Previsão da Frota Municipal de Veículos Elétricos no Brasil usando Machine Learning

Mariana Ciervo Borges[1], Paula Donaduzzi Rigo[1], Julio Cezar Mairesse Siluk[1], Anderson dos Santos Cezário[1], Gabriel Machado Lunardi[2]

[1]NIC – Núcleo de Inovação e Competitividade – Universidade Federal de Santa Maria (UFSM)

[2]NADIA – Núcleo de Análise de Dados e Inteligência Artificial – UFSM

{mariana.ciervo,anderson.cezario}@acad.ufsm.br,

{paula.rigo,jsiluk,gabriel.lunardi}@ufsm.br

Resumo. *Este trabalho apresenta uma abordagem prática para predizer a frota municipal de veículos elétricos (VEs) no Brasil. O objetivo é avaliar algoritmos de aprendizado de máquina na predição por município, relacionando precisão estatística (RMSE e R^2) e custo computacional (CPU). Assim, busca-se apoiar o planejamento energético e a seleção de algoritmos de Machine Learning.*

1. Introdução

A transição energética global tem impulsionado uma mudança significativa no setor de transportes, com os veículos elétricos (VEs) emergindo como uma alternativa aos combustíveis fósseis [Smith et al. 2023]. Esta ascensão é motivada por uma crescente conscientização ambiental e políticas de incentivo que visam reduzir as emissões de gases de efeito estufa e a dependência de fontes não renováveis [Chhetri et al. 2025].

Porém, a inserção regional dos VEs nos municípios brasileiros apresenta desafios e oportunidades que demandam uma compreensão aprofundada sobre a distribuição e o perfil dos usuários. Compreensão que é vital para toda a cadeia produtiva e de serviços que se beneficia da eletrificação, incluindo a indústria automotiva, o setor de energia, empresas de tecnologia e prestadores de serviços de manutenção e abastecimento [Chhetri et al. 2025].

O uso de AutoML permite a análise sistemática e reprodutível, mesmo que o volume de dados de 5.570 municípios seja manejável. Para lidar com essa complexidade, a aplicação de Machine Learning (ML) e AutoML é essencial, permitindo a comparação de algoritmos de regressão enquanto se avalia o desempenho de CPU e a escalabilidade do modelo, tornando possível a identificação de padrões de consumo por município e estimativas de frota.

2. Metodologia

O banco de dados foi construído a partir da convergência de indicadores econômicos, sociais e tecnológicos. Os dados foram coletados através de três fontes primárias: o Instituto Brasileiro de Geografia e Estatística [IBGE 2026] para indicadores demográficos; a Agência Nacional de Energia Elétrica [ANEEL, 2026a, 2026b] para dados do setor elétrico e bases técnicas de irradiação solar (como as do INPE); e a Secretaria Nacional de Trânsito [SENATRAN 2025] para os dados da frota de veículos por município e combustível.

A Tabela 1 detalha os preditores (features) utilizados, suas respectivas unidades de medida e as fontes consultadas. Para viabilizar a análise a nível municipal, foi realizado um processo de data merging utilizando o código identificador de 7 dígitos do IBGE de cada município como chave primária. Este procedimento permitiu a junção de variáveis heterogêneas em um dataset consolidado, garantindo a relação entre os dados técnicos e os perfis socioeconômicos.

Tabela 1. Preditores do dataset, unidades e fontes de dados.

