UNIVERSIDADE FEDERAL DE PERNAMBUCO
CAMPUS AGRESTE
NÚCLEO DE GESTÃO
CURSO DE CIÊNCIAS ECONÔMICAS

JOSÉ CORREIA DE FARIAS FILHO

INTEGRAÇÃO DO MODELO RANDOM FOREST E DA ANÁLISE ESPACIAL NOS DETERMINANTES DO IFDM NOS MUNICÍPIOS DE PERNAMBUCO (2013-2021)

Caruaru
2025

JOSÉ CORREIA DE FARIAS FILHO

INTEGRAÇÃO DO MODELO RANDOM FOREST E DA ANÁLISE ESPACIAL NOS DETERMINANTES DO IFDM NOS MUNICÍPIOS DE PERNAMBUCO (2013-2021)

Trabalho de Conclusão de Curso apresentado à Coordenação do Curso de Ciências Econômicas do Campus Agreste da Universidade Federal de Pernambuco - UFPE no formato de monografia, como requisito parcial para a obtenção do grau de bacharel em Ciências Econômicas.

Área de concentração: Economia Regional e Urbana

Orientador: Prof. Dr. Denis Fernandes Alves

Caruaru

2025

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

FILHO, JOSÉ CORREIA DE FARIAS FILHO.
INTEGRAÇÃO DO MODELO RANDOM FOREST E DA ANÁLISE ESPACIAL NOS DETERMINANTES DO IFDM NOS MUNICÍPIOS DE PERNAMBUCO (2013-2021) / JOSÉ CORREIA DE FARIAS FILHO FILHO. - Caruaru, 2025.
62

Orientador(a): DENIS FENANDES ALVES ALVES
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro Acadêmico do Agreste, Ciências Econômicas, 2025.

1. Economia regional. I. ALVES, DENIS FENANDES ALVES. (Orientação). II. Título.

330 CDD (22.ed.)

JOSÉ CORREIA DE FARIAS FILHO

INTEGRAÇÃO DO MODELO RANDOM FOREST E DA ANÁLISE ESPACIAL NOS DETERMINANTES DO IFDM NOS MUNICÍPIOS DE PERNAMBUCO (2013-2021)

Trabalho de Conclusão de Curso apresentado à Coordenação do Curso de Ciências Econômicas do Campus Agreste da Universidade Federal de Pernambuco – UFPE no formato de monografia, como requisito parcial para a obtenção do grau de bacharel em Ciências Econômicas.

Aprovado em: 15/12/2025.

BANCA EXAMINADORA

_____

Prof. Dr. Denis Fernandes Alves
Universidade Federal de Pernambuco (Orientador)

_____

Prof. Dr. José Sergio Casé de Oliveira (Examinador Interno)
Universidade Federal de Pernambuco

_____

Prof. Dr. Joebson Maurilio Alves Dos Santos (Examinador Externo)
Universidade de Pernambuco

AGRADECIMENTOS

Primeiramente agradeço a Deus por tudo as pessoas que colocou em meu caminho os momentos que serviram de ensinamentos. Meu guia espiritual nos momentos bons e principalmente nos momentos difíceis que passei. Agradeço a minha mãe que me apoia em toda minha jornada e sempre acredita e faz suas orações me deixam motivado e com força todos os dias. Agradeço imensamente o meu orientador/mentor, Prof. Dr. Denis Fernandes que me orienta e acredita mais em mim do que as vezes eu mesmo acredito aceito me orientar e comprou todos os momentos de orientação com muito animo e força de vontade. Além disso, tem muitos ensinamentos que aprendi com ele e com certeza irão me ajudar na caminhada partindo daqui ao futuro fora a inspiração e exemplo para vida.

Agradeço a todas as pessoas e momentos que perpassaram por minha caminhada até o momento de agora houveram vários momentos diferentes que me trouxeram até hoje. Sou grato a Universidade Federal de Pernambuco por ter entrado no ensino superior e por crescer como pessoa aos poucos dentro desse ambiente é um ambiente de estudo, mas também social de vivência aprendizado e adaptação.

RESUMO

O estudo tem o objetivo de aplicar o modelo *Random Forest* em conjunto com técnicas de análise exploratória para entender e prever o IFDM dos municípios de Pernambuco em 2021. Para tanto utilizou-se dados oriundos do Atlas e no portal da Firjan nos anos de 2013 a 2021, considerando uma trajetória importante dentro da conjuntura econômica para prever o IFDM de 2021 e com isso testar para autocorrelação espacial. Na etapa de modelagem, empregou-se o algoritmo Random Forest, escolhido por sua capacidade de lidar com múltiplas variáveis, pela robustez e por ser um método de *ensemble* que reduz o sobreajuste ao combinar as predições de diversas árvores de decisão. Os resultados indicam que o Random Forest se mostrou adequado ao conjunto de dados, apresentando desempenho estatisticamente significativo. Ao testar o valor predito para autocorrelação espacial por meio do I de Moran, observou-se uma estatística positiva e significativa. A AEDE também evidenciou padrões espaciais relevantes. Além disso, a combinação entre o Random Forest e as análises espaciais contribuiu para suavizar resultados apresentados pelo Índice de Moran, reduzindo ruídos e controlando possíveis endogeneidades das variáveis. Como consequência, foi possível obter maior clareza na identificação de correlações espaciais, potenciais *clusters* e efeitos de transbordamento. Entre os achados, destaca-se a formação de um *cluster* de Baixo-Alto desenvolvimento em torno de Moreno, na Região Metropolitana do Recife, município que apresenta o menor IFDM de sua mesorregião. Por outro lado, Caruaru, uma das dez cidades com melhor IFDM do estado, e seus municípios vizinhos compõem um *cluster* de desenvolvimento positivo (Alto-Alto). O estudo também identificou *outliers*, como Afrânio, no Oeste Pernambucano, que apresenta baixo desenvolvimento mesmo estando cercado por municípios com indicadores mais elevados (Baixo-Alto), indicando a necessidade de análises adicionais para compreender esse comportamento. Por fim, os resultados demonstram-se especialmente relevantes para a formulação de políticas públicas, pois evidenciam a importância de prever variáveis de desenvolvimento e compreender o comportamento regional do crescimento econômico dos municípios. Isso permite direcionar esforços a *clusters* de subdesenvolvimento e às variáveis mais influentes identificadas pelo modelo, além de potencializar regiões com correlações espaciais positivas e ampliar os efeitos de transbordamento do desenvolvimento.

Palavras-chave: Random Forest, Análise Exploratória de Dados Espacial, Pernambuco, Desenvolvimento Municipal, IFDM.

ABSTRACT

The study aims to apply the Random Forest model in combination with exploratory analysis techniques to understand and predict the IFDM of municipalities in Pernambuco in 2021. To this end, data from the Atlas and the Firjan portal from 2013 to 2021 were used, considering a relevant trajectory within the economic context to forecast the 2021 IFDM and subsequently test for spatial autocorrelation. In the modeling stage, the Random Forest algorithm was employed due to its ability to handle multiple variables, its robustness, and its ensemble nature, which reduces overfitting by combining predictions from multiple decision trees. The results indicate that Random Forest was well suited to the dataset, showing statistically significant performance. When testing the predicted value for spatial autocorrelation using Moran's I, a positive and significant statistic was observed. The Exploratory Spatial Data Analysis (ESDA) also revealed relevant spatial patterns. Moreover, the combination of Random Forest with spatial analyses helped improve the prediction of Moran's Index by reducing noise and controlling for variable endogeneity. As a result, it became possible to more clearly identify spatial correlations, potential *clusters*, and *spillover* effects. Among the findings, the formation of a Low-High development *cluster* around Moreno, in the Metropolitan Region of Recife, the municipality with the lowest IFDM in its mesoregion, stands out. Conversely, Caruaru, one of the ten municipalities with the highest IFDM in the state, along with its neighboring municipalities, forms a positive development *cluster* (High-High). The study also identified outliers, such as Afrânio, in the Western region of Pernambuco, which shows low development despite being surrounded by municipalities with higher indicators (Low-High), indicating the need for additional analyses to better understand this behavior. Finally, the results are especially relevant for public policy formulation, as they highlight the importance of predicting development variables and understanding the regional dynamics of municipal economic growth. This enables targeted action toward underdeveloped clusters and the key variables identified by the model, while also strengthening regions with positive spatial correlations and enhancing development spillover effects.

Keywords: Random Forest, Exploratory Spatial Data Analysis, Pernambuco, Municipal Development, IFDM.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O uso de técnicas estatísticas, em especial as ferramentas da econometria, constitui um dos pilares da análise econômica. Paralelamente, o avanço tecnológico tem transformado de forma acelerada a sociedade, especialmente na era digital e dos contínuos progressos científicos. Nesse contexto, a área de inteligência artificial tem se destacado como uma das principais fronteiras do conhecimento, com ênfase nas metodologias de aprendizado de máquina (*machine learning*), que oferecem novas possibilidades para a investigação e compreensão de fenômenos econômicos complexos e as questões espaciais que os envolvem.

Segundo Athey (2019) técnica de *machine learning* ganha cada vez mais espaço para previsões na economia, por ter como características previsões mais precisas quando se tratando de grandes quantidades de dados e encontrar de forma ágil padrões em dados complexos. Para Athey (2019) a junção de novos conjuntos de dados produzidos pela economia com técnicas de aprendizado de máquina mudará a economia de muitas maneiras fundamentais que podem estar conectadas a novas questões novas abordagens ou no envolvimento de economistas na engenharia e implementações de políticas.

O campo de estudo de inteligência artificial é amplo e tem base no campo da estatística. A abordagem teórica em *machine learning* pode-se dividir os modelos em i) supervisionado, os quais utilizam dados rotulados; ii) não supervisionados, que utilizam dados não rotulados. Além disso, entende-se por dados rotulados aqueles que já tem uma classificação ou resultado pré-estabelecido, por outro lado os não rotulados são dados onde o próprio algoritmo encontrará padrões sem um resultado já definido (Athey, p. 509, 2019).

O presente estudo adapta o método do Índice De Desenvolvimento Humano (IDH) e sua variação local IDHM com o Índice de Firjan de Desenvolvimento Municipal IFDM. De acordo com Tobaigy, Alamoudi e Bafail (2023) esses Índices são essenciais para medir o progresso humano e classificar áreas geográficas com base em seus níveis de desenvolvimento. A análise de Tobaigy, Alamoudi e Bafail indicam que ao incorporar variáveis como, educacional, econômicas e de saúde, fornece uma melhor visão da qualidade de vida da população.

No entanto, para além da predição, a análise espacial dos dados assume papel fundamental na identificação de *clusters* de municípios que apresentam níveis relativamente inferiores de desenvolvimento humano (municipal). Essa abordagem permite compreender não apenas os valores do indicador, mas também como o desenvolvimento se distribui

territorialmente, revelando padrões de concentração e dependência espacial (Anselin, 1995; Almeida, 2012).

No contexto brasileiro, a integração entre técnicas preditivas e análise espacial ainda é incipiente na literatura de economia regional. Poucos estudos têm buscado combinar o potencial das metodologias de *machine learning* com ferramentas de análise exploratória de dados espaciais (AEDE) para investigar os determinantes e a distribuição do desenvolvimento socioeconômico entre os municípios, embora isso não seja indicativo de lacunas na literatura este estudo busca identificar se há uma possível suavização e melhora da AEDE com essa integração.

Outro aspecto relevante refere-se ao uso do Índice FIRJAN de Desenvolvimento Municipal (IFDM) como variável de análise. Diferentemente do Índice de Desenvolvimento Humano Municipal (IDHM), calculado pelo Programa das Nações Unidas para o Desenvolvimento (PNUD) apenas a cada dez anos, o IFDM apresenta atualização anual e baseia-se em indicadores públicos oficiais das áreas de emprego e renda, educação e saúde (FIRJAN, 2023). Essa periodicidade e abrangência tornam o IFDM uma alternativa valiosa para estudos temporais e espaciais sobre o desenvolvimento local.

A aplicação de técnicas de análise espacial, como o Indicador Local de Associação Espacial (LISA), mostra-se útil para identificar agrupamentos (*clusters*) de municípios com desempenho semelhante, sejam eles de alto ou baixo desenvolvimento. Tais análises permitem compreender a dinâmica espacial do processo desenvolvimentista ao longo do tempo, fornecendo evidências sobre a persistência ou a mudança desses padrões regionais (Anselin, 1995).

Além do contexto da Análise univariada, temos também a bivariada que relaciona duas variáveis diferentes no espaço. Segundo Rodrigues et al. (2015), a autocorrelação espacial global bivariada permite verificar se uma variável observada em determinada região tem alguma associação com outra variável em regiões vizinhas.

Dessa forma, alinhar o estudo preditivo do IFDM por meio do modelo *Random Forest* com as ferramentas da AEDE constitui a principal contribuição deste trabalho, pois integra a capacidade do *machine learning* à compreensão espacial do desenvolvimento, oferecendo uma visão mais completa e dinâmica das desigualdades regionais com a possível suavização integrando os dois modelos.

Posto isto, este estudo emprega o método de *Random Forest*, que é uma técnica de *machine learning,* para prever o Índice FIRJAN de Desenvolvimento Municipal (IFDM) nos municípios de Pernambuco, utilizando dados rotulados para o período de 2013 a 2021. A

predição do IFDM é relevante, pois pode suavizar e melhorar o ajuste do modelo de AEDE fornecendo subsídios para a formulação de políticas públicas mais eficazes e direcionadas em regiões menos favorecidas. Além disso, a aplicação de técnicas de aprendizado de máquina na economia, embora ainda incipiente no contexto brasileiro, tem apresentado crescimento expressivo e consolidado sua importância como ferramenta complementar às abordagens econométricas tradicionais, justificando a ampliação de seu uso em análises regionais.

O estado de Pernambuco localizado na região do Nordeste tem um dos IFDMs mais baixos comparado aos outros estados mesmo com políticas de desenvolvimento feitas no ano de 2010. Dessa forma, o Estado de Pernambuco foi escolhido como alvo para verificar-se, pois encontram-se cidades com índices de desenvolvimento muito baixos, como exemplo de Afrânio e Moreno. Por conseguinte, utilizou-se os anos de 2013 a 2021 que são os mais recentes encontrados no Atlas e foi feita uma junção com o IFDM dos mesmos anos das variáveis disponíveis no Firjan a escolha dessas fontes de dados dar-se por causa da relação de cálculo dos próprios índices utilizando variáveis de Educação, renda e Saúde.

