UNIVERSIDADE ESTADUAL DO MARANHÃO CENTRO DE CIÊNCIAS TECNOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS MESTRADO PROFISSIONAL EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS

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

PHILIPE SAMPAIO LIMA

Dissertação apresentada ao curso de Mestrado Profissional em Engenharia da Computação e Sistemas na Universidade Estadual do Maranhão como requisito para obtenção do título de Mestre sob orientação do Prof. Dr. Ewaldo Eder Carvalho Santana.

2025

UNIVERSIDADE ESTADUAL DO MARANHÃO CENTRO DE CIÊNCIAS TECNOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS MESTRADO PROFISSIONAL EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS

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

PHILIPE SAMPAIO LIMA

2025

Lima, Philipe Sampaio
Estimativas de arrecadação do ICMS do Estado do Maranhão usando algoritmos de machine learning / Philipe Sampaio Lima. – São Luis, MA, 2025.
44 fl.

Dissertação (Mestrado Profissional em Engenharia da Computação e Sistemas) - Universidade Estadual do Maranhão, 2025.

Orientador: Prof. Dr. Ewaldo Eder Carvalho Santana.

1.Arrecadação de ICMS. 2.Machine Learning. 3.Previsão de Séries Temporais. 4.Finanças Públicas. 5.Maranhão. I.Titulo.

CDU: 336.225.62:004.8

Elaborado por Cássia Diniz - CRB 13/910

Philipe Sampaio Lima

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

Dissertação apresentada ao curso de Mestrado Profissional em Engenharia da Computação e Sistemas na Universidade Estadual do Maranhão como requisito para obtenção do título de Mestre sob orientação do Prof. Dr. Ewaldo Eder Carvalho Santana.

Documento assinado digitalmente
gov.br
EWALDO EDER CARVALHO SANTANA
Data: 22/12/2025 13:30:45-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Ewaldo Eder Carvalho Santana.

Documento assinado digitalmente
gov.br
MAURO SERGIO SILVA PINTO
Data: 22/12/2025 19:36:05-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Mauro Sérgio Silva Pinto

Documento assinado digitalmente
gov.br
THIAGO CARDOSO FERREIRA
Data: 29/12/2025 16:20:21-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Thiago Cardoso Ferreira.

São Luís – MA 2025

RESUMO

A previsão da arrecadação de tributos é um pilar para o planejamento fiscal e a gestão pública eficiente. O Imposto sobre Circulação de Mercadorias e Serviços (ICMS) constitui a principal fonte de receita para os estados brasileiros, e sua projeção acurada é fundamental para a alocação de recursos em áreas estratégicas. No entanto, a complexidade de sua dinâmica, influenciada por variáveis macroeconômicas não lineares, e a carência de estudos aplicados à realidade do estado do Maranhão representam um desafio para os gestores públicos. Este trabalho se propõe a resolver essa lacuna, investigando como técnicas de Machine Learning podem aprimorar a precisão na previsão da arrecadação mensal do ICMS no Maranhão. O objetivo geral é desenvolver e validar modelos computacionais avançados, utilizando uma série histórica de dados econômicos e sociais de janeiro de 1997 a abril de 2024. A pesquisa, de natureza quantitativa e aplicada, adotou o framework CRISP-DM. Os dados foram coletados de fontes públicas como a SEFAZ-MA, IBGE e Banco Central. Inicialmente, dezenove variáveis independentes foram consideradas, e um modelo de Regressão Linear Múltipla será utilizado para selecionar as mais relevantes, como o PIB, o consumo de diesel e indicadores de consumo de energia elétrica. Serão implementados, comparados e validados quatro algoritmos de Machine Learning: Random Forest, Decision Tree, Regressão Linear e XGBoost. A avaliação de desempenho será realizada com as métricas RMSE, MAE, MAPE, SMAPE e R^2, utilizando a técnica de validação cruzada (k-fold com k=10) e uma divisão de dados em 80% para treino e 20% para teste. Este estudo contribui com um modelo prático e validado que pode ser integrado ao processo de planejamento orçamentário do estado, promovendo uma gestão fiscal mais transparente, eficiente e baseada em dados.

Palavras-Chave: Arrecadação de ICMS; Machine Learning; Previsão de Séries Temporais; Finanças Públicas; Maranhão.

ABSTRACT

Tax collection forecasting is a cornerstone of fiscal planning and efficient public management. The Tax on Circulation of Goods and Services (ICMS) constitutes the main source of revenue for Brazilian states, and its accurate projection is crucial for allocating resources to strategic areas. However, the complexity of its dynamics, influenced by non-linear macroeconomic variables, and the lack of studies applied to the reality of the state of Maranhão pose a challenge for public administrators. This work aims to address this gap by investigating how machine learning techniques can improve the accuracy of forecasting monthly ICMS revenue in Maranhão. The overall objective is to develop and validate advanced computational models using a historical series of economic and social data from January 1997 to April 2024. This quantitative and applied research adopted the CRISP-DM framework. Data were collected from public sources such as SEFAZ-MA, IBGE, and the Central Bank. Initially, nineteen independent variables were considered, and a Multiple Linear Regression model was used to select the most relevant ones, such as GDP, diesel consumption, and electricity consumption indicators. Four machine learning algorithms were implemented, compared, and validated: Random Forest, Decision Tree, Linear Regression, and XGBoost. Performance evaluation was performed using the RMSE, MAE, MAPE, SMAPE, and R^2 metrics, using the k-fold cross-validation technique (with k=10) and a data split of 80% for training and 20% for testing. This study contributes a practical and validated model that can be integrated into the state's budget planning process, promoting more transparent, efficient, and data-driven fiscal management.

Keywords: ICMS Revenue; Machine Learning; Time Series Forecasting; Public Finance; Maranhão.

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ARIMA – Auto-Regressive Integrated Moving Average (Média Móvel Integrada Autoregressiva)
BCB – Banco Central do Brasil
CBS – Contribuição sobre Bens e Serviços
CNN – Convolutional Neural Network (Rede Neural Convolucional)
CRISP-DM – Cross-Industry Standard Process for Data Mining (Processo Padrão Inter-Indústrias para Mineração de Dados)
ELM – Extreme Learning Machine (Máquina de Aprendizagem Extrema)
GARCH – Generalized Autoregressive Conditional Heteroskedasticity (Heterocedasticidade Condicional Autoregressiva Generalizada)
GLM – Generalized Linear Model (Modelo Linear Generalizado)
HAR – Heterogeneous Autoregressive model (Modelo Autoregressivo Heterogêneo)
IBGE – Instituto Brasileiro de Geografia e Estatística
IBS – Imposto sobre Bens e Serviços
ICMS – Imposto sobre Circulação de Mercadorias e Serviços INPC – Índice Nacional de Preços ao Consumidor
IPCA – Índice Nacional de Preços ao Consumidor Amplo LightGBM – Light Gradient Boosting Machine
LIME – Local Interpretable Model-agnostic Explanations LSTM – Long Short-Term Memory (Memória de Longo Prazo)
MA – Maranhão
MAE – Erro Médio Absoluto
MAPE – Erro Percentual Médio Absoluto
MDIC – Ministério da Indústria, Comércio Exterior e Serviços
Mtb – Ministério do Trabalho
N-BEATS – Neural Basis Expansion Analysis for Time Series
OLS – Ordinary Least Squares (Mínimos Quadrados Ordinários)
P&C – Perturb and Combine (Perturbar e Combinar)
PCA – Principal Component Analysis (Análise de Componentes Principais)

PECS – Programa de Pós-Graduação em Engenharia de Computação e Sistemas

PIB – Produto Interno Bruto

R^2 – Coeficiente de Determinação

RMSE – Root Mean Squared Error (Raiz do Erro Quadrático Médio)

SARIMA – Seasonal Auto-Regressive Integrated Moving Average (ARIMA Sazonal)

SEFAZ-MA – Secretaria da Fazenda do Estado do Maranhão

Secex – Secretaria de Comércio Exterior

SHAP – SHapley Additive exPlanations

SMAPE – Erro Percentual Médio Absoluto Simétrico

SVR – Support Vector Regression (Regressão de Vetores de Suporte)

UEMA – Universidade Estadual do Maranhão

VAR – Vector Autoregressive (Vetor Autoregressivo)

XGBoost – Extreme Gradient Boosting

AGRADECIMENTOS

Agradeço a Deus pela oportunidade de tudo; à minha família; à minha esposa, Keila Karla Matos Setenta, que tanto me apoiou; à minha mãe, Fernanda Sampaio; ao meu orientador, Ewaldo Éder; ao meu amigo e professor, Kassyo Augusto; ao meu professor de longa data, Thiago Cardoso; à minha amiga e também chefe, Luciana Gehlen, pelo incentivo e motivação; e a toda a minha equipe de serviço, que segurou as pontas quando não pude estar presente. Reconheço que este é um agradecimento simples, se comparado à contribuição de todos que, de forma direta ou indireta, me ajudaram a chegar até aqui. Essa conquista é um troféu que não ergo sozinho.

