

**ESTIMATIVAS DE ARRECADAÇÃO DO ICMS DO ESTADO DO MARANHÃO
USANDO ALGORITMOS DE MACHINE LEARNING**

**ESTIMATES OF ICMS COLLECTION IN THE STATE OF MARANHÃO USING
MACHINE LEARNING ALGORITHMS**

Área temática: Finanças, Contabilidade e Controladoria

LIMA, Philipe
SANTANA, Ewaldo
CAMPOS, Lúcio
SANTANA, Tiago
Universidade Estadual do Maranhão - UEMA
SOUSA, Kassyo
Universidade FUMEC

Resumo

Este trabalho analisa alguns algoritmos de *machine learning* para estimar a receita mensal de ICMS no Maranhão, empregando dados de janeiro de 1997 até abril de 2024. A base abrange variáveis econômicas como índices de preços, a taxa SELIC e o valor do dólar. Os dados foram submetidos a uma análise descritiva para detectar *outliers* e escalonados com *MinMaxScaler* para padronização. Modelos como *XGBoost*, Regressão Linear, *Random Forest* e *Decision Tree* foram validados por meio de *k-fold* ($k=10$) e analisados com métricas *RMSE*, *MAE*, *MAPE*, *sMAPE* e R^2 . Ensaios estatísticos avaliaram a performance dos modelos. Com a implementação em Python, a pesquisa disponibiliza recursos para aprimorar o planejamento tributário e a elaboração de políticas públicas no Maranhão.

Palavras-chave: ICMS; Previsão; Machine Learning; Algoritmos.

Abstract

This paper analyzes some machine learning algorithms to estimate monthly ICMS revenue in Maranhão, using data from January 1997 to April 2024. The database includes economic variables such as price indexes, the SELIC rate, and the value of the dollar. The data were subjected to descriptive analysis to detect outliers and scaled with *MinMaxScaler* for standardization. Models such as *XGBoost*, Linear Regression, *Random Forest*, and *Decision Tree* were validated using *k-fold* ($k=10$) and analyzed with *RMSE*, *MAE*, *MAPE*, *sMAPE*, and R^2 metrics. Statistical tests evaluated the performance of the models. With the implementation in Python, the research provides resources to improve tax planning and public policy development in Maranhão.

Keywords: ICMS; Forecast; Machine Learning; Algorithms.

1. Introdução

O ICMS é o principal tributo estadual no Brasil, regulamentado pela Constituição Federal e legislações estaduais, como a Lei nº 7.799/2002 no Maranhão. Sua arrecadação impacta diretamente políticas públicas e gestão administrativa, sendo influenciada por variáveis

econômicas, fiscais e sociais, além de fatores externos como inflação, preços de commodities e mudanças na legislação. A projeção precisa da receita tributária é essencial para o planejamento orçamentário, conforme estabelecido na Lei nº 4.320/1964 e na Lei de Responsabilidade Fiscal (LRF). No entanto, a crescente complexidade dos dados e a necessidade de maior precisão demandam novas abordagens, como o uso de algoritmos de *machine learning* para otimizar as estimativas.

Dada a relevância do ICMS na composição das receitas públicas, diversas pesquisas têm explorado métodos para a estimativa de sua arrecadação. Entre elas, destacam-se: a) Modelos estatísticos e econométricos, como o estudo de Black (2021), que utilizou modelos estruturais de séries temporais com variáveis econômicas e fiscais para prever a arrecadação do ICMS; b) Algoritmos de aprendizado de máquina, conforme demonstrado por Murta (2024), que aplicou esses métodos para prever o ICMS no estado de Minas Gerais; c) Redes neurais profundas, exploradas por Dornelles, Schwartz e Braatz (2022), os quais empregaram modelos univariados de redes neurais recorrentes (RNN) para comparar estimativas automatizadas com aquelas produzidas pela Secretaria da Fazenda do Rio Grande do Sul.

A arrecadação do ICMS, vinculada à circulação de bens e serviços, é impactada por variáveis sociais, econômicas, fiscais e fatores internacionais, como equilíbrio fiscal, inflação, oscilações em preços de commodities, mudanças na legislação tributária e políticas econômicas, que influenciam o consumo e o desempenho do tributo (DIEESE, 2022). Paralelamente, o avanço na disponibilidade de dados públicos, por meio de iniciativas como os Portais da Transparência (BRASIL, 2019), o Sistema de Gestão de Séries Temporais do Banco do Brasil (Banco do Brasil, 2020) e dados do IBGE (2021), possibilita análises detalhadas sobre finanças públicas, reforçando a transparência e o estudo da arrecadação.

Nesse contexto, a combinação de avanços tecnológicos, especialmente em algoritmos de aprendizado de máquina, com o acesso ampliado a dados e a necessidade de maior eficiência na gestão pública, apresenta oportunidades para o desenvolvimento de novos conhecimentos na estimativa tributária. Assim, o presente estudo tem como objetivo comparar o desempenho de algoritmos de aprendizado de máquina para previsão de arrecadação do ICMS no estado do Maranhão usando variáveis de consumo, sociais e econômicas.

A pesquisa torna-se impar por ser pioneira no contexto do Maranhão, ampliar o escopo metodológico ao acrescentar variáveis sociais aos modelos e por comparar o desempenho de diferentes algoritmos, incluindo Regressão Linear Múltipla, *Extreme Gradient Boost (XGBoost)*, *Random Forest* e *Decision Tree* (DUCHESNAY, LOFSTEDT E YOUNES, 2021), utilizando uma série temporal extensa (mensal de janeiro de 1997 a abril de 2024) e incluir variáveis sociais

e econômicas específicas do estado. Entre as variáveis analisadas, destacam-se o Produto Interno Bruto (PIB), a taxa de juros SELIC, a cotação do dólar, o número de novos empregos gerados no Maranhão, o saldo da balança comercial maranhense e o consumo de energia elétrica residencial e total no Brasil e no Nordeste.