O presente estudo busca realizar predições por meio de técnicas de aprendizado de máquina, que vêm ganhando destaque na estimação de variáveis socioeconômicas. Além disso, pretende-se identificar os fatores mais relevantes para prever o desenvolvimento humano regional, contribuindo para a formulação de políticas públicas mais eficazes. Ademais, não se tem a intenção de esgotar todos os conceitos sobre aplicação de técnicas de *machine learning* para o estudo o IFDM do estado de Pernambuco, mas sim contribuir como um estudo que aumenta as aplicações dessa nova abordagem no campo da economia a fim de trazer novas perspectivas e contribuindo como uma base para estudos futuros que venham a aprofundar o conhecimento.

Os anos de abordagem desse estudo serão de 2013 a 2021. Dessa forma, foram encontrados resultados similares aos estudos de Tobaigy, Alamoudi e Bafail (2023) ao utilizar-se o *Random Forest*, além disso os resultados estatísticos na análise de AEDE foram significantes. Por fim, a integração dos modelos mostrou que é possível suavizar e melhorar a visualização de *clusters* espaciais combinando a predição do modelo e a capacidade de análise de correlação espacial da AEDE.

Além desta introdução, o estudo está estruturado em cinco seções. A segunda apresenta a revisão da literatura e as principais contribuições teóricas que fundamentam o trabalho, incluindo uma visão geral sobre os dados e o modelo adotado. A terceira descreve a base de dados e a metodologia, abordando a exploração e a análise descritiva inicial. A quarta seção reúne o processamento e a análise dos dados, o treinamento e a aplicação do modelo *Random*

Forest, bem como os resultados obtidos. Nessa mesma seção, realiza-se a Análise Exploratória de Dados Espaciais (AEDE), com a identificação de padrões de correlação espacial a partir das bases da FIRJAN e do Atlas. Por fim, a quinta seção apresenta as considerações finais do estudo.

2 REVISÃO DE LITERATURA

2.1 O IFDM como indicador de desenvolvimento municipal

O IFDM é uma adaptação do (IDH), originalmente concebido pela PNUD, para refletir as condições de vida nos municípios brasileiros. O IFDM é composto por três dimensões fundamentais: longevidade, educação e renda. No entanto, ele é calculado a partir de dados disponíveis em nível municipal (Firjan, 2025). A metodologia descrita pela Firjan mostra que o IFDM permite comparações entre municípios e regiões, que oferece um diagnóstico objetivo das desigualdades sociais e estruturais que afetam diretamente a qualidade de vida dos indivíduos.

A sistematização do indicador contribui para o planejamento e formulação de políticas públicas mais eficazes, uma vez que fornece subsídios técnicos para a gestão municipal. Com base nas informações da PNUD e do IBGE o índice reflete as transformações socioeconômicas ocorridas ao longo do tempo (Firjan, 2023). Tobaigy, Alamoudi e Bafail (2023), analisaram indicadores compostos como o IDHM, destacam a importância dar padronização metodológica dar acesso às comparações significativas entre localidades. Os autores ressaltam que os fatores como o acesso à saúde, à educação e às oportunidades econômicas são determinantes para a variação no índice, influenciando diretamente a mobilidade social e o planejamento público.

Ao apresentar o IFDM, reforça-se o valor dos indicadores sintéticos na análise da gestão pública. Embora tenha periodicidade anual e utilize uma metodologia distinta, o IFDM também confirma a correlação entre o desenvolvimento humano e o investimento em políticas públicas estruturantes (Firjan, 2023). De forma geral, a sua literatura destaca a relevância como instrumento de avaliação e gestão do desenvolvimento local, sendo amplamente utilizado em estudos acadêmicos, relatórios técnicos e análises institucionais voltadas à promoção de bem-estar e redução de desigualdades.

Muitos estudos têm utilizado o Índice FIRJAN de Desenvolvimento Municipal (IFDM) para compreender desigualdades regionais no Brasil sob diferentes recortes geográficos e temáticos. Avelino, Bressan e Cunha (2013) analisaram as capitais brasileiras entre 2005 e 2010 e identificaram que fatores contábeis da gestão pública influenciam significativamente o desempenho do IFDM, especialmente nas dimensões de Emprego, Renda e Educação. Corrêa e Duarte (2017), ao estudarem 295 municípios de Santa Catarina, constataram que, embora o estado apresente resultados superiores à média nacional, persistem disparidades internas marcantes entre regiões mais industrializadas e áreas rurais. Já Kruger e Bourscheidt (2021), ao mapearem o IFDM no Paraná, observaram padrões espaciais de concentração, com municípios

de alto desenvolvimento nas regiões Norte Central e Metropolitana de Curitiba, e baixos índices nas regiões Centro-Sul e Norte, evidenciando a persistência de desigualdades estruturais. Esses trabalhos, cobrindo o período de 2010 a anos recentes, demonstram a versatilidade do IFDM como ferramenta para avaliar a dinâmica socioeconômica brasileira em múltiplas escalas.

Para realidade brasileira o Atlas Brasil adota uma metodologia similar para o cálculo do IDHM, adaptando os indicadores para o nível local. O IDHM também é composto por três dimensões que são longevidade, educação e renda, mas com variáveis mais específicas à realidade municipal (IPEA, 2014). A longevidade é obtida através da *proxy* da expectativa de vida ao nascer já a educação é calculada com base no fluxo escolar da população jovem e na escolaridade da população adulta, e a renda é representada pela renda per capita municipal. Tal como no IDH global, cada dimensão é normalizada entre os valores mínimo e máximo fixados pelo PNUD e o índice final resulta da média geométrica dos três subíndices. Essa adaptação permite comparações entre os municípios brasileiros e o monitoramento de desigualdades regionais (IPEA, 2014).

Com o fito de analisar o comportamento das variáveis econômicas é necessário compreender os fatores que podem afetá-las. Ademais, a abordagem econômica frequentemente dispõe de modelos matemáticos em específico os modelos econométricos e mais recentemente modelos de aprendizado de máquina, conforme Athey (2019) afirmam em seu estudo. A seguir será apresentado e explicado o modelo utilizado no presente estudo e na literatura.

2.2 Aplicação de índices e utilização dos modelos de *machine learning* na economia

Na economia, é crescente o uso de modelos de *machine learning* para auxiliar nas análises e estudos empíricos. A fim de compreender o comportamento das variáveis econômicas, torna-se necessário identificar os fatores que podem influenciá-las de maneira direta ou indireta. Ademais, a abordagem econômica tradicionalmente se apoia em modelos matemáticos, em especial os modelos econométricos, e, mais recentemente, tem incorporado técnicas de aprendizado de (Athey, 2019).

Segundo Athey (2019) o uso de técnicas de *machine learning* na economia amplia as possibilidades de análise e previsão, especialmente em contextos com grandes volumes de dados e múltiplas variáveis. No entanto, ela enfatiza que, diferentemente da econometria comumente utilizada, voltada à identificação de relações causais, o *machine learning* prioriza a capacidade preditiva. A autora defende, assim, a integração entre ambas as abordagens, que combina a possibilidade de identificar causalidade de alguns modelos econométricos com a

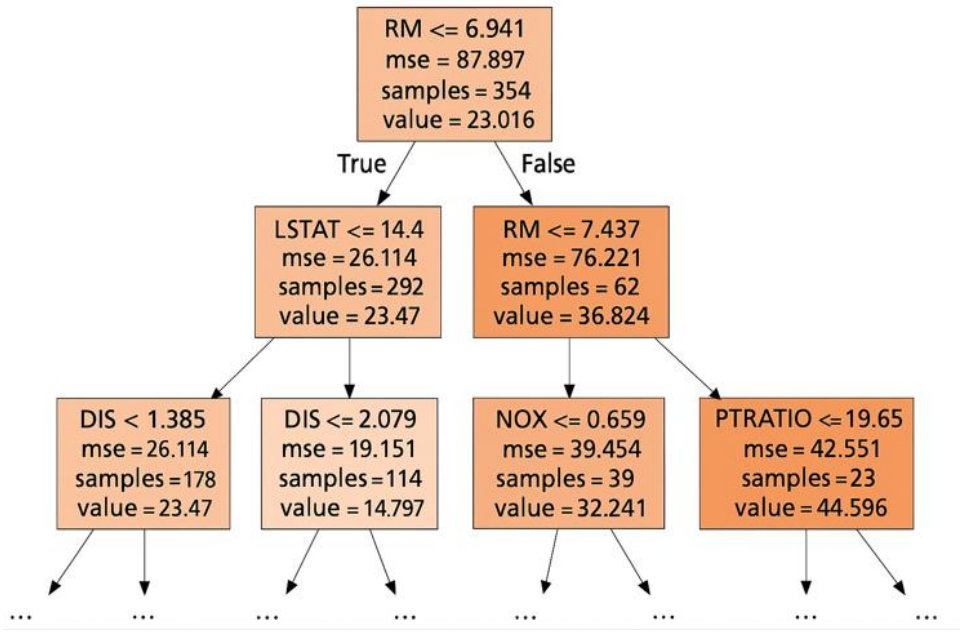
flexibilidade e o poder computacional do *machine learning*, de modo a aprimorar tanto a precisão das análises quanto a compreensão dos fenômenos econômicos.

O funcionamento a princípio como do modelo *Random Forest* se baseia em entender como funciona o modelo de árvore de decisão. Uma árvore de decisão é representada pela estrutura de árvore, onde possui um determinado número de caminhos possíveis de decisão, o que será formalmente explicado adiante, e há o resultado de cada um deles. Parte do princípio de buscar solucionar um problema através de perguntas para prever seu resultado, que podem ser por dados categóricos ou dados numéricos (GRUS, 2021).

Desde a criação do algoritmo ID3, por J. Ross Quinlan em 1986, as árvores de decisão tornaram-se uma das principais técnicas de aprendizado de máquina supervisionado. Pouco antes, em 1984, Breiman e colaboradores já haviam desenvolvido o algoritmo CART (*Classification and Regression Trees*), também amplamente utilizado. Ambos os métodos funcionam de forma semelhante: eles dividem o conjunto de dados em partes menores, passo a passo, até chegar a grupos mais homogêneos, processo conhecido como “divisão e conquista”. Essa abordagem hierárquica, que se assemelha a uma árvore com ramos e folhas, torna o processo de classificação e previsão mais claro e eficiente, facilitando a interpretação dos resultados. A árvore de decisão possui vários pontos positivos como sua interpretação fácil, além de sua previsão ter um modelo de simples visualização. Como é possível verificar através da Figura 1 (GRUS, 2021).

No entanto, existe algumas ressalvas para a utilização de árvores de decisão que devem ser levadas em consideração ao se escolher esse modelo para aplicar nos dados. Esse modelo acaba sendo totalmente dependente dos dados de treinamento, já que uma pequena variação pode acabar modificando a estrutura da árvore por completo (Harrison, 2020).

Figura 1: Modelo de Árvore de Decisão



Fonte: Adaptado de Harrison (2020, p. 194).

Outro aspecto a se observar através da Figura 1 é que os erros estão representados pelo MSE (Erro quadrático médio), então a "folha" que apresentar o menor MSE será a escolhida. Sendo assim os modelos de árvore de decisão são afetados pelo ajuste dos dados de treinamento que pode levar a um superajuste, na literatura comumente chamado de overfitting[1] (GRUS, 2021). O algoritmo de *Random Forest* é um conjunto de árvores de decisões que buscam corrigir a tendência das árvores à super adequação. Quando é criada uma série de árvores e são treinadas com subamostras e atributos aleatórios (Harrison, 2020).

A fim de corrigir a tendência da árvore de decisão é usado o *bootstrap*. Segundo Grus (2021) o *bootstrap* é feito por meio de uma amostra cuja média é 100 dessa forma, caso todos os pontos fiquem próximos a 100 então a mediana será próxima a este valor. Contudo, caso a distribuição de metade dos dados são próximos a 0 e a outra metade dos dados próximos a 200, a mediana não parece ser bem determinada e eficiente.

De acordo com Efron (1979) o *bootstrap* é um método para conseguir estimar a função de distribuição sem depender de suposições paramétricas. Nesse sentido o método utiliza dos próprios dados para estimar uma aproximação boa o suficiente da função *F* da distribuição da

[1] De acordo com Athey (2019), a definição de *overfitting* é um ajuste de forma excessiva ao ruído ou às coincidências específicas do conjunto de dados de treinamento, captando correlações espúrias que não são encontradas em novos dados, o que prejudica sua capacidade de generalização e compromete sua robustez preditiva em diferentes contextos.

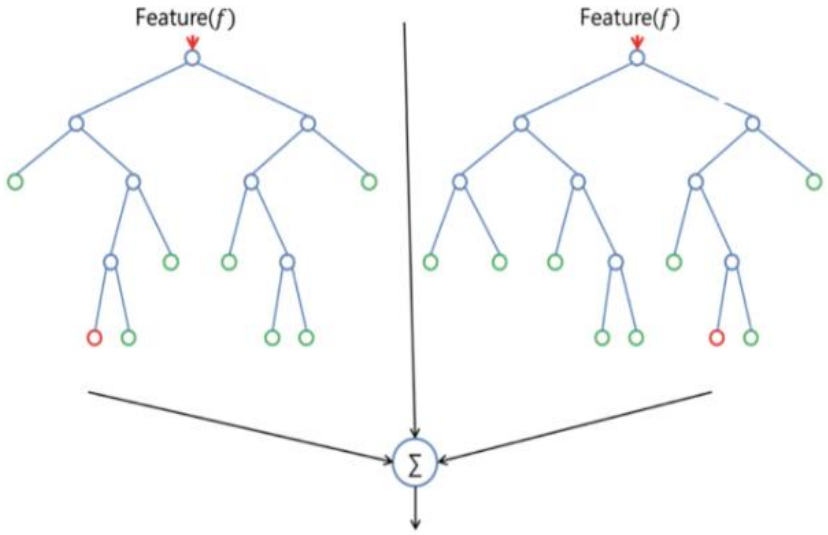
população. Queremos estimar um parâmetro estatístico $\theta = t(F)$ como F é desconhecida, não podemos calcular diretamente θ, então usamos o estimador amostral $\hat{\theta} = t(\hat{F})$. Dessa maneira podemos gerar B amostras $\{X^1, X^2, \dots, X^b\}$ de tamanho n cada, por reposição amostral com reposição de X. Sendo assim, para cada amostra de X^b podemos calcular o estimador $\hat{\theta}^b = t(\hat{F}^b)$. Posto isso, temos $\{\hat{\theta}^1, \hat{\theta}^2, \dots, \hat{\theta}^b\}$ que representa a distribuição *bootstrap*.

No caso da *Random Forest*, o *bootstrap* é usado para corrigir essa tendência da utilização de uma árvore específica. Quando são criadas várias árvores, elas são diferentes umas das outras, trazendo como benefício da viabilidade do uso do conjunto total dos dados (GRUS, 2021). No entanto, existe algumas ressalvas para a utilização de árvores de decisão que devem ser levadas em consideração ao se escolher esse modelo para aplicar nos dados. Esse modelo acaba sendo totalmente dependente dos dados de treinamento, já que uma pequena variação pode acabar modificando a estrutura da árvore por completo (Harrison, 2020).