Sumário

1. INTRODUÇÃO

A previsão de receitas tributárias constitui elemento fundamental para o planejamento fiscal estratégico e a gestão pública eficiente, particularmente em economias complexas como a brasileira, caracterizada por significativa heterogeneidade regional e estrutura tributária multifacetada. No contexto federativo brasileiro, o Imposto sobre Circulação de Mercadorias e Serviços (ICMS) representa a principal fonte de arrecadação dos entes subnacionais, configurando- se como instrumento essencial para o financiamento de políticas públicas e investimentos em setores prioritários como educação, saúde e infraestrutura (MURTA, 2024; SECRETARIA DA FAZENDA DO ESTADO DO RIO GRANDE DO SUL, 2021).

A literatura especializada evidencia que as projeções orçamentárias elaboradas por órgãos oficiais frequentemente apresentam viés otimista sistemático, caracterizado pela superestimação do crescimento econômico e das receitas futuras, especialmente durante ciclos econômicos expansivos (FRANKEL, 2011). Este fenômeno pode induzir à adoção de políticas fiscais procíclicas, onde os governos subnacionais falham na constituição de reservas durante períodos favoráveis, comprometendo sua capacidade de resposta a choques econômicos adversos. Adicionalmente, a arrecadação do ICMS é influenciada por um conjunto multidimensional de variáveis macroeconômicas — incluindo taxa de juros, inflação, câmbio e indicadores setoriais — cujas interações apresentam características dinâmicas e não-lineares, conferindo elevada complexidade aos modelos de previsão (GOMES, 2023; PEDROSA et al., 2023).

Historicamente, a modelagem preditiva de receitas tributárias tem se fundamentado em métodos econométricos tradicionais e modelos de séries temporais, particularmente a família de modelos ARIMA (Auto-Regressive Integrated Moving Average). Estas abordagens demonstram eficácia na captura de padrões sazonais e tendências temporais, tendo sido aplicadas com êxito na projeção da arrecadação de ICMS em diversos estados brasileiros (SOUSA et al., 2019). Todavia, embora adequados para cenários de relativa estabilidade econômica, tais modelos apresentam limitações na incorporação de múltiplas

variáveis exógenas e na modelagem de relações complexas, especialmente em contextos de choques estruturais (MARTINS; GALEGALE, 2021).

A evolução das técnicas analíticas, impulsionada pelo incremento da capacidade computacional e pela crescente disponibilidade de grandes volumes de dados (big data), tem propiciado o desenvolvimento de abordagens inovadoras. Nesse contexto, as técnicas de Aprendizado de Máquina (Machine Learning) emergem como alternativa metodológica promissora para a previsão fiscal (AKINRINOLA et al., 2024). Algoritmos como Random Forest, Gradient Boosting e Redes Neurais Artificiais demonstram capacidade superior na identificação de padrões complexos e relações não-lineares em dados multidimensionais, prescindindo de premissas rígidas sobre distribuições estatísticas e frequentemente resultando em ganhos significativos de acurácia preditiva (AL- KARKHIA; RZĄDKOWSKI, 2025).

A aplicação de técnicas de aprendizado de máquina na previsão de indicadores econômicos, ciclos recessivos e arrecadação tributária tem se consolidado como fronteira de pesquisa em economia aplicada e finanças públicas, com estudos empíricos demonstrando seu potencial para aprimorar processos de gestão pública e tomada de decisão baseada em evidências (SANTOS, 2022; DÖPKE; FRITSCHE; PIERDZIOCH, 2017). O presente trabalho insere-se nesta vertente metodológica, propondo a aplicação e validação de modelos computacionais avançados para a previsão da receita de ICMS no estado do Maranhão.

1.1 Contextualização

O ICMS caracteriza-se como tributo indireto de competência estadual, cuja arrecadação apresenta dependência funcional de múltiplos fatores macroeconômicos e setoriais, incluindo nível de atividade econômica, variações cambiais, política monetária e medidas de política fiscal. Esta multiplicidade de determinantes confere elevada complexidade à modelagem preditiva, constituindo desafio permanente para gestores públicos na elaboração de projeções confiáveis e no estabelecimento de metas orçamentárias realistas.

As incertezas inerentes à conjuntura econômica, tanto em âmbito nacional quanto regional, tornam a tarefa de previsão de variáveis econômicas fundamental e desafiadora, especialmente em cenários de crise, devido à volatilidade e incertezas políticas e econômicas (ZANAZZI DORNELLES et al., 2022). Nesse contexto, demanda-se o desenvolvimento de instrumentos preditivos robustos e sofisticados, capazes de incorporar a volatilidade característica dos mercados contemporâneos e de fornecer suporte técnico adequado aos processos decisórios. A ausência de ferramentas analíticas robustas, que incorporem os avanços metodológicos atuais, compromete significativamente a acurácia das projeções e, consequentemente, a tomada de decisão (AMARAL et al., 2023).

Tradicionalmente, a literatura tem empregado métodos estatísticos e econométricos convencionais para a previsão de arrecadação tributária. Contudo, os avanços em tecnologias de análise de dados e a expansão da disponibilidade de informações históricas têm evidenciado a superioridade de técnicas de aprendizado de máquina na modelagem de fenômenos caracterizados por não-linearidade e dinamicidade, como a arrecadação do ICMS (DORNELLES, SCHWARTZER; BRAATZ, 2022; BLACK, 2021).

Os modelos contemporâneos possibilitam a incorporação sistemática de variáveis macroeconômicas — como índices de preços ao consumidor, taxa SELIC e cotação cambial — além de indicadores setoriais relacionados ao comércio e à produção industrial, resultando em incremento substantivo da precisão das estimativas. A implementação de técnicas de validação cruzada e a utilização de métricas quantitativas de desempenho permitem avaliação objetiva dos algoritmos, elevando a confiabilidade dos resultados obtidos e sua aplicabilidade prática.

No contexto específico do Maranhão, a aplicação de técnicas avançadas de Machine Learning para previsão de arrecadação do ICMS permanece incipiente, representando oportunidade significativa para contribuições tanto acadêmicas quanto para o aprimoramento da gestão pública estadual. O desenvolvimento de modelos adaptados às especificidades regionais pode proporcionar melhor compreensão das variáveis determinantes da arrecadação e servir como fundamento para políticas públicas mais eficientes.

1.2 Problema de Pesquisa

Não obstante a relevância estratégica do ICMS para as finanças estaduais, observa-se escassez de estudos quanto à previsão de sua arrecadação com o uso de algoritmos de machine learning, especialmente no contexto do estado do Maranhão. Uma pesquisa realizada nas bases ScienceDirect e ACM Digital Library, utilizando termos como “forecasting tax”, “forecasting revenue”, “machine learning tax forecasting” e outros correlatos, revelou a ausência de trabalhos específicos voltados ao Maranhão e um número limitado de estudos aplicados ao Brasil como um todo. Essa lacuna contrasta com a crescente aplicação de técnicas de aprendizado de máquina em previsões fiscais em outros contextos regionais, como demonstrado por estudos que utilizaram modelos como LSTM, Random Forest e Gradient Boosting para estimar receitas tributárias em estados como Rio de Janeiro (SILVA; FIGUEIREDO, 2020), Minas Gerais (MURTA, 2024), Espírito Santo (CARMO; BOLDT; KOMATI, 2023) e o Nordeste brasileiro de forma agregada (PEDROSA et al., 2023).

Esta lacuna metodológica impede o desenvolvimento de previsões mais precisas da receita tributária, comprometendo negativamente os processos de planejamento orçamentário e a execução eficiente de políticas públicas.

Adicionalmente, o processo de reforma tributária em curso no Brasil, que prevê a substituição gradual do ICMS por novos tributos — como o Imposto sobre Bens e Serviços (IBS) e a Contribuição sobre Bens e Serviços (CBS) — demanda avaliações prospectivas do comportamento histórico do ICMS para fundamentar tecnicamente as mudanças estruturais propostas no sistema tributário nacional (BRASIL, 2023). Neste contexto, torna-se imperativa a disponibilização de ferramentas analíticas capazes de simular diferentes cenários econômicos e projetar impactos futuros com base em padrões históricos identificados.

Diante deste quadro problemático, formula-se a seguinte questão de pesquisa: **Como técnicas de Machine Learning podem ser aplicadas para aprimorar a precisão na previsão da arrecadação do ICMS no estado do Maranhão, considerando variáveis econômicas e sociais como determinantes preditivos**?

1.3 Objetivos

A presente investigação científica tem como finalidade precípua aplicar técnicas computacionais avançadas de Machine Learning para a previsão da arrecadação do ICMS no estado do Maranhão, utilizando variáveis econômicas e sociais como insumos para modelos preditivos robustos, de forma a identificar qual o melhor algoritmo para esta previsão.

