplicação dessas abordagens na gestão eficiente do consumo energético, destacando ganhos consistentes em precisão preditiva (ADNANE, *et al.*, 2023).

Paralelamente, pesquisas na área de mobilidade e sistemas energéticos têm explorado o uso de técnicas avançadas de modelagem na análise de desempenho de veículos, incluindo sistemas híbridos e elétricos. Nesse contexto, Arıkuşu*et al.* (2023) demonstram que métodos de *machine learning* contribuem significativamente para a estimativa do consumo energético em veículos híbridos, promovendo melhorias na eficiência energética e no suporte à tomada de decisão.

Diante desse cenário, o presente estudo tem como objetivo comparar o desempenho preditivo dos modelos de Regressão Linear Múltipla e Random Forest na estimativa da eficiência de combustível de veículos.

MATERIAL E MÉTODOS

Base de dados e área de estudo

O presente estudo utilizou o conjunto de dados *Auto MPG Dataset*, amplamente empregado em pesquisas de *machine learning* voltadas à análise da eficiência energética de veículos. Os dados foram obtidos a partir do *UCI Machine Learning Repository*, um dos principais repositórios utilizados na literatura científica para desenvolvimento e validação de modelos preditivos (DUA; GRAFF, 2019).
O dataset foi originalmente compilado por Quinlan (1993) a partir de informações da indústria automotiva e contém aproximadamente 398 observações referentes a características mecânicas e de desempenho veicular. Entre asvariáveis disponíveis destacam-se: número de cilindros, deslocamento do motor, potência, peso do veículo, aceleração, ano do modelo e origem, além da variável respostarepresentada pelo consumo de combustível (MPG).

O problema foi formulado como uma tarefa de regressão supervisionada, na qual o consumo de combustível é estimado com base em variáveis explicativas associadas às características físicas e mecânicas dos veículos. Essa abordagem é amplamente adotada em estudos de eficiência energética, modelagem automotiva e aplicações de aprendizado de máquina voltadas à predição do consumo de combustível.

A variável MPG foi definida como variável dependente, enquanto as demais variáveis foram consideradas preditoras. A Tabela 1 apresenta a descrição das variáveis utilizadas.

Tabela 1. Descrição das variáveis do conjunto de dados Auto MPG.

Variável	**Descrição**
Consumo de combustível (MPG)	Consumo de combustível do veículo
Cilindros	Número de cilindros do motor
Descolacamento (in^3)	Volume do motor
Potência (HP)	Potênciadesenvolvida pelomotor
Peso (libras)	Peso do veículo
Aceleração (s)	Tempo deaceleração
Ano do modelo	Ano de fabricação
Origem (país)	País de origem

Fonte: Adaptado de UCI Machine Learning Repository; Ross Quinlan (1993).

Pré-processamento dos dados

Após a obtenção do conjunto de dados, foi realizada uma etapa de pré-processamento com o objetivo de garantir a qualidade dos dados e melhorar o desempenho dos modelos. Essa etapa incluiu tratamento de valores ausentes, codificaçãode variáveis categóricas, normalização dos dados e a divisão do conjunto de dados em subconjuntos de

treinamento e teste, procedimentos amplamente utilizados em pipelines de modelagem preditiva (KOUKARAS; TJORTJIS, 2025).

Inicialmente, foi realizada uma verificação de valores ausentes, sendo observada a presença de dados faltantes na variável potência. os quais foram substituídos pela mediana da variável. A utilização da mediana foi adotada por ser uma medida de tendência central robusta, menos sensível à presença de *outliers*. Esse procedimento é comum em etapas de pré-processamento, uma vez que conjuntos de dados reaisfrequentemente apresentam valores ausentes e requerem técnicas de imputação antes da modelagem (JOEL *et al.*, 2025).

A variável *origin* foi tratada como categórica, sendo convertida em fatores representando as categorias Estados Unidos, Europa e Japão, permitindo melhor interpretação e utilização nos modelos.

Análise exploratória dos dados

A análise exploratória dos dados (EDA) foi conduzida com o objetivo de compreender a estrutura do conjunto de dados, identificar padrões e avaliar relações entre as variáveis. Foram utilizadas estatísticas descritivas, análise de correlação e visualizações gráficas, conforme recomendações clássicas da literatura (HAN*et al.*, 2022).

