APLICAÇÃO DE TÉCNICAS DE MACHINE LEARNING NA IDENTIFICAÇÃO DE MUNICÍPIOS DO NORDESTE BRASILEIRO COM MAIOR POTENCIAL PARA ATRAÇÃO DE INVESTIMENTOS A PARTIR DE DADOS SOCIOECONÔMICOS

Application of Machine Learning Techniques to Identify Municipalities in the Brazilian Northeast with the Highest Potential to Attract Investments Based on Socioeconomic Data

Arthur Lins da Gama[1]

Guilherme Fernando Cavalcanti Pereira[2]

Resumo

Este trabalho investiga a aplicação de técnicas de Machine Learning (ML) para identificar o potencial de atração de investimentos nos municípios do Nordeste brasileiro, utilizando uma base de dados socioeconômicos e fiscais públicos (IBGE, SICONFI) referente ao período de 2018 a 2021. O objetivo principal foi superar as limitações dos modelos econométricos tradicionais na captura das complexas dinâmicas regionais, comparando modelos de Regressão Linear Múltipla (OLS) com algoritmos de ensemble baseados em árvores (Random Forest e XGBoost). A metodologia envolveu etapas de engenharia de variáveis para correção de multicolinearidade e assimetria, seleção de atributos para remoção de viés e validação temporal. Os resultados demonstraram a superioridade do modelo XGBoost, que alcançou um R^2 de 0,882 na previsão do PIB per capita, superando significativamente o baseline linear (R^2 0,761) e resolvendo paradoxos estruturais observados no modelo OLS, como a relação entre investimento público e riqueza. A análise de interpretabilidade revelou que a receita corrente per capita, a taxa de emprego formal e a estrutura industrial são os determinantes mais críticos do potencial econômico, em detrimento de indicadores estáticos de gestão fiscal. Adicionalmente, um modelo de classificação (F1-Score 0,84) permitiu mapear a estabilidade do potencial de investimento, confirmando a natureza estrutural das desigualdades regionais. Conclui-se que o uso de ML oferece uma ferramenta robusta e acionável para investidores privados e gestores públicos, permitindo a identificação precisa de polos de desenvolvimento fora das capitais e orientando a alocação eficiente de recursos.

Palavras-chave: Machine Learning; Desenvolvimento Regional; Economia do Nordeste; XGBoost; Atração de Investimentos.

[1] Graduando no curso de Ciência da Computação da faculdade CESAR School, 2025

[2] Orientador do Trabalho de Conclusão de Curso

Abstract

This work investigates the application of Machine Learning (ML) techniques to identify the potential for attracting investments in municipalities of the Brazilian Northeast, using a database of public socioeconomic and fiscal data (IBGE, SICONFI) from 2018 to 2021. The main objective was to overcome the limitations of traditional econometric models in capturing complex regional dynamics by comparing Multiple Linear Regression (OLS) models with tree-based ensemble algorithms (Random Forest and XGBoost). The methodology involved feature engineering steps to correct multicollinearity and skewness, feature selection to remove bias, and temporal validation. The results demonstrated the superiority of the XGBoost model, which achieved an R^2 of 0.882 in predicting GDP per capita, significantly outperforming the linear baseline (R^2 0.761) and resolving structural paradoxes observed in the OLS model, such as the relationship between public investment and wealth. Interpretability analysis revealed that current revenue per capita, formal employment rate, and industrial structure are the most critical determinants of economic potential, rather than static fiscal management indicators. Additionally, a classification model (F1-Score 0.84) allowed mapping the stability of investment potential, confirming the structural nature of regional inequalities. It is concluded that the use of ML offers a robust and actionable tool for private investors and public managers, enabling the precise identification of development poles outside the capitals and guiding the efficient allocation of resources.

Keywords: Machine Learning; Regional Development; Northeast Economy; XGBoost; Investment Attraction.

1 Introdução

Nos últimos anos, a tecnologia tem alterado de maneira significativa a forma como lidamos com dados e tomamos decisões estratégicas, especialmente no campo econômico. Nesse contexto, técnicas de Machine Learning (ML) vêm ganhando destaque pela capacidade de identificar padrões complexos e realizar previsões a partir de grandes volumes de dados, mesmo em cenários caracterizados por relações não lineares e alta dimensionalidade. Diferentemente das abordagens econométricas tradicionais, cujo foco central está frequentemente na inferência causal e na interpretação dos coeficientes estimados, os métodos de ML priorizam o desempenho preditivo e a identificação de estruturas recorrentes nos dados.

Nesse sentido, este trabalho adota uma perspectiva preditiva e exploratória. O objetivo não é estabelecer relações de causalidade econômica entre os indicadores analisados, mas sim identificar padrões socioeconômicos associados ao maior potencial de atração de investimentos nos municípios do Nordeste brasileiro. A

ênfase recai sobre a capacidade dos modelos em aprender, a partir dos dados observados, combinações de características que historicamente se associam a melhores desempenhos econômicos, representados neste estudo pelo PIB per capita.

Embora a análise de importância das variáveis permita interpretações alinhadas à literatura de desenvolvimento regional, tais resultados devem ser compreendidos como associações estatísticas e padrões aprendidos pelos modelos, e não como evidências causais. Dessa forma, o uso de ML neste trabalho busca complementar, e não substituir abordagens explicativas tradicionais, oferecendo uma ferramenta orientada à previsão e ao apoio à tomada de decisão em contextos de elevada complexidade econômica.

1.1 Justificativa

O crescimento econômico regional continua sendo um desafio no Brasil, e o Nordeste ocupa posição central nesse debate. Embora a região tenha avançado em áreas como infraestrutura, agricultura, energias renováveis e turismo, ainda persiste um desequilíbrio entre seu potencial econômico e o volume de investimentos recebidos, tanto públicos quanto privados. Parte desse descompasso decorre da dificuldade de identificar oportunidades a partir de análises quantitativas capazes de representar adequadamente a complexidade das economias locais.

A maior parte dos estudos regionais ainda se apoia predominantemente em modelos econométricos tradicionais, que, embora adequados para análises explicativas e inferências causais, apresentam limitações na captura de relações não lineares e interações complexas entre variáveis socioeconômicas. Esse aspecto é particularmente relevante no contexto municipal, marcado por heterogeneidade estrutural, diferenças de escala e dinâmicas produtivas diversas.

Ao empregar técnicas de Machine Learning, especialmente modelos de ensemble baseados em árvores como o Random Forest e o XGBoost, o trabalho busca explorar uma abordagem metodológica alternativa, mais adequada à identificação de padrões em dados socioeconômicos complexos. Esses algoritmos permitem lidar

com alta dimensionalidade e não linearidades, oferecendo maior flexibilidade na análise do potencial econômico dos municípios.