1.3.1 Objetivo Geral

Desenvolver modelos computacionais de Machine Learning para prever a arrecadação mensal do ICMS no estado do Maranhão, fundamentados em variáveis econômicas e sociais, visando contribuir para o aprimoramento do planejamento fiscal estratégico e da gestão pública baseada em evidências.

1.3.2 Objetivos Específicos

- Identificar e selecionar variáveis econômicas e sociais estatisticamente relevantes para explicar a variabilidade da arrecadação do ICMS no Maranhão, mediante aplicação de técnicas de seleção de variáveis;
- Realizar tratamento e pré-processamento rigoroso dos dados coletados, assegurando qualidade, consistência e adequação das informações utilizadas nos modelos preditivos;
- Desenvolver, implementar e comparar sistematicamente diferentes algoritmos de Machine Learning — incluindo XGBoost, Random Forest, Regressão Linear e Decision Tree — para identificar o modelo com superior desempenho preditivo;
- Avaliar quantitativamente o desempenho dos modelos mediante aplicação de métricas estatísticas consolidadas: RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), sMAPE (Symmetric Mean Absolute Percentage Error) e R^2 (Coeficiente de Determinação);

- Analisar e interpretar os resultados obtidos à luz da literatura especializada, discutindo suas implicações práticas para a gestão fiscal e formulação de políticas públicas no estado do Maranhão.

1.4 Justificativa

O presente estudo justifica-se pela importância estratégica da previsão acurada da arrecadação do ICMS para o planejamento financeiro e orçamentário do Estado do Maranhão. Em contexto de crescente volatilidade econômica e dos desafios enfrentados pelo setor público na gestão eficiente de recursos escassos, modelos preditivos robustos e tecnologicamente atualizados tornam-se instrumentos indispensáveis para fundamentar decisões informadas e promover alocação otimizada de recursos públicos.

A aplicação de técnicas de Machine Learning para previsão da arrecadação do ICMS representa uma abordagem metodologicamente inovadora no contexto estadual, especialmente diante da predominância de modelos estatísticos tradicionais — como séries temporais ARIMA ou regressões lineares — na literatura especializada. Embora amplamente utilizados, tais modelos frequentemente assumem linearidade, estacionariedade e relações fixas entre variáveis, o que pode limitar sua capacidade de capturar dinâmicas não lineares, interações complexas e choques estruturais comuns nas séries de receita tributária (PEDROSA et al., 2023; SOUSA et al., 2019).

Essas limitações podem se traduzir em previsões enviesadas ou superotimistas, com implicações concretas para a gestão orçamentária: superestimativas de receita levam a comprometimentos orçamentários insustentáveis, enquanto subestimativas podem resultar em subutilização de recursos ou alocação ineficiente de políticas públicas (FRANKEL, 2011; RODGERS; JOYCE, 1996).

Nesse sentido, algoritmos como Random Forest e XGBoost — capazes de modelar relações não lineares, interações entre variáveis macroeconômicas e lags heterogêneos sem pressupostos restritivos — demonstram desempenho superior em termos de acurácia e robustez (MURTA, 2024; DORNELLES; SCHWARTZER; BRAATZ,

2022). A adoção desses modelos avançados pode, portanto, contribuir diretamente para a melhoria da qualidade do planejamento fiscal, a redução de riscos orçamentários e o fortalecimento da Lei de Responsabilidade Fiscal (LRF), ao oferecer estimativas mais confiáveis das receitas esperadas. Para o Estado do Maranhão — que, como demonstrado nesta pesquisa, carece de estudos específicos nessa área —, essa inovação metodológica representa não apenas um avanço técnico, mas uma ferramenta estratégica para a melhoria da governança fiscal e a otimização do gasto público.

Do ponto de vista metodológico, a construção de modelos de previsão de arrecadação de ICMS específicos para o Maranhão é essencial para capturar as particularidades estruturais e conjunturais desse estado. Diferentemente de abordagens genéricas ou nacionais, um modelo local pode incorporar variáveis que refletem com maior fidelidade a composição setorial da economia maranhense — marcada pela relevância da agropecuária, da indústria extrativa (notadamente o complexo de alumínio e minério de ferro em São Luís e região) e dos serviços portuários —, bem como choques regionais, como sazonalidade climática, dinâmicas demográficas e políticas fiscais locais. Ignorar essas especificidades pode resultar em previsões enviesadas, com implicações diretas para o planejamento orçamentário, a alocação de recursos em áreas críticas (saúde, educação, segurança) e o cumprimento da Lei de Responsabilidade Fiscal (LRF).

Além disso, esta pesquisa contribui para preencher uma lacuna significativa na literatura, já que, conforme identificado em buscas nas bases ScienceDirect e ACM Library com termos como “forecasting tax”, “machine learning tax forecasting” e “revenue prediction Brazil”, não há estudos empíricos aplicados ao Maranhão e apenas um número reduzido voltado a outros estados brasileiros (PEDROSA et al., 2023; MURTA, 2024; CARMO; BOLDT; KOMATI, 2023). Ao demonstrar a viabilidade e o ganho de precisão da aplicação de algoritmos avançados de machine learning — como Random Forest e XGBoost — em um contexto de dados fiscais estaduais limitados e não estacionários, este trabalho expande o repertório técnico disponível para gestores públicos e fortalece a ciência de dados aplicada ao setor público no Brasil, com potencial de

replicação em outros estados com perfis econômicos semelhantes. Assim, a pesquisa não apenas avança academicamente, mas também oferece ferramentas concretas para a melhoria da governança fiscal e da transparência orçamentária em um dos estados mais desiguais e fiscalmente vulneráveis do país.

Finalmente, os resultados desta pesquisa podem subsidiar tecnicamente processos de modernização da gestão fiscal estadual, promovendo maior transparência, eficiência e fundamentação científica nas decisões de política pública, com potencial impacto positivo na qualidade dos serviços públicos oferecidos à população maranhense.

2. REVISÃO DE LITERATURA

A consulta da literatura disponível, apresenta os fundamentos teóricos e conceituais relacionados à previsão da arrecadação do ICMS, com ênfase nos modelos estatísticos e algoritmos de Machine Learning aplicados a esta problemática. Esta etapa visa contextualizar teoricamente o objeto de pesquisa, identificar lacunas metodológicas na literatura especializada e fundamentar cientificamente a abordagem metodológica proposta. Para tanto, são examinados os conceitos fundamentais sobre o tributo em questão, os modelos historicamente utilizados para previsão fiscal, além dos avanços recentes no emprego de técnicas computacionais avançadas para análise preditiva em contextos tributários.

2.1 Fundamentos Conceituais sobre ICMS

O Imposto sobre Circulação de Mercadorias e Serviços (ICMS) constitui um dos principais tributos do sistema federativo brasileiro, instituído pela Constituição Federal de 1988 e regulamentado pelas legislações estaduais específicas. Sua base de incidência está fundamentalmente associada à dinâmica econômica, particularmente às operações de circulação de mercadorias e prestação de serviços, conferindo-lhe elevada sensibilidade às variações da atividade econômica regional (BRASIL, 2023). No estado do Maranhão, o ICMS representa parcela substancial da receita estadual, constituindo fonte de financiamento

fundamental para políticas públicas essenciais nos setores de educação, saúde e infraestrutura.

A arrecadação do ICMS apresenta dependência funcional de múltiplos fatores macroeconômicos, incluindo variações cambiais, política monetária, crescimento do PIB e medidas de política fiscal. Adicionalmente, mudanças no marco regulatório e reformas tributárias exercem impacto direto sobre sua dinâmica arrecadatória. Esta complexidade multidimensional demanda o desenvolvimento de modelos preditivos robustos que considerem não apenas dados históricos de arrecadação, mas também variáveis explicativas economicamente relevantes (MURTA, 2024).

A compreensão aprofundada do comportamento do ICMS constitui requisito essencial para o planejamento orçamentário eficiente e a gestão pública baseada em evidências. Em contextos de elevada incerteza econômica e social, modelos preditivos progressivamente mais sofisticados têm sido desenvolvidos para auxiliar gestores públicos na elaboração de projeções confiáveis e tecnicamente fundamentadas, contribuindo para processos decisórios informados e alocação otimizada de recursos públicos (SECRETARIA DA FAZENDA DO ESTADO DO RIO GRANDE DO SUL, 2021).

2.2 Modelos de Previsão de Arrecadação Tributária

A previsão de arrecadação tributária tem sido objeto de extensa investigação acadêmica, com desenvolvimento de metodologias que abrangem desde técnicas econométricas tradicionais até algoritmos computacionais avançados de modelagem estatística. Historicamente, técnicas econométricas estruturais e modelos lineares têm sido empregados para prever receitas públicas, incluindo o ICMS (BLACK, 2021). Estes modelos possibilitam a incorporação sistemática de variáveis macroeconômicas e setoriais, oferecendo fundamentação teórica sólida para análise preditiva. Contudo, a crescente disponibilidade de dados e avanços tecnológicos têm impulsionado a utilização de técnicas mais avançadas, incluindo modelos de séries temporais e algoritmos de aprendizado de máquina.