2. Referencial Teórico

O planejamento orçamentário constitui uma atividade central para os gestores públicos, envolvendo aspectos como fiscalização, prestação de contas, racionalização dos recursos e definição de sua aplicação. Nesse contexto, a estimativa precisa das receitas públicas torna-se essencial para a fixação das despesas, em conformidade com os marcos legais estabelecidos, como a Lei nº 4.320/1964 (BRASIL, 1964), a Constituição da República Federativa do Brasil de 1988 (BRASIL, 1988) e a Lei de Responsabilidade Fiscal (LRF, Lei Complementar nº 101/2000) (BRASIL, 2000).

Dessa forma, a projeção de receitas, especialmente do ICMS tem sido avaliada através de Modelos tradicionais como o ARIMA têm sido eficazes para identificar padrões históricos de arrecadação (Azevedo, Silva e Gatsios, 2015), auxiliando na previsão em cenários incertos. Recentemente, avanços tecnológicos, como redes neurais artificiais associadas à seleção de variáveis (Carmo, Boldt e Komati, 2019), permitiram lidar com dados complexos e relações não lineares, modernizando a gestão tributária e possibilitando previsões precisas em contextos fiscais desafiadores. Tais pesquisas são viáveis devido a integração de aprendizado de máquina com Big Data que está revolucionando a administração tributária, permitindo antecipar receitas, identificar fraudes e otimizar sistemas fiscais. Conforme Migalinas (2023), a inteligência artificial fortalece a gestão tributária e apoia o planejamento orçamentário em cenários de volatilidade econômica.

Adicionalmente, a literatura tem destacado o uso de modelos não lineares sofisticados, como redes neurais recorrentes (RNN) e algoritmos baseados em *Decision Trees*. Dornelles, Schwartzer e Braatz (2022) evidenciaram a eficácia das RNNs na previsão da arrecadação do ICMS no Rio Grande do Sul, superando métodos tradicionais em termos de precisão preditiva. Paralelamente, o *Extreme Gradient Boosting (XGBoost)*, discutido por Chen e Guestrin (2016), tem se consolidado como uma abordagem robusta para problemas preditivos que envolvem grandes volumes de dados, unindo alta acurácia e eficiência computacional.

Os algoritmos de *machine learning* têm sido utilizados para análise de séries temporais em diversas áreas, como finanças, energia, saúde e previsão climática. Esses modelos permitem

identificar padrões complexos e realizar previsões precisas, mesmo em cenários com alta variabilidade e grandes volumes de dados. Conforme demonstrado a seguir:

Quadro 1 – Pesquisas com aplicação de algoritmos para previsão

Métodos	Resumo e aplicação	Autores
<i>Ensemble methods, Deep Learning models, SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations)</i>	Algoritmos como <i>ensemble methods</i> e <i>deep learning</i> melhoram a previsão, mas enfrentam desafios como dados desbalanceados e seleção de variáveis em planejamento econômico e estratégias empresariais.	Mustafa I. Al-Karkhi, Grzegorz Rządowski (2024)
<i>N-BEATS (Neural Basis Expansion Analysis for Time Series), Temporal Fusion Transformers</i>	Modelos como <i>N-BEATS</i> e <i>Temporal Fusion Transformers</i> superam benchmarks ingênuos, mesmo com variações de energia renovável em simulações de mercado de energia e transição energética.	Felix Nitsch, Christoph Schimeczek, Valentin Bertsch (2024)
Redes Neurais <i>LSTM (Long Short-Term Memory)</i> , <i>CNN-LSTM (Convolutional Neural Network - Long Short-Term Memory)</i> , <i>GARCH (Generalized Autoregressive Conditional Heteroskedasticity)</i> , <i>HAR (Heterogeneous Autoregressive model)</i>	Redes neurais (<i>LSTM</i> , <i>CNN-LSTM</i>) superam modelos tradicionais, com destaque para <i>CNN-LSTM</i> em previsões de curto prazo em finanças e criptomoedas.	Zih-Chun Huang, Ivan Sangiorgi, Andrew Urquhart (2024)
<i>LightGBM (Light Gradient Boosting Machine)</i> , <i>PCA (Principal Component Analysis)</i> , <i>Dimensionality Reduction</i> (Técnicas de redução de dimensionalidade)	<i>LightGBM</i> se destaca na previsão, exceto em eventos inesperados, com melhorias ao integrar dados de mobilidade em redes 5G e automação de recursos.	Ana Almeida, Pedro Rito, Susana Brás, Filipe Cabral Pinto, Susana Sargento (2024)
Redes Neurais (Redes Neurais Artificiais), <i>Random Forest</i> (Floresta Aleatória), <i>ELM (Extreme Learning Machine)</i> , <i>SVR (Support Vector Regression)</i> , <i>LSTM (Long Short-Term Memory)</i>	Redes Neurais superam outros modelos na previsão de geração de energia operações de sistemas fotovoltaicos flutuantes e integração à rede elétrica.	Mohd Herwan Sulaiman et al. (2024)

Fonte: elaborado pelos autores.

3. Metodologia

Para alcançar o objetivo de estimar a arrecadação do ICMS no estado do Maranhão a partir de variáveis econômicas e sociais, foram empregadas técnicas de aprendizado de máquina em séries temporais, utilizando recursos computacionais baseados na linguagem de programação Python versão 3.12.8, Sistema Operacional Windows 10 e bibliotecas como *Scikit-learn*, *Pandas*, *xgboost*, *matplotlib* e *NumPy*, reconhecidas por sua robustez em aplicações analíticas e preditivas (PEDREGOSA et al., 2011) e a escrita do código contou com ajuda parcial do uso de inteligência artificial, com foco em redução de redundância e clareza nos parâmetros.

Coleta e Pré-processamento de Dados

Os dados utilizados no estudo foram obtidos de diversas fontes, respeitando a mesma granularidade temporal (mensal), com uma série histórica abrangendo o período de janeiro de 1997 a abril de 2024. Destaca-se que, até a submissão deste estudo, não houve divulgação ou atualização adicional da arrecadação de ICMS pelo Conselho Nacional Fazendário (CONFAZ).