Do ponto de vista técnico, a pesquisa contribui ao investigar essa lacuna metodológica em estudos regionais aplicados ao Nordeste brasileiro. Do ponto de vista prático, o modelo desenvolvido pode funcionar como ferramenta de apoio à decisão para investidores e gestores públicos, auxiliando a identificação de municípios com maior potencial de atração de investimentos com base em evidências empíricas.

1.2 Problema

Quais são os principais determinantes socioeconômicos que explicam o potencial de investimento dos municípios do Nordeste? E como modelos de ML podem utilizar esses fatores para prever e classificar os municípios mais promissores?

1.3 Objetivos

1.3.1 Objetivo geral

Avaliar e comparar modelos de ML capazes de identificar municípios do Nordeste brasileiro com maior potencial de atração de investimentos, utilizando dados socioeconômicos, setoriais e geográficos.

1.3.2 Objetivos específicos

- OE1: Levantar, tratar e organizar uma base de dados socioeconômicos, fiscais e demográficos dos municípios do Nordeste.
- OE2: Aplicar e comparar algoritmos de ML para duas abordagens:
  - **Regressão:** Para prever o valor do potencial de investimento (*proxy*: PIB per capita).
  - **Classificação:** Para classificar os municípios em faixas de potencial (ex: Baixo, Médio, Alto).
- OE3: Validar o desempenho dos modelos por meio de métricas estatísticas e validação temporal, identificar as variáveis mais relevantes e

gerar um ranking dos municípios com maior potencial previsto, como apoio à tomada de decisão.

1.4 Estrutura do Artigo

O presente artigo está organizado da seguinte forma: na Seção 2 é apresentada a fundamentação teórica, reunindo conceitos de ML aplicados ao potencial de investimento e aspectos socioeconômicos do Nordeste brasileiro; na Seção 3 descreve-se a metodologia empregada, incluindo a natureza da pesquisa, procedimentos de coleta e tratamento de dados, bem como os algoritmos utilizados; na Seção 4 são expostos o desenvolvimento da pesquisa e a análise dos resultados obtidos; finalmente, na Seção 5 são apresentadas as conclusões, limitações do estudo e sugestões para trabalhos futuros.

2 Fundamentação Teórica

2.1 Desenvolvimento Regional e Fatores Socioeconômicos

O desenvolvimento econômico regional envolve processos complexos voltados à redução de disparidades e à melhoria das condições de vida fora dos grandes centros. Estudos clássicos, como os de Hissa-Teixeira (2018), mostram que o crescimento no Nordeste não ocorre de forma homogênea, sendo influenciado por fatores espaciais, infraestrutura e concentração de serviços.

Compreender o que impulsiona o potencial econômico municipal é essencial para orientar investimentos públicos e privados. Pesquisas recentes, como Antunes (2024), destacam variáveis como estrutura produtiva, formalização do trabalho, capacidade fiscal e dinâmica demográfica. Esses fatores ajudam a explicar diferenças de desempenho entre municípios, frequentemente captadas por indicadores como o PIB per capita, e fundamentam a seleção das variáveis utilizadas nos modelos preditivos deste estudo.

2.2 Aplicação de ML na Análise Socioeconômica

O Machine Learning é um campo da Inteligência Artificial voltado ao desenvolvimento de algoritmos capazes de aprender a partir de dados e identificar padrões complexos. Em estudos socioeconômicos e de políticas públicas, essas

técnicas têm sido empregadas para prever indicadores de desenvolvimento e classificar municípios, apresentando vantagens frente a modelos estatísticos tradicionais (Mitchell, 1997; Páscoa, 2020).

Entre os algoritmos mais utilizados destacam-se os modelos de ensemble baseados em árvores, como o Random Forest (Breiman, 2001), e métodos de Gradient Boosting, como o XGBoost (Chen; Guestrin, 2016). Esses modelos apresentam bom desempenho em dados tabulares heterogêneos, pois capturam relações não lineares e interações complexas entre variáveis, justificando sua adoção neste trabalho.

2.3 Aspectos Econômicos do Nordeste Brasileiro

O Nordeste brasileiro possui uma economia diversa e marcada por contrastes regionais. Apesar do potencial em setores como turismo, agricultura e energias renováveis, a região ainda recebe menos investimentos do que áreas mais desenvolvidas do país (Ribeiro; Domingues; Perobelli, 2021).

Parte desse desequilíbrio está associada à carência de infraestrutura e à concentração histórica das atividades produtivas, fatores apontados como centrais para a manutenção das desigualdades regionais (Araújo; Souza; Lima, 2019). Nesse contexto, a escassez de instrumentos analíticos integrados limita a precisão das decisões estratégicas (Paixão; Nogueira, 2018), reforçando a necessidade de abordagens quantitativas como a proposta neste estudo.

2.4 Avaliação de Modelos Preditivos

A avaliação de modelos de Machine Learning depende das métricas escolhidas, pois elas refletem a capacidade de generalização dos modelos para novos dados (Páscoa, 2020). Em problemas de regressão, são utilizadas métricas como RMSE, MAE e R^2, enquanto tarefas de classificação empregam Precisão, Recall e F1-score (Mitchell, 1997).

A escolha dessas métricas está alinhada ao objetivo preditivo da pesquisa e permite a comparação objetiva entre diferentes modelos, auxiliando na seleção daquele com melhor desempenho.

2.5 Limitações do PIB per capita como Proxy de Desenvolvimento

O PIB per capita é amplamente utilizado como indicador sintético de desempenho econômico, mas apresenta limitações como proxy de desenvolvimento. Ele não capta aspectos distributivos, informalidade ou qualidade de vida, podendo superestimar o desempenho de municípios com atividades concentradas.

Apesar disso, o indicador permanece associado a variáveis estruturais relevantes, como formalização do trabalho, arrecadação e estrutura produtiva. Neste estudo, o PIB per capita é utilizado como uma proxy operacional do potencial econômico, adequada ao objetivo preditivo e interpretada em conjunto com outros indicadores socioeconômicos.

3 Metodologia

Nesta seção, é descrita a metodologia empregada para a execução da pesquisa, abordando sua classificação e os procedimentos técnicos de coleta, tratamento e modelagem dos dados.

3.1 Classificação da Pesquisa

A pesquisa possui natureza aplicada, pois busca desenvolver uma solução prática para identificar municípios do Nordeste com maior potencial de atração de investimentos. A abordagem é quantitativa, fundamentada na coleta e análise de indicadores socioeconômicos, utilizando técnicas estatísticas e algoritmos de Machine Learning.

Quanto aos objetivos, o estudo apresenta caráter:

- **exploratório**, ao investigar padrões pouco analisados entre variáveis econômicas e sociais;
- **descritivo**, ao caracterizar os municípios segundo emprego formal, estrutura produtiva, arrecadação e PIB per capita;

- **explicativo**, por empregar modelos preditivos para identificar os fatores que mais influenciam o desempenho econômico municipal.