Além disso, outra forma de fazer a seleção do melhor resultado da *Random Forest* pode ser buscar em uma seleção aleatória a melhor divisão em cada nó de cada árvore. Para k-ésima árvore, um vetor aleatório Θk é gerado, independente dos vetores aleatórios anteriores $\Theta^1, \dots, \Theta^{k-1}$, mas com a mesma distribuição, e uma árvore é desenvolvida usando o conjunto de treinamento e Θk, resultando em um classificador $h(x, \Theta k)$ onde x é um vetor de entrada.

Sendo assim, segundo Breiman (2001) Θ consiste em um número de inteiros aleatórios independentes entre 1 e K. A natureza e a dimensionalidade de Θ dependem de seu uso na construção de árvores. Após a geração de um grande número de árvores, elas votam na classe mais popular. Esses procedimentos são denominados de *Random Forests* a Figura 2 exemplifica visualmente esse conceito (Breiman, p. 2, 2001).

Figura 2: Funcionamento do modelo de *Random Forest*



Fonte: Hudson, 2018.

O modelo *Random Forest* é um método de *ensemble learning* baseado em árvores de decisão, no qual diversas árvores são construídas a partir de subconjuntos dos dados e, em seguida, combinadas para gerar uma previsão mais precisa e estável. O resultado é obtido por meio de votação, quando se trata de classificação, ou pela média dos valores previstos, no caso de regressão (Santos, 2022). Cada árvore que compõe a *Random Forest* possui uma estrutura hierárquica formada por nós, onde são realizadas decisões binárias baseadas nas variáveis analisadas (*features*). O nó raiz representa o ponto inicial de decisão, enquanto os nós folha indicam os resultados, sem novas subdivisões. Dessa forma, o modelo combina a simplicidade das árvores individuais com o poder coletivo do conjunto, resultando em previsões mais robustas e menos suscetíveis a erros (Santos, 2024).

De acordo com os estudos de Santos (2024) o modelo consiste em um espaço $X \in R^p$ de respostas preditoras onde p representa um conjunto de p variáveis numéricas. A árvore de decisão divide o espaço de dados em partes menores chamadas regiões e essas regiões são homogêneas. Cada região pode ser R^j para $j = 1, \dots, J$. Desta forma ao definirmos o regressor de uma Árvore de Decisão é criada a partição das preditoras definida pelas regiões $R^1, R^2, \dots, R^J$ no espaço (Santos, 2024).

Além disso, a profundidade de cada modelo pode-se dar pelo máximo do número de divisões feitas a partir do nó inicial para se chegar a cada um dos nós terminais da Árvore. A equação apresentada no estudo de Santos (2024) para exemplificar e definir esse conceito de regressor é dada pela Equação (1):

$$T(x, \Theta) = \sum_{j=1}^{j} \gamma_j \, I \, \{x \in R_j\} \quad (1)$$

em que $\Theta = \{(Rj, \gamma j)\} \, para \, j = 1, \ldots, J$ onde Rj indica as J regiões no espaço das preditoras e γ_j indica as previsões, pode-se afirmar que Θ contém todas as informações de aplicação do método (Santos, 2024).

A subseção a seguir apresenta os principais estudos que versam a respeito de aprendizado de máquina para prever o IFDM ou índices semelhantes e estudos relevantes na área de Análise espacial a AEDE, como também algumas de suas variáveis componentes, como a expectativa de vida ao nascer. Ademais, é feita uma revisão empírica com esses estudos apontando os resultados dos estudos na área e suas contribuições com a intenção de proporcionar uma base teórica para o estudo em questão.

2.3 Revisão empírica

Os estudos que analisam o IFDM ou índices similares que se utilizam de econometria e *machine learning* na economia propõe uma abordagem de entendimento destes índices, pois auxilia decisões de políticas públicas. Não obstante, os estudos também visam fazer previsões futuras do índice baseado em dados anteriores com a finalidade de sua projeção.

Na área da econometria tem se explorado o uso de técnicas de *Machine learning* aprimorando e expandindo as possibilidades de análise dos índices econômicos. Nesse contexto, o estudo de Sherman et al. (2023) faz uso de dados provenientes de satélites para prever o IDH do mundo todo em alta resolução. Os autores combinam imagem de satélites com alta resolução e as técnicas de *Random Forest* para fazer estimativas em mais de 190 países validando o IDH de cada país o estudo permite análises geográficas detalhadas e determinações mais precisas de áreas com baixo desenvolvimento humano.

Na mesma direção segue o estudo de Arumnisaa (2023) que em seu estudo aborda três modelos de machine learning, *Random Forest, Support Vector Machine (SVM) e AdaBoost,* seu estudo é definido a nível local das províncias da Indonésia, utilizando dados como renda per capita, taxas de matrículas brutas e líquidas entre outros. Esse estudo demonstra a eficiência dessas técnicas para o entendimento e previsão do indicador econômico.

Segundo o estudo de Tobaigy, Alamoudi e Bafail (2023) são utilizados dados para determinar e classificar os fatores significativos que afetam o IDH na Indonésia. Ele utiliza dados como a educação, expectativa de vida e alguns outros dados para definir a previsão do

IDH da Indonésia. Seu modelo foi capaz de prever com alta precisão o valor do índice e quais ações priorizar partindo das varáveis mais relevantes encontradas para aumentar o nível de IDH do país (Tobaigy, Alamoudi e Bafail, 2023).

Outro estudo que analisa indicadores econômicos é o estudo de Mcbride (2015) o qual verifica as variáveis que determinam o nível de pobreza em países emergentes utilizando-se de aprendizado de máquina e comparando aos métodos tradicionais já consolidados a fim de verificar a eficácia desses modelos mais atualizados para previsão de índices dentre os modelos analisados destaca-se o *Random Forest*.

Kaur et al. (2019) analisaram a eficácia de diferentes algoritmos de aprendizado de máquina supervisionado, como regressão logística, *Random Forest*, Cubist, Elastic Net, redes neurais e máquinas de vetores de suporte, na previsão do Índice de Qualidade de Vida em diversos países. O estudo apontou a eficiência e viabilidade para esse tipo de previsão, avaliadas por indicadores estatísticos como o coeficiente de determinação (R^2), o erro quadrático médio (RMSE) e a acurácia. Os resultados destacam o potencial dessas abordagens para apoiar a formulação de políticas públicas e o planejamento estratégico voltado ao desenvolvimento socioeconômico.

Özden e Guleryuz (2021) analisaram a relação entre desenvolvimento econômico e capital humano utilizando técnicas de regressão, classificação e métodos de ensemble, identificando padrões complexos e não lineares que explicam como variáveis de capital humano influenciam o desenvolvimento socioeconômico. O estudo concluiu que modelos preditivos bem ajustados possuem alto poder de predição e interpretabilidade, fornecendo subsídios para políticas públicas voltadas ao crescimento sustentável. Baseando-se em Hansan et al. (2016), que aplicaram técnicas como *Random Forest*, Gaussian Process e Multilayer Perceptron para prever o PIB per capita de Bangladesh, os autores reforçam o potencial do aprendizado de máquina na análise e predição de indicadores econômicos.

Segundo Lopes (2021) que faz a análise da AEDE no IDHM para o Brasil há uma correlação espacial e formação de *clusters* entre municípios que tem os melhores IDHM e que se encontram predominantemente nas regiões do Sudeste, Sul e Centro-Oeste. Além disso, em contraponto o estudo aponta que há *clusters* Baixo-Baixo entre municípios com baixo IDHM que predominam nas regiões Norte e Nordeste. Portanto o estudo conclui que é necessário um redimensionamento de políticas públicas priorizando essas regiões.

Outro artigo relevante para o estudo empírico, é o de Marconato (2019) que faz uma análise AEDE traz uma discussão do IDHM para o Brasil nos anos de 1991 e 2010 verificando as diferenças nas formações de *clusters*. O estudo encontra que pouco se mudou na formação

de agrupamentos de baixo e alto IDHM o qual 90,7 % dos grupos que tinham baixo desenvolvimento em 1997 permaneceram em 2010. Sendo assim o estudo aponta que a educação embora tenha sido o que percentualmente mais cresceu em relação a renda e saúde ainda tinha índices muito baixos com relação as regiões de alto desenvolvimento. Dessa forma, conclui-se no estudo que políticas direcionadas a melhora da educação impactam positivamente o desenvolvimento socioeconômico dos municípios.

Ademais, um estudo que propõe o AEDE para o IFDM é apresentado por Dalchiavon (2017) em sua tese de dissertação. Este estudo analisa o desenvolvimento econômico dos municípios paranaenses entre 2005 e 2013 via IFDM. Os resultados revelam crescimento heterogêneo, com destaque para as mesorregiões Norte Central e Oeste, que concentraram os maiores índices. A AEDE confirmou autocorrelação espacial positiva, identificando *clusters* positivos nessas regiões e *clusters* negativos no Centro-Sul. A análise do emprego mostrou crescimento de 47,98% no período, com os setores de Construção Civil, Calçados e Mecânica liderando. Conclui-se que as regiões com melhor desempenho no IFDM coincidem com os *clusters* positivo e com o maior crescimento do emprego.

Ademais, no contexto brasileiro, ainda não há um acervo amplo de estudos sobre o tema. Destaca-se apenas o trabalho de Oliveira (2023), que buscou prever o IDHM no território nacional utilizando dados do Atlas do Desenvolvimento Humano e avaliando a eficácia do modelo *Random Forest* na estimativa desse índice. No entanto, não foram identificadas pesquisas que enfoquem o IFDM e empreguem simultaneamente métodos de *machine learning* e análise espacial (estratégia descrita na seção seguinte). Diante disso, o presente estudo propõe-se a utilizar-se dessa integração para analisar o caso específico do estado de Pernambuco, no período de 2013 a 2021 e verificar se é viável essa suavização por meio o *Random Forest* e se a melhoras nas análises da AEDE.

3 METODOLOGIA

A metodologia utilizada nesse estudo tem como objetivo analisar o IFDM do estado de Pernambuco por meio de métodos de *Machine learning* e AEDE e mostrar variáveis que são relevantes e impactam o índice como também possam ser futuramente alvo de políticas públicas com estudos aprofundados. Posto isso, o estudo utiliza-se de análises descritivas para entender melhor os dados e a união dos métodos *Random Forest* e AEDE com intenção de verificar a viabilidade dos resultados. Ademais, Através do I de Moran é verificado se há correlação espacial, o *Random Forest* faz a predição do IFDM como também a verificação das variáveis que mais o impactam. A base de dados foi obtida de duas fontes oficiais, Atlas do Desenvolvimento Humano no Brasil, mantido pelo PNUD, e a base de dados do sistema FIRJAN.

3.1 Métodos de *machine learning*

Após obter os dados, foram submetidos a um processo de organização e submetidas a procedimentos de limpeza e tratamento, tais como, tratamento de valores ausentes e padronização das variáveis explicativas. Além da transposição do formato de apresentação dos dados para um modelo em painel, pois essa transposição aumentou o poder de predição do modelo.

O processo de modelagem preditiva foi aplicado o *Random Forest*. A escolha dessa abordagem se justifica pela abordagem apontadas na revisão empírica como no estudo apontado por McBride (2015) que utiliza este método e sua robustez frente à multicolinearidade, capacidade em lidar com muitas variáveis e captura relações não lineares como também tolerância a ruídos nos dados. A variável dependente foi o IFDM, e as variáveis independentes incluíram indicadores socioeconômicos, educacionais.

O modelo foi e dividido em conjuntos de treinamento os quais foram 70% de teste e 30% para avaliar a performance preditiva: R^2, o Erro Médio Absoluto *Mean Absolute Error* (MAE) e RMSE. Esses serviram para escolha de disposição das variáveis e acurácia da predição do modelo, como também seu grau de dispersão em relação aos valores reais observados (SANTOS, 2024). Além disso foi feita uma comparação entre modelos e *machine learning* e econométricos, como *Long Short Term memory* (LSTM) XGBosst e uma regressão linear múltipla. A regressão linear foi feita com o método padrão de Mínimos Quadrados Ordinários (MQO).

Ademais, o modelo LSTM é uma camada de neurônios artificiais que segundo o estudo de Santos (2022) tem maior capacidade de lidar com dados não lineares e a diversidade de dados como séries temporais e classificação dentro de uma mesma estimação. Outro modelo aplicado foi o XGBoost que se utiliza de árvores de decisão, mas de uma forma diferente aplicando ou método alternativo ao *bagging* o *boosting* que em geral tem maior acurácia pois incorpora o erro de cada árvores para um função de aprendizado *Loss function* (SANTOS, 2022).

No entanto, utilizou-se o modelo *Random Forest* como opção, pois diferente dos outros métodos, como por exemplo o XG*Boost* que embora apresente resultados levemente melhores, a suavização proporcionada pela média das árvores de decisão para o modelo *Random Forest* torna o modelo menos suscetível a hiper ajustes incorporando possíveis ruídos nos resultados, além disso essas características auxiliam na união da AEDE, a qual sua predição se mostra com melhores resultados para a correlação.

3.2 Análise Exploratória de Dados Espaciais

A definição de Análise exploratória de dados espaciais (AEDE) é descrita como um método para inferir a distribuição e associação de variáveis no espaço entre as unidades avaliadas, além de perceber padrões e formas de instabilidade espacial que pode se identificar possíveis *outliers*. Por conseguinte, também é possível identificar uma autocorrelação espacial por meio dessa exploração espacial (Almeida, 2012; Alves, 2020).

A autocorrelação espacial é definida como a auto comparação de um atributo em diferentes localizações. Ademais, a AEDE utiliza-se em grande parte de indicadores de intensidade como per capita ou a nível de área, pois melhora a comparabilidade de munícipios com tamanhos diferentes pode isolar o viés de escala. Para Alves (2020) a aplicação da AEDE utiliza-se a adoção de uma matriz de vizinhança que tem o objetivo de verificar a semelhança das variáveis no espaço e tem a capacidade de ponderar essa variável em diferentes localizações construindo e inferindo se há ou não correlação espacial. Dessa forma, aplica-se uma matriz n por n e segue-se um critério de contiguidade como *queen* ou *rook* e em alguns casos por distância geográfica.