Os modelos estatísticos clássicos, particularmente regressão linear múltipla

e modelos ARIMA, mantêm relevância devido à sua simplicidade interpretativa e transparência metodológica. Tais abordagens demonstram eficácia em contextos caracterizados por baixa volatilidade e relações predominantemente lineares entre variáveis. Todavia, em ambientes caracterizados por elevada não-linearidade e dinamicidade, como a arrecadação do ICMS em estados com economias heterogêneas, estas abordagens podem apresentar limitações significativas (DORNELLES; SCHWARTZER; BRAATZ, 2022).

Os modelos baseados em séries temporais têm demonstrado eficácia na captura de padrões sazonais e tendências temporais. Técnicas como SARIMA (Seasonal ARIMA) e suavização exponencial (Holt-Winters) são frequentemente aplicadas na previsão de receitas tributárias, permitindo análises fundamentadas em dados históricos com periodicidade mensal ou trimestral. Não obstante suas contribuições metodológicas, estes modelos também apresentam limitações quanto à capacidade de incorporar múltiplas variáveis externas de forma sistêmica (LIMA et al., 2025).

Nos últimos anos, técnicas de Aprendizado de Máquina têm emergido como alternativa metodológica promissora para previsão fiscal, superando muitas das restrições inerentes aos métodos tradicionais. Algoritmos como XGBoost, Random Forest e Decision Tree têm demonstrado superior acurácia na previsão de arrecadação de ICMS, especialmente quando aplicados a conjuntos de dados caracterizados por riqueza de variáveis econômicas e sociais (MURTA, 2024). Estes modelos não-paramétricos demonstram capacidade de capturar relações complexas e não-lineares entre variáveis, resultando em estimativas de superior qualidade.

A validação cruzada e a utilização de métricas quantitativas de desempenho, incluindo RMSE, MAE, MAPE e R^2, possibilitam avaliação objetiva e comparativa do desempenho dos modelos. Esta abordagem metodológica incrementa a confiabilidade das previsões e permite ajustes contínuos para melhor adaptação a diferentes cenários econômicos e políticos (LIMA et al., 2025).

2.2.1 Modelos Estatísticos e Econométricos

Os modelos estatísticos e econométricos têm sido extensivamente utilizados

na previsão de arrecadação tributária, particularmente em contextos onde as relações entre variáveis apresentam características relativamente lineares e teoricamente bem fundamentadas. Entre as técnicas mais consolidadas destacam-se a regressão linear múltipla, modelos VAR (Vector Autoregressive) e sistemas de equações simultâneas, que possibilitam a incorporação sistemática de variáveis explicativas como PIB, inflação, taxa de juros e outros indicadores macroeconômicos relevantes (BLACK, 2021).

Estes modelos apresentam particular utilidade na compreensão das relações causais entre variáveis e na provisão de explicações teoricamente consistentes sobre os fatores determinantes da arrecadação. Adicionalmente, possuem sólida fundamentação teórica e ampla aceitação na literatura econômica especializada (DORNELLES; SCHWARTZER; BRAATZ, 2022). Contudo, uma de suas principais limitações metodológicas reside na pressuposição de linearidade nas relações, o que pode resultar em previsões imprecisas em cenários caracterizados por alta volatilidade ou mudanças estruturais significativas.

Estudos empíricos, como o desenvolvido por Black (2021), aplicaram modelos estruturais de séries temporais para prever a arrecadação do ICMS em Minas Gerais, obtendo resultados satisfatórios em períodos de relativa estabilidade econômica. Entretanto, em momentos de crise ou retração econômica, os erros de previsão tendem a aumentar substancialmente, evidenciando a necessidade de complementar estas abordagens com técnicas metodologicamente mais flexíveis.

No contexto brasileiro, trabalhos pioneiros exploraram a aplicação de modelos econométricos para previsão fiscal, como o estudo conduzido pela Secretaria da Fazenda do Rio Grande do Sul (2021), que desenvolveu modelo estrutural baseado em regressão múltipla para simular diferentes cenários de arrecadação do ICMS. Embora tenha obtido resultados satisfatórios em termos de consistência teórica, o modelo apresentou limitações em cenários com choques econômicos imprevistos.

A literatura contemporânea aponta que os modelos estatísticos tradicionais carecem de flexibilidade para capturar relações não-lineares e interações complexas entre variáveis. Consequentemente, observa-se tendência crescente de combinação destas técnicas com métodos de aprendizado de máquina, criando modelos híbridos que aproveitam as vantagens metodológicas de ambas as abordagens (LIMA et al.,

2025).

2.2.2 Modelos Baseados em Séries Temporais

Os modelos baseados em séries temporais constituem instrumental metodológico amplamente consolidado na previsão de fenômenos cujo comportamento evolui temporalmente, como a arrecadação tributária. Técnicas como ARIMA, SARIMA e modelos de suavização exponencial (Holt-Winters) são comumente empregadas para capturar tendências, sazonalidades e ciclos presentes nos dados históricos (LIMA et al., 2025).

Uma das principais vantagens destes modelos reside na capacidade de extrapolação de padrões observados historicamente para períodos futuros, prescindindo de conhecimento explícito sobre variáveis explicativas. Esta característica os torna particularmente úteis em situações onde os fatores determinantes da arrecadação não são completamente compreendidos ou não estão disponíveis (BLACK, 2021).

No contexto específico do ICMS, estudos como o de Murta (2024) destacam a eficácia do modelo SARIMA na previsão da arrecadação mensal em Minas Gerais, especialmente para captura de sazonalidades anuais e tendências de longo prazo. Contudo, estes modelos também enfrentam limitações significativas quando ocorrem mudanças estruturais, como reformas tributárias ou crises econômicas súbitas.

A integração de modelos de séries temporais com variáveis exógenas tem constituído abordagem metodológica promissora. O modelo ARIMAX, por exemplo, permite incluir variáveis econômicas como entrada adicional, combinando modelagem temporal com fatores explicativos externos (DORNELLES; SCHWARTZER; BRAATZ, 2022). Este tipo de abordagem tem demonstrado maior robustez em cenários economicamente dinâmicos.

Não obstante os avanços metodológicos, os modelos de séries temporais ainda enfrentam dificuldades na manipulação de grandes volumes de dados e na modelagem de relações não-lineares complexas. Neste sentido, a literatura contemporânea aponta para convergência gradual entre técnicas tradicionais e métodos de Machine Learning, visando superar as limitações individuais de cada

abordagem metodológica (NITZSCHE; SCHIMECZEK; BERTSCH, 2024).

2.2.3 Aprendizado de Máquina Aplicado à Previsão Fiscal

Com o avanço das tecnologias de análise de dados e a crescente disponibilidade de bases históricas estruturadas, o aprendizado de máquina tem se consolidado como ferramenta metodológica poderosa na previsão de arrecadação tributária. Algoritmos como XGBoost, Random Forest, Regressão Linear e Decision Tree têm demonstrado desempenho superior em relação aos modelos tradicionais, especialmente na identificação de padrões complexos e relações não-lineares (MURTA, 2024).

Estes modelos apresentam capacidade de processamento de grandes volumes de dados e incorporação sistemática de múltiplas variáveis independentes, incluindo índices de preços, taxa SELIC e cotação cambial, possibilitando modelagem mais precisa e adaptável a diferentes cenários econômicos (LIMA et al., 2025). Adicionalmente, técnicas de validação cruzada e métricas de avaliação quantitativas, como RMSE, MAE e R^2, permitem comparação e refinamento sistemático dos modelos.

Estudos empíricos, como o de Dornelles, Schwartzer e Braatz (2022), demonstraram a eficácia de redes neurais recorrentes na previsão da arrecadação do ICMS no Rio Grande do Sul, superando modelos tradicionais em termos de precisão preditiva. Outras investigações confirmam que algoritmos de ensemble, como XGBoost e Random Forest, tendem a obter os melhores resultados em cenários caracterizados por alta volatilidade e não-linearidade (NITZSCHE; SCHIMECZEK; BERTSCH, 2024).

A principal vantagem metodológica do Machine Learning reside na sua flexibilidade e capacidade de aprendizagem a partir dos dados, sem dependência de pressuposições rígidas sobre distribuição das variáveis ou forma funcional das relações. Esta característica os torna especialmente adequados para contextos como o do Maranhão, onde a economia apresenta heterogeneidade e sofre influências de múltiplos fatores (MURTA, 2024).

A implementação destes modelos em linguagens como Python, utilizando

bibliotecas especializadas como Scikit-learn, Pandas e XGBoost, permite automatização do processo de previsão e integração de atualizações em tempo real, incrementando a eficiência e utilidade prática para a gestão pública (PEDREGOSA et al., 2011).