Os dados foram analisados descritivamente para examinar suas distribuições e detectar *outliers*, que poderiam prejudicar as análises preditivas. Em seguida, foram normalizados com o método *MinMaxScaler*, ajustando-os ao intervalo [0, 1]. Essa etapa é essencial, pois algoritmos de *machine learning* são sensíveis a diferenças de escala entre variáveis, podendo ter desempenho comprometido sem normalização adequada (HAN et al., 2011).

Modelos Preditivos

Os modelos selecionados para a análise foram: *XGBoost*, *Random Forest*, *Decision Tree* e Regressão Linear.

A escolha desses algoritmos justifica-se por sua capacidade de capturar tendências em séries temporais e modelar relações não lineares entre variáveis. Estudos na literatura corroboram sua eficácia em contextos de dados complexos e heterogêneos.

Em tarefas preditivas com conjuntos de dados e variáveis extensos, o *XGBoost* é conhecido por sua regularização (que ajuda a reduzir o sobreajuste, que se refere a quando o modelo adotado se adapta demais as previsões realizadas) e robustez em ambientes desafiadores de alta dimensão e correlação (Chen & Guestrin, 2016). A capacidade de tratamento de dados ausentes e o uso de paralelização o tornam promissor para séries temporais de arrecadação de impostos.

Em relação à *Decision Tree*, apesar de sua simplicidade, ela pode ser útil em termos de interpretabilidade, facilitando a visualização de regras de decisão que podem ser utilizadas na comunicação com o gestor público e os tomadores de decisão (Quinlan, 1986). Ela também pode superar significativamente modelos isolados em termos de poder preditivo quando combinada com abordagens *ensemble* (Breiman, 2001), como demonstrado pelos métodos baseados em *Random Forest*.

Este último utiliza múltiplas árvores de decisão para gerar previsões consistentes, minimizando o risco de sobreajuste ao empregar amostras aleatórias de dados e atributos em cada iteração (Breiman, 2001). Sua habilidade em capturar interações entre variáveis e fornecer estimativas de relevância o torna particularmente útil em análises fiscais e econômicas, conforme destacado por Chen e Guestrin (2016). Esse algoritmo é amplamente aplicado em

previsões de séries temporais econômicas, onde diversos fatores externos influenciam os resultados.

Por fim, a Regressão Linear geralmente é empregada como um modelo de referência devido à sua simplicidade e habilidade em oferecer percepções iniciais sobre as relações lineares entre variáveis, sendo abordada na literatura como uma ferramenta de análise preliminar em investigações tributárias (Montgomery et al., 2012).

Validação e Avaliação dos Modelos

A validação dos modelos foi realizada utilizando a técnica de *k-fold cross-validation* com $k = 10$. Esse método divide o conjunto de dados em 10 partes iguais, treinando os modelos em nove partes e testando na parte restante, repetindo o processo 10 vezes. Essa abordagem é amplamente reconhecida por equilibrar viés e variância nas estimativas preditivas (KOHAVI, 1995).

As métricas de desempenho empregadas para avaliar os modelos foram o Erro Quadrático Médio (*RMSE*): Mede o desvio quadrático médio entre os valores observados e previstos, sendo particularmente sensível a grandes erros (HARRISON, 2020); Erro Médio Absoluto (*MAE*): Avalia o erro absoluto médio, sendo menos influenciado por *outliers* do que o *RMSE* (WILLMOTT; MATSUURA, 2005); Erro Percentual Absoluto Médio (*MAPE*): Mede o erro médio em termos percentuais, facilitando a interpretação em variáveis de diferentes escalas (ARMSTRONG; COLLOPY, 1992); Erro Percentual Absoluto Médio Simétrico (*sMAPE*): Variante do *MAPE* que ajusta as discrepâncias causadas por valores extremos (MAKRIDAKIS, 1993) e Coeficiente de Determinação (R^2): Representa a proporção da variação explicada pelo modelo em relação aos dados observados, sendo valores próximos a 1 indicativos de maior precisão (KVALSETH, 1985).

Seleção de Variáveis

Além dos modelos preditivos, foi aplicada a Regressão Linear Múltipla para identificar as variáveis independentes mais relevantes na explicação da arrecadação do ICMS. Segundo Hair *et al.* (2010), essa técnica é adequada para reduzir a complexidade dos modelos ao selecionar variáveis com maior poder explicativo.

No início da análise, a base de dados consolidada contava com 20 variáveis, sendo uma dependente (ICMS Total) e 19 independentes, como pode ser visto no Quando 1. O processo de seleção visou identificar as variáveis com maior relevância estatística para o modelo preditivo, facilitando a interpretação e a aplicabilidade dos resultados.

Justificativa e Limitações

Embora os aspectos matemáticos dos modelos sejam relevantes, o foco desta pesquisa está na aplicação prática e no uso de técnicas avançadas para melhorar as projeções tributárias, em vez de um aprofundamento teórico em estatística ou álgebra linear, dada a disponibilidade de ferramentas computacionais para tais análises. Para simplificar os modelos, adotou-se a regressão linear múltipla, visando identificar as variáveis que mais contribuem para explicar a arrecadação, conforme destacado por Hair *et al.* (2010). Inicialmente, a base de dados continha 20 variáveis, sendo uma dependente (ICMS Total) e as demais independentes.

Quadro 2 – Lista de Todas Variáveis.