A avaliação da matriz por contiguidade baseia-se na partilha de fronteiras por duas regiões vizinhas. Cria-se os mapas de acordo com sua interação no espaço que pode ser compartilhando apenas com uma aresta de vizinhança ou vários vértices que no geral essas matrizes fazem alusão ao movimento de peças do tabuleiro de xadrez, a convenção de

contiguidade pode ser rainha (*queen*), torre (*rook*) e/ou bispo (*bishop*). Há também a matriz que pondera pontos e distância no espaço chamado k-vizinhos próximos, assim pode-se definir k, como a quantidade de vizinhos mais próximos se optar por exemplo 5, 10 ou 15. Nesse sentido existem outras formas também de definir a matriz como por peso das distâncias, definindo uma distância que se utilizará e ponderando-a, ou por matriz do inverso das distâncias.

A estatística I de Moran ou Índice de Moran utiliza-se dessa matriz de pesos para fazer a inferência da correlação espacial global e existem também a possibilidade de se inferir informações locais por meio do LISA (*Local Indicators of Spatial Association*). O presente estudo utiliza-se do I de Moran e do LISA por serem bem difundidos na literatura[2] (Almeida, 2012). O I de Moran é indicador de correlação espacial de um atributo comparado entre regiões diferentes dentro de uma análise espacial. Tem seu índice calculado semelhante a correlação de Pearson com suas características a seguinte equação define o cálculo do Índice de Moran como:

$$I = \frac{\sum\sum w_{ij}(y_i - \hat{y})(y_j - \hat{y})}{\sum(y - \mu)^2} \quad (2)$$

Onde há a verificação se existe correlação de uma mesma variável em diferentes pontos no espaço que é ponderada pela matriz espacial w_{ij} da Equação 2 (Romero, 2006).

Com base nos Resultados do I de Moran optou-se por utilizar a matriz de vizinhança rainha, pois apresentou uma autocorrelação espacial maior na análise dos dados. Dessa forma, as variáveis que foram utilizadas para fazer as correlações espaciais univariadas foram, IFDM e o PIB per capita dos anos de 2013 e 2021. Já a análise bivariada foi aplicada para o ano de 2021, o qual é o mais recente com base nos dados utilizados, com a intenção de validar as correlações mais próximas a atualidade. Uma alternativa para visualização de autocorrelação espacial que é no diagrama de dispersão espacial ilustrado na Figura 3 a seguir:

[2] O I de Moran proposto em 1948 por Patrick A. P. Moran mede a correlação espacial, usando uma medida de auto covariância na forma de produto cruzado procurando testar a significância estatística e rejeitar a hipótese nula de que a correlação é um efeito estocástico. Utiliza-se o método para entendimento de formação de clusters espaciais e identificação de outliers (Alves, 2020). Cliff (1981) aponta outros modelos de correlação espacial como, Geary's C, *Join Count Statistics, Getis-Ord G e G*,* além dos modelos abordados por Anselim (1995) o I de Moran e o *Local Indicators of Spatial Association* (LISA) a diferença desses métodos é a abordagem matemática utilizada em cada um deles é melhor em determinadas ocasiões. Dessa forma, o estudo abordado utiliza o I de Moran e o Lisa, pois a intenção é localizar clusters significativos e os métodos tem interpretações mais claras e intuitivas o que se adequa melhor a este estudo.

Figura 3: Diagrama da representação da associação espacial, segundo quadrante

Fonte: Romero, 2006.

De acordo com o índice de Moran é possível mostrar uma dispersão no plano cartesiano, o qual decompõe a associação espacial em quatro quadrantes, cada quadrante da Figura 3 enfatiza a interação do local da variável e seus vizinhos. O quadrante I se localizam os dados que tem alto valor e seus vizinhos apresentam as mesmas características, no quadrante III é o inverso os quais apresentam baixo valor assim como seus vizinhos.

Destarte permanecer nos quadrantes I e III a estatística de Moran terá valor positivo e apresentará uma ideia de possível correlação espacial de semelhança nos vizinhos. No caso que os valores no diagrama prevaleçam nos quadrantes II e IV o valor do estatístico de Moran será negativo, indicando que as relações prevalecentes são aquelas onde os valores de um ponto alto encontra vizinhos com valores baixos o contrário também sendo possível (Alves, 2020).

Ademais, a AEDE deste estudo faz uso das do método de análise espacial citado, propondo os testes de autocorrelação espacial global e local de forma univariada e bivariada, em conjunto com a apresentação de mapas de significância, diagramas de dispersão e formação de *clusters* ou padrões de associação espacial local.

A matriz espacial do I de Moran Bivariado é representada por WZ_2 que é a defasagem da variável Z_2. Dessa forma a presença de autocorrelação espacial positiva indica uma associação dos valores das variáveis que está sendo estudada e de suas localizações. Sendo assim, a autocorrelação positiva mostra que municípios com uma alta (baixa) renda per capita são rodeados por municípios com um alto (baixo) infraestrutura, por exemplo. O coeficiente I de Moran bivariado é dado pela Equação 3 (Alves, 2020).

$$I^{Z1Z2} = \frac{n}{S0} \frac{\acute{z}_1 1 W Z_2}{\acute{z}_1 Z_1} \quad (3)$$

Posto isto, o estudo aplica a análise do I de Moran bivariado na variável IFDM utilizando o PIB per capita. A utilização do IFDM como variável Y e o PIB per capita com variável X mostraram-se relevantes como aplicação para o estudo de correlação espacial.

3.3 O IFDM predito e análises espaciais

Portanto, este estudo busca integrar o poder preditivo dos modelos de *machine learnin*g às inferências espaciais da AEDE, propondo uma predição baseada na variável IFDM. Essa abordagem contribui para reduzir problemas de vieses e inconsistências analíticas, resultando em estimativas mais robustas e confiáveis.

A integração entre *machine learning* e AEDE justifica-se pela necessidade de obter predições mais precisas do IFDM ao mesmo tempo em que se reconhece a importância da estrutura espacial no comportamento desse indicador. Enquanto o *machine learning* aprimora o desempenho preditivo ao captar padrões complexos nos dados, a AEDE permite identificar dependências e heterogeneidades espaciais que influenciam diretamente o desenvolvimento municipal. A combinação dessas abordagens reduz vieses analíticos, evita interpretações isoladas e contribui para estimativas mais consistentes e alinhadas à realidade territorial de Pernambuco (Barros, 2020).

3.4 Base de Dados

A base de dados utilizada e tratada é retirada do Atlas do Desenvolvimento Humano (ADH) em conjunto com os dados do IFDM do site do FIRJAN dos anos de 2013 a 2021, foram utilizados esses dados, pois são os dados mais recentes disponíveis do Atlas e também o índice de Firjan tem uma atualização municipal contínua e sem dados faltantes. Por conseguinte, a escolha desses dados tem o intuito de identificar quais variáveis mais explicam o IFDM no estado de Pernambuco nos últimos anos. Além disso, todos os métodos abordados nesse artigo foram implementados pela ferramenta *python* o motivo de uso desta ferramenta se deu pela facilidade de implementar técnicas de *machine learning* por meio dela.

Ademais, para explorar os dados e suas características, caso existam dados ausentes, executar o tratamento desses dados para melhor compreensão das estatísticas descritivas e

melhor resultados ao utilizar o modelo *Random Forest*. Dessa forma, para uma melhor visualização do IFDM e a inferência de hipóteses baseadas em estudos recentes da área como Simões (2019) foi feita a nível de mesorregião uma agregação dos dados dos municípios.

Aliado ao pré-processamento de dados também foi preciso para variáveis utilizadas na predição do modelo *Random Forest* são elas descritas na Tabela 1 a seguir.

Tabela 1: Descrição das variáveis utilizadas

Siglas	Descrição da variável
IDEB_FI	Índice que calcula o desempenho médio dos alunos no fundamental nos anos iniciais
IDEB_FF	Índice que calcula o desempenho médio dos alunos no fundamental nos anos finais
TX_F_LAB	Índice de acesso das escolas a laboratórios de informática no fundamenal
TX_M_LAB	Índice da taxa de acesso das escolas a laboratórios de informática no Ensino médio
TX_NET_F	Índice de acesso à internet no fundamental
TX_NET_M	Índice de acesso à internet no ensino médio
FORM_D_F	Índice da formação dos docents no fundamental de escolas no geral
FORM_D_M	Índice da formação dos docents no Ensino médio de escolas no geral
FORMPUB_D_F	Índice da formação dos docents no fundamental apenas na rede pública
FORMPUB_D_M	Índice da formação dos docents no Ensino médio apenas na rede pública
PIB_Per	PIB per capita
TX_MORTE_MUN	Índice do nível de mortalidade dos municípios
TX_Doença_Amb	Índice de doenças causadas por saneamento básico
TX_AT_PRIM	Índice de internações que poderiam ser resolvidas na atenção primaria e não foram
TX_ABS_ÁGUA	Índice da população com acesso abastecimento de água
TX_HOMIC	Índice de mortalidade por homicídios
IFDM	Índice de Firjan do desenvolvimento municipal

Fonte: Retirados do Firjan e Atlas do Desenvolvimento Humano, de 2013 a 2021.

As variáveis utilizadas como explicativas são todas retiradas do ADH e por se tratar de dados nível municipal verifica-se a ocorrência de dados ausentes em diversas dessas variáveis. Posto isto, a estatística descritiva dessas variáveis está no apêndice A para melhor compreensão.

Nesse sentido, foi aplicado a média dos anos anteriores em decorrência da falta de coleta ou lacunas censitárias em anos sequenciais. Além disso, a seção seguinte apresenta-se e conduz uma análise de sensibilidade com os dados para avaliar a robustez das predições e verificar o efeito nos dados e desempenho do modelo.

4 RESULTADOS

4.1 Análise descritiva

Esta seção explora a natureza dos dados, com o objetivo de proporcionar uma compreensão aprofundada por meio da utilização de estatísticas descritivas que possibilitem identificar o comportamento do índice de Firjan na média das mesorregiões.

Nesse sentido, Pernambuco é divido em 5 mesorregiões as quais são a Região metropolitana de Recife, que capital do estado a qual tem forte base em serviços, indústria, comércio, turismo e o Porto de Suape, a Mata Pernambucana a qual é uma zona da Mata úmida, antiga região canavieira, tem-se a região do Agreste Pernambucano região de transição entre a mata úmida e o sertão semiárido, a região do Região de transição entre a Mata úmida e o Sertão semiárido a qual é localizada no oeste do estado, às margens do Rio São Francisco e por fim o Sertão Pernambucano a qual é a maior mesorregião em área, clima semiárido, parte do Polígono das Secas.

Ademais, com base na estatística descritiva para auxiliar no entendimento melhor dos dados, foi aplicado médias que melhoram o entendimento da variável em questão, desvios padrões que melhoram o entendimento da dispersão e do nível de variação dos dados e medianas que diferente da média não é afetada por valores extremos proporcionando uma medida sem ruídos de tendência central, para melhor visualizações gerou-se um *boxplot* que dispõe de todas essas interpretações de forma visual dividindo-os em quantis os dados, além disso se fez agregações por mesorregião para entendimento da variável de forma geral nas regiões.

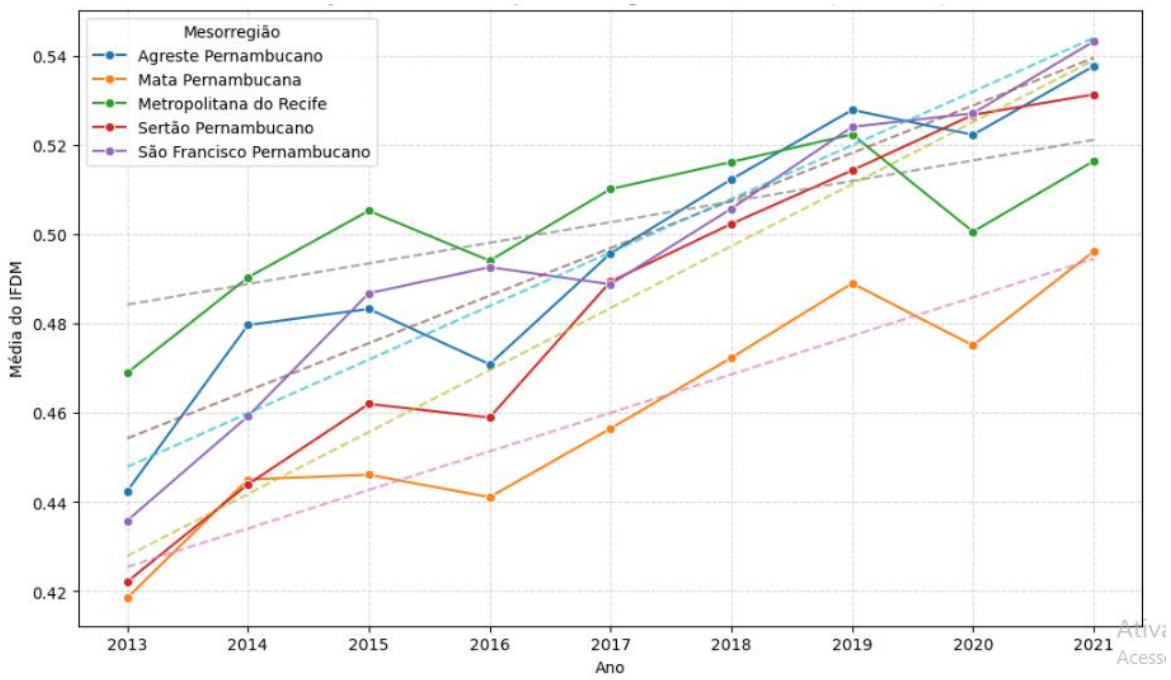
A Figura 4 apresenta a evolução da média do referido índice no período de 2013 a 2021, agregada por mesorregiões do estado de Pernambuco. Os pontos de 2014 a 2016 no Brasil foram anos de recessão Brasileira que afetam diversos estados Brasileiros o PIB brasileiro nesse período retraiu em torno de 8,6% (COLOMBO, 2017). Além disso o ano de 2020 foi marcado pela crise pandêmica que foi o corona vírus sendo assim esses dois períodos de crise no Brasil são vistos também no IFDM quando se analisa o valor médio do índice por mesorregião nestes anos.

Pode-se visualizar que a Mesorregião do Agreste composta segundo o Segundo Sobel (2009) pelas cidades de Caruaru, Garanhuns, Santa Cruz do Capibaribe, Toritama, Pesqueira, que são as cidades com maiores índices, é a mesorregião que na média tem o segundo melhor índice de desenvolvimento. Dessa forma pode-se inferir que essas regiões tendem a ter melhor

qualidade de vida que nas demais regiões seja em nível de educação, renda ou saúde ou até as três dimensões em conjunto.

Simões (2019) aponta que a explicação da região Agreste ter melhor índice de Desenvolvimento deve-se ao fato de que essa região assim como o Sertão entre os anos 2000 e 2010 tiveram políticas públicas voltadas ao desenvolvimento mais especificamente voltadas ao avanço de serviços básicos nesse caso pode-se atribuir a melhora do desenvolvimento dessas mesorregiões de Pernambuco a esses avanços de serviços básicos.