A Figura 1 apresenta o boxplot das variáveis numéricas após a padronização dos dados, procedimento realizado para permitir a comparação entre variáveis originalmente mensuradas em diferentes escalas.

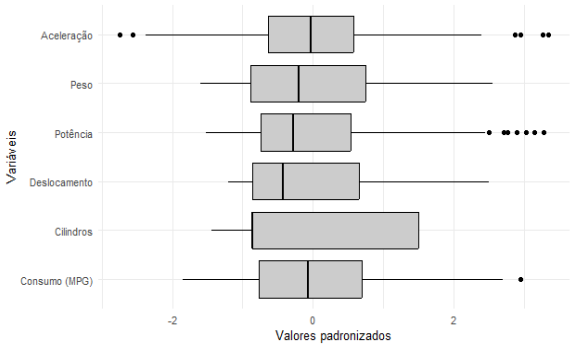

Figura 1. Boxplot das variáveis numéricas padronizadas.

O uso de boxplots é amplamente empregado na análise exploratória de dados para a visualização da distribuição e identificação de valores discrepantes (TUKEY, 1977).

Os boxplots das variáveis padronizadas evidenciaram diferenças significativas na dispersão dos dados, com destaque para as variáveis peso, potência e deslocamento, que apresentaram maior variabilidade. Além disso, foram identificados valores discrepantes (*outliers*), especialmente nas variáveis potência e aceleração, indicando a presença de observações com características extremas.

A matriz de correlação revelou associações relevantes entre as variáveis quantitativas do conjunto de dados (Figura 2).

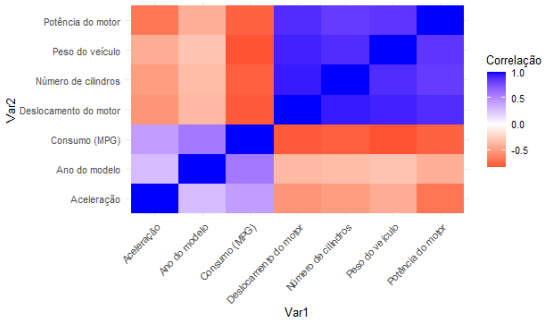

Figura2. Matriz de correlação entre as variáveis numéricas do conjunto de dados.

Observou-se correlação negativa entre MPG e variáveis como peso, potência e deslocamento, indicando que veículos mais pesados e com motores maiores tendem a apresentar menor eficiência energética. Em contrapartida, o ano do modelo apresenta correlação positiva com o consumo de combustível, sugerindo que veículos mais recentes tendem a ser mais eficientes.

Análises semelhantes são encontradas na literatura, que utiliza matrizes de correlação para investigar relações entre variáveis do conjunto de dados Auto MPG (SARKAR; HASAN, 2025).

A Figura 3 apresenta a relação entre o peso do veículo e o consumo de combustível (MPG).

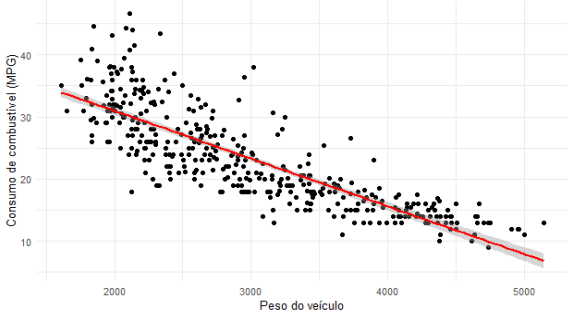

Figura 3. Relação entre o peso do veículo e o consumo de combustível (MPG).

A análise de dispersão entre peso e MPG evidenciou uma relação inversa bem definida, reforçando a

influência significativa do peso do veículo no consumo de combustível.

Esse comportamento é consistente com a análise exploratória realizada com o conjunto de dados Auto MPG, nas quais gráficos de dispersão mostram evidencia relações negativas entre características do veículo e o consumo de combustível (DEAN; FRANCIS, 2024).

Modelo de Regressão Linear Múltipla

A regressão linear múltipla foi utilizada como modelo de referência para comparação com técnicas de *machine learning*. Esse modelo assume uma relação linear entre a variável resposta e um conjunto de variáveis explicativas, sendo expresso por:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad (1)$$

Em que:$y_i$representa o valor observado da variável resposta para a i-ésima observação,x_{ij}Ccorresponde ao valor da j-ésima variável explicativa para a i-ésima observação,β_0 É o intercepto do modelo, $\beta_1, \beta_2, \ldots, \beta_p$são os coeficientes de regressão associados às variáveis explicativas e ε_iRrepresenta o termo de erro aleatório do modelo (SUN et al., 2023).