3.2 Procedimentos Técnicos

Os procedimentos técnicos envolveram três etapas principais: coleta e preparação dos dados, análise exploratória e modelagem preditiva.

Coleta e Tratamento de Dados (ETL): Inicialmente, foi realizado o levantamento documental e a extração de dados de bases públicas, abrangendo o período de 2018 a 2021.

As fontes utilizadas foram:

- **Instituto Brasileiro de Geografia e Estatística (IBGE):** Para dados de PIB municipal, Estimativas de População e Área Territorial.
- **Tesouro Nacional (SICONFI e Tesouro Transparente):** Para dados fiscais de Receitas e Despesas Municipais (FINBRA) e a nota de Capacidade de Pagamento (CAPAG).
- **Base dos Dados (via BigQuery):** Para dados de vínculos de emprego formal (RAIS/CAGED).

No tratamento dos dados, realizaram-se a padronização das siglas de UF, a agregação de informações fiscais por município e ano e o tratamento de valores ausentes, utilizando imputação pela mediana e o Random Forest Classifier para estimar valores faltantes da variável CAPAG. A escolha desse modelo se justifica por sua capacidade de lidar com variáveis categóricas e não lineares, além de capturar interações complexas entre indicadores fiscais e econômicos, sem impor pressupostos paramétricos. Essa abordagem permite uma imputação mais consistente com a estrutura observada nos dados, em comparação a métodos univariados ou estritamente lineares.

Análise Exploratória e Modelagem: A análise exploratória (EDA) foi conduzida para identificar correlações e multicolinearidade, através de matrizes de correlação (heatmaps), e para entender a distribuição e os outliers das variáveis, por meio de boxplots. Na etapa de modelagem, o trabalho foi dividido em duas abordagens:

- **Regressão:** Para prever o valor do PIB per capita, foram treinados e comparados os modelos de Regressão Linear Múltipla, Random Forest Regressor e XGBoost Regressor.
- **Classificação:** Para classificar os municípios em faixas de potencial, foram aplicados modelos como a Regressão Logística, o Random Forest Classifier e o XGBoost Classifier.

Os modelos de Machine Learning foram inicialmente treinados com hiperparâmetros padrão como baseline. Em seguida, realizou-se ajuste de hiperparâmetros (tuning) por meio de validação cruzada e busca em grade, com o objetivo de melhorar o desempenho preditivo e reduzir risco de overfitting. Os parâmetros finais foram selecionados com base nas métricas definidas para cada tarefa, preservando a comparabilidade entre os modelos avaliados.

Avaliação e Interpretação: Para a validação dos modelos, foram utilizadas métricas estatísticas distintas para cada abordagem, conforme definido na Seção 2.4. Por fim, foi realizada a análise da importância das variáveis e a geração de um ranking de municípios com o maior potencial previsto pelo modelo de melhor performance.

3.3 Variáveis e Indicadores Utilizados

O conjunto de variáveis foi organizado em dimensões econômicas, fiscais, demográficas e sociais (Quadro 1). O PIB per capita foi adotado como variável dependente nos modelos de regressão, enquanto nos modelos de classificação definiu-se uma variável categórica baseada no quartil superior.

Quadro 1 - Dicionário de variáveis

Nome da coluna	Descrição	Papel no modelo
id_municipio	Código IBGE do município	Identificador único
municipio	Nome do município	Identificação

ano	Ano de referência (2018–2021)	Eixo temporal
uf	Unidade da Federação (sigla)	Controle regional
populacao	População residente estimada	Controle demográfico
area	Área territorial do município (km²)	Controle demográfico
classificacao_capag	Nota da Capacidade de Pagamento (1–4)	Indicador fiscal
pib	Produto Interno Bruto total (R$ milhões)	Indicador econômico
impostos_liquidos	Impostos líquidos de subsídios	Indicador econômico
va	Valor adicionado total	Indicador econômico agregado
va_agropecuaria	Valor adicionado da agropecuária	Estrutura produtiva
va_industria	Valor adicionado da indústria	Estrutura produtiva
va_servicos	Valor adicionado dos serviços	Estrutura produtiva

va_adespss	Valor adicionado de assistência e previdência social	Indicador social
total_vinculos	Total de vínculos formais de trabalho	Indicador de mercado de trabalho
despesa_corrente	Despesa corrente total do município	Indicador fiscal
despesa_capital	Despesa de capital total do município	Indicador fiscal / investimento
despesa_geral	Despesa total (corrente + capital)	Indicador fiscal agregado
receita_corrente	Receita corrente total do município	Indicador fiscal / arrecadação
is_capital	Indicador binário de capital (1 = capital)	Controle categórico
densidade_demo	Densidade demográfica (população/área)	Variável demográfica
taxa_emprego_formal	Relação entre vínculos formais e população total	Indicador econômico

crescimento_pib_abs	Crescimento absoluto do PIB em relação ao ano anterior	Indicador de dinâmica econômica
crescimento_pib_perc	Crescimento percentual do PIB em relação ao ano anterior	Indicador de dinâmica econômica
pib_per_capita	PIB per capita (R$ mil/hab)	Variável dependente (alvo)
crescimento_pib_pc_abs	Crescimento absoluto do PIB per capita	Indicador de variação de renda
crescimento_pib_pc_perc	Crescimento percentual do PIB per capita	Indicador de variação de renda

Fonte: Elaborado pelo autor (2025).

A etapa de preparação dos dados foi crucial para o sucesso da modelagem, envolvendo quatro ações principais:

Engenharia de Variáveis (Normalização por Razão): Conforme identificado na Análise Exploratória (Seção 4.1), as variáveis brutas (fiscais, econômicas e demográficas) sofriam de alta multicolinearidade e eram distorcidas pelo porte de cada município. Para mitigar esse problema e focar na eficiência e estrutura (em vez de volume), as variáveis foram normalizadas:

- **Estrutura Produtiva:** Variáveis absolutas como va_agropecuaria, va_industria e va_servicos foram convertidas em indicadores percentuais (perc_agro, perc_industria, perc_servicos), calculados como a razão entre o VA setorial e o VA total.
- **Variáveis Fiscais:** A receita_corrente foi convertida em um indicador *per capita* (receita_corrente_pc). Para as despesas, foi criada a capex_share, a

razão entre a despesa_capital (investimento) e a despesa_geral, medindo o foco real em investimento.

- **Outras Razões:** A populacao foi normalizada pela área (criando densidade_demo) e os impostos_liquidos pelo pib (perc_impostos).

O acréscimo de uma unidade ao denominador em alguns cálculos foi uma técnica de suavização para garantir a estabilidade numérica e evitar divisões por zero.