Variáveis	Identificação	Fonte
ICMS Total	ICMS	Conselho Nacional Fazendário
Cotação do Dólar	DOLAR_COT	Instituto de Pesquisa Econômica Aplicada - devido a temporalidade da cotação ser diário, a agregação de dados foi feita com base na média mensal a fim de que se obtivesse a mesma temporalidade que as demais variáveis.
188 - Índice nacional de preços ao consumidor (INPC) - Var. % mensal Fonte: IBGE	INPC	SGS - Sistema Gerenciador de Séries Temporais - v2.1
433 - Índice nacional de preços ao consumidor amplo (IPCA) - Var. % mensal Fonte: IBGE	IPCA	
4380 - PIB mensal - Valores correntes (R\$ milhões) - R\$ (milhões) - Fonte: BCB-Depec	PIB	
13244 - Exportação de bens - Maranhão - US\$ (mil) - Fonte: MDIC/Secex	EXPOR_MA	
13245 - Importação de bens - Maranhão - US\$ (mil) - Fonte: MDIC/Secex	IMPOR_MA	
13247 - Saldo da balança comercial - Maranhão - US\$ (mil) - Fonte: MDIC/Secex	BALANCA_MA	
13011 - Empregos formais gerados - Maranhão - Unidades - Fonte: MTb	GER_EMP_MA	
1402 - Brasil - Comercial GWh	ENERGIA_ELETRICA_CONSUMO_BRASIL	
1403 - Brasil - Residencial GWh	ENERGIA_ELETRICA_RESIDENCIAL_BRASIL	
1404 - Consumo de energia elétrica - Brasil - Industrial - GWh	ENERGIA_ELETRICA INDUSTRIAL_BRASIL	
1412 - Consumo de energia elétrica - Região Nordeste - Comercial - GWh	ENERGIA_ELETRICA_CONSUMO_NORDESTE	

1413 - Consumo de energia elétrica - Região Nordeste - Residencial - GWh	ENERGIA_ELETRICA_RESIDENCIAL_NORDESTE	Instituto de Pesquisa Econômica Aplicada
1414 - Consumo de energia elétrica - Região Nordeste - Industrial - GWh	ENERGIA_ELETRICA_INDUSTRIAL_NORDESTE	
1396 - Óleo diesel	DIESEL_CONSUMO_BAR_DIA_MIL	
1393 - Consumo de derivados de petróleo - Gasolina	GASOLINA_CONSUMO_BAR_DIA_MIL	
Taxa SELIC	SELIC	
Salário Mínimo	SM	
Cotação do Barril do Petróleo	PETRO_COT	

Fonte: elaborado pelos autores.

Após a aplicação do método de Regressão Linear Múltipla, verificou-se que nem todas as variáveis inicialmente consideradas apresentavam contribuição significativa para a explicação do modelo preditivo. O critério adotado para exclusão foi o p-valor $> 0,05$, indicando que os coeficientes das respectivas variáveis não eram estatisticamente diferentes de zero ao nível de significância de 5%. Como resultado, foram selecionadas as variáveis que demonstraram relevância estatística para aplicação aos algoritmos preditivos.

Quadro 3 – Lista de Variáveis Selecionadas.

Variáveis	Identificação	Fonte
4380 - PIB mensal - Valores correntes (R\$ milhões) - R\$ (milhões) - Fonte: BCB-Depec	PIB	SGS - Sistema Gerenciador de Séries Temporais - v2.1
1402 - Brasil - Comercial GWh	ENERGIA_ELETRICA_CONSUMO_BRASIL	
1403 - Brasil - Residencial GWh	ENERGIA_ELETRICA_RESIDENCIAL_BRASIL	
1412 - Consumo de energia elétrica - Região Nordeste - Comercial - GWh	ENERGIA_ELETRICA_RESIDENCIAL_NORDESTE	
1413 - Consumo de energia elétrica - Região Nordeste - Residencial - GWh	ENERGIA_ELETRICA_INDUSTRIAL_NORDESTE	
1396 - Óleo diesel	DIESEL_CONSUMO_BAR_DIA_MIL	

Fonte: elaborado pelos autores.

4. Resultados

Inicialmente, realizou-se uma análise estatística descritiva dos dados, incluindo a aplicação do teste de normalidade de *Shapiro-Wilk* (Shapiro & Wilk, 1965), para verificar se as variáveis seguem uma distribuição normal. Nesse teste, p-valores menores que 0,05 indicam a rejeição da hipótese nula de normalidade dos dados.

Tabela 1 – Estatísticas Descritivas

Variável	Média	Mediana	Desvio-Padrão	n	Shapiro p-valor
ICMS milhões	342,443	250,222	293,061	328	0,000
SELIC	1,068	1,020	0,511	328	0,000
INPC	0,503	0,455	0,446	328	0,000
IPCA	0,496	0,450	0,400	328	0,000
PIB bilhões	375,436	330,607	248,320	328	0,000
EXPOR_MA milhares	200,973	182,745	134,548	328	0,000
IMPOR_MA milhares	248,865	191,415	223,105	328	0,000
BALANCA_MA milhares	-47,891	-1,132	179,516	328	0,000
GER_EMP_MA	592,607	530,000	2,319	328	0,000
SM	586,073	510,000	375,886	328	0,000
DOLAR_COT	2,833	2,374	1,255	328	0,000
PETRO_COT	60,899	59,403	31,593	328	0,000
Variável	Média	Mediana	Desvio-Padrão	n	Shapiro p-valor
GASOLINA_CONSUMO_BAR_DI milhares	415,741	386,000	112,231	328	0,000
DIESEL_CONSUMO_BAR_DIA milhares	832,482	839,000	189,490	328	0,000
ENERGIA_ELETRICA_CONSUMO _BRASIL (GWh) milhares	5.768,000	5.769,000	1.694,000	328	0,000
ENERGIA_ELETRICA_RESIDENCI AL_BRASIL (GWh) milhares	9.219,000	8.957,000	2.543,000	328	0,000
ENERGIA_ELETRICA_INDUSTRIA L_BRASIL (GWh) milhares	13.440,000	13.982,000	1.994,000	328	0,000
ENERGIA_ELETRICA_CONSUMO _NORDESTE (GWh)	871.226,00 0	862.500,000	298.331,000	328	0,000
ENERGIA_ELETRICA_RESIDENCI AL_NORDESTE (GWh)	1.695,000	1.607,000	666,098	328	0,000
ENERGIA_ELETRICA_INDUSTRIA L_NORDESTE (GWh)	2.071,000	1.998,000	268,395	328	0,000

Fonte: elaborado pelos autores.