A seleção do modelo foi realizada com base no Critério de Informação de Akaike (AIC), que considera simultaneamente a qualidade do ajuste e a complexidade do modelo (AKAIKE, 1974) (Equação 2).

$$\mathrm{AIC} = -2\ln(\mathrm{L}) + 2k \ (2)$$

Em que *L*representa a função de verossimilhança do modelo e *k*corresponde ao número de parâmetros estimados.

Foi aplicado o procedimento*backwardelimination*, no qual as variáveis menos significativassão removidas progressivamente, até a obtenção de um modelo parcimonioso (DAJLES; CAVANAUGH, 2024; RIANSUT, 2022).

Modelos de *machine learning*: Random Forest

O Random Forest é um algoritmo de aprendizado de máquina baseado em conjuntos de árvores de decisão, amplamente utilizado em problemas de regressão e classificação. Proposto por Breiman (2001), o método pertence à classe de técnicas de *ensemble learning*, nas quais múltiplos modelos são combinados com o objetivo de melhorar a capacidade preditiva e a robustez das previsões.

O algoritmo constrói diversas árvores de decisão a partir de subconjuntos do conjunto de dados gerados por meio de amostragem bootstrap. Durante a construção das árvores, apenas um subconjunto aleatório das variáveis explicativas é considerado em cada divisão, o que reduz a correlação entre as árvores e aumenta a diversidade do modelo. As previsões individuais são então agregadas para produzir a estimativa final, contribuindo para a redução da variância e tornando o modelo menos suscetível ao sobreajuste (JAGATHA; SCHNEIDER; SAUTER, 2024).

De forma geral, a predição do modelo Random Forest pode ser expressa como a média das previsões individuais geradas pelas árvores de decisão que compõem o conjunto (GUEZOULI et al., 2024) (Equação 3):

$$\hat{y}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad (3)$$

Em que $T_b(x)$representa a previsão da b-ésima árvore de decisão para a observação x, e B corresponde ao número total de árvores no modelo.
Estratégia de validação

Para avaliar a capacidade preditiva e a generalização dos modelos propostos, foi adotada uma estratégia de validação baseada na separação dos dados em conjuntos de treinamento e teste (*hold-out validation*). O conjunto de dados foi particionado aleatoriamente, sendo 70% das observações destinadas ao treinamento dos modelos e os 30% restantes reservados para a avaliação externa (*test set*), conforme prática consolidada em estudos de machine learning para avaliação da capacidade de generalização dos modelos (JAMES *et al.*, 2021).

O conjunto de treinamento foi utilizado para o ajuste dos modelos de Regressão Linear Múltipla e Random Forest, permitindo a identificação dos padrões subjacentes aos dados. Com intuito de aumentar a robustez das estimativas e reduzir a variabilidade associada à seleção amostral, foi empregada validação cruzada repetida do tipo *k-fold*, com 10 partições e 3 repetições. Essa abordagem é amplamente recomendada na literatura por proporcionar estimativas mais estáveis e confiáveis do desempenho preditivo (XU; GOODACRE, 2018).

Após o processo de treinamento e validação interna, os modelos ajustados foram avaliados no conjunto de teste, composto por observações independentes não utilizadas durante o treinamento. Essa etapa permite uma avaliação imparcial da capacidade de

generalização dos modelos, evitando viés otimista nas estimativas de desempenho.

O desempenho preditivo foi então quantificado por meio de métricas estatísticas de erro e ajuste, possibilitando a comparação entre os modelos considerados (MONTGOMERY; PECK; VINING, 2012).

Diagnóstico estatístico

Após o ajuste do modelo de Regressão Linear Múltipla, foram realizados procedimentos de diagnóstico estatístico para verificar o atendimento dos pressupostos fundamentais do modelo, condição essencial para a validade das inferências e da interpretação dos coeficientes estimados.

A análise dos resíduos foi realizadapor meio do gráfico de resíduos *versus* valores ajustados, visando identificar padrões indicativos de não linearidade ou heterocedasticidade. A suposição de normalidade dos resíduos foi avaliada por meio do gráfico Q-Q plot (*quantile-quantileplot*) e da inspeção do histograma dos resíduos (LI *et al.*, 2024).