Preditores (Nome da Feature)	Unidade	Fonte Consultada
Potência Instalada	kW	ANEEL (2026a)
Tarifa Elétrica	R$	ANEEL (2026b)
Irradiância Solar	Wh/m^2	Tiba (2000)
Frota Municipal	N/A	SENATRAN (2026)
PIB	R$	IBGE (2026)
IDH Educação / Saúde / Renda	%	PNUD (2023)
População Total	N/A	IBGE (2026)
Área Territorial	km^2	IBGE (2026)
Densidade Demográfica	hab/km^2	IBGE (2026)
Educação Superior	%	IBGE (2026)
Número de Famílias	N/A	IBGE (2026)
Verticalização das Cidades	%	IBGE (2026)
Famílias em Área Rural	%	IBGE (2026)
Casas Próprias / Alugadas	N/A	IBGE (2026)
Residentes por Família	N/A	IBGE (2026)
Famílias com mais de 3 Salários Mínimos	%	IBGE (2026)
Média Salários Mínimos	R$	IBGE (2026)
Número de Empresas/Empregados	N/A	IBGE (2026)

Para a modelagem preditiva, utilizou-se o PyCaret 3.4, uma biblioteca de AutoML que automatiza o ciclo do aprendizado de máquina. A escolha se deve pela necessidade de avaliar múltiplos algoritmos de regressão sob condições idênticas de pré-processamento, facilitando a medição do desempenho computacional.

Na etapa de pré-processamento, na função setup, foram removidas variáveis identificadoras para evitar overfitting e aplicada a eliminação de multicolinearidade com limiar de 0,9, diminuindo a redundância e o custo computacional. Além disso, os dados brutos passaram por limpeza automática, tratando inconsistências como valores não numéricos no PIB e normalizando os tipos de dados.

Para a modelagem, empregou-se a função compare_models para treinar simultaneamente algoritmos de diferentes famílias, como Regressão Linear e Random Forest. A validação foi realizada com K-Fold de 10 dobras, assegurando que as métricas de erro (RMSE) e precisão (R^2) fossem estatisticamente robustas e representativas diante da variância dos dados. Essa abordagem automatizada permitiu uma comparação massiva e confiável do desempenho dos modelos através da engenharia de dados.

3. Resultados

A análise comparativa dos resultados da Tabela 2, baseada em métricas de erro (RMSE), poder explicativo (R^2) e custo computacional (tempo total de treinamento e validação).

Tabela 2. Comparação de desempenho dos modelos de regressão

Sigla	Modelo	RMSE (n° VE)	R^2 (%)	TT (Sec)
omp	Orthogonal Matching Pursuit	507,44	80,20	0,04
br	Bayesian Ridge	620,42	80,55	0,04
en	Elastic Net	622,01	80,56	0,07
ard	Automatic Relevance Determination	632,55	79,62	0,13
lasso	Lasso Regression	632,62	80,17	0,08
llar	Lasso Least Angle Regression	632,62	80,17	0,04
ridge	Ridge Regression	639,19	79,60	0,09
kr	Kernel Ridge	639,46	79,69	0,54
lr	Linear Regression	645,21	78,19	1,09
lar	Least Angle Regression	648,02	77,76	0,04
et	Extra Trees Regressor	711,22	82,32	1,75
rf	Random Forest Regressor	794,67	80,18	7,45
par	Passive Aggressive Regressor	872,26	72,27	0,04
knn	K Neighbors Regressor	883,75	79,25	0,05
ada	AdaBoost Regressor	912,01	56,19	0,22
gbr	Gradient Boosting Regressor	958,12	59,42	1,91
dt	Decision Tree Regressor	1104,94	22,68	0,12

O modelo Orthogonal Matching Pursuit (OMP) apresenta a melhor performance em termos de erro absoluto, com o menor RMSE (507,44 VE) e o menor tempo de processamento (0,046 segundos). Ainda que todos os modelos tenham tempos aceitáveis, o OMP mantém a performance estatística com a estrutura mais simples possível (parcimônia), sugerindo que algoritmos esparsos capturam as relações lineares nos dados eficientemente e sem custo elevado.