Os dados analisados apresentam alta variabilidade, evidenciada pelos valores de desviopadrão, que podem ser atribuídos a fatores econômicos, sociais ou específicos de consumo. A análise inicial permitiu identificar que a série temporal analisada cobre 328 meses, compreendendo mudanças significativas no cenário econômico.

Modelagem por Regressão Linear Múltipla

A aplicação da Regressão Linear Múltipla teve como objetivo avaliar a contribuição de cada variável independente na explicação da arrecadação do ICMS. Inicialmente, foram incluídas todas as variáveis no modelo.

Tabela 2 – Regressão Linear Múltipla com todas as variáveis

CONFIGURAÇÃO DO MODELO DE REGRESSÃO LINEAR MÚLTIPLA				
Dep .Variable: ICMS		R-squared:		0,971
Model: OLS		R-squared adj		0,969
Method: Least Squares		F-statistic:		542,1
Prob (F-statistic):		0,000		
Variáveis	coef	Std err	t	P> t
const	964,084	5.068,101	0,190	0,849
SELIC	-0,017	0,026	-0,662	0,508
INPC	-0,031	0,073	-0,429	0,668
IPCA	-0,008	0,074	-0,105	0,916
PIB	1,048	0,099	10,579	0,000
EXPOR_MA	556,521	2.925,696	0,190	0,849
IMPOR_MA	-1.062,544	5.586,032	-0,190	0,849
BALANCA_MA	-1.270,303	6.677,751	-0,190	0,849
GER_EMP_MA	0,012	0,022	0,540	0,590
SM	-0,111	0,090	-1,238	0,217
DOLAR_COT	0,108	0,028	3,881	0,000
PETRO_COT	0,020	0,017	1,226	0,221
GASOLINA_CONSUMO_BAR_DIA_MIL	-0,019	0,029	-0,653	0,514
DIESEL_CONSUMO_BAR_DIA_MIL	-0,163	0,037	-4,394	0,000
ENERGIA_ELETRICA_CONSUMO_BRASIL	-0,610	0,091	-6,728	0,000
ENERGIA_ELETRICA_RESIDENCIAL_BRASIL	0,750	0,120	6,238	0,000
ENERGIA_ELETRICA_INDUSTRIAL_BRASIL	0,037	0,043	0,857	0,392
ENERGIA_ELETRICA_CONSUMO_NORDESTE	0,586	0,091	6,460	0,000
ENERGIA_ELETRICA_RESIDENCIAL_NORDESTE	-0,752	0,138	-5,453	0,000
ENERGIA_ELETRICA_INDUSTRIAL_NORDESTE	-0,055	0,030	-1,820	0,070

Fonte: elaborado pelos autores.

Variáveis com coeficientes não significativos (p-valor > 0,05) foram removidas, resultando em um modelo mais parcimonioso e estatisticamente relevante.

Tabela 3 – Regressão Linear Múltipla com variáveis significativas

CONFIGURAÇÃO DO MODELO DE REGRESSÃO LINEAR MÚLTIPLA				
Dep .Variable: ICMS		R-squared:		0,967
Model: OLS		R-squared Adj.		0,966
Method: Least Squares		F-statistic:		1564
Prob (F-statistic):		0,000		
Variáveis	coef	std err	t	P> t
const	-0,005	0,007	-0,651	0,516
PIB	1,051	0,049	21,306	0,000
DIESEL_CONSUMO_BAR_DIA_MIL	-0,157	0,027	-5,819	0,000
ENERGIA_ELETRICA_CONSUMO_BRASIL	-0,595	0,080	-7,476	0,000
ENERGIA_ELETRICA_RESIDENCIAL_BRASIL	0,573	0,114	5,018	0,000
ENERGIA_ELETRICA_CONSUMO_NORDESTE	0,486	0,087	5,608	0,000
ENERGIA_ELETRICA_RESIDENCIAL_NORDESTE	-0,536	0,130	-4,129	0,000

Fonte: elaborado pelos autores.

O modelo final apresentou um R^2 ajustado de 0,966, indicando que aproximadamente 96,6% da variação na arrecadação do ICMS é explicada pelas variáveis selecionadas. A estatística F reforçou a relevância do modelo ($p < 0,05$), demonstrando sua adequação.

Avaliação dos Algoritmos de *Machine Learning*

Os algoritmos foram avaliados com base em métricas de desempenho, conforme apresentado abaixo:

Tabela 4 – Métricas de Teste dos Algoritmos de *Machine Learning*

Algoritmos	RMSE	MAE	MAPE	sMAPE	R ²
<i>Random Forest</i>	50.591.014,595	27.934.023,370	8,339	8,305	0,975
<i>Decision Tree</i>	59.016.180,263	31.503.947,218	9,090	9,231	0,965
Regressão Linear	44.988.544,764	27.491.262,645	12,691	13,814	0,980
<i>XGBoost</i>	56.894.670,788	30.852.072,873	8,203	8,169	0,968

Fonte: elaborado pelos autores.

A Regressão Linear apresentou o melhor poder explicativo geral, com o maior valor de R^2 (0,980), e os menores erros absolutos (*RMSE* e *MAE*). Por outro lado, o *Decision Tree* obteve o pior desempenho, com os maiores valores de erro.