2.3 Estado da Arte

O estado da arte na previsão de arrecadação do ICMS revela transição metodológica gradual de modelos estatísticos tradicionais para técnicas computacionais avançadas de Machine Learning. Investigações recentes têm demonstrado que algoritmos como XGBoost, Random Forest e redes neurais recorrentes oferecem superior precisão nas previsões, especialmente quando aplicados a dados históricos enriquecidos com variáveis macroeconômicas e setoriais (DORNELLES; SCHWARTZER; BRAATZ, 2022).

No contexto brasileiro, trabalhos como o de Murta (2024) aplicaram estas técnicas para prever o ICMS em Minas Gerais, obtendo resultados superiores aos modelos convencionais em termos de acurácia e robustez estatística. Adicionalmente, estudos conduzidos no Rio Grande do Sul (SECRETARIA DA FAZENDA DO RS, 2021) corroboram a superioridade dos modelos de aprendizado de máquina em cenários de alta volatilidade e mudança estrutural.

Não obstante os avanços metodológicos, persiste escassez de estudos específicos para o estado do Maranhão, evidenciando necessidade de investigações que explorem modelos adaptados à realidade regional. O desenvolvimento de algoritmos treinados com dados locais pode proporcionar ganhos significativos de acurácia e utilidade prática para a gestão fiscal estadual (LIMA et al., 2025).

2.3.1 Trabalhos Relacionados no Brasil

Diversas investigações têm sido desenvolvidas no Brasil com o objetivo de prever a arrecadação do ICMS, concentrando-se principalmente em estados com maior disponibilidade de dados e infraestrutura tecnológica avançada. Um dos trabalhos pioneiros foi desenvolvido por Black (2021), que utilizou modelos

estruturais de séries temporais para prever a arrecadação em Minas Gerais, obtendo resultados satisfatórios em períodos de estabilidade econômica.

Subsequentemente, Murta (2024) aplicou técnicas de Machine Learning para prever o ICMS no mesmo estado, alcançando desempenho superior em termos de acurácia e robustez estatística. Seus resultados demonstraram que algoritmos como XGBoost e Random Forest superaram significativamente modelos estatísticos tradicionais, especialmente em cenários de alta volatilidade econômica. No Rio Grande do Sul, Dornelles, Schwartzer e Braatz (2022) desenvolveram modelo baseado em redes neurais recorrentes para prever a arrecadação do ICMS, alcançando resultados expressivos em termos de precisão preditiva. Estes autores enfatizam a importância da integração de variáveis macroeconômicas e setoriais para aprimoramento da qualidade das previsões.

A literatura contemporânea tem evidenciado a aplicação crescente de técnicas avançadas de deep learning e outros métodos computacionais inovadores, conforme sistematizado na Tabela 1, que demonstra a diversidade de abordagens metodológicas empregadas por diferentes pesquisadores.

Tabela 1 – Trabalhos Correlatos

Métodos	Resumo e aplicação	Autores
Ensemble methods, Deep Learning models, SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations)	Algoritmos como *ensemble methods* e *deep learning* melhoram a previsão, mas enfrentam desafios como dados desbalanceados e seleção de variáveis em planejamento econômico e estratégias empresariais.	Mustafa I. Al-Karkhi, Grzegorz Rządkowski (2024)
N-BEATS (Neural Basis Expansion Analysis for Time Series), Temporal Fusion Transformers	Modelos como *N-BEATS* e *Temporal Fusion Transformers* 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)

Métodos	Resumo e aplicação	Autores
Redes Neurais *LSTM (Long Short-Term Memory), CNN-LSTM (Convolutional Neural Network - Long Short-Term Memory), GARCH (Generalized Autoregressive Conditional Heteroskedasticity), HAR (Heterogeneous Autoregressive Model*	Redes neurais (*LSTM, CNN-LSTM*) superam modelos tradicionais, com destaque para *CNN-LSTM* em previsões de curto prazo em finanças e criptomoedas.	Zih-Chun Huang, Ivan Sangiorgi, Andrew Urquhart (2024)
LightGBM (Light Gradient Boosting Machine), PCA (Principal Component Analysis), Dimensionality Reduction (Técnicas de redução de dimensionalidade)	*LightGBM* 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), *Random Forest* (Floresta Aleatória), *ELM (Extreme Learning Machine), SVR (Support Vector Regression), LSTM (Long Short-Term Memory)*	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 pelo autor.

2.3.2 Técnicas Utilizadas: XGBoost, Random Forest, Regressão Linear e Decision Tree

Entre os algoritmos de Machine Learning mais utilizados na previsão de arrecadação do ICMS, destacam-se o XGBoost, Random Forest, Regressão Linear e Decision Tree. Cada algoritmo possui características metodológicas distintas que os tornam adequados para diferentes cenários e tipos de dados.

O XGBoost constitui algoritmo de boosting que combina múltiplas árvores de decisão para otimizar a precisão preditiva. É reconhecido por sua eficiência computacional e capacidade de processamento de dados esparsos, sendo amplamente utilizado em competições de ciência de dados (CHEN; GUESTRIN, 2016). Na previsão de ICMS, tem demonstrado eficácia na captura de padrões complexos e relações não-lineares (LIMA et al., 2025).

O Random Forest constitui modelo de ensemble que constrói múltiplas árvores de decisão e combina seus resultados para reduzir o risco de overfitting. É

especialmente útil em conjuntos de dados com grande número de variáveis, demonstrando capacidade de processamento de redundâncias e interações complexas (BREIMAN, 2001). Em estudos como o de Murta (2024), o Random Forest obteve resultados satisfatórios na previsão da arrecadação do ICMS em Minas Gerais.

A Regressão Linear, apesar de constituir método relativamente simples, mantém ampla utilização devido à sua facilidade interpretativa e capacidade de identificação das variáveis mais relevantes no modelo (HAIR et al., 2010). Serve também como fundamentação para técnicas mais avançadas, como regressão Ridge e Lasso, que adicionam regularização para evitar problemas de multicolinearidade. O Decision Tree constitui técnica intuitiva e facilmente visualizável, que segmenta dados em grupos baseados em condições lógicas. Embora apresente menor precisão que modelos de ensemble, é útil para identificação de padrões claros e interpretáveis nos dados (QUINLAN, 1986). Em contextos de gestão pública, esta característica é valiosa para tomada de decisão transparente e tecnicamente fundamentada.

3. METODOLOGIA

A metodologia empregada neste estudo tem como objetivo descrever o processo sistemático de desenvolvimento e validação dos modelos de Machine Learning para previsão da arrecadação do ICMS no estado do Maranhão. A pesquisa foi estruturada em etapas que abrangem desde a definição do tipo de estudo, coleta e pré-processamento dos dados, até a modelagem computacional e avaliação dos resultados. Essa abordagem visa garantir a reprodutibilidade, a precisão e a aplicabilidade prática dos modelos construídos, com base em técnicas reconhecidas na literatura científica.

3.1 Tipo de Pesquisa

Este trabalho caracteriza-se como uma pesquisa aplicada, de natureza quantitativa, fundamentada em métodos computacionais avançados. Segue uma abordagem experimental, com foco na construção de modelos preditivos utilizando

técnicas de Machine Learning, tendo como base uma série temporal histórica da arrecadação mensal do ICMS no Maranhão (MURTA, 2024). O objetivo é gerar conhecimento útil para a gestão pública, contribuindo tanto teoricamente quanto pràticamente para o planejamento fiscal estadual.

3.2 Fontes e Coleta de Dados

Os dados utilizados foram obtidos de fontes públicas e institucionais, como a Secretaria da Fazenda do Estado do Maranhão (SEFAZ-MA), Instituto Brasileiro de Geografia e Estatística (IBGE) e Banco Central do Brasil. Foram coletadas informações mensais referentes à arrecadação do ICMS no período de janeiro de 1997 a abril de 2024, bem como variáveis econômicas e sociais potencialmente explicativas dessa receita tributária (DORNELLES; SCHWARTZER; BRAATZ, 2022).

Além disso, foram compilados indicadores macroeconômicos relevantes, tais como taxa SELIC, índice de preços ao consumidor amplo (IPCA), Produto Interno Bruto (PIB) estadual, valor médio do dólar comercial e outros fatores setoriais relacionados ao comércio e à indústria maranhense. Essas informações foram agregadas em uma base única, permitindo a análise conjunta entre as variáveis independentes e a variável alvo — a arrecadação mensal do ICMS.

A coleta de dados foi realizada em duas etapas principais: primeiramente, obteve-se os valores históricos da arrecadação líquida do ICMS, já corrigidos por inflação e ajustados sazonalmente; posteriormente, foram reunidas as séries temporais das variáveis exógenas, buscando compatibilizar as datas e periodicidades para garantir a consistência dos registros. Todo o processo foi documentado para facilitar a replicação futura.