Tratamento de Assimetria (Transformação Logarítmica): Variáveis com forte assimetria (como PIB per capita e receita per capita) passaram pela transformação logarítmica (log1p) para estabilizar variância e melhorar a performance dos modelos.

Limpeza de Dados: Registros com valores inválidos após transformações foram convertidos em *NaN* e posteriormente removidos, garantindo consistência no conjunto final.

Seleção Final de Variáveis (Remoção de Viés): A variável perc_adespss foi excluída após diagnóstico realizado durante a etapa de modelagem com Random Forest. Embora apresentasse alta importância estatística, verificou-se que seu comportamento funcionava como um proxy inverso de riqueza municipal, uma vez que municípios com maior participação relativa de assistência e previdência social tendem a apresentar menor dinamismo econômico estrutural. A permanência dessa variável fazia com que o modelo aprendesse predominantemente um efeito compensatório de gasto social, desviando o foco da análise do potencial econômico produtivo para características associadas à dependência de transferências. Dessa forma, sua exclusão buscou mitigar um viés econômico, e não apenas estatístico, permitindo que os modelos capturassem de forma mais adequada fatores estruturais ligados à geração de renda, emprego formal e capacidade produtiva.

3.4 Validação Temporal e Estratégia de Treino–Teste

A validação dos modelos seguiu um critério temporal explícito, respeitando a ordem cronológica dos dados. O conjunto de treinamento foi composto por observações referentes aos anos de 2018 a 2020, enquanto o ano de 2021 foi reservado exclusivamente para teste. Essa estratégia busca simular um cenário real de

previsão, no qual informações históricas são utilizadas para estimar o desempenho econômico futuro dos municípios.

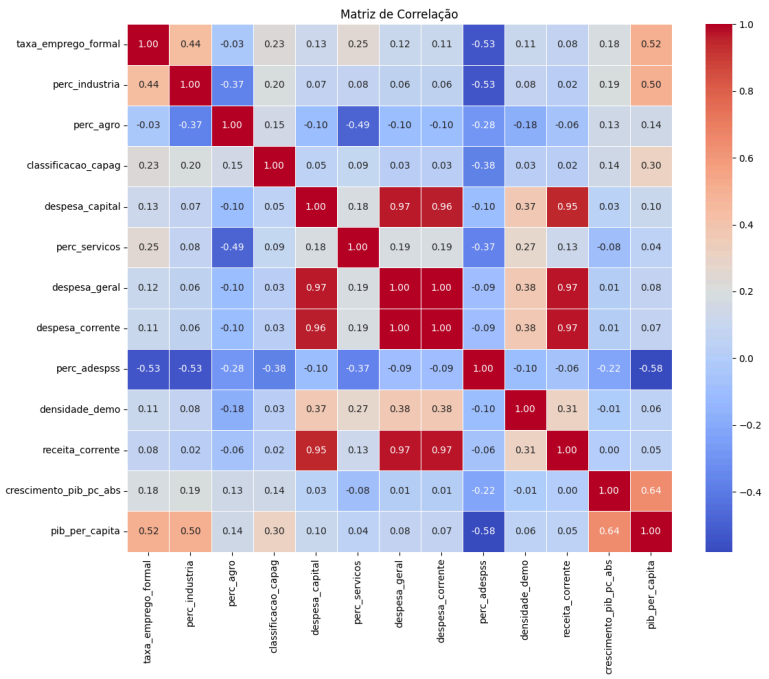
Para mitigar o risco de data leakage, o split temporal foi realizado antes do treinamento dos modelos. Todas as etapas de transformação, normalização e engenharia de atributos foram conduzidas utilizando apenas informações disponíveis até o período de treino, garantindo que o desempenho observado no conjunto de teste reflita a capacidade de generalização temporal dos modelos.

4 Desenvolvimento

4.1 Análise Exploratória dos Dados (EDA)

A etapa inicial de EDA permitiu compreender a estrutura dos dados e as relações entre os principais indicadores socioeconômicos e fiscais. O primeiro enfoque foi a correlação entre as variáveis considerando o conjunto completo de municípios do Brasil.

Figura 1 - Heatmap de correlação das variáveis - Brasil



Fonte: Elaborado pelo autor (2025).

A matriz de correlação (Figura 1) indicou elevado grau de multicolinearidade entre variáveis fiscais, como receita_corrente e despesa_corrente. Esse comportamento reflete o próprio funcionamento orçamentário municipal, no qual maior capacidade de arrecadação tende a estar associada a maiores níveis de gasto. Do ponto de vista prático, esse padrão evidencia que variáveis absolutas capturam principalmente o porte do município, e não sua eficiência econômica, reforçando a necessidade de normalização por população ou por razão, conforme adotado na metodologia.

A análise também revelou correlação positiva entre PIB per capita e taxa_emprego_formal, sugerindo que municípios com maior nível de formalização do mercado de trabalho apresentam maior capacidade de geração de renda. Esse resultado está alinhado à literatura de desenvolvimento regional e indica que a formalização funciona como um indicador estrutural relevante do dinamismo econômico local. Por outro lado, a fraca correlação linear entre PIB per capita e densidade demográfica sugere que o impacto dessa variável ocorre de forma indireta ou não linear, dependendo da interação com outros fatores produtivos e institucionais.

A avaliação das distribuições das variáveis evidenciou forte assimetria na variável alvo (PIB per capita) e a presença de outliers, especialmente em municípios com atividades altamente concentradas, como mineração, energia ou polos industriais específicos. Do ponto de vista econômico, esses outliers não representam erros, mas características estruturais relevantes, que precisam ser tratadas de forma adequada para não distorcer o treinamento dos modelos.

Diante desse cenário, foi aplicada a transformação logarítmica (log1p) no PIB per capita e em variáveis associadas, com o objetivo de reduzir a heterocedasticidade e estabilizar a variância. Essa transformação permitiu preservar a informação econômica contida nos extremos da distribuição, ao mesmo tempo em que tornou o conjunto de dados mais adequado para a modelagem preditiva, especialmente nos algoritmos de Machine Learning utilizados nas etapas seguintes.

4.2 Modelagem Estatística – Regressão Linear Múltipla (OLS)

Como baseline para a modelagem preditiva, foi ajustado um modelo de Regressão Linear Múltipla (OLS). Inicialmente, realizou-se a análise diagnóstica de multicolinearidade por meio do fator de inflação da variância (VIF). Os resultados indicaram que, após a etapa de engenharia de variáveis descrita na Seção 3, todos os preditores apresentaram valores de VIF abaixo dos limites recomendados, garantindo a estabilidade dos coeficientes estimados.