Figura 4: IFDM por mesorregião Pernambuco - 2013-2021



Fonte: Elaboração própria com base nos Dados do Atlas e Firjan.

Além disso, a região de acordo com os dados do IFDM o Sertão de Pernambuco composto pelas cidades de Serra Talhada, Salgueiro, Arcoverde, Afogados da Ingazeira, Ouricuri, entre outras, no ano de 2019 passou a ser a terceira melhor na média de desenvolvimento passando a Região Metropolitana de Recife (RMR) que é composta por 14 municípios como Recife, Olinda, Jaboatão dos Guararapes, Paulista, Cabo de Santo Agostinho, Igarassu, Abreu e Lima, Araçoiaba, Ilha de Itamaracá, São Lourenço da Mata, Moreno, Camaragibe.

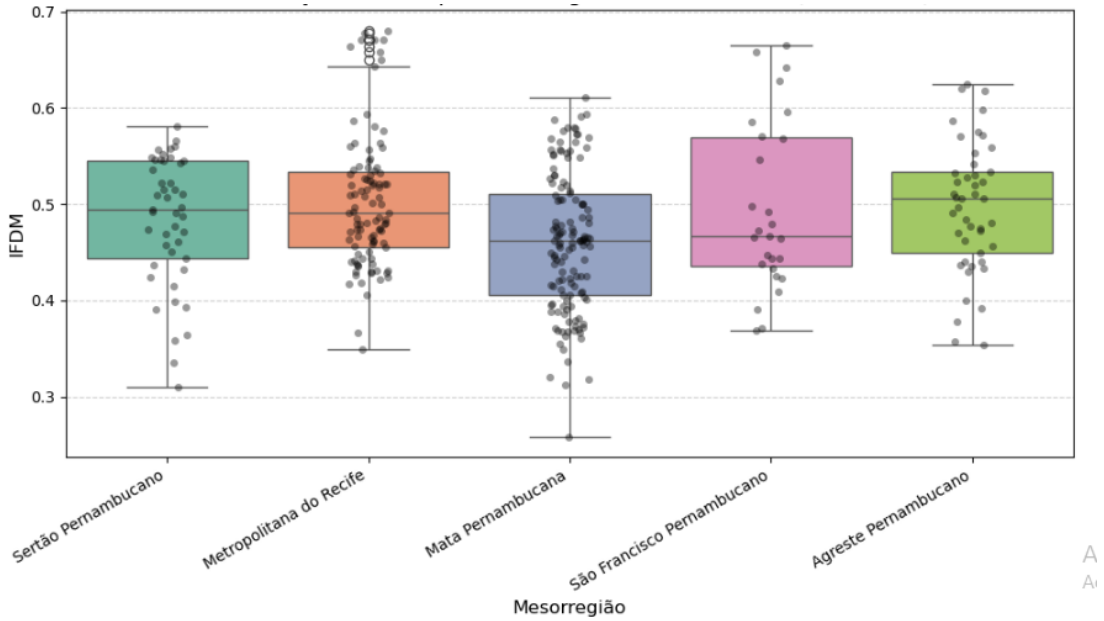
O melhor índice em 2021 é a mesorregião do São Francisco composta, de acordo com Simões 2019, pelas cidades de Petrolina Cabrobó, Belém do São Francisco, que tem os maiores índices. Destarte, corroborando como o estudo de Simões (2019), parte dessa melhora no indicador capitada pela média da evolução do IFDM ao longo dos anos pode-se atribuir aos

investimentos de políticas públicas de educação básica, aumento de renda para os mais pobres e vulneráveis além de infraestrutura básica como o projeto luz para todos (SIMÕES, 2019)

Outro aspecto para se entender melhor os dados é verificar a divisão da mediana se há *outliers*, que se distanciam muito do padrão dos dados do grupo. A figura 5 a seguir exemplifica esses dados através de uma *boxplot* para visualização da distribuição dos dados do IFDM para mesorregiões de 2013 a 2019.

A Figura 5 apresenta um *boxplot* e mostra que há poucos *outliers* com dispersão positiva por mesorregião, com exceção da Metropolitana de Recife, Olinda e Cabo de Santo Agostinho onde são *outliers* positivos, mas existem *outliers* com dispersão negativa considerável principalmente na região metropolitana de Recife como Araçoiaba, São Lorenço da Mata e Moreno. No estudo de Simões 2019 o Nordeste teve uma melhora significativa de seu índice de desenvolvimento humano, no entanto no desenvolvimento proporcional do índice entre as regiões do Brasil a região Nordeste é a que menos cresceu se comparado com a outras regiões (Simões, 2019).

Figura 5: Distribuição do IFDM por Mesorregião de Pernambuco – 2013-2021



Fonte: Elaboração própria.

O baixo crescimento da região no geral pode ser uma possível explicação para esses municípios de baixo IFDM impactados pelas variáveis que serão apresentadas na seção em sequência. A visualização mostra que os dados estão em sua maioria dentro do padrão de dispersão e a seguir algumas estatísticas para entender melhor com dados numéricos do índice

agregado. A Tabela 2 apresenta a mediana, média e desvio padrão, da média geral do IFDM das mesorregiões. Pode-se notar que a Região Metropolitana de Recife teve a maior média, contudo após 2019 teve seu índice passado pelo Sertão de Pernambuco no ano de 2019.

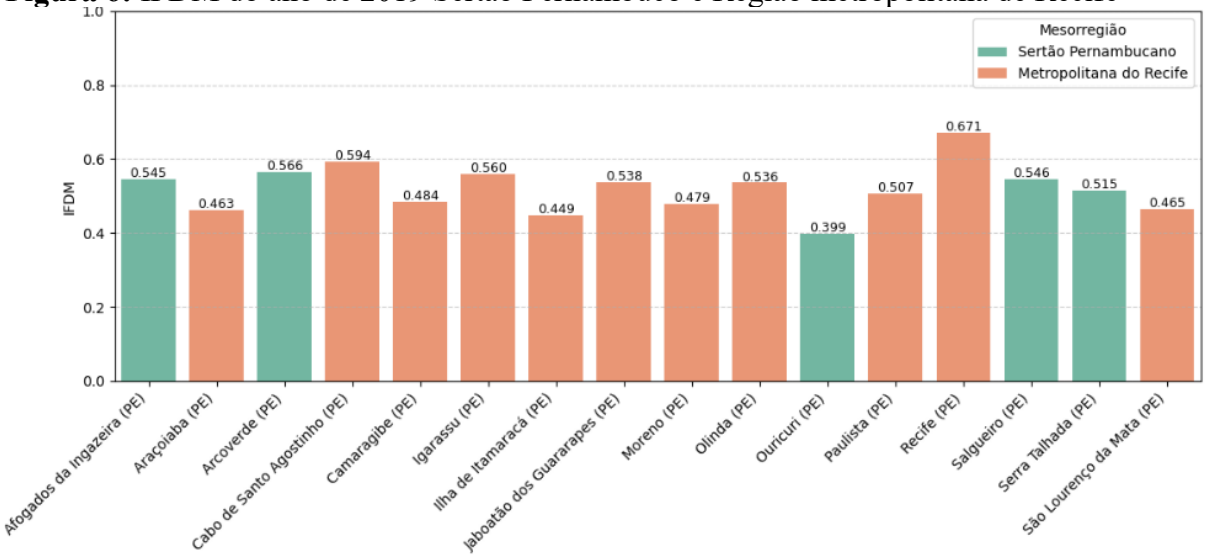
Tabela 2: Estatística descritiva IFDM, média geral (2013 - 2021)

Mesorregião	Média	Desvio Padrão	Mediana
Agreste Pernambucano	0,496	0,067	0,505
Mata Pernambucana	0,460	0,071	0,462
Metropolitana do Recife	0,502	0,069	0,490
Sertão Pernambucano	0,483	0,067	0,494
São Francisco Pernambucano	0,495	0,088	0,466

Fonte: Elaboração própria.

Dessa forma embora Recife seja um município que o valor de seu IFDM em 0.67 está acima da média de 0.502 os municípios da sua mesorregião estão abaixo da média. Por fim, listou-se com o fito de contribuição descritiva um *ranking* de melhores municípios com maiores IFDMs e acima da média de 0.59 mencionada como uma melhora da região Nordeste pelo estudo de Simões tem-se respectivamente em ordem decrescente os municípios de Fernando de Noronha, Recife, Petrolina, Ipojuca, Caruaru, Rio Formoso e Primavera. A seguir temos a Figura 6 que nos ilustra os municípios da Mesorregião Metropolitana de Recife e Sertão de Pernambuco para o ano de 2019 que ilustra os valores de IFDMs e nos ajuda a entender as médias do IFDM.

Figura 6: IFDM do ano de 2019 Sertão Pernambuco e Região metropolitana de Recife



Fonte: Elaboração própria.

A seção seguinte apresenta o treinamento do modelo com o objetivo de prever o IFDM para o ano de 2021 e compará-lo aos valores observados. Essa proposta contribui para a literatura ao reconhecer não apenas o espaço crescente que o *machine learning* vem conquistando na economia, mas também sua utilidade para antecipar indicadores relevantes para a gestão pública, como o IFDM, auxiliando na avaliação do desenvolvimento municipal. Além disso, a predição mostra-se particularmente justificável diante da demora na disponibilização de dados estatísticos pelos órgãos de fomento nas esferas municipal, estadual e federal, reforçando a importância de métodos capazes de fornecer estimativas atualizadas e confiáveis.

4.2 Treinando modelo e ajustando estimativas do *Random Forest*

O modelo utilizado foi treinado com uma divisão de 70% dos dados coletados para treino e 30% para teste. Essa divisão comumente utilizada na literatura, outro parâmetro ajustado é a quantidade de árvores que o modelo utiliza fixada em 700 árvores, a literatura do modelo aponta que acima de 500 árvores é ideal para aplicações do *Random Forest* (DURÃES, p- 780, 2022). Além dessa divisão, foi utilizado um padrão para aleatóriedade do treino, frequentemente utilizado em algoritmos de *machine learn* o *Random state* igual a 10 com intuito de deixar os resultados replicáveis. Nesse sentindo, o estudo aplica um teste de robustez do modelo com o método de *cross-validation que* segundo Cunha (2019) é um bom método para se calcular risco esperado e verificar as estimativas preditas.

Os resultados da Tabela 3 indicam que o modelo Random Forest apresentou desempenho satisfatório tanto na etapa de treinamento quanto na etapa de teste. O erro absoluto médio (MAE) de 0,009 no treino e 0,025 no teste, assim como o erro quadrático médio (RMSE) de 0,011 e 0,033, respectivamente, demonstram que os valores preditos se aproximam bem dos valores observados, caracterizando boa precisão preditiva e boa aplicabilidade para prever o IFDM de Pernambuco.

Tabela 3: Resultados do Random Forest

Treinamento	**Teste**
MAE: 0.009	MAE: 0.025
RMSE: 0.011	RMSE: 0.033
R^2: 0.975	R^2: 0.787

Fonte: Elaboração própria com base nos resultados.

O principal ponto a se notar é o decaimento de explicabilidade do modelo comum R^2 de 97% no treino e cai próximo a 79% no teste é esperado que modelos como Random Forest caiam ao fazer um teste para dados generalizados, mas mostra que o modelo de treino se ajustou a ruídos que pode estar ligado as variáveis de controle, embora não foi uma queda muito expressiva apresentado na Tabela 2.

A Tabela 4 a seguir apresenta os resultados obtidos pela técnica do *Cross-Validation* pode-se notar que o teste é feito utilizando toda a base de dados como o método e ainda apresentam uma boa generalização do modelo *Random Forest* em média com 80 por cento de acurácia. Além disso nesse método a média de erro das predições ainda se encontram próximas a 0.031.

Tabela 4: *Cross-Validation* K-fold 10

K folds	1	2	3	4	5	6	7	8	9	10
R^2:	0.760	0.836	0.758	0.737	0.829	0.849	0.847	0.826	0.843	0.797
RMSE	0.033	0.031	0.033	0.031	0.029	0.029	0.031	0.032	0.027	0.029
Média R^2:	0.808	Média RMSE:	0.031							

Fonte: Elaboração própria com base nos resultados.

Nesse sentido pode-se inferir a princípio que o modelo *Random Forest* é um bom modelo para predição do IFDM corroborando com os estudos de Tobaigy, Alamoudi e Bafail, (2023) feito com dados de satélite para predição do Índice. Além disso, os modelos de aprendizado de máquina em geral demandam quantidades de dados exorbitantes os chamados *Big Datas,* no entanto seguindo os estudos não Tobaigy, Alamoudi e Bafail, (2023) e Santos (2024) que afirmam que a técnica de *Random Forest* é não paramétrica e robusto a ruídos em comparação com Técnicas de *deep learning* como redes neurais ou até mesmo comparado ao XG*Boost.* Desse modo, o modelo não necessita de um base da dados longa para ser efetivo corroborando com os estudos os resultados apontam que é um bom modelo para a predição do Índice de Desenvolvimento Humano.

4.3 Comparação dos modelos

Utilizou-se outros modelos com a intenção de comparar o modelo *Random Forest* com os resultados de outros métodos de *machine learning*. A comparação servirá como base e compara-se não só o poder de predição como também a eficiência computacional e o sobreajuste dos outros modelos. Ademais, os resultados de uma regressão linear múltipla aplicada sobre a mesma variável tem os resultados apresentados na Tabela 5 a seguir:

Tabela 5: Resultados Regressão linear múltipla

Estatísticas	**Valor**
R-squared:	0.522
Adj. R-squared:	0.518
F-statistic:	112.5
Prob (F-statistic):	0.000
Nº de observações:	1.664
Df Model:	16
Df Residuals:	1647

Fonte: Elaboração própria com base nos resultados.

Assim como no modelo *Random Forest* a regressão linear teve variáveis como PIB per capita, IDEB do ensino fundamental, Taxa bruta de mortalidade significativos com p valor abaixo de 5% de confiança. No entanto, a regressão linear obteve uma explicação pior frente a

Random Forest com um R^2 de 0,52 o que significa que as variáveis explicam 52% do resultado do IFDM, contudo essa comparação e apenas para predição do modelo existem outras métricas quando se considera eficiência e propósitos do modelo que não foram consideradas com relação a este estudo. Aplicou-se também um modelo de rede neural *Long Short Term Memory* (LSTM) que é uma variação robusta dos modelos de rede neurais recorrentes, pois sua caracterísca é carregar os dados em uma memória de longo prazo em conjunto com análise de curto prazo. A Tabela 6 a seguir apresenta-se os dados.