3.2.1 Variáveis Utilizadas

Para a modelagem preditiva, foram inicialmente consideradas cerca de vinte variáveis econômicas e sociais potencialmente correlacionadas à arrecadação do ICMS no Maranhão. Entre elas, destacaram-se indicadores como o PIB estadual, o Índice Nacional de Preços ao Consumidor Amplo (IPCA), a taxa Selic, o preço médio

do barril do petróleo, o câmbio real-dólar, além de variáveis setoriais como o volume de vendas no varejo, a produção industrial e o consumo de energia elétrica (MURTA, 2024).

Após a realização de uma análise exploratória e a aplicação de técnicas estatísticas, como regressão linear múltipla, foi possível identificar quais dessas variáveis possuíam maior poder explicativo sobre a variação da arrecadação do ICMS. As variáveis com p-valor superior a 0,05 serão descartadas, pois não apresentarão significância estatística suficiente para influenciar o modelo (HAIR et al., 2010).

As variáveis selecionadas incluem: (i) PIB do Maranhão; (ii) IPCA acumulado nos últimos 12 meses; (iii) taxa Selic média mensal; (iv) cotação média do dólar comercial; (v) volume de vendas no varejo do estado; (vi) produção industrial total no Maranhão; e (vii) consumo médio de energia elétrica residencial e comercial. Essas variáveis foram então normalizadas e utilizadas como insumo para os algoritmos de Machine Learning.

A inclusão de variáveis sociais representa uma inovação metodológica importante deste estudo, diferenciando-o de trabalhos anteriores que se limitaram a indicadores econômicos tradicionais. A hipótese é que aspectos demográficos e socioeconômicos também podem influenciar indiretamente na dinâmica da arrecadação tributária, especialmente em contextos regionais específicos (BRASIL, 2023).

Por fim, todas as variáveis foram transformadas em formato numérico padronizado e organizadas em um conjunto de dados estruturado, pronto para ser utilizado nas etapas subsequentes de modelagem e treinamento dos algoritmos preditivos.

Além dos modelos preditivos, será 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. A seguir, a lista de variáveis e suas fontes:

Quadro 1 – Variáveis Utilizadas

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	
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	

Variáveis	Identificação	Fonte
Taxa SELIC	SELIC	Instituto de Pesquisa Econômica Aplicada
Salário Mínimo	SM	
Cotação do Barril do Petróleo	PETRO_COT	

Fonte: Elaborado pelo autor.

3.2.2 Tratamento e Pré-processamento dos Dados

O tratamento e o pré-processamento dos dados serão realizados com o objetivo de garantir a qualidade e a confiabilidade das informações utilizadas nos modelos de Machine Learning. Inicialmente, foi feita uma análise descritiva para identificar valores ausentes, inconsistências e outliers, que poderiam comprometer a eficácia dos algoritmos preditivos (PEDREGOSA et al., 2011).

Para lidar com valores ausentes, optou-se pela interpolação linear ou eliminação de registros incompletos, dependendo da quantidade e relevância da informação. Em seguida, serão aplicadas transformações matemáticas, como logaritmação e normalização, com o intuito de reduzir a assimetria dos dados e equalizar a escala entre variáveis. A técnica MinMaxScaler será utilizada para ajustar os valores ao intervalo [0,1], essencial para algoritmos sensíveis à magnitude das variáveis (HAN et al., 2011).

Como parte do processo de limpeza, também será verificada a multicolinearidade entre as variáveis independentes, evitando redundâncias que pudessem prejudicar a interpretação e a performance dos modelos. Com isso, garantiu-se que os dados fossem adequados para a etapa de modelagem computacional, aumentando a robustez e a capacidade preditiva dos algoritmos escolhidos.

3.3 Modelagem Computacional

A modelagem computacional envolverá a implementação de diferentes algoritmos de Machine Learning para prever a arrecadação mensal do ICMS com base nas variáveis selecionadas. Para isso, foi adotada uma abordagem de aprendizado supervisionado, com divisão dos dados em conjuntos de treino e teste, seguindo a

proporção 80/20. A linguagem Python, versão 3.12.8, juntamente com bibliotecas como Scikit-learn, Pandas, NumPy e XGBoost, foi utilizada para o desenvolvimento dos modelos (PEDREGOSA et al., 2011).

O processo de modelagem inicia-se com a definição de hiperparâmetros e configurações iniciais para cada algoritmo, seguido pelo treinamento com os dados históricos. Após esse estágio, será realizada a validação cruzada com k=10 folds, visando avaliar a estabilidade e a generalização dos modelos em diferentes subconjuntos de dados. Além disso, técnicas de otimização, como GridSearchCV e RandomizedSearchCV, foram aplicadas para encontrar as melhores combinações de parâmetros (CHEN; GUESTRIN, 2016).

Ao final, os modelos serão comparados com base em métricas quantitativas como RMSE, MAE, MAPE, sMAPE e R^2, permitindo identificar aquele com melhor desempenho preditivo. A interpretação dos resultados será complementada com gráficos de importância de variáveis e análise residual, visando compreender quais fatores mais influenciaram nas previsões e onde ocorreram maiores erros de estimativa.

3.3.1 Linguagens e Ferramentas Utilizadas

A implementação dos modelos será realizada utilizando a linguagem de programação Python, versão 3.12.8, devido à sua ampla aceitação na comunidade científica e disponibilidade de bibliotecas especializadas em análise de dados e Machine Learning. Entre as ferramentas utilizadas, destacam-se o Scikit-learn para implementação dos algoritmos, o Pandas para manipulação e preparação dos dados, o NumPy para operações matemáticas e o Matplotlib e Seaborn para visualização dos resultados (PEDREGOSA et al., 2011).

O ambiente de desenvolvimento foi configurado no sistema operacional Windows 10, com uso de Jupyter Notebook para facilitar a execução interativa e a documentação do código. Além disso, a biblioteca XGBoost foi utilizada especificamente para o treinamento dos modelos Gradient Boosting, devido ao seu desempenho superior em competições de ciência de dados e aplicações preditivas (CHEN; GUESTRIN, 2016).

Também será adotado o uso de técnicas de automação e versionamento com Git e GitHub, garantindo rastreabilidade, colaboração e reprodutibilidade do projeto. O uso dessas tecnologias permitiu organizar o fluxo de trabalho de forma eficiente, além de facilitar a manutenção e atualização dos modelos ao longo do tempo.

3.3.2 Algoritmos de Machine Learning Selecionados

Foram selecionados quatro algoritmos de Machine Learning para a previsão da arrecadação do ICMS no Maranhão: XGBoost, Random Forest, Regressão Linear e Decision Tree. Cada um desses modelos possui características distintas que os tornam adequados para diferentes cenários e tipos de dados. O XGBoost, por exemplo, é um método de boosting altamente eficiente e amplamente utilizado em competições de ciência de dados, devido à sua capacidade de lidar com dados esparsos e capturar relações complexas (CHEN; GUESTRIN, 2016).

O Random Forest é um modelo ensemble baseado na combinação de múltiplas árvores de decisão, proporcionando maior robustez e menor risco de overfitting. Ele é especialmente útil quando há grande número de variáveis e interações complexas entre elas (BREIMAN, 2001). Já a Regressão Linear é uma técnica simples e interpretável, frequentemente utilizada como baseline em problemas de previsão, servindo como ponto de partida para comparação com modelos mais avançados (HAIR et al., 2010).

O Decision Tree, por sua vez, é uma técnica intuitiva e fácil de visualizar, sendo útil para identificar padrões claros nos dados. Embora seja menos preciso que os modelos de ensemble, ele pode revelar relações importantes entre as variáveis e auxiliar na interpretação dos resultados (QUINLAN, 1986). A combinação desses algoritmos permite uma análise abrangente e comparativa, ajudando a identificar qual modelo apresenta melhor desempenho na previsão da arrecadação do ICMS.

3.3.3 Métricas de Avaliação dos Modelos

Na avaliação de modelos preditivos baseados em regressão e aprendizado de máquina, o uso adequado de métricas de desempenho é essencial para garantir a

precisão, a robustez e a generalização das previsões (Plebris et al., 2025). Dentre as diversas métricas disponíveis, algumas se destacam por suas propriedades estatísticas e sua ampla aplicação prática: o Erro Médio Absoluto (MAE), o Erro Percentual Médio Absoluto (MAPE), o Erro Percentual Médio Absoluto Simétrico (sMAPE), o coeficiente de correlação (R), o coeficiente de determinação (R^2) e o Erro Quadrático Médio da Raiz (RMSE).