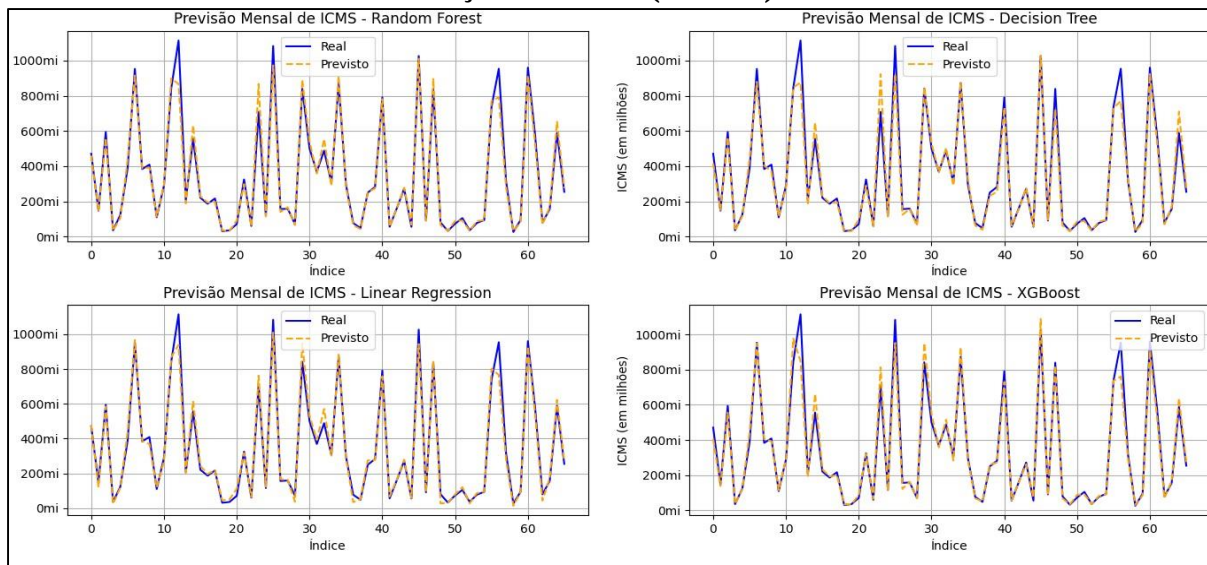
Fórmula Final do Modelo

Com base nos resultados da regressão linear múltipla, a fórmula final para estimar a arrecadação do ICMS foi definida como:

$$\text{ICMS} = -0.0048 + 1.0513 * \text{PIB} + -0.1568 * \text{DIESEL_CONSUMO_BAR_DIA_MIL} + -0.5945 * \text{ENERGIA_ELETTRICA_CONSUMO_BRASIL} + 0.5729 * \text{ENERGIA_ELETTRICA_RESIDENCIAL_BRASIL} + 0.4856 * \text{ENERGIA_ELETTRICA_CONSUMO_NORDESTE} + -0.5357 * \text{ENERGIA_ELETTRICA_RESIDENCIAL_NORDESTE}$$

O gráfico a seguir ilustra a previsão da arrecadação do ICMS, de acordo com cada algoritmo adotado. Representado por 20% de toda a amostra, ou seja, os últimos 66 meses.

Gráfico 1 – Previsão de Arrecadação do ICMS (milhões)



Fonte: elaborado pelos autores.

Embora visualmente não se observem grandes discrepâncias entre os valores reais e previstos, as diferenças entre os modelos se tornam evidentes ao analisar as métricas de avaliação, que revelam nuances não perceptíveis na representação gráfica.

5. Conclusões

Os resultados da pesquisa destacam a importância do uso de técnicas preditivas baseadas em aprendizado de máquina para analisar a arrecadação do ICMS no Maranhão. A investigação mostrou a eficácia de modelos que integram variáveis econômicas e sociais, oferecendo novas perspectivas para a gestão tributária. Entre os algoritmos avaliados, a Regressão Linear apresentou o maior coeficiente de determinação (R^2) e os menores erros ($RMSE$ e MAE), evidenciando sua adequação para identificar tendências e realizar previsões precisas, especialmente em contextos onde simplicidade e transparência são essenciais para a tomada de decisão na gestão pública.

Por outro lado, modelos mais avançados, como *Random Forest* e *XGBoost*, demonstraram desempenho competitivo, particularmente em métricas como $sMAPE$ e $MAPE$, evidenciando seu potencial para identificar interações não lineares e padrões complexos nos dados. Esses algoritmos se destacam como ferramentas para aplicações futuras, especialmente em contextos onde a riqueza de dados e a complexidade das relações entre variáveis são significativas. Em contrapartida, a *Decision Tree* apresentou resultados menos expressivos,

ênfatizando a importância de avaliar cuidadosamente as características do problema e as limitações a cada técnica ao selecionar o algoritmo mais adequado.

Este estudo contribui para o entendimento do uso de dados históricos e técnicas de aprendizado de máquina para projeção de receitas tributárias, apoiando gestores na formulação de políticas mais eficazes. Os resultados também sugerem caminhos para futuras pesquisas, como a exploração de novos algoritmos, ajustes de hiperparâmetros, inclusão de variáveis explicativas e análise de fatores externos, como mudanças legislativas e eventos econômicos globais, que podem influenciar as previsões de arrecadação.

Esta pesquisa ganha mais relevância diante das mudanças fiscais em curso no Brasil, como a substituição progressiva do ICMS por novos tributos a partir de 2026, conforme a Sugestão de Alteração Constitucional (SAC) nº 45/2019, que propõe o IBS (Impostos sobre bens e serviços) e a CBS (contribuições sobre bens e serviços) para simplificar a tributação sobre consumo (BRASIL, 2023). Essa transição exigirá ajustes de projeção de receitas, dado o reordenamento dos sistemas de arrecadação regional. Nesse contexto, avaliações sobre o desempenho do ICMS antes da reforma são essenciais para subsidiar mudanças eficazes e minimizar impactos na implementação do novo modelo (Ministério da Economia, 2022).

Apesar dos modelos terem apresentado resultados satisfatórios, a precisão e robustez poderiam ser melhoradas com a inclusão de variáveis como consumo de energia elétrica, consumo de combustíveis, taxa de desemprego estadual e volume de vendas no comércio varejista, todas relacionadas à economia estadual. No entanto, a dificuldade em obter esses dados de forma fidedigna e atualizada em nível estadual limitou sua incorporação ao estudo, sendo a temporalidade o maior desafio.

Por fim, considera-se que as contribuições deste estudo podem embasar a adoção de soluções mais eficientes, fundamentadas em dados e alinhadas às demandas da gestão pública, promovendo maior precisão na alocação de recursos e no planejamento fiscal de longo prazo.