Para avaliar possíveis diferenças regionais, comparou-se o comportamento do modelo entre os conjuntos Brasil e Nordeste. Observou-se uma redução no R^2 ajustado e uma inversão de sinal da variável capex_share, positiva no modelo nacional e negativa no modelo estimado apenas para o Nordeste. Esse resultado indica que a dinâmica econômica regional não é plenamente capturada por uma especificação linear única.

Tabela 1 - Modelo OLS final - Nordeste

Variável	**Valores**
R^2	0.761
taxa_emprego_formal	2.315028
perc_industria	1.969565
perc_agro	1.494511
log_receita_corrente_pc	0.176562
log_densidade	0.029120
capex_share	-0.196941

Fonte: Elaborado pelo autor (2025).

No modelo OLS estimado para o Nordeste, apresentado na Tabela 1, obteve-se R^2 ajustado de 0,761. Entretanto, a variável capex_share apresentou coeficiente negativo (−0,197), caracterizando um resultado contraintuitivo do ponto de vista econômico. Em termos teóricos, espera-se que uma maior participação de despesas de capital, associadas a investimento público, contribua positivamente para o desempenho econômico. O sinal negativo sugere que, no contexto nordestino, municípios com maior proporção relativa de investimento tendem a ser aqueles com

menor base econômica, utilizando o investimento como mecanismo compensatório, e não como reflexo de dinamismo estrutural.

Esse comportamento evidencia uma limitação importante do modelo linear. O OLS assume relações médias e lineares entre as variáveis, não sendo capaz de capturar efeitos condicionais ou interações complexas, como situações em que o impacto do investimento depende do nível de arrecadação, da estrutura produtiva ou do grau de formalização do mercado de trabalho. Assim, o coeficiente negativo do capex_share não deve ser interpretado como uma relação causal direta, mas como um indício de que o efeito econômico do investimento público varia de acordo com o contexto estrutural do município.

Dessa forma, embora o OLS seja útil como referência inicial e ferramenta interpretativa, seus resultados reforçam a necessidade de métodos mais flexíveis. Modelos não lineares, como os baseados em árvores, permitem capturar interações e regimes distintos de comportamento econômico, justificando a adoção de técnicas de Machine Learning nas etapas seguintes da análise.

4.3 Modelo de Machine Learning: Random Forest

Após o OLS ter estabelecido o baseline de desempenho (R^2 ajustado = 0.761) e indicado a presença de relações não lineares, a modelagem avançou para algoritmos baseados em árvores. O primeiro método testado foi o Random Forest (Breiman, 2001).

Conforme detalhado na Seção 3.3, foi nessa fase quando o viés que a variável perc_adespss introduzia no modelo foi diagnosticado. Uma iteração inicial, utilizando o conjunto de features que incluía essa variável, obteve métricas de teste aparentemente excelentes (R^2 = 0.908).

Já o modelo Random Forest final treinado no conjunto de dados ajustado, alcançou os seguintes resultados no conjunto de teste:

- R^2: 0.862
- RMSE: 0.171
- MAPE: 4,99%

Esse desempenho supera o baseline OLS e sugere maior capacidade de capturar relações não lineares do dataset. Além disso, a distribuição das importâncias das variáveis mostrou-se mais equilibrada e alinhada com interpretações econômicas plausíveis.

4.4 Modelo de Machine Learning: XGBoost

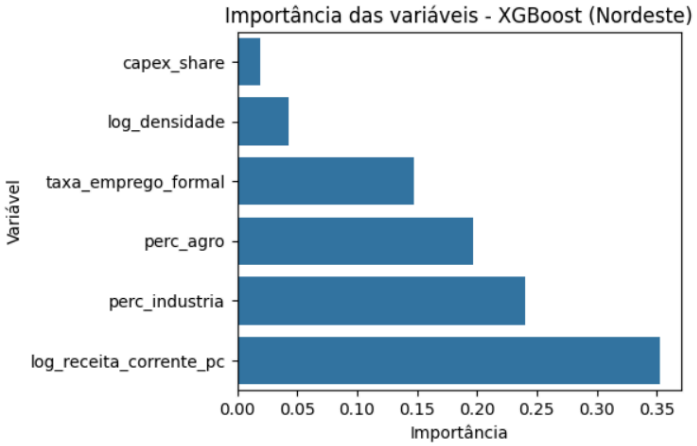
O algoritmo XGBoost (Extreme Gradient Boosting) (Chen;Guestrin, 2016) foi testado em seguida. Trata-se de um método de ensemble que constrói árvores de forma sequencial, ajustando novos modelos para corrigir os erros das iterações anteriores.

O XGBoost foi treinado com o mesmo conjunto final de dados utilizado no Random Forest, buscando melhorar o desempenho preditivo. Os resultados no conjunto de teste foram:

- R^2: 0.882
- RMSE: 0.158
- MAPE: 4,6%

O modelo apresentou desempenho superior ao Random Forest e forneceu uma estrutura de importância das variáveis mais consistente, conforme ilustrado na Figura 2.

Figura 2 - Importância das variáveis - XGBoost (Nordeste)



Fonte: Elaborado pelo autor (2025).

A análise da Figura 2 permite identificar os principais determinantes do PIB per capita segundo o modelo. A log_receita_corrente_pc e a taxa_emprego_formal permanecem entre as variáveis de maior influência, reforçando o papel desses indicadores na explicação do desenvolvimento municipal.

Um ponto a destacar é o comportamento do capex_share. Enquanto no OLS essa variável apresentou sinal negativo, no XGBoost ela aparece como preditor positivo e com relevância intermediária. Esse resultado sugere que a relação entre investimento e PIB per capita envolve padrões não lineares que o modelo linear não capturou.

4.5 Comparação e Seleção do Modelo Preditivo Final

A etapa de regressão avaliou três modelos, OLS, Random Forest e XGBoost. A Tabela 2 apresenta os resultados obtidos no conjunto de teste.

Tabela 2 - Comparativo de desempenho dos modelos de Regressão

Model	R^2	RMSE	MAE	MAPE (%)
Linear Regression (Nordeste)	0.761137	0.225915	0.157966	6.103624
Random Forest (Nordeste)	0.862062	0.171063	0.127406	4.987552
XGBoost (Nordeste)	0.882309	0.158010	0.117689	4.610549

Fonte: Elaborado pelo autor (2025).

Os resultados mostram que os modelos de ML superaram o OLS nas métricas avaliadas. O modelo linear serviu como referência inicial, mas apresentou menor capacidade de explicar a variação do PIB per capita no Nordeste.

Entre os métodos de ensemble, o XGBoost obteve o melhor desempenho, com maior R^2 e menores erros. Além disso, representou de forma mais consistente os efeitos das variáveis explicativas, o que contribuiu para sua escolha como modelo final para as etapas de análise espacial e classificação.