Adicionalmente, foram aplicados testes estatísticos formais. A presença de heterocedasticidade foi investigada pelo teste de Breusch-Pagan, enquanto a independência dos resíduos foi avaliada pelo teste de Durbin-Watson, para detectar autocorrelação dos resíduos (ZHANG et al., 2020).
Análise de *overfitting*

A análise de *overfitting* foi conduzidacom o objetivo de avaliar a capacidade de generalização dos modelos ajustados. O *overfitting* ocorre quando um modelo se ajusta excessivamente aos dados de treinamento, capturando inclusive ruídos presentes no conjunto de dados, o que compromete seu desempenho em dados não observados. Esse fenômeno constitui um dos principais desafios na construção de modelos preditivos confiáveis (GOODFELLOW; BENGIO; COURVILLE, 2016).

Para investigar a ocorrência desse fenômeno, foi realizada a comparação entre o desempenho dos modelos nos conjuntos de treinamento e teste, utilizando o coeficiente de determinação (R^2) como métrica principal. Diferenças substanciais entre os valores obtidos nesses conjuntos foram interpretadas como indicativas de possível sobreajuste.

Essa abordagemé amplamente utilizada para verificar a robustez e a capacidade de generalização de modelos preditivos, sendo particularmente relevante em algoritmos baseados em árvores, como o Random Forest, que podem apresentar elevado desempenho no conjunto de treinamento caso não sejam devidamente validadosem dados independentes (TETKO et al., 2024).

Métricas de Avaliação

Foram utilizadas métricas estatísticas amplamente utilizadas para avaliar e comparar o desempenho de modelos preditivos (HODSON, 2022; DIAMANTOPOULOU; PAPAMICHAIL, 2024).São elas: o coeficiente de determinação (R^2), o erro quadrático médio (MSE), a raiz do erro quadrático médio (RMSE), o erro absoluto médio (MAE), o erro percentual absoluto médio (MAPE), o erro médio de viés (MBE), a estatística t associada ao viés (tStat) e coeficiente de eficiência de Nash-Sutcliffe (NSE) (Equações 4 a 11, respectivamente).

$$R^2 = \left[\frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}\right]^2 \quad (4)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (7)$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (8)$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \quad (9)$$

$$t\ Stat = \frac{MBE}{RMSE/\sqrt{n}} \quad (10)$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (11)$$

Onde $\hat{y}_i$ é o valor predito pelo modelo, y_i é o valor observado, n é o número de observações,$\bar{y}$ é a média dos valores observados e $\bar{\hat{y}}$representa a média dos valores previstos.

Todas as análises foram realizadas utilizando o pacote de *software* R versão 4.4.1 (R CORE TEAM, 2025) com o RStudio IDE.

RESULTADOS E DISCUSSÃO

Regressão Linear Múltipla

A Tabela 2 apresenta as métricas de desempenho do modelo de regressão linear múltipla no conjunto de teste. Observa-se que o modelo apresentou coeficiente de determinação R^2=0,8057, indicando que aproximadamente 80,6% da variabilidade do consumo de combustível foi explicada pelas variáveis incluídas no modelo. Os valores de RMSE (3,4128) e MAE (2,7146) indicam erros de predição relativamente baixos, sugerindo desempenho satisfatório do modelo estatístico.

Tabela 2. Métricas de desempenho do modelo de regressão linear múltipla no conjunto de teste.

Modelo	Valores
R^2	0,8057
RMSE	3,4128
MAE	2,7146
MAPE	12,6881
MBE	0,1485
MSE	11,6472
t Stat	0,4710
NSE	0,8052

R^2: coeficiente de determinação; RMSE: raiz do erro quadrático médio; MAE: erro absoluto médio; MAPE: erro percentual absoluto médio; MBE: erro médio de viés; MSE: erro quadrático médio; t Stat: estatística t do viés; NSE: coeficiente de eficiência de Nash-Sutcliffe.

A Figura 5 apresenta a comparação entre os valores observados (Real) e os valores previstos (Previsto) pelo modelo de regressão linear múltipla para o consumo de combustível no conjunto de teste.