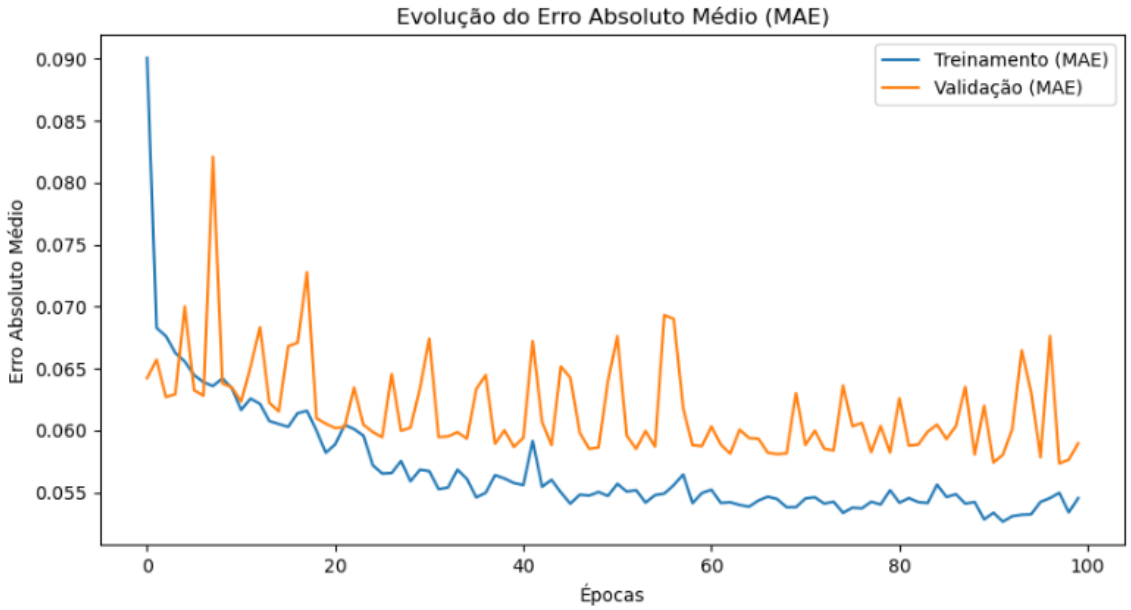
Tabela 6: Resultados *Long Short Term Memory*

Estatísticas	**Valores**
R^2:	0.655
RMSE	0.066
MAE	0.053

Fonte: Elaboração própria com base nos resultados.

Dessa forma, o modelo LSTM apresentou um resultado de 65% de explicação com os mesmos dados utilizados no *Random Forest* inferior ao método de árvores, mas melhor que a regressão linear múltipla. Ademais foi utilizado 100 *epochs* para treinar o modelo como padrão de análise utilizada para esse modelo, o resultado pode ser modificado se existir uma base de dados maior, ideal para o LSTM, e aumentando o número de *epochs* com um poder computacional maior. A seguir na Figura 7 uma ilustração da evolução do erro do modelo.

Figura 7: Resultados do LSTM



Fonte: Elaboração própria com base nos resultados.

Além disso implementou-se mais um modelo para comparação nesse sentido, o modelo aplicado foi o XG*Bosst* que semelhante ao Random Forest trabalha com as árvores de decisões como parâmetros de decisão do método, mas diferentemente o XG*Bosst* utiliza as árvores como ajuste isso permite uma maior acurácia dos dados, mas pode gerar maior *overfiting*.

Tabela 7: Resultados do XG*Boost*

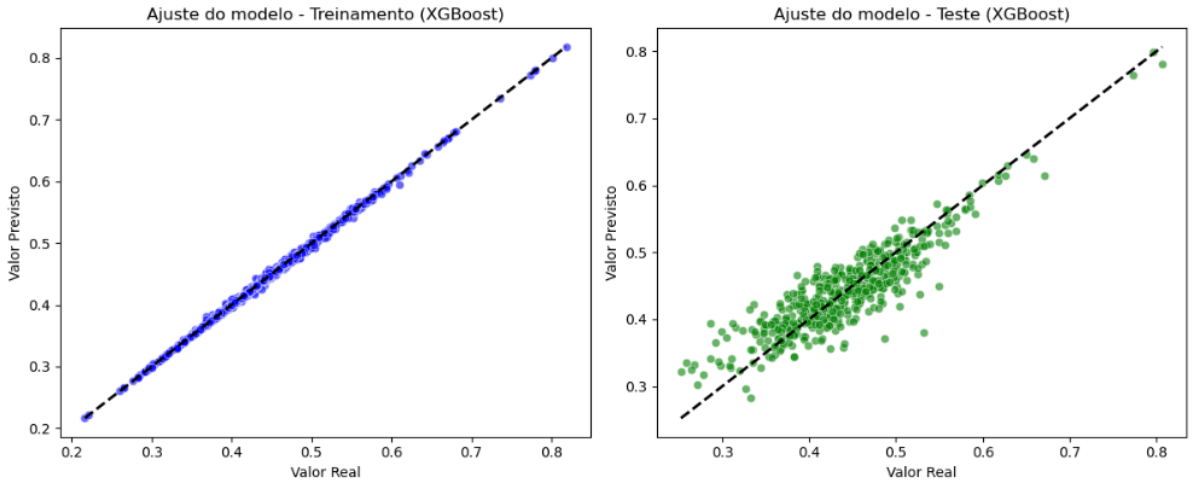
Treinamento	**Teste**
MAE: 0.001	MAE: 0.024
RMSE: 0.002	RMSE: 0.032
R^2: 0.998	R^2: 0.800

Fonte: Elaboração própria com base nos resultados.

Desta maneira, os resultados do *XGBosst* apresentam um acuraria levemente melhor que o Random Forest utilizado como demonstra a Figura 8, no entanto o método *XGBosst* demanda u maior poder computacional e necessita de maior controle e ajustes de seus hiperparâmetros, que são os controles do método para análise dos dados aplicou-se os hiperparâmetros padrão da

documentação, além disso o *XGBoost* por ajustar o erro baseado nas árvores anteriores corre maior risco de *overfitng* caso não seja ajustado corretamente.

Figura 8: Ajuste do modelo *XGBoost no* Treino e Teste

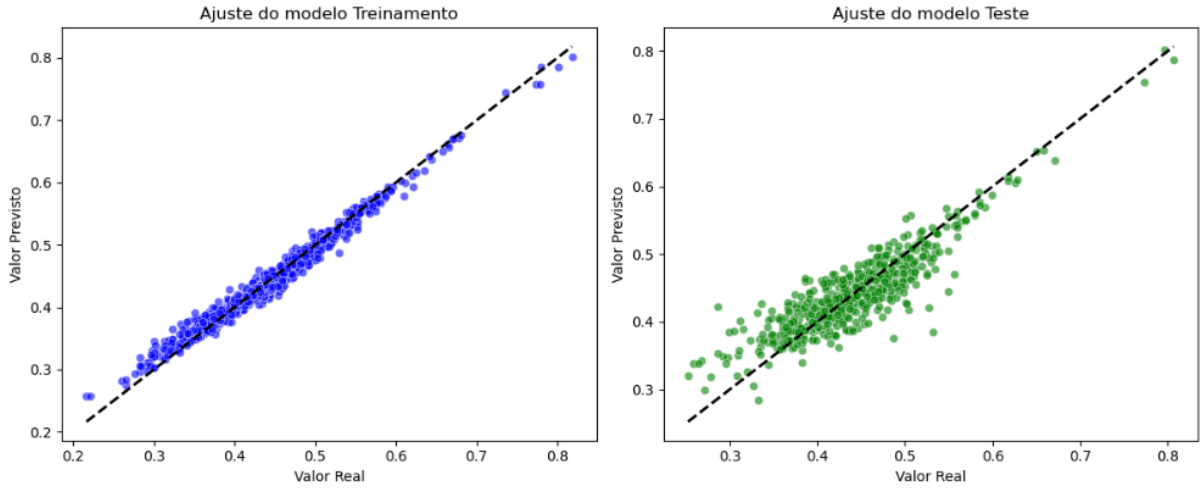


Fonte: Elaboração própria com base nos resultados.

O modelo XG*Boost* frente ao modelo *Random Forest* possui resultados levemente melhores contudo a característica de se ajustar aos dados baseados em erros pode incorporar ruídos a variável predita. Nesse sentido, para o objetivo do estudo não se mostrou um bom modelo para suavizar as análises da AEDE com os dados analisados.

Na Figura 9 estão os gráficos de como o modelo comportou-se no treino e como se comportou no teste, de forma visual, que aponta um certo superajuste. Desta maneira, o modelo apresenta leve *overfiting* no treino e a generalização dos dados de teste perdem parte da capacidade de explicação. Pode-se atribuir esse resultado às várias variáveis utilizadas e a poucos dados de coleta dessas variáveis como por exemplo o IDEB com lacunas vazias de coleta de dados para alguns municípios.

Figura 9: Ajuste do modelo *Random Forest no* Treino e Teste



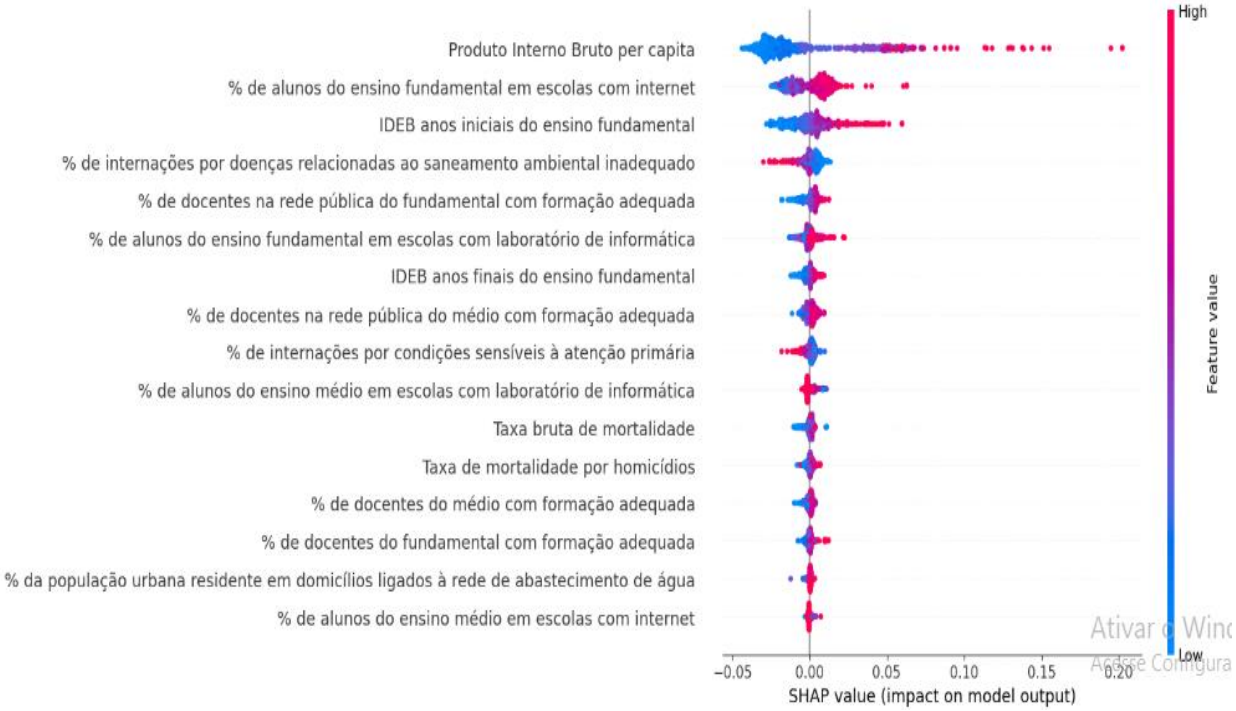
Fonte: Elaboração própria com base nos resultados.

Fez-se a aplicação de *Cross-Validation* kfold para verificar a robustez do modelo. Segundo Cunha 2019 o kfold divide os dados em k partes e utiliza-se k-1 fold para treino e testa-se com a parte restante, é feita a divisão de toda a base de dados repete-se o treino e teste k vezes em cada divisão. Dessa forma cada ponto de divisão dos dados é utilizado para teste e verificar a estabilidade do modelo (CUNHA, 2019).

No entanto ao se comparar com o teste de Cross-validation que tem uma média de R^2 de 81% muito próximo ao modelo gerado com menor valor 76% e maior de 84% tem-se que apesar do leve *overfiting* é um bom modelo para a proposta de predição do IFDM além disso o Erro quadrático médio e o erro absoluto médio no kfold estão quase que similares ao modelo gerado indicando, que mesmo moderadamente superajustado o modelo tem boa generalização. Ademais, para verificar o efeito das variáveis foi utilizado.

Ademais, para verificar o efeito das variáveis foi utilizado a Figura 10 que representa um gráfico violino para verificar o impacto das variáveis do modelo. De acordo com Trevisan (2022), o eixo horizontal representa um valor SHAP, enquanto a cor do ponto nos mostra se aquela observação tem um valor maior ou menor, quando comparada a outras observações. As cores na Figura 10 representam os valores médios das características naquela posição. As regiões vermelhas indicam que a maioria dos valores são altos, enquanto as regiões azuis indicam que a maioria dos valores são baixos (Bravaresto, 2024).

Figura 10: Impacto da *features* no modelo *Random Forest*



Fonte: Elaboração própria com base nos resultados.

Os dados apresentam uma noção das variáveis que tem mais correlação com o IFDM, de acordo com a seleção do *Random Forest*, dos municípios de Pernambuco e o grau de impacto dessas variáveis usa-se os métodos SHAP *e feature importance*, que são ferramentas para averiguar como as variáveis impactam no modelo. Nota-se que o Produto interno bruto per capita tem maior impacto, a porcentagem de acesso à internet de escolas do fundamental e a nota do desempenho do IDEB para o fundamental.

Figura 11: Importância das Variáveis

Fonte: Elaboração própria com base nos resultados.

Por ordem de impacto da correlação respectivamente as variáveis que mais se destacam no modelo são: PIB per capita, taxa de escolas no fundamental com acesso à internet, o nível de notas do IDEB do fundamental, a taxa de doenças por saneamento ambiental básico, taxa de escolas do fundamental com acesso à laboratórios de informática. Avalia-se o PIB per capita com correlação forte e positiva, assim como Taxa de escolas do fundamental com acesso à internet, IDEB do fundamental e a taxa de escolas do fundamental com laboratórios de informática. Já variáveis como taxa de doença por falta de saneamento ambiental têm uma correlação forte e negativa no modelo, que se espera algo relacionado, pois índices maiores de internações por doenças de saneamento básico podem estar ligados com uma infraestrutura pior na cidade.