Apesar do desempenho superior apresentado pelos modelos de Machine Learning, sua adoção envolve um trade-off entre performance preditiva e interpretabilidade. Modelos lineares, como a Regressão Linear Múltipla (OLS), oferecem maior transparência na interpretação dos coeficientes, mas apresentam limitações na

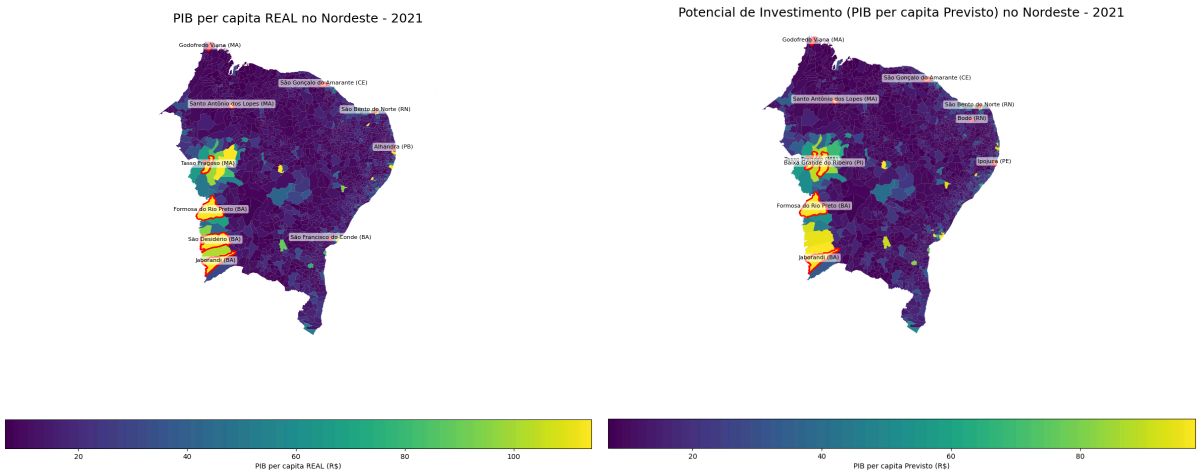
captura de relações não lineares e interações complexas entre variáveis socioeconômicas. Por outro lado, modelos baseados em árvores, como o Random Forest e o XGBoost, ampliam significativamente a capacidade preditiva ao explorar estruturas mais flexíveis, ainda que com menor interpretabilidade econômica direta.

Neste trabalho, optou-se por utilizar o OLS como baseline interpretativo e os modelos de Machine Learning como ferramentas complementares orientadas à previsão. O ajuste de hiperparâmetros foi conduzido de forma controlada, sem a aplicação de estratégias exaustivas de tuning, buscando equilibrar desempenho, estabilidade e transparência metodológica, além de preservar a comparabilidade entre os modelos avaliados. Essa escolha reforça o caráter preditivo e exploratório da pesquisa, em consonância com os objetivos propostos.

4.6 Análise Espacial e Ranking

A etapa seguinte consistiu em aplicá-lo para estimar o potencial de investimento no ano mais recente disponível (2021). Para avaliar sua capacidade de representar padrões territoriais, foram produzidos mapas comparando o PIB per capita observado e o previsto pelo modelo.

Figura 3 – Distribuição Espacial do PIB per capita no Nordeste (2021): (A) Real vs. (B) Previsto



Fonte: Elaborado pelo autor (2025).

A comparação apresentada na Figura 3 sugere uma boa correspondência espacial entre os valores observados e as previsões. O mapa previsto reproduz os principais padrões identificados no mapa real, como:

1. A concentração litorânea: Especialmente nas regiões metropolitanas de Salvador, Recife e Fortaleza.
2. Os polos de interior: Como o eixo Petrolina-Juazeiro e o oeste baiano.
3. Áreas de menor dinamismo econômico, sobretudo no semiárido.

Essa validação visual complementa o desempenho estatístico do modelo (R^2 = 0.882) e indica que o XGBoost capturou, de forma consistente, estruturas territoriais relevantes da economia nordestina. Com isso, torna-se possível utilizá-lo na identificação de áreas com maior potencial relativo de investimento, incluindo regiões fora dos principais centros urbanos.

4.6.1 Ranking de Potencial de Investimento (Top 5)

Com o modelo validado também do ponto de vista espacial, foi elaborado um ranking dos municípios com maior potencial previsto. A Tabela 3 apresenta os cinco municípios que se destacaram em 2021 segundo os fundamentos econômicos capturados pelo modelo.

Tabela 3 – Top 5 Municípios do Nordeste por Potencial Previsto (2021)

Município	UF	PIB per capita (Real)	PIB per capita (Previsto)
Godofredo Viana	MA	219.637670	207.097458
Santo Antônio dos Lopes	MA	210.483640	202.387024
São Bento do Norte	RN	204.331473	193.071777
Formosa do Rio Preto	BA	176.485026	165.812805
São Gonçalo do Amarante	CE	175.099621	160.109558

Fonte: Elaborado pelo autor (2025).

A leitura do ranking reforça a interpretabilidade observada no XGBoost (Seção 4.4). Os municípios melhor posicionados apresentam características econômicas consistentes com o que o modelo aprendeu durante o treinamento. Entre os padrões identificados, destacam-se:

- Mineração e Energia (Gás): Godofredo Viana (MA) e Santo Antônio dos Lopes (MA) são ranqueados no topo devido à sua altíssima participação industrial (ex: 0.933 em S. A. dos Lopes).
- Energia Eólica/Polo de Petróleo: São Bento do Norte (RN) é identificado pela combinação de alta indústria e alta taxa_emprego_formal.
- Polos Industriais/Petroquímicos: Municípios que não aparecem no Top 5, mas figuram entre os dez primeiros, como São Francisco do Conde (BA) e Ipojuca (PE), destacam-se pela elevada participação industrial e pelos níveis mais altos de emprego formal.

É importante ressaltar que alguns municípios do ranking apresentam PIB per capita elevado em função de atividades econômicas altamente específicas e concentradas, como mineração, petróleo ou energia, o que limita a generalização dos resultados. Assim, o ranking deve ser interpretado como uma medida de potencial estrutural econômico, e não como indicação direta de oportunidades de investimento diversificado.

Esses resultados mostram que o modelo foi capaz de identificar padrões coerentes com a literatura sobre desenvolvimento regional e com os indicadores apresentados na Figura 2, confirmando o papel da estrutura produtiva e da formalização do trabalho como principais determinantes do desempenho municipal.

4.7 Classificação Binária

Após a seleção do XGBoost como melhor modelo de regressão para prever o log_pib_pc, a etapa seguinte consistiu em desenvolver um modelo de classificação. A proposta aqui é prática: em vez de estimar um valor contínuo, busca-se categorizar os municípios em dois grupos, “Alta Oportunidade”, definidos como aqueles acima do terceiro quartil (P75) do PIB per capita, e “Baixa Oportunidade”, que reúne os demais.