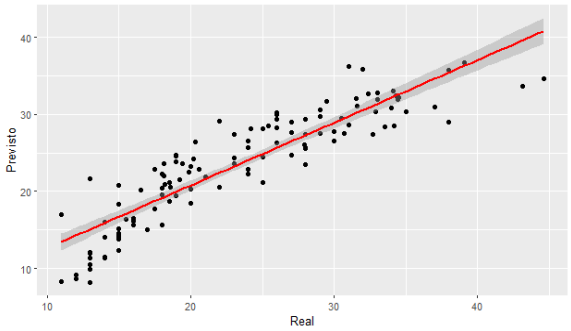

Figura 5. Valores observados versus valores previstos pelomodelo de regressão linear múltipla.

Observa-se que a maior parte dos pontos se distribui próxima à linha de tendência, sugerindo boa concordância entre os valores observados e estimados. Esse comportamento indica que o modelo apresenta capacidade satisfatória de ajuste e boa capacidade preditiva para estimar o consumo de combustível a partir das características mecânicas dos veículos.

Além disso, não se observam padrões sistemáticos ou desvios extremos relevantes, sugerindo que o modelo representa adequadamente a relação entre as variáveis analisadas.

Estudos recentes que utilizam o conjunto de dados Auto MPG também empregam esse tipo de representação gráfica para analisar a capacidade de previsão de modelos de regressão e aprendizado de máquina aplicada à eficiência energética de veículos, apresentando resultados consistentes com os observados neste estudo (ÇEKEN; GEMCI, 2025).

A Figura 6mostra o gráfico dos resíduosemfunção dos valoresajustados pelo modelo de regressãolinearmúltipla.

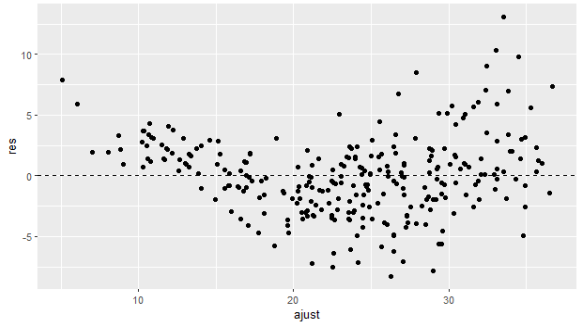

Figura 6. Resíduos em função dos valores ajustados para o modelo de regressão linear múltipla.

A análise dos resíduosindicouque os resíduos se distribuem de forma relativamentealeatóriaemtorno da linha horizontal emzero, sem a presença de padrõessistemáticosevidentes.

Esse comportamento sugere que o modelo atende razoavelmente ao pressuposto de homocedasticidade. Embora alguns pontos apresentem maior dispersão, especialmenteemvalores mais elevados, não se observa presença de estruturasclaras que indiqueminadequação do modelo.

A análisegráfica dos resíduos constitui um procedimento importante na avaliação da adequação de modelos de regressão, pois permite verificar possíveis violações dos pressupostos do modelo (KUTNER *et al.*, 2005).

A Figura 7 apresenta o gráfico Q-Q dos resíduos do modelo de regressãolinearmúltipla, utilizado para avaliar a normalidade dos erros.

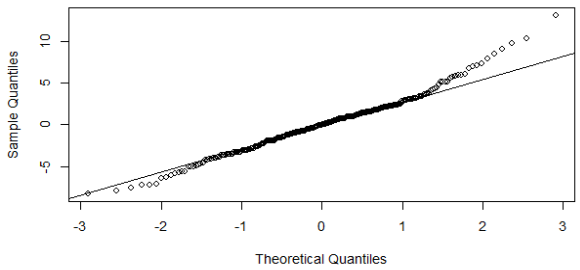

Figura 7. Q-Q plot dos resíduos do modelo de regressãolinearmúltipla.

Observa-se que a maioria dos pontos está próxima à linha de referência, indicando distribuição aproximadamente normal.

Pequenos desvios nas extremidades são observados, mas não comprometem a adequação do modelo. A análise por meio do gráfico Q-Q é amplamente utilizada para identificar desvios da normalidade (ADEBISI *et al.*, 2024).

De forma geral, esses resultados indicam adequação global do modelo, embora evidenciem possíveis limitações associadas à linearidade da abordagem.

ModelosRandom Forest

O modelo Random Forest foi ajustado com o objetivo de capturar relações não lineares e interações complexas entre as variáveis explicativas. No conjunto de teste, o modelo apresentou R^2=0,8390, indicando melhoria em relação ao modelo linear, com aproximadamente 83,9% da variabilidadedo consumo de combustívelexplicada pelo modelo.