O Erro Médio Absoluto (MAE) mede a média das magnitudes dos erros em um conjunto de previsões, sem considerar a direção dos erros. Ele é simples de interpretar, sendo expresso na mesma unidade da variável-alvo, o que facilita a compreensão prática do erro cometido pelo modelo (Steurer et al., 2020). Porém, o MAE não penaliza erros maiores de forma proporcional ao seu impacto real, limitando sua sensibilidade a valores extremos. A métrica é expressa da seguinte forma:

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

Já o Erro Percentual Médio Absoluto (MAPE) expressa o erro como uma porcentagem, permitindo comparações entre séries com escalas diferentes. No entanto, essa métrica apresenta limitações quando os valores reais são próximos de zero, podendo gerar distorções ou valores infinitos (Botchkarev, 2024). Para mitigar esse problema, o Erro Percentual Médio Absoluto Simétrico (sMAPE) foi introduzido, equilibrando os erros positivos e negativos e oferecendo maior estabilidade em cenários com baixa magnitude de observações. As expressões matemáticas das métricas são:

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

$$sMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}$$

O coeficiente de correlação (R) mede a força da relação linear entre os valores previstos e os valores reais. Valores próximos a 1 indicam alta correlação positiva, sugerindo que o modelo captura bem a tendência dos dados. Entretanto, R não informa sobre a magnitude dos erros, apenas sobre a direção da associação entre as variáveis (Abo El Nasr et al., 2024). A seguir, tem-se a forma de cálculo:

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

O coeficiente de determinação (R^2) complementa essa análise, representando a proporção da variância explicada pelo modelo. Um R^2 próximo a 1 indica que o modelo consegue explicar grande parte da variabilidade dos dados, embora ele também possa ser otimista em amostras com ajuste excessivo (overfitting). Por isso, deve sempre ser usado em conjunto com outras métricas de validação. Apresenta-se a fórmula correspondente à métrica:

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

Por fim, o Erro Quadrático Médio da Raiz (RMSE) é uma das métricas mais utilizadas, especialmente em contextos onde erros maiores devem ser mais penalizados. Isso ocorre porque o RMSE eleva os erros ao quadrado antes de calcular a média, o que aumenta o peso dos grandes desvios. Essa característica o torna útil em aplicações críticas, como previsão financeira e diagnóstico médico, onde a precisão é fundamental (Magazzino et al., 2025; Ramadan et al., 2023). A medida é definida pela seguinte equação:

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

Em síntese, a escolha das métricas de avaliação deve considerar o contexto do problema, a natureza dos dados e os objetivos do modelo. O uso combinado de MAE, MAPE, sMAPE, R, R^2 e RMSE permite uma visão abrangente e crítica do desempenho dos modelos, contribuindo para decisões mais informadas tanto no processo de seleção quanto na interpretação dos resultados (Elmert-Streib & Dehmer, 2024; Steurer et al., 2020).

Essas métricas serão calculadas tanto no conjunto de treino quanto no de teste, garantindo que os modelos sejam avaliados em condições reais de generalização. Com base nesses critérios, foi possível identificar o modelo com melhor desempenho e recomendar sua aplicação prática na gestão fiscal estadual.

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.

4 CRONOGRAMA

Mês/Período	Atividades Previstas
Setembro (2ª quinzena)	• Finalização da Análise de Dados e Modelagem: execução da versão final dos algoritmos (Random Forest, Decision Tree, Regressão Linear e XGBoost); coleta e organização das métricas de desempenho (RMSE, MAE, MAPE, SMAPE e R^2); geração de gráficos e tabelas comparativas.• Início da Redação do Capítulo de Resultados: descrição da análise exploratória e dos resultados da modelagem computacional.
Outubro	• Conclusão da Redação dos Capítulos de Resultados e Discussão: finalização do capítulo 4, interpretação dos resultados à luz da literatura e discussão das implicações práticas para a gestão fiscal do Maranhão.• Redação do Capítulo de Conclusões e Trabalhos Futuros: síntese das contribuições, limitações e propostas de novas pesquisas.• Consolidação da Versão Preliminar: integração de todos os capítulos em um único documento; revisão completa das referências segundo ABNT; entrega da versão preliminar ao orientador.
Novembro	• Revisão e Ajustes Finais: análise do feedback do orientador; realização de ajustes metodológicos, textuais e de formatação.• Formatação e Submissão para a Banca: adequação final às normas do PECS/UEMA; submissão oficial da dissertação para avaliação da banca.
Dezembro	• Preparação para a Defesa: elaboração e ensaio da apresentação de slides.• Defesa da Dissertação: apresentação perante a banca examinadora.• Correções Finais e Depósito: implementação das alterações solicitadas pela banca; depósito da versão final na biblioteca da UEMA para homologação do título.

REFERÊNCIAS

AKINRINOLA, O. et al. Application of machine learning in tax prediction: A review with practical approaches. Global Journal of Engineering and Technology Advances, v. 18, n. 2, p. 102-117, 2024. Disponível em: https://doi.org/10.30574/gjeta.2024.18.2.0028 . Acesso em: 12 jul. 2025.

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.

BAI, J.; NG, S. Boosting diffusion indices. Journal of Applied Econometrics, v. 24, p. 607-629, 2009. Disponível em: https://doi.org/10.1002/jae.1063 . Acesso em: 12 jul. 2025.

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

BRASIL. Ministério da Economia. Proposta de Reforma Tributária – SAC nº 45/2019. Disponível em https://www.gov.br/fazenda/pt-br/acesso-a-informacao/acoes-e-programas/reforma-tributaria/propostas/reforma-tributaria-final.pdf . Acesso em: 30 nov. 2024.

BREIMAN, L. Arcing classifiers. The Annals of Statistics, v. 26, n. 3, p. 801–824, 1998. Disponível em: http://www.jstor.org/stable/120055 . Acesso em: 16 jul. 2025.

BREIMAN, L. Random forests. Machine Learning, v. 45, p. 5–32, out. 2001. Disponível em: https://doi.org/10.1023/A:1010933404324 . Acesso em: 16 jul. 2025.

CARMO, Marcelo Magalhães do; BOLDT, Francisco de Assis; KOMATI, Karin Satie. Previsão de receitas de ICMS do estado do Espírito Santo através de Seleção de Características em Cascata e técnicas de Aprendizado de Máquina. In: [s.l.], 2023. Disponível em: https://sol.sbc.org.br/index.php/eniac/article/view/9277/9179

CHEN, T.; GUESTRIN, C. XGBoost: a scalable tree boosting system. In: PROCEEDINGS OF THE 22nd ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (KDD '16), 2016, New York. Anais [...]. New York: Association for Computing Machinery, 2016. p. 785–794. Disponível em: https://doi.org/10.1145/2939672.2939785 . Acesso em: 17 jul. 2025.

DALTOÉ DE FREITAS , E. .; SMITH SCHNEIDER , P. . IMPACTOS NA ARRECADAÇÃO DE ICMS DO RIO GRANDE DO SUL COM A DIFUSÃO DA GERAÇÃO DISTRIBUÍDA . Revista Brasileira de Energia Solar, [S. l.], v. 11, n. 2, p. 172–181, 2021. DOI: 10.59627/rbens.2020v11i2.322. Disponível em: https://rbens.emnuvens.com.br/rbens/article/view/322 . Acesso em: 16 jul. 2025.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, v. 7, p. 1–30, 2006. Disponível em: https://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf . Acesso em: 17 jul. 2025.

DIEBOLD, F. X.; MARIANO, R. S. Comparing Predictive Accuracy. Journal of Business & Economic Statistics, v. 20, n. 1, p. 134-144, 2002. Disponível em: http://dx.doi.org/10.1198/073500102753410444 . Acesso em: 12 jul. 2025.

DÖPKE, J.; FRITSCHE, U.; PIERDZIOCH, C. Predicting recessions with boosted regression trees. International Journal of Forecasting, v. 33, p. 745-759, 2017. Disponível em: http://dx.doi.org/10.1016/j.ijforecast.2017.02.003 . Acesso em: 12 jul. 2025.

DORNELLES, G. Z.; SCHWARTZER, F. R.; BRAATZ, J. Redes neurais aplicadas na previsão de receita de ICMS no Rio Grande do Sul. (Texto para Discussão TE/RS nº 19). Porto Alegre: Secretaria da Fazenda do Estado do Rio Grande do Sul, 2022. Disponível em: https://tesouro.fazenda.rs.gov.br/upload/1643376127_Artigo%20Modelo%20Redes%20Neurais_Texto%20de%20Discussao.pdf . Acesso em: 3 dez. 2024.

DUCHESNAY, E.; LOFSTEDT, T.; YOUNES, F. Statistics and machine learning in Python. France: Engineering School, 2021. HAL Id: hal-03038776. Disponível em: https://hal.science/hal03038776v3 . Acesso em: 4 dez. 2024.

ELMERT-STREIB, Frank; DEHMER, Matthias. Evaluation of regression models: model assessment, model selection and generalization error. Ain Shams Engineering Journal, 2024. Disponível em: https://www.mdpi.com/2504-4990/1/1/32 . Acesso em: 17 jul. 2025.