Três algoritmos foram treinados e comparados utilizando o mesmo conjunto final de variáveis:

1. Regressão Logística: Como baseline linear para a classificação.

2. Random Forest Classifier: O ensemble de árvores por bagging.
3. XGBoost Classifier: O ensemble de árvores por boosting.

A avaliação considerou Acurácia, Precisão, Recall e F1-score. Os resultados no conjunto de testes estão apresentados na Tabela 4.

Tabela 4 - Comparativo de desempenho dos modelos de classificação

Modelo	Acurácia	Precisão	Recall	F1-score
XGBoost	0.918467	0.876056	0.809896	0.841678
Random Forest	0.910801	0.857542	0.799479	0.827493
Logistic Regression	0.892683	0.873377	0.700521	0.777457

Fonte: Elaborado pelo autor (2025).

4.8 Análise de Estabilidade Temporal

Além da avaliação pontual apresentada na Tabela 4, buscou-se verificar se o classificador XGBoost mantém consistência ao longo do tempo. A expectativa é que um modelo desse tipo seja capaz de capturar fundamentos econômicos relativamente estáveis, e não apenas variações anuais.

Para essa análise, considerou-se o período de 2018 a 2021. Os municípios foram agrupados em três categorias, tanto com base nos valores observados do PIB per capita quanto nas classificações produzidas pelo modelo:

- Nunca alta oportunidade: Municípios que nunca estiveram no quartil superior (P75) no período.
- Sempre alta oportunidade: Municípios que estiveram no quartil superior (P75) em todos os anos.
- Instável: Municípios que flutuaram, entrando ou saindo da categoria "Alta Oportunidade".

Tabela 5 - Comparação da Distribuição de Estabilidade (Real vs. Previsto, 2018-2021)

Categoria de Estabilidade	Distribuição Real (%)	Distribuição Prevista pelo Modelo (%)
Nunca alta oportunidade	69,6%	68,4%
Sempre alta oportunidade	20,0%	19,3%
Instável	10,4%	12,3%

Fonte: Elaborado pelo autor (2025).

Os resultados mostram que o modelo reproduziu de maneira próxima a distribuição observada nas três categorias. A proporção prevista de municípios classificados como “Sempre alta oportunidade” (19,3%) praticamente coincide com o valor real (20,0%). O mesmo ocorre no grupo “Nunca alta oportunidade”, com 68,4% previstos frente a 69,6% reais. A categoria “Instável” aparece ligeiramente ampliada nas previsões, o que pode refletir a sensibilidade do modelo a variações marginais entre anos.

Figura 4 – Mapa de Estabilidade de Oportunidade Prevista pelo Modelo (2018-2021)

Fonte: Elaborado pelo autor (2025).

O mapa apresentado na Figura 4 sintetiza espacialmente essas classificações. Ele permite identificar áreas que mantiveram desempenho consistentemente elevado ao

longo do período, bem como regiões classificadas como “Instáveis”, que podem sinalizar trajetórias econômicas em transformação. A observação dos municípios que alteraram de categoria contribui para a interpretação de tendências e possíveis oportunidades emergentes.

4.9 Discussão dos Resultados

Os resultados obtidos ao longo das etapas de modelagem indicam que o uso de técnicas de Machine Learning foi eficaz para identificar padrões associados ao potencial econômico dos municípios do Nordeste. A comparação entre os modelos evidenciou que abordagens lineares apresentam limitações na representação da heterogeneidade regional, enquanto modelos baseados em árvores mostraram maior capacidade de capturar relações não lineares e interações relevantes entre variáveis socioeconômicas.

A análise da importância das variáveis revelou que fatores ligados à estrutura produtiva, à arrecadação e à formalização do trabalho exercem papel central na explicação do desempenho econômico municipal. Esses achados sugerem que o potencial de investimento está mais associado à vitalidade econômica instalada do que a indicadores fiscais isolados, reforçando interpretações presentes na literatura de desenvolvimento regional.

Do ponto de vista territorial, a consistência entre os padrões espaciais observados e previstos pelo modelo indica que parte das desigualdades regionais possui caráter estrutural e persistente. Municípios que se destacam de forma recorrente tendem a apresentar bases produtivas consolidadas, enquanto regiões de menor dinamismo mantêm desempenho reduzido ao longo do tempo, o que limita o impacto de variações conjunturais de curto prazo.

A classificação temporal reforça essa interpretação ao mostrar que a maior parte dos municípios permanece estável em suas categorias de potencial, com uma fração menor apresentando comportamento instável. Esses casos podem sinalizar territórios em transição, nos quais mudanças na estrutura produtiva ou no mercado de trabalho podem alterar trajetórias econômicas futuras.

Do ponto de vista aplicado, os resultados oferecem subsídios para decisões públicas e privadas ao permitir a identificação de municípios com fundamentos econômicos mais sólidos. No entanto, as estimativas devem ser interpretadas como indicadores de potencial estrutural, e não como previsões determinísticas ou recomendações automáticas de investimento, devendo ser complementadas por análises setoriais específicas.

5 Conclusão

Este Trabalho de Conclusão de Curso avaliou a aplicação de técnicas de Machine Learning a um conjunto abrangente de dados socioeconômicos e fiscais para identificar municípios do Nordeste brasileiro com maior potencial para atração de investimentos. Os resultados obtidos ao longo das etapas de análise exploratória, modelagem e validação temporal indicam que essa abordagem é viável e fornece estimativas consistentes, especialmente quando comparada a modelos lineares tradicionais.

A comparação entre os métodos evidenciou que abordagens econométricas oferecem limitações diante da heterogeneidade regional, enquanto modelos baseados em árvores, em particular o XGBoost, apresentaram desempenho superior tanto na regressão quanto na classificação. A análise de importância das variáveis indicou que fatores relacionados à estrutura produtiva, à formalização do trabalho e à capacidade de arrecadação exercem papel central na explicação do desempenho econômico municipal, em consonância com a literatura de desenvolvimento regional.

Do ponto de vista aplicado, o modelo proposto funciona como uma ferramenta de apoio à tomada de decisão, auxiliando na identificação de municípios com fundamentos econômicos mais sólidos. No entanto, seus resultados não devem ser interpretados como substitutos de análises qualitativas locais, uma vez que características institucionais, setoriais e territoriais específicas podem influenciar de forma decisiva a atratividade de investimentos em cada município.

Como limitações, destacam-se a periodicidade anual e a defasagem dos dados públicos, que restringem a capacidade de capturar choques recentes ou mudanças

estruturais de curto prazo. Como perspectivas futuras, sugere-se a incorporação de indicadores mais granulares, bem como o desenvolvimento de ferramentas interativas que permitam o acompanhamento contínuo do desempenho municipal.