Tabela 3. Métricas de desempenho do modelo Random Forest no conjunto de teste.

Modelo	Valores
R^2	0,8390
RMSE	3,1678
MAE	2,0021
MAPE	8,1156
MBE	-0,4410
MSE	10,0349
t Stat	-1,5206
NSE	0,8322

R^2: coeficiente de determinação; RMSE: raiz do erro quadrático médio; MAE: erro absoluto médio; MAPE: erro percentual absoluto médio; MBE: erro médio de viés; MSE: erro quadrático médio; t Stat: estatística t do viés; NSE: coeficiente de eficiência de Nash-Sutcliffe.

Adicionalmente, foram observadas reduções consistentes nas métricas de erro (RMSE = 3,1678; MAE = 2,0021; MAPE = 8,1156), evidenciando maior precisão preditiva do modelo Random Forest.

A Figura 8apresenta a comparação entre os valores observados e previstos pelo modelo Random Forest.

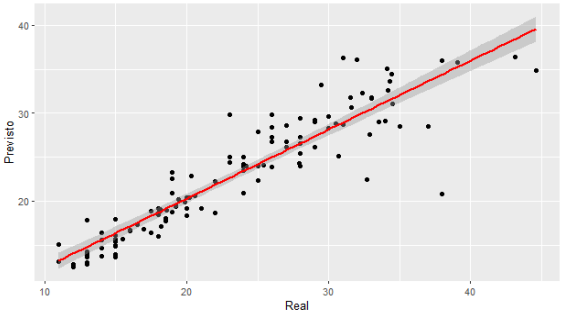

Figura 8. Valores observados versus previstos pelo modelo Random Forest.

Observa-se que os pontos se encontram mais próximos da linha de referência, com menor dispersão em relação ao modelo de regressão linear, indicando um melhor ajuste global. Esse comportamento sugere que o modelo apresenta maior capacidade de generalização e melhor desempenho preditivo.

Apesar do desempenho superior, ainda é possível observar alguns desvios em valores mais elevados de consumo, o que indica que, embora o modelo seja mais flexível, ainda existem limitações na captura completa da variabilidade dos dados.

De forma geral, os resultados evidenciam que o Random Forest apresenta desempenho superior à regressão linear múltipla, especialmente devido à sua capacidade de modelar relações não lineares e interações entre variáveis, conforme discutido por AurélienGéron (2022), tornando-se uma abordagem mais adequada para a predição do consumo de combustível de veículos.

Esse resultado está em concordância com estudos recentes que demonstram a superioridade de modelos de aprendizado de máquina em relação a abordagens lineares em problemas com relações complexas entre variáveis (GUO*et al.*, 2024).

Os menores valores de erro do Random Forest em comparação ao modelo estatístico, estão alinhados com a literatura, na qual métodos de *machine learning* frequentemente apresentam menores erros de predição em comparação com modelos tradicionais de regressão (TZOU*et al.*, 2023).

Capacidade de Generalização

A capacidade de generalização foi avaliada por meio da comparação do desempenho nos conjuntos de treinamento e teste. Observou-se que o modelo Random Forest apresentou valores de R^2 = 0,8390 semelhantes entre os conjuntos, indicando boa capacidade de generalização e ausência de overfitting significativo.

Esses resultados sugerem que o modelo é capaz de realizar previsões consistentes em dados não observados, conforme discutido na literatura sobre aprendizado de máquina (GOODFELLOW*et al.*, 2016).

Análise de superfície de resposta

A análise de superfícies de resposta foi empregada para investigar o efeito conjunto das variáveis explicativas sobre o consumo de combustível.

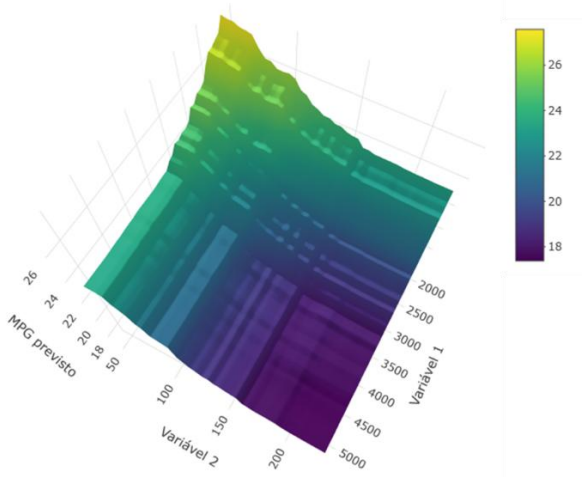

Figura 9. Superfície de resposta do consumo de combustível (MPG) em função do peso e da potência do motor, estimada pelo modelo Random Forest.