Referencial Bibliográfico

Al-Karkhi, M. I., & Rządowski, G. (2024). Innovative machine learning approaches for complexity in economic forecasting and SME growth: A comprehensive review. *Journal of Economy and Technology*, 3 , 109-122. <https://doi.org/10.1016/j.ject.2025.01.001>. Acesso em 20 de março de 2024.

Almeida, A., Rito, P., Brás, S., Pinto, F. C., & Sargento, S. (2024). A machine learning approach to forecast 5G metrics in a commercial and operational 5G platform: 5G and mobility.

Computer Communications, 228, 107974. <https://doi.org/10.1016/j.comcom.2024.107974>. Acesso em 20 de março de 2024.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, 8(1), 69–80. Disponível em <https://marketing.wharton.upenn.edu/wp-content/uploads/2022/02/149-JSAError-Measures-for-Generalizing-about-Forecasting-Methods-Empirical-Comparisons.pdf>. Acesso em 09 de dezembro de 2024.

Azevedo, R. R., Silva, J. M., & Gatsios, R. C. (2015). Comparação de modelos de previsão de série temporal com base no ICMS estadual. Anais do Congresso USP de Controladoria e Contabilidade. Disponível em: <https://congressousp.fipecafi.org/anais/artigos152015/35.pdf>. Acesso em: 06 de janeiro de 2025.

Banco do Brasil. (2020). Sistema de Gestão de Séries Temporais. Banco do Brasil. <https://www3.bcb.gov.br/sgspub/localizarseries/localizarSeries.do?method=prepararTelaLocalizarSeries>. Acesso em 03 de dezembro de 2024.

Black, C. (2021, novembro). *Modelo estrutural de previsão para o ICMS no Rio Grande do Sul*. Apresentado no 10º Encontro de Economia Gaúcha, PUCRS. Disponível em https://tesouro.fazenda.rs.gov.br/upload/1643376139_Artigo%20Modelo%20Estrutural_Texto%20de%20Discussao.pdf. Acesso em 9 de dezembro de 2024.

Brasil. (1988). Constituição da República Federativa do Brasil de 1988. Brasília, DF: Presidência da República. Disponível em: http://www.planalto.gov.br/ccivil_03/Constituicao/Constituicao.htm. Acesso em 03 de dezembro de 2024.

Brasil. (2019). Portais da Transparência. Governo do Brasil. <https://portaldatransparencia.gov.br/>. Acesso em 03 de dezembro de 2024.

Brasil. Constituição da República Federativa do Brasil de 1988. (1988). Presidência da República, Casa Civil. Recuperado em 17 de dezembro de 2024, de https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm

Brasil. Lei de Responsabilidade Fiscal - LRF - Lei Complementar nº 101/2000. (2000). Presidência da República, Casa Civil. Recuperado em 17 de dezembro de 2024, de https://www.planalto.gov.br/ccivil_03/leis/lcp/lcp101.htm

Brasil. Lei No 4.320, de 17 de março de 1964. (1964). Presidência da República, Casa Civil. Recuperado em 17 de dezembro de 2024, de https://www.planalto.gov.br/ccivil_03/leis/l4320.htm#:~:text=LEI%20No%204.320%2C%20DE%2017%20DE%20MAR%C3%87O%20DE%201964&text=Estatui%20Normas%20Gerai s%20de%20Direito,Munic%C3%ADpios%20e%20do%20Distrito%20Federal

Brasil. Câmara dos Deputados. PEC nº 45/2019: Reforma Tributária. Brasília, 2023. Disponível em: <https://www12.senado.leg.br/publicacoes/estudos-legislativos/tipos-deestudos/boletinslegislativos/bol105#:~:text=Este%20boletim%20legislativo%20tem%20o,7%20de%20julho%20de%202023>. Acesso em 9 de março de 2025.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>. Acesso em 30 de novembro de 2024.

Carmo, M. M., Boldt, F. A., & Komati, K. S. (2019). Previsão de receitas de ICMS do estado do Espírito Santo através de Seleção de Características em Cascata e técnicas de Aprendizado de Máquina. Disponível em <https://sol.sbc.org.br/index.php/eniac/article/view/9277>. Acesso em 9 de dezembro de 2024.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. Disponível em: <https://dl.acm.org/doi/pdf/10.1145/2939672.2939785> . Acesso em: 06 de janeiro de 2025.

Departamento Intersindical de Estatística e Estudos Socioeconômicos. (2022). Redução do ICMS dos combustíveis, energia elétrica, transportes e comunicação (Nota Técnica, n.º 270). Departamento Intersindical de Estatística e Estudos Socioeconômicos. <https://www.dieese.org.br/notatecnica/2022/notaTec270ICMS.pdf>. Acesso em 30 de novembro de 2024.

Dornelles, G. Z., Schwartz, F. R., & Braatz, J. (2022). Redes Neurais Aplicadas na Previsão de Receita de ICMS no Rio Grande do Sul (Texto para Discussão TE/RS nº 19). Secretaria da Fazenda do Estado do Rio Grande do Sul. https://tesouro.fazenda.rs.gov.br/upload/1643376127_Artigo%20Modelo%20Redes%20Neurais_Texto%20de%20Discussao.pdf. Acesso em 03 de dezembro de 2024.

Duchesnay, E., Lofstedt, T., & Younes, F. (2021). *Statistics and Machine Learning in Python*. Engineering School, France. HAL Id: hal-03038776. Disponível em <https://hal.science/hal03038776v3>. Acessado em 04 de dezembro de 2024.

Governo do Estado do Maranhão. (2002). Lei n.º 7.799, de 19 de dezembro de 2002: Dispõe sobre o Sistema Tributário do Estado do Maranhão. Diário Oficial do Estado do Maranhão. <https://sistemas1.sefaz.ma.gov.br/portalsefaz/files?codigo=13942>. Acessado em 03 de março de 2025.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2010). *Multivariate data analysis* (7th ed.). Pearson Education Limited. Disponível em <https://www.drnishikantjha.com/papersCollection/Multivariate%20Data%20Analysis.pdf>. Acesso em 9 de dezembro de 2024.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann. Recuperado em 17 de dezembro de 2024, de <https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Seriesin-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf>.