Em contrapartida, o Extra Trees Regressor (ET) obteve o maior coeficiente de determinação ($R^2 = 82{,}32\%$), indicando superioridade na explicação da variância dos dados. Contudo, esse desempenho acarreta um custo computacional 38 vezes superior ao do OMP. Vale ressaltar que ambos os tempos (0,04s e 1,75s) são negligenciáveis diante da escala de 5.570 municípios e da baixa frequência de atualização dos dados demográficos. Por outro lado, o RMSE superior do ET mostra que modelos ensemble podem trazer complexidade desnecessária quando visam a redução do erro absoluto. Para a aplicação prática, o RMSE consolida-se como a métrica prioritária, visto que o erro médio de 711 veículos impacta diretamente na infraestrutura de recarga. Essa disparidade revela um trade-off fundamental entre poder explicativo e precisão para o planejamento energético.

Outros modelos lineares, como Lasso Least-Angle Regression (LLAR) e Bayesian Ridge (BR), também apresentaram bom desempenho, com RMSE inferior a 630 e tempos de processamento abaixo de 0,1 segundos. Já modelos ensemble como Random Forest (RF) e Gradient Boosting (GBR) mostraram-se menos eficientes para o escopo do problema, com tempos de treinamento de até 7 segundos e RMSE superior a 790. A análise das variáveis mais representativas nos modelos de melhor performance (OMP e

BR) identificou o PIB, o IDH e a capacidade instalada de energia solar como os principais fatores explicativos para a adoção de VEs nos municípios.

4. Conclusão

Do ponto de vista da engenharia de dados, a escolha do modelo deve considerar o ambiente de produção. Para cenários que exigem baixa latência e re-treinamentos frequentes, o modelo OMP é o mais indicado devido à sua eficiência, parcimônia e precisão absoluta (R^2). Para estudos exploratórios, o modelo ET é valioso para validar e explicar a relevância teórica das variáveis socioeconômicas no fenômeno da eletrificação.

Conclui-se que o uso de AutoML permite não apenas a seleção do algoritmo mais veloz, mas a identificação de modelos que equilibram o erro absoluto (RMSE) com a realidade prática do planejamento energético municipal. Este trabalho fornece, portanto, uma base sólida para que gestores utilizem dados socioeconômicos na antecipação da demanda por eletromobilidade no Brasil.

Agradecimentos

Os autores agradecem ao PROBITI (Programa Institucional de Bolsas de Iniciação Tecnológica e Inovação) e à FAPERGS (Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul) pelo apoio financeiro e institucional concedido para a realização deste trabalho.

Referências

ANEEL, Agência Nacional de Energia Elétrica (2026a). Geração distribuída. https://dadosabertos.aneel.gov.br/group/geracao-distribuida. Acesso em: 20 jan. 2026.

ANEEL, Agência Nacional de Energia Elétrica (2026b). Ranking das tarifas de eletricidade. https://portalrelatorios.aneel.gov.br/luznatarifa/rankingtarifas. Acesso em: 20 jan. 2026.

Chhetri, J. et al. (2025). Predicting electric vehicle diffusion and its impact on the productive chain. *Forthcoming / Technical Report*.

IBGE, Instituto Brasileiro de Geografia e Estatística (2026). Sidra, sistema ibge de recuperação automática. https://sidra.ibge.gov.br/home/pmc/brasil. Acesso em: 21 jan. 2026.

PNUD, United Nations Development Programme (2023). Atlas do desenvolvimento humano no brasil. http://www.atlasbrasil.org.br/. Acesso em: 21 out. 2023.

PyCaret (2026). An open source, low-code machine learning library in python. https://github.com/pycaret/pycaret.

SENATRAN, Secretaria Nacional de Trânsito (2025). Frota de veículos 2025. https://www.gov.br/transportes/pt-br/assuntos/transito/conteudo-Senatran/frota-de-veiculos-2025. Acesso em: 21 jan. 2026.

Smith, A. et al. (2023). Inclusive innovation in just transitions: the case of smart local energy systems in the uk. *Environmental Innovation and Societal Transitions*, 47:100719.

Tiba, C. (2000). *Atlas Solarimétrico do Brasil: Banco de Dados Terrestres*, volume 1. Atlas Solarimétrico Do Brasil.