A Figura 9 apresenta a superfície gerada pelo modelo Random Forest para a variável peso do veículo e potência do motor.

Observa-se que o consumo de combustível (MPG) diminui de forma consistente com o aumento do peso e da potência do motor, evidenciando uma relação inversa entre essas variáveis e a eficiência energética dos veículos.

Além disso, a superfície apresenta comportamento não linear, com variações mais acentuadas em determinadas regiões, indicando que o efeito das variáveis não ocorre de maneira uniforme ao longo de seus domínios.

A inclinação da superfície também evidencia a presença de interação entre peso e potência, sugerindo que o impacto de uma variável sobre o consumo depende do nível da outra.

Estes resultados reforçam a capacidade do modelo Random Forest em capturar padrões complexos e não lineares, justificando seu desempenho superior em comparação à regressão linear múltipla.

CONCLUSÃO

Os resultados demonstraram que ambos os modelos avaliados foram capazes de prever o consumo de combustível de forma satisfatória. No entanto, o modelo Random Forest apresentou desempenho superior, com maior capacidade explicativa e menores erros de predição.

A superioridade do Random Forest está associada à sua habilidade de capturar relações não lineares e interações entre variáveis, enquanto a regressão linear múltipla se destaca pela simplicidade e interpretabilidade.

Os achados indicam que variáveis como peso e potência exercem influência significativa na eficiência de combustível, sendo melhor representadas por modelos mais flexíveis. Como limitação, destaca-se o uso de um único conjunto de dados, o que pode restringir a generalização dos resultados.

Como perspectivas futuras, recomenda-se a aplicação de modelos mais avançados, bem como o uso de técnicas modernas de interpretabilidade (*e.g.*, SHAP), visando aprofundar a compreensão dos padrões identificados e ampliar a robustez das análises.

AGRADECIMENTOS

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Conselho Nacional de DesenvolvimentoCientífico e Tecnológico (CNPq), pela concessão de bolsas de pesquisa.

À Universidade Estadual da Paraíba (UEPB), pelosuporte institucional.

REFERÊNCIAS

ADEBISI, O. D**.** *et al.* Residual diagnostics and regression assumptions analysis. *Open Journal of Social Sciences*, 12, 548–573, 2024.

ADNANE, M.; KHOUMSI, A.; TROVÃO, J. P. F. Efficient management of energy consumption of electric vehicles using machine learning: a

systematic and comprehensive survey. *Energies*, 16(13), 4897, 2023.

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723, 1974.

ANANTHA, S**.** *et al.* Fuel efficiency analysis using machine learning approach. *International Journal of Vehicle Structures & Systems*, 17(3), 532–538, 2025.

ARIKUŞU, Y. S.; BAYHAN, N.; TIRYAKI, H. Determinação da economia de energia por meio da estimativa do consumo de combustível com métodos de aprendizado de máquina. *Energy*, 272, 127182, 2023.

ARÉVALO CORRERA, P. *et al.* A Systematic Review on the Integration of Artificial Intelligence into Energy Management Systems for Electric Vehicles: Recent Advances and Future Perspectives. *World Electric Vehicle Journal*, 15(8), 364, 2024. Disponível em: https://www.mdpi.com/2032-6653/15/8/364. Acesso em: 30 mar. 2026.

BREIMAN, L. Random forests. *Machine Learning*, 45(1), 5–32, 2001.

ÇEKEN, H.; GEMCI, C. Fuel efficiency and emission prediction using machine learning. *Journal of Sustainable Computing and Artificial Intelligence*, 2025.

DAJLES, A.; CAVANAUGH, J. Bootstrap approximation of model selection probabilities for multimodel inference frameworks. *Entropy*, 26(7), 2024.

DEAN, J.; FRANCIS, P. Predicting fuel efficiency using the Auto MPG dataset. *Theoretical and Engineering*, 2024.

DIAMANTOPOULOU, M. J.; PAPAMICHAIL, D. M. Performance evaluation of regression-based machine learning models. *Hydrology*, 11, 2024.