Em síntese, este trabalho contribui para a área ao demonstrar como técnicas modernas de Machine Learning podem complementar abordagens tradicionais na análise do desenvolvimento regional, oferecendo um instrumento quantitativo robusto para apoiar decisões estratégicas e ampliar o uso de métodos preditivos na avaliação do potencial econômico municipal.

REFERÊNCIAS

ANTUNES, Raryson Miletto Câmara. **Indicadores socioeconômicos determinantes da condição financeira nos municípios brasileiros**. 2024. Dissertação (Mestrado em Ciências Contábeis) – Universidade Federal do Rio Grande do Norte, Natal, 2024. Disponível em: https://repositorio.ufrn.br/server/api/core/bitstreams/26a034e8-e9a4-4c60-a35d-dd112eb319a2/content. Acesso em: 23 ago. 2025.

ARAÚJO, Tarcisio Patricio de; SOUZA, Aldemir do Vale; LIMA, Roberto Alves de. Nordeste: economia e mercado de trabalho. **Estudos Avançados**, São Paulo, v. 33, n. 97, p. 135-154, set./dez. 2019. Disponível em: http://scielo.br/j/ea/a/gRvbkXwjxKcvjxHmqzCGWcc. Acesso em: 15 mar. 2025.

BASE DOS DADOS. **Cadastro Geral de Empregados e Desempregados (Novo Caged)**. Disponível em: https://basedosdados.org/dataset/562b56a3-0b01-4735-a049-eeac5681f056. Acesso em: 10 set. 2025.

BASE DOS DADOS. **Relação Anual de Informações Sociais (RAIS)**. Disponível em: https://basedosdados.org/dataset/3e7c4d58-96ba-448e-b053-d385a829ef00. Acesso em: 10 set. 2025.

BISHOP, Christopher M. **Pattern Recognition and Machine Learning**. New York: Springer, 2006.

BREIMAN, Leo. Random forests. **Machine Learning**, New York, v. 45, n. 1, p. 5-32, 2001. Disponível em: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf. Acesso em: 26 jul. 2025.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. *In*: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., 2016, New York. **Proceedings** [...]. New York: ACM, 2016. p. 785–794. Disponível em: https://doi.org/10.1145/2939672.2939785. Acesso em: 26 jul. 2025.

HISSA-TEIXEIRA, Keuler. Uma análise da estrutura espacial dos indicadores socioeconômicos do nordeste brasileiro (2000-2010). **EURE (Santiago)**, Santiago, v. 44, n. 131, p. 101-124, jan. 2018. Disponível em: http://dx.doi.org/10.4067/S0250-71612018000100101. Acesso em: 23 ago. 2025.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Áreas Territoriais**. Rio de Janeiro: IBGE, 2021. Disponível em: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/15761-areas-dos-municipios.html. Acesso em: 10 set. 2025.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Estimativas de população**. Rio de Janeiro: IBGE, 2021. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html. Acesso em: 10 set. 2025.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Produto Interno Bruto dos Municípios**. Rio de Janeiro: IBGE, 2021. Disponível em: https://www.ibge.gov.br/estatisticas/economicas/contas-nacionais/9088-pib-dos-municipios.html. Acesso em: 10 set. 2025.

KELLEHER, John D.; MAC NAMEE, Brian; D'ARCY, Aoife. **Fundamentals of machine learning for predictive data analytics**: algorithms, worked examples, and case studies. Cambridge: MIT Press, 2015. Disponível em: https://mitpress.mit.edu/9780262029445/fundamentals-of-machine-learning-for-predictive-data-analytics/. Acesso em: 14 jul. 2025.

MITCHELL, Tom M. **Machine Learning**. New York: McGraw-Hill, 1997.

PAIXÃO, Márcia Cristina Silva; NOGUEIRA, Jorge Madeira. Investimento estrangeiro direto no Nordeste brasileiro: relação com o meio ambiente, inovações e potencial de spillovers. **Revista Econômica do Nordeste**, Fortaleza, v. 48, n. 2, p. 45-60, out. 2017. Disponível em: https://www.bnb.gov.br/revista/ren/article/view/727. Acesso em: 15 mar. 2025.

PAIXÃO, Márcia Cristina Silva; NOGUEIRA, Jorge Madeira. Investimento estrangeiro direto no Nordeste brasileiro: vetor de desenvolvimento?. **Cadernos do Desenvolvimento**, Rio de Janeiro, v. 13, n. 22, p. 55-80, jan./jun. 2018. Disponível em: https://www.cadernosdodesenvolvimento.org.br/ojs-2.4.8/index.php/cdes/article/view/26. Acesso em: 15 mar. 2025.

RIBEIRO, Luiz Carlos de Santana; DOMINGUES, Edson Paulo; PEROBELLI, Fernando Salgueiro. **Investimentos estruturantes e desigualdades regionais na região Nordeste do Brasil**. Fortaleza: Banco do Nordeste, 2015. Disponível em: https://bnb.gov.br/documents/45787/671973/Investimentos+estruturantes+e+desigualdades+regionais+na+regi%C3%A3o+nordeste+do+Brasil.pdf/8ed43c55-5db2-fd26-0c37-fb4d58ba9481?version=1.0&t=1638534571272&download=true. Acesso em: 05 jun. 2025.

SANTOS, C. R. S. *et al.* Aplicação do Machine Learning na Previsão de Índices Econômicos: O IDHM com o Modelo Random Forest de 2012 à 2021. *In*: ENCONTRO DE ECONOMIA DA REGIÃO SUL, 16., 2023, Florianópolis. **Anais [do] 16º Encontro de Economia da Região Sul**. Porto Alegre: ANPEC Sul, 2023. p. 1-17. Disponível em: https://www.researchgate.net/publication/375992831_APLICACAO_DO_MACHINE_LEARNING_NA_PREVISAO_DE_INDICES_ECONOMICOS_O_IDHM_COM_O_MODELO_RANDOM_FOREST_DE_2012_A_2021. Acesso em: 23 ago. 2025.

SILVEIRA NETO, Raul da Mota; AZZONI, Carlos Roberto. **Non-spatial government policies and regional income inequality in Brazil**. São Paulo: FEA-USP, 2008. Disponível em: https://ideas.repec.org/p/ekd/000238/23800008.html. Acesso em: 13 ago. 2025.

TESOURO NACIONAL. **Capacidade de Pagamento (CAPAG)**. Brasília. Disponível em: https://www.tesourotransparente.gov.br/ckan/dataset/capag-municipios. Acesso em: 10 set. 2025.

TESOURO NACIONAL. **Siconfi**: sistema de informações contábeis e fiscais do setor público brasileiro. Brasília. Disponível em: https://siconfi.tesouro.gov.br/siconfi/index.jsf. Acesso em: 10 set. 2025.