Avalia-se possíveis causas dessas variáveis afetarem mais o modelo. Segundo Cunha (2019) há indícios que melhorar a renda de uma região assim como seu nível de educação proporciona uma melhora no índice de desenvolvimento humano. Dessa maneira as variáveis que estão se destacando tem relação com esse impacto já que o PIB per capita é uma boa proxy para renda na literatura econômica e de acordo com o modelo escolas com acesso à internet assim como o desempenho desses alunos apresentam impactos significativos.

Os dados apontam algumas variáveis como o Produto interno Bruto per capita como o mais influente no IFDM com alto impacto no modelo, mas isso não significa que as outras variáveis não são importantes, pois elas servem como controle do modelo e para interações agregando valor ao combiná-las com outras. Contudo, ao aplicar esses métodos temos a visualização de variáveis importantes que podem ser alvo de políticas públicas com estudos da forma correta, pois não é preciso aprofundamento no estudo de cada variável e importância não

significa causalidade. Desse modo, será utilizado o PIB per capita como proxy para o crescimento dos municípios e análise conjunta com o IFDM mais a frente.

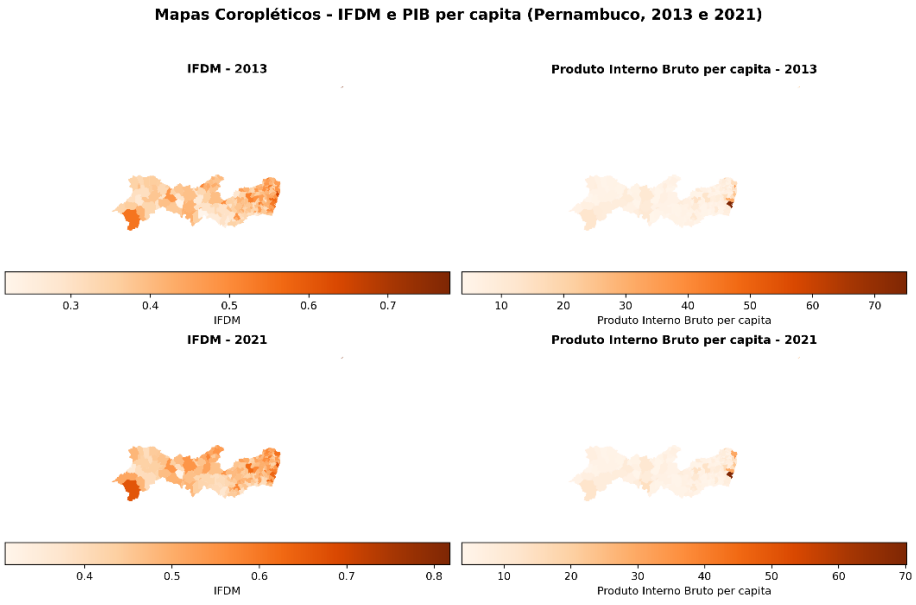
Destarte, é possível notar que o nível de renda do município tem uma forte correlação no Índice de desenvolvimento humano e também o nível de educação base, o acesso à informação através da internet. Além disso, saneamento básico tem um efeito, de acordo com os dados, é negativo, pois aponta que quanto mais internações provenientes de doenças por falta de saneamento ambiental impacta no modelo negativamente.

A subseção seguinte apresenta a aplicação da AEDE para o IFDM e o PIB per capita nos anos de 2013 e 2021, explorando seus padrões de distribuição espacial e possíveis aglomerações. Já na subseção 4.5, realiza-se uma comparação entre os resultados da AEDE para o IFDM observado em 2021 e aqueles obtidos a partir do IFDM predito pelo modelo Random Forest, permitindo avaliar a consistência espacial das estimativas geradas pelo modelo.

4.4 Análise Exploratória de Dados Espaciais

Esta subseção tem o objetivo de avaliar por meio da AEDE o IFDM no comparativo com os índices de 2013 e 2021, bem como a interação do PIB per capita com crescimento econômico. Em primeiro lugar avalia-se se há algum indício de correlação espacial A Figura 12 a seguir ilustra em aspectos espaciais as duas variáveis para os dois anos em medida no mapa de calor, evidenciando alto IFDM em regiões como Petrolina e nas mesorregiões Metropolitana do Recife e Mata Pernambucana em 2013.

Figura 12: Análise Espacial do IFDM e PIB per capita - 2013 e 2021



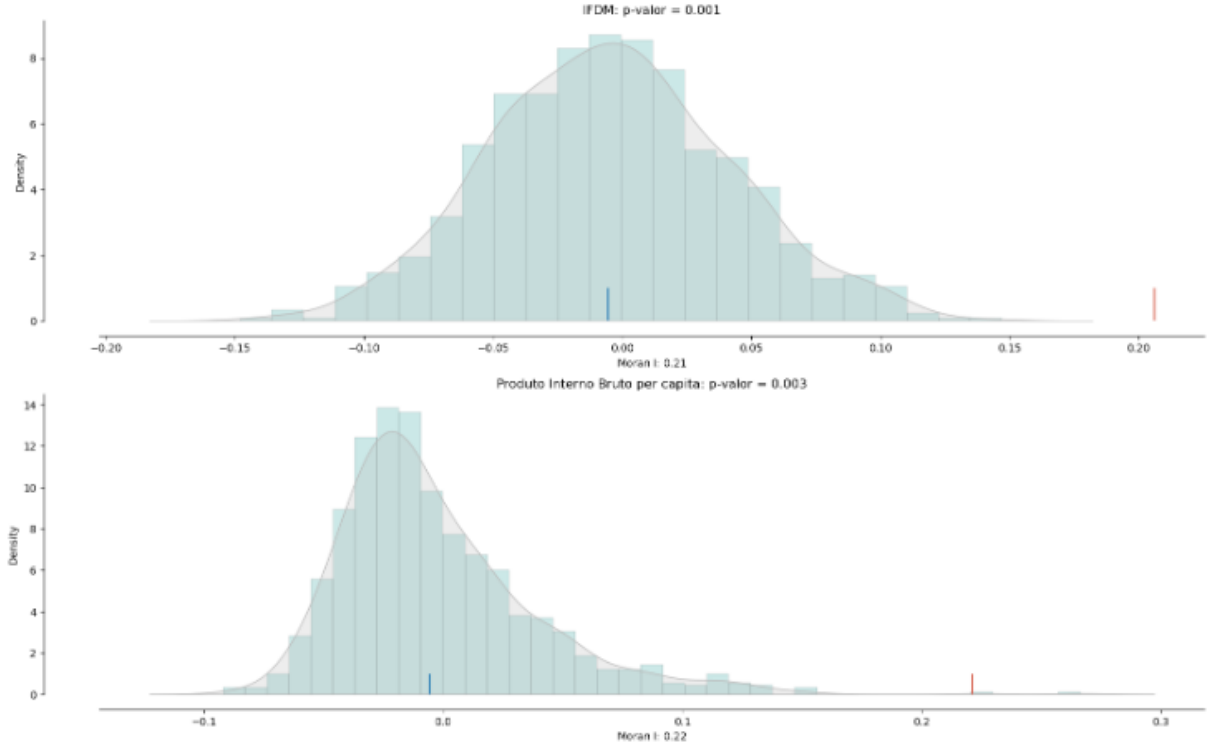
Fonte: Elaboração própria com base nos resultados.

Em 2021, o IFDM apresentou avanços significativos tanto no Sertão pernambucano quanto no Agreste, refletindo melhorias importantes nas áreas de emprego, renda, saúde e educação nesses territórios. No litoral Norte do estado, o destaque ficou para o município de Goiana, cujo desempenho foi impulsionado, entre outros fatores, pelo crescimento do PIB per capita. Goiana vem consolidando-se como um polo industrial e de serviços, beneficiado pela presença de grandes empreendimentos e por sua localização estratégica próxima às fronteiras com a Paraíba e à Região Metropolitana do Recife. Esse dinamismo econômico contribuiu para que o município se sobressaísse no cenário estadual em 2021 (Calife, 2021).

No estudo dos Mapas utilizou-se a matriz do tipo Queen, pois é a matriz que se mostra com melhores resultados quando se comparado aos outros tipos e também porque ela se configura ideal para os dados pela relação de contiguidade. Pode-se notar que a Região Metropolitana de Recife tem um grau mais forte de IFDM como aponta a Figura 12 assim como Petrolina e Serra Talhada. Além disso, na parte do PIB per capita não temos fortes níveis desse indicador com exceção de algumas cidades na Região Metropolitana de Recife. A seguir será feito uma análise detalhada espacial para extrair informações e se há significância estatística utilizando o índice de Moran.

A Figura 13 a seguir fez-se uma análise de distribuição da confiança estatística do I de Moran com relação ao IFDM e PIB per capita de correlação espacial com significância estatística a 95% de confiança. Desta forma, é possível afirmar que há correlação espacial dessas variáveis. A partir da análise espacial, infere-se que municípios com baixo nível de desenvolvimento tendem a replicar esse padrão em sua vizinhança, disseminando efeitos negativos para os municípios adjacentes. De modo análogo, áreas com alto desenvolvimento exercem influência positiva sobre seus entornos, elevando os índices dos municípios vizinhos. Esse comportamento espacial reforça a evidência de dependência espacial apresentada por Cunha (2019).

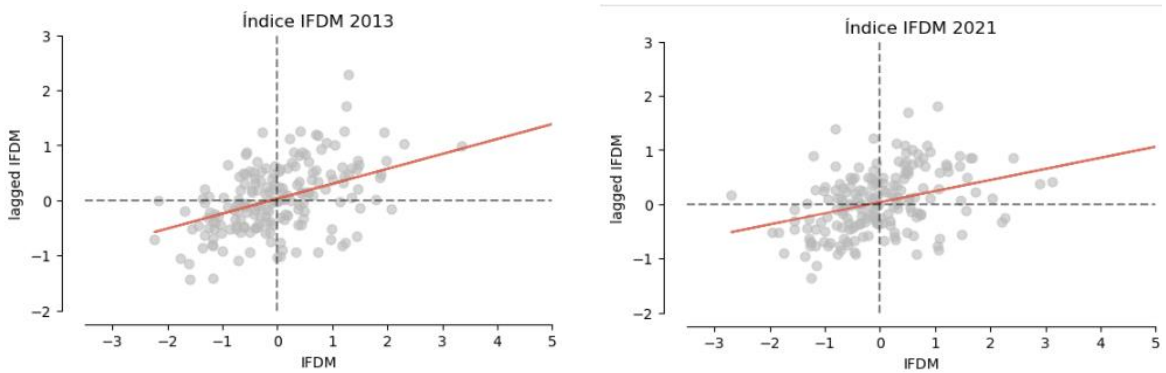
Figura 13: Distribuição de Probabilidade Índice de Moran IFDM e PIB per capita



Fonte: Elaboração própria com base nos resultados.

Para ilustrar esse comportamento, tem-se o mapa de dispersão de Moran Global, o qual evidencia a presença de autocorrelação espacial positiva. A Figura 14 a seguir mostra os quadrantes do diagrama permitem identificar padrões de associação espacial, como municípios de alto desenvolvimento cercados por áreas igualmente desenvolvidas (*High-High*) e municípios de baixo desenvolvimento vizinhos a áreas semelhantes (*Low-Low*). Esse padrão é sintetizado no gráfico global exibido na a seguir. A Figura 14 a apresenta o mapa de dispersão para o IFDM em 2013 e 2021.

Figura 14: Dispersão de I de Moran Global



(A) IFDM - 2013

(B) IFDM - 2021

Fonte: Elaboração própria com base nos resultados.

Ao verificar a Tabela 8 a seguir temos fortes indícios que há uma correlação espacial para as variáveis encontradas. Rejeita-se a hipótese nula de que a correlação espacial dessas variáveis é aleatória a 1 % de significância com IFDM de 2013 com correlação espacial de 0.18 e PIB per capita de 0.32, já para 2021 temos respectivamente Correlações de 0.24 e 0.23. A seguir as Figuras 15 e 16 apresentam uma Análise de dependência espacial do PIB per capita e IFDM para os anos 2013 e 2021.

Tabela 8: Índice Global de Moran 2013 e 2021

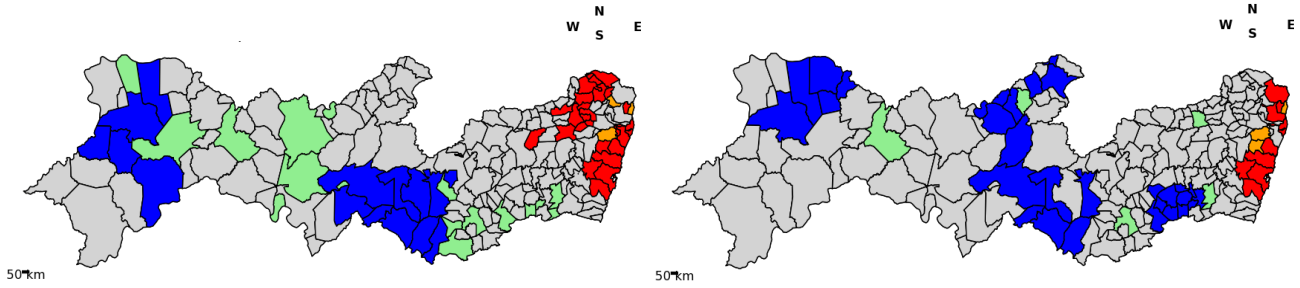
Variáveis	**Ano**	**Índice de Moran**	**p-valor:**
IFDM	2013	0.178	0.001
PIB per capita		0.321	0.001
IFDM	2021	0.243	0.001
PIB per capita		0.230	0.002

Fonte: Elaboração própria com base nos resultados.

Segundo Almeida (2012), no LISA é feito a estatística através de uma avaliação de uma série de valores obtidos por meio de permutações dos valores atribuídos aos vizinhos, o número de permutações pode ser definido por quem aplica a técnica, o intuito é gerar uma distribuição normal e rejeitar a hipótese da aleatoriedade espacial. Dessa forma, é gerado correlações locais nos municípios e seus vizinhos e os locais com significância pintados no mapa podem ser definidos com dinâmica espacial própria e merecem atenção individual em estudos futuros (Anselin, 1988; Alves, 2020).

A Figura 15 a seguir, apresenta o LISA referente ao PIB per capita e IFDM de 2013, revela que alguns pontos destoantes em relação ao IFDM que surgem nos quadrantes Alto-Baixo e Baixo-Alto. Os municípios pintados configuram potenciais *clusters* espaciais, apresentando níveis de desenvolvimento distintos de seus vizinhos imediatos. Embora esse comportamento sugira dinâmicas locais específicas. No caso do PIB per capita, há *cluster* Baixo-Baixo fortes nas mesorregiões do agreste e sertão de Pernambucano. Além disso cidades como Salgueiro apresenta uma característica de Alto-Baixo o que sugere que os seus vizinhos, embora com baixo PIB não afeta a cidade que tem alto PIB per capita para entender os motivos dessa característica é preciso estudos focados e aprofundados que não está no escopo deste estudo. Os quadrantes estão divididos em cores vermelho para Alto-Alto, azul para Baixo-Baixo laranja para Alto-Baixo e por fim verde para Baixo-Alto.