Estimando a Arrecadação da Dívida Ativa da União com Machine Learning: Uma análise baseada nos dados de arrecadação do período de 2015 a 2021. Revista da CGU, [S. l.], v. 14, n. 26, 2022. DOI: 10.36428/revistadacgu.v14i26.529. Disponível em: https://revista.cgu.gov.br/Revista_da_CGU/article/view/529 . Acesso em: 16 jul. 2025

FRANKEL, Jeffrey. Over-optimism in forecasts by oficial budget agencies and its implications. Oxford Review of Economic Policy, Oxford, v. 27, n. 4, p. 536–562, 2011. Disponível em: http://oxrep.oxfordjournals.org . Acesso em: 10 out. 2025.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. The Annals of Statistics, v. 29, n. 5, p. 1189–1232, 2001. Disponível em: http://www.jstor.org/stable/2699986 . Acesso em: 17 jul. 2025.

GOVERNO DO ESTADO DO MARANHÃO. 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, 2002. Disponível em: https://sistemas1.sefaz.ma.gov.br/portalsefaz/files?codigo=13942 .

HAIR, J. F. et al. Multivariate data analysis. 7. ed. Upper Saddle River: Pearson Education, 2010. Disponível em: https://www.drnishikantjha.com/papersCollection/Multivariate%20Data%20Analysis.pdf. Acesso em: 17 jul. 2025.

HAN, J.; KAMBER, M.; PEI, J. Data mining: concepts and techniques. 3. ed. Morgan Kaufmann, 2011. Disponível em: http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf . Acesso em: 17 jul. 2025.

HUANG, Z.-C.; SANGIORGI, I.; URQUHART, A. Forecasting Bitcoin volatility using machine learning techniques. Journal of International Financial Markets, Institutions and Money, v. 97, 102064, 2024. Disponível em: https://doi.org/10.1016/j.intfin.2024.102064 . Acesso em: 17 jul. 2025.

HYNDMAN, R. J.; ATHANASOPOULOS, G. Forecasting: principles and practice. 2. ed. OTexts, 2021. Disponível em: https://otexts.com/fpp2/ses.html . Acesso em: 17 jul. 2025..

JAMES, G. et al. An Introduction to Statistical Learning. 2. ed. Cham: Springer, 2023. (Springer Texts in Statistics). Disponível em: https://doi.org/10.1007/978-3- 031-38747-0_8 . Acesso em: 12 jul. 2025.

KVÅLSETH, T. O. Cautionary Note about R^2. The American Statistician, v. 39, n. 4, p. 279-285, 1985. Disponível em: https://doi.org/10.1080/00031305.1985.10479448 . Acesso em: 12 jul. 2025.

MARTINS, E.; GALEGALE, N. V. Uma análise comparativa entre os métodos tradicionais e MARTINS, Emerson; GALEGALE, Napoleão Verardi. Uma análise comparativa entre os métodos tradicionais e algoritmos de aprendizado de máquina para previsão de vendas no segmento varejista. In: XVI Simpósio dos Programas de Mestrado Profissional, 24 e 25 nov. 2021. Disponível em: http://www.pos.cps.sp.gov.br/files/artigo/file/1127/ae89f165eab8981a0543ec3a8863dc46.pdf . Acesso em: 17 jul. 2025.

MIGALINAS, S. Revolução fiscal: IA e Big Data na modernização da fiscalização. Migalhas, 2023. Disponível em: https://www.migalhas.com.br/depeso/417201/revolucao-fiscal-ia-e-big-data-na-modernizacao-da-fiscalizacao . Acesso em: 11 out. 2024.

MINISTÉRIO DA ECONOMIA. Reforma Tributária: contexto, mudanças e impactos. Estudo Especial nº 19, 2024. Disponível em: https://www2.senado.leg.br/bdsf/bitstream/handle/id/647648/EE19_2024.pdf .

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. Introduction to linear regression analysis. Vol. 821. Hoboken: John Wiley & Sons, 2012. Disponível em: https://statanaly.com/wp-content/uploads/2023/05/IntroductiontoLinearRegressionAnalysisbyDouglasC.Mo

ntgomeryElizabethA.PeckG .GeoffreyViningz-lib.org .pdf . Acesso em: 17 jul. 2025.

MURTA, J. V. R. (2024). Modelos de Machine Learning para previsão de arrecadação de ICMS em Minas Gerais . Dissertação de Mestrado, Universidade de Brasília. Disponível em: http://icts.unb.br/jspui/bitstream/10482/50888/1/2024_JoaoVitorRoqueMurta_DISSERT.pdf . Acesso em: 30 nov. 2024.

NITZSCH, F.; SCHIMECZEK, C.; BERTSCH, V. Applying machine learning to electricity price forecasting in simulated energy market scenarios. Energy Reports , v. 12, p. 5268–5279. https://doi.org/10.1016/j.egyr.2024.11.013 . Acesso em 20 de março de 2024.

PEDREGOSA, F. et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011. Disponível em: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf . Acesso em: 17 jul. 2025.

PEDROSA, L. S. et al. 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, Rio de Janeiro, v. 55, p. 43-82, 2023. Disponível em: https://doi.org/10.12957/cadest.2023.83132 . Acesso em: 12 jul. 2025.

QUINLAN, J. R. Induction of decision trees. Machine Learning, v. 1, p. 81–106, 1986. Disponível em: https://doi.org/10.1007/BF00116251 . Acesso em: 17 jul. 2025.

REIMAN, L. Random forests. Machine Learning, v. 45, n. 1, p. 5–32, 2001. Disponível em: https://doi.org/10.1023/A:1010933404324 . Acesso em: 12 jul. 2025.

RODGERS, Robert; JOYCE, Philip. The Effect of Underforecasting on the Accuracy of Revenue Forecasts by State Governments. Public Administration Review, Hoboken, v. 56, n. 1, p. 48-56, jan./fev. 1996. Disponível em: http://www.jstor.org/stable/3110053 . Acesso em: 10 out. 2025.

SANTOS, R. Q. Estimando a Arrecadação da Dívida Ativa da União com Machine Learning: Uma análise baseada nos dados de arrecadação do período de 2015 a 2021. Revista da CGU, v. 14, n. 26, 2022. Disponível em: https://doi.org/10.36428/revistadacgu.v14i26.529 . Acesso em: 12 jul. 2025.

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, apresentado no 10º Encontro de Economia Gaúcha. Disponível em: https://tesouro.fazenda.rs.gov.br/upload/1643376139_Artigo%20Modelo%20Estrutural_Texto%20de%20Discussao.pdf . Acesso em: 03 dez. 2024.

SILVA, Priscila; FIGUEIREDO, Karla. Aprendizado Profundo Aplicado na Previsão de Receita Tributária Utilizando Variáveis Endógenas. In: ENCONTRO NACIONAL DE

INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC), 17. , 2020, Evento Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020 . p. 414-425. ISSN 2763-9061. DOI: https://doi.org/10.5753/eniac.2020.12147.

STEURER, M.; HILL, R. J.; PFEIFER, N. Metrics for evaluating the performance of machine learning based automated valuation models. Journal of Property Research, v. 38, n. 2, p. 99–129, 2021. Disponível em: https://doi.org/10.1080/09599916.2020.1858937 . Acesso em: 17 jul. 2025.

SULAIMAN, M. H.; JADIN, M. S.; MUSTAFFA, Z.; AZLAN, M. N. M.; DANIYAL, H. Short-term forecasting of floating photovoltaic power generation using machine learning models. Cleaner Energy Systems, v. 9, p. 100137, 2024. Disponível em: https://doi.org/10.1016/j.cles.2024.100137 . Acesso em: 3 dez. 2024.

TRUNFIO, T. A. et al. Multiple regression model to analyze the total LOS for patients undergoing laparoscopic appendectomy. BMC Medical Informatics and Decision Making, v. 22, n. 141, 2022. Disponível em: https://doi.org/10.1186/s12911-022-01884-9 . Acesso em: 12 jul. 2025.

ZUANAZZI, Gabriel; SCHWARTZER, Fernando Roberto; BRAATZ, Jacó. Redes Neurais Aplicadas na Previsão de Receita de ICMS no Rio Grande do Sul. Textos para Discussão TE/RS, n. 19, p. 1–25, jan. 2022. Disponível em: https://tesouro.fazenda.rs.gov.br/upload/arquivos/202508/29094521-1643376127-artigo-modelo-redes-neurais-texto-de-discussao.pdf . Acesso em: 12 jul. 2025.

Emitido em 13/01/2026

DOCUMENTOS COMPROBATÓRIOS Nº 53/2026 - DPARQ (11.14.68.07.05)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 13/01/2026 14:24)
CRISTIANE DE JESUS PEREIRA GASPAR
SECRETARIO III
866500

Para verificar a autenticidade deste documento entre em https://sis.sig.uema.br/documentos/ informando seu número: **53**, ano: **2026**, tipo: **DOCUMENTOS COMPROBATÓRIOS**, data de emissão: **13/01/2026** e o código de verificação: **111dd73995**