Harrison, M. *Machine learning: guia de referência rápida: trabalhando com dados estruturados em Python*. São Paulo: Novatec, 2020.

Huang, Z.-C., Sangiorgi, I., & Urquhart, A. (2024). Forecasting Bitcoin volatility using machine learning techniques. *Journal of International Financial Markets, Institutions and*

Money, 97 , 102064. <https://doi.org/10.1016/j.intfin.2024.102064>. Acesso em 20 de março de 2025.

Instituto Brasileiro de Geografia e Estatística (IBGE). (2021). IBGE - Estatísticas e pesquisas. IBGE. <https://www.ibge.gov.br/indicadores>. Acesso em 03 de dezembro de 2024.

Kvalseth, T. O. (1985). Cautionary note about R^2 . *The American Statistician*, 39(4), 279–285. <https://doi.org/10.1080/00031305.1985.10479448>. Acesso em 03 de março de 2025.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)* (pp. 1137–1145). Morgan Kaufmann. Disponível em <https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf>. Acesso em 03 de março de 2025.

IPEA. (2024). Preço por barril do petróleo bruto Brent (FOB) (EIA366_PBRENT366). Disponível em <http://www.ipeadata.gov.br/ExibeSerie.aspx?module=m&serid=1650971490&oper=view>. Acesso em 9 de dezembro de 2024.

IPEA. (2024). Salário mínimo vigente (MTE12_SALMIN12). Disponível em <http://www.ipeadata.gov.br/ExibeSerie.aspx?serid=1739471028>. Acesso em 9 de dezembro de 2024.

IPEA. (2024). Taxa de câmbio comercial para compra: real (R\$) / dólar americano (US\$) - média (GM366_ERC366). Disponível em <http://www.ipeadata.gov.br/ExibeSerie.aspx?stub=1&serid=38590&module=M>. Acesso em 9 de dezembro de 2024.

Makridakis, S. (1993). Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting*, 9(4), 527–529. [https://doi.org/10.1016/0169-2070\(93\)90079-3](https://doi.org/10.1016/0169-2070(93)90079-3). Acesso em 03 de dezembro de 2024

Migalhas. (2023). Revolução fiscal: IA e Big Data na modernização da fiscalização. Disponível em: <https://www.migalhas.com.br/depeso/417201/revolucao-fiscal-ia-e-big-data-namodernizacao-da-fiscalizacao> . Acesso em 06 de janeiro de 2025.

Ministério da Economia. Reforma Tributária: contexto, mudanças e impactos . Estudo Especial nº 19, 4 mar. 2024. Disponível em: https://www2.senado.leg.br/bdsf/bitstream/handle/id/647648/EE19_2024.pdf . Acesso em 9 de março de 2025.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis. Wiley. [https://www.kwcsangli.in/uploads/3--Introduction to Linear Regression Analysis 5th ed. Douglas C. Montgomery Elizabeth A. Peck and G. .pdf](https://www.kwcsangli.in/uploads/3--Introduction%20to%20Linear%20Regression%20Analysis%205th%20ed.%20Douglas%20C.%20Montgomery%20Elizabeth%20A.%20Peck%20and%20G.%20.pdf). Acesso em 03 de dezembro de 2024

Murta, J. V. R. (2024). *Modelos de Machine Learning para previsão de arrecadação de ICMS em Minas Gerais* [Dissertação de Mestrado, Universidade de Brasília]. Repositório Institucional da Universidade de Brasília. Disponível em http://icts.unb.br/jspui/bitstream/10482/50888/1/2024_JoaoVitorRoqueMurta_DISSERT.pdf. Acesso em 30 de novembro de 2024.

Nitsch, F., Schimeczek, C., & Bertsch, V. (2024). Applying machine learning to electricity price forecasting in simulated energy market scenarios. *Energy Reports*, 12 , 5268-5279. <https://doi.org/10.1016/j.egy.2024.11.013>. Acesso em 20 de março de 2024.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research*, 12, 2825-2830. Disponível em <https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>. Acesso em 9 de dezembro de 2024.

Pedrosa, L., Moura, F. R. de, Braatz, J., & Marino, C. E. dos S. (2024). O EFEITO DE VARIÁVEIS MACROECONÔMICAS NACIONAIS NAS RECEITAS PÚBLICAS: UM ESTUDO PARA A PREVISÃO DO ICMS DOS ESTADOS DO NORDESTE. *Cadernos Do IME - Série Estatística*, 55, 43–82. <https://doi.org/10.12957/cadest.2023.83132>. Acesso em 03 de dezembro de 2024

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106. <https://doi.org/10.1007/BF00116251>. Acesso em 03 de dezembro de 2024.

Secretaria da Fazenda do Estado do Rio Grande do Sul. (2021). Modelo estrutural de previsão para o ICMS no Rio Grande do Sul [Texto de discussão]. Artigo apresentado no 10º Encontro de Economia Gaúcha, promovido pela PUCRS. Disponível em https://tesouro.fazenda.rs.gov.br/upload/1643376139_Artigo%20Modelo%20Estrutural_Texto%20de%20Discussao.pdf . Acesso em 03 de dezembro de 2024.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4), 591–611. <https://doi.org/10.1093/biomet/52.3-4.591>. Acesso em 03 de dezembro de 2024.

Sulaiman, M. H., Jadin, M. S., Mustaffa, Z., Azlan, M. N. M., & Daniyal, H. (2024). Shortterm forecasting of floating photovoltaic power generation using machine learning models. *Cleaner Energy Systems*, 9 , 100137. <https://doi.org/10.1016/j.cles.2024.100137>. Acesso em 03 de dezembro de 2024

Willmott, C. J., & Matsuura, K. (2005). Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in assessing average model performance. *Climate Research*, 30, 79–82. <https://doi.org/10.3354/cr030079>. Acesso em 03 de dezembro de 2024.