DUA, D.; GRAFF, C. UCI Machine Learning Repository. Irvine: University of California, 2019.

GÉRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Sebastopol, 2022.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge: MIT Press, 2016.

GUO, F**.** *et al.* Machine learning vs. statistical model for prediction modeling. *Journal of Hazardous Materials*, 469, 2024.

GUEZOULI, L. *et al.* IoT and AI for real-time water monitoring and leak detection. *Revue des Energies Renouvelables*, 2024.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 4. ed. Burlington: Morgan Kaufmann, 2022.

HODSON, T. O. RMSE or MAE: when to use them or not. *Geoscientific Model Development*, 15, 5481–5487, 2022.

JAGATHA, J. V.; SCHNEIDER, P.; SAUTER, D. Random-forest-based regression model using sensors. *Sensors*, 24(13), 4193, 2024.

JAMES, G. *et al.* An Introduction to Statistical Learning: with Applications in R. 2. ed. New York: Springer, 2021.

JOEL, L. O.; DOORSAMY, W.; PAUL, B. S. Comparative study of imputation techniques. *Applied Artificial Intelligence*, 2025.

KOUKARAS, P.; TJORTJIS, C. Data preprocessing and feature engineering. *AI*, 6(10), 257, 2025.

KUTNER, M. H. et al. *Applied Linear Statistical Models*. 5. ed. Boston: McGraw-Hill, 2005.

LI, W. *et al.* A plot is worth a thousand tests. *Journal of Computational and Graphical Statistics*, 2024.

MAHMOOD, S.; SUN, H.; ALI ALHUSSAN, A. *et al.* Abordagem de aprendizado de máquina baseada em aprendizado ativo para aprimorar a sustentabilidade ambiental no consumo de energia em edifícios verdes. *Scientific Reports*, 14, 19894, 2024. https://doi.org/10.1038/s41598-024-70729-4.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to Linear Regression Analysis*. 5. ed. Hoboken: Wiley, 2012.

MUTALE, B.; WITHANAGE, N. C.; MISHRA, P. K.; SHEN, J.; ABDELRAHMAN, K.; FNAIS, M. S. A performance evaluation of random forest, artificial neural network, and support vector machine learning algorithms to predict spatio-temporal land use-land cover dynamics: a case from Lusaka and Colombo. *Frontiers in Environmental Science*, 12, 1431645, 2024. doi: 10.3389/fenvs.2024.1431645.

NGUYEN, V. N. *et al.* IoT-driven approach integrated with machine learning. *Alexandria Engineering Journal*, 118, 664–680, 2025.

PEINADO, L. B. *et al.* Statistical and Machine Learning Models for Air Quality: comparison and relevance of classical statistical models. *Algorithms*, 18(12), 783, 2025. Disponível em: https://www.mdpi.com/1999-4893/18/12/783. Acesso em: 30 mar. 2026.

QUINLAN, R. Combining instance-based and model-based learning. *Proceedings of ICML*, 1993.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing, 2025. Disponível em: https://www.R-project.org/. Acesso em: 27 jan. 2026.

RIANSUT, W. Effectiveness of model selection criteria. *Recent Science and Technology*, 2022.

SARKAR, M. A. B.; HASAN, M. J. Predicting fuel efficiency using Auto MPG dataset, 2025.

SUN, J. et al. Energy efficiency modeling using machine learning, 2023.

TETKO, I. V. et al. Be aware of overfitting. *Journal of Cheminformatics*, 16, 2024.

TUKEY, J. W. *Exploratory Data Analysis*. Reading: Addison-Wesley, 1977.

TZOU, S. J. et al. Comparison between linear regression and machine learning methods. *Journal of the Chinese Medical Association*, 86, 1028–1036, 2023.

WANG, Z. *et al.* Vehicular fuel consumption estimation using machine learning. *Energies*, 17(6), 1410, 2024.

XU, Y.; GOODACRE, R. On splitting training and validation set. *Journal of Analysis and Testing*, 2, 249–262, 2018.

YOO, S.; SHIN, J.; CHOI, S. H. Vehicle fuel efficiency prediction using machine learning. *Scientific Reports*, 15, 14815, 2025.

IEA. *Eficiência Energética 2023*. IEA, Paris, 2023. Disponível em: https://www.iea.org/reports/energy-efficiency-2023. Acesso em: 30 mar. 2026.