Figura 15: Análise de Dependência Espacial Local (LISA) do IFDM e do PIB per capita nos Municípios de Pernambuco em 2013



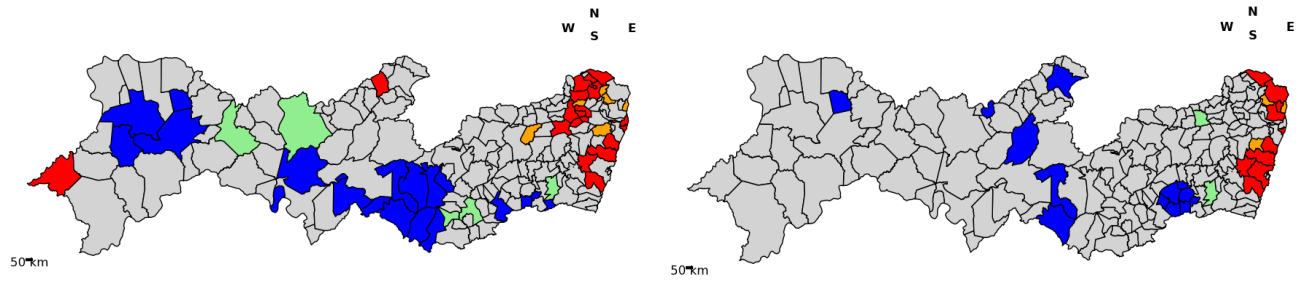
(A) IFDM - 2013 (B) PIB Per Capita 2013

Fonte: Elaboração própria com base nos resultados.

Por outro lado e com fortes *clusters* espaciais temos o IFDM apresentando formações de subdesenvolvimento nas regiões do Agreste e sertão assim como na análise do PIB per capita, mas também com regiões do São Francisco com alguns municípios próximos que parecem não serem afetados como Parnamirim, Salgueiro, Serra Talhada, Floresta e outros. Vale destacar que existem *clusters* Alto-Alto concentrados na Mesorregião Metropolitana de Recife e na Mata Pernambucana.

A Figura 16, correspondente a análise espacial do IFDM e PIB per capita de 2021. Para o ano de 2021 no que se trata do PIB tem-se uma correlação espacial menor e com menos áreas consideradas significativas a exceção da região Metropolitana de Recife que se mostra um *cluster* espacial Alto-Alto com leve aumento em relação a 2013. Para o IFDM vemos uma alteração similar contudo surgiu um *cluster* espacial positivo na cidade de Afrânio que não demonstrou nem uma significância na análise do ano de 2013 indicando uma possível mudança espacial dentro desses anos.

Figura 16: Análise de Dependência Espacial Local (LISA) do IFDM e do PIB per capita nos Municípios de Pernambuco em 2021



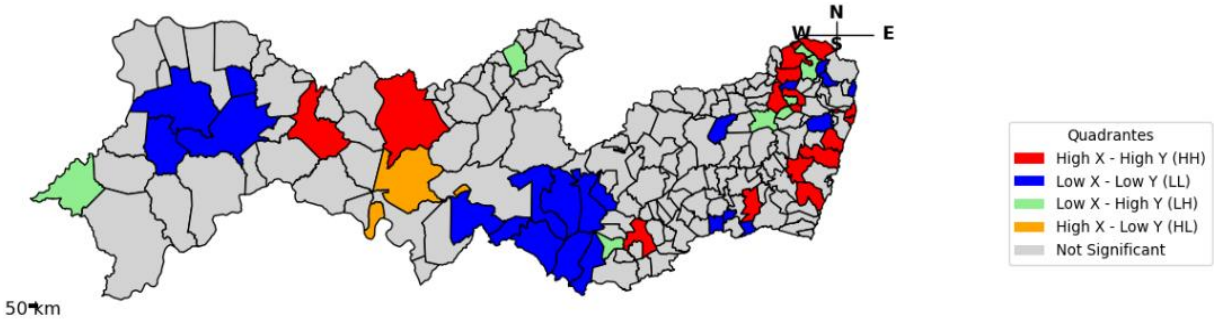
(A) IFDM - 2021 (B) PIB Per Capita 2021

Fonte: Elaboração própria com base nos resultados.

A transição dos anos de 2013 a 2021 apresentam algumas mudanças visuais aparentes. Os *clusters* de regiões com Baixo- Baixo diminuíram utilizando de base o estudo de Cunha (2019) parte dessa evolução se deu possivelmente as políticas públicas na região tanto em investimento de educação quanto programas sociais, como interiorização do ensino superior entre esse período e intervenções em termos de infraestrutura provenientes do governo federal e estadual. Em contrapartida a região metropolitana de recife ampliou *clusters* do tipo Alto-Alto corroborando um possível transbordamento.

Ademais, a Figura 17 apresenta-se a análise do LISA bivariado entre o IFDM e o PIB per capita, em que o IFDM foi definido como variável Y e o PIB per capita como variável X. A partir dessa relação, observa-se que os pontos Alto-Alto indicam *clusters* de desenvolvimento econômico, nos quais altos valores locais do PIB per capita estão associados a elevados índices de IFDM em municípios vizinhos. Em contraste, os pontos Baixo-Baixo evidenciam áreas de vulnerabilidade socioeconômica, caracterizadas pela presença de baixo IFDM local associado a baixos níveis de PIB per capita em seu entorno (ALVES, 2020).

Figura 17: Análise de dependência espacial local (LISA) Bivariado do IFDM e PIB per capita dos municípios de Pernambuco 2021



Fonte: Elaboração própria com base nos resultados.

A análise espacial bivariada se mostra importante para ver a dinâmica dessas duas variáveis IFDM e PIB per capita. Posto isto, o mapa apresenta alguns pontos com correlações espaciais significativas alguns *clusters* espaciais com Baixo-Alto um exemplo é o município de Afrânio, que, apesar de fazer divisa com Dormentes e Petrolina, não acompanha o nível de desenvolvimento de seus vizinhos o que significa que alto PIB per capita dessas cidades não está se refletindo na cidade de Afrânio. Em contra partida, temos regiões que apresentam a característica de ter altos IFDMs com vizinhos com altos PIB per capita como parte da Região Metropolitana de Recife que mais uma vez se mostra significativo apresentando esse efeito transbordamento outro importante apontado é Floresta que se mostra um *cluster* Alto-Baixo para a análise do Bivariado representa uma área com alto desenvolvimento local, porém sem efeito positivo sobre os municípios do entorno.

4.5 Análise espacial utilizando-se a Predição do *Random Forest*

Esta subseção tem como objetivo integrar as análises do modelo *Random Forest (RF)* e do método de AEDE, de modo a aprofundar o exame das estruturas espaciais no estado de Pernambuco. Para isso, utiliza-se a variável do IFDM predita pelo modelo *Random Forest* já treinado, o que permite construir uma análise baseada em predição e correlacioná-la como uma proxy para desenvolvimento econômico utilizada. Vale ressaltar que o modelo RF pode suavizar e melhorar a previsão espacial, mas não indica causas profundar desses padrões além disso *clusters* aparentes podem ser artefatos estatísticos da distribuição do erro do modelo em conjunto com o AEDE, e pode não indicar alguns padrões reais subjacentes o que sugere necessidade de se aprofundar a abordagem. Nesse contexto, o uso da predição do RF visa não apenas suavizar os possíveis ruídos encontrados na análise da AEDE, mas também mitigar potenciais endogeneidades associados ao IFDM observado em cada localidade.

Posto isto, para entender melhor em média como estão as predições do modelo RF foi feito uma divisão na Tabela 9 a seguir dos resultados do IFDM Predito com o IFDM real por mesorregião apontando a média dos valores e o seu desvio padrão com intuído de se comparar as predições do modelo com os resultados reais.

Tabela 9: Estatística descritiva IFDM e IFDM Predito por Mesorregião

Mesorregião	**IFDM**		**IFDM Predito**	
	Média	**Desvio Padrão**	**Média**	**Desvio Padrão**

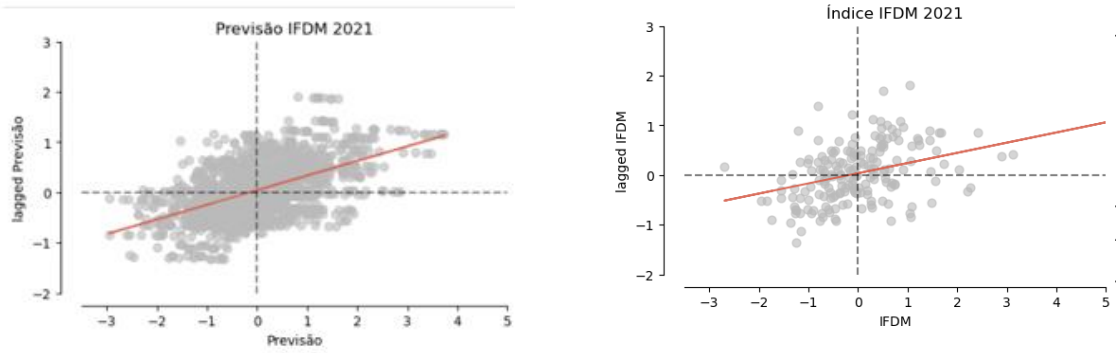
Agreste Pernambucano	0,496	0,067	0.495	0.065
Mata Pernambucana	0,460	0,071	0.461	0.066
Metropolitana do Recife	0,502	0,069	0.50	0.068
Sertão Pernambucano	0,483	0,067	0.48	0.062
São Francisco Pernambucano	0,495	0,088	0.529	0.081

Fonte: Elaboração própria com base nos resultados.

Pode-se inferir da Tabela 8 que em relação aos dados reais há uma boa aproximação média dos valores do IFDM todas as mesorregiões, além disso o Sertão tem o erro maior entre todas de de 0.483 a 0.48 cerca de 0.03 pontos o modelo está na média acertando bem a predição dos dados, contudo ainda sim uma previsão não chega a substituir os dados reais, mas pode auxiliar na suavização e melhora da AEDE.

Ao gerar os Gráficos de dispersão do I de Moran global utilizando a variável do IFDM predito se obteve resultados similares ao método espacial com a utilização da variável real Apontado na Figura 18 a seguir. Pode-se notar que há uma tendência dos pontos se concentrarem nos quadrantes Baixo-Baixo e Alto-Alto.

Figura 18: Dispersão de I de Moran Global IFDM Predito e IFDM Real 2021



(A) IFDM Predito 2021 (B) IFDM - 2021

Fonte: Elaboração própria com base nos resultados.

A primeira figura aponta os pontos gerados pelo IFDM Predito os quais estão com uma dispersão entre os quadrantes Alto-Alto e Baixo-Baixo com mais pontos com significância estatística comparado ao IFDM que mostra uma dispersão de menos pontos. Contudo, as retas estão com inclinação positiva e com a concentração dos pontos similares nos mesmos quadrantes.

Ademais, ao visualizar a significância do I de Moran temos a rejeição da hipótese nula de que os resultados espaciais são aleatórios. Além disso, a aplicação das predições aumentou

o nível de força do I de Moran de 0.243 para 0.29 aumentando a correlação espacial que pode indicar que a predição pode ter suavizado efeitos estruturais, deste modo a suavização pode ter melhorado a correlação espacial deixando mais visíveis na AEDE. Dessa forma, pode-se indicar que o modelo tem indícios para aprimorar a capacidade de identificar possíveis *spillovers* espaciais e de compreender o desenvolvimento municipal com base em critérios de vizinhança. Além disso, com estudos mais aprofundados, pode contribuir para prever impactos da dinâmica espacial futura a partir de dados recentes, oferecendo um instrumento interessante para análise territorial de suavização e melhor visualização de correlações espaciais.

A Tabela 10 A seguir compara os valores do IFDM constatado o Real e do IFDM fruto da predição do modelo RF o Predito. Desse modo podemos ver que os dois tem significância estatística o que apresenta o resultado de que o modelo RF em suas predições conseguiu apresentar correlação espacial rejeitando a hipótese nula de que a correlação espacial gerada é aleatória.

Tabela 10: Índice Global de Moran IFDM Predito e PIB per capita 2021

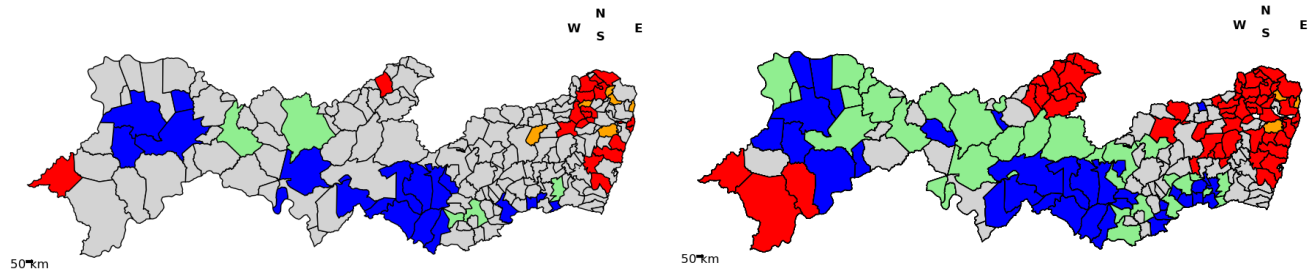
Variáveis	Índice de Moran	p-valor:
IFDM predito	0.290	0.001
IFDM Real	0.243	0.001

Fonte: Elaboração própria com base nos resultados.

Destarte, pode-se perceber também que os resultados de correlação espacial aumentam significativamente. Houve um aumento evidenciando que a junção das duas técnicas pode ser viável para se propor relações espaciais com melhores visualizações e resultados.

A Figura 19 a seguir mostra que ao controlar o modelo pelas predições tornou-se possível identificar novas formações de possíveis *clusters* espaciais, que pode indicar que o modelo suavizou as análises melhorando a visualização desses *clusters*, para os quadrantes Alto-Alto, Baixo-Alto e Baixo-Baixo. Dessa forma temos uma comparação do modelo do *Random Forest* em comparação com os dados reais.

Figura 19: Análise de dependência espacial local (LISA) do IFDM Real e IFDM Predito dos municípios de Pernambuco 2021



(A) IFDM - 2021 (B) IFDM Predito 2021

Fonte: Elaboração própria com base nos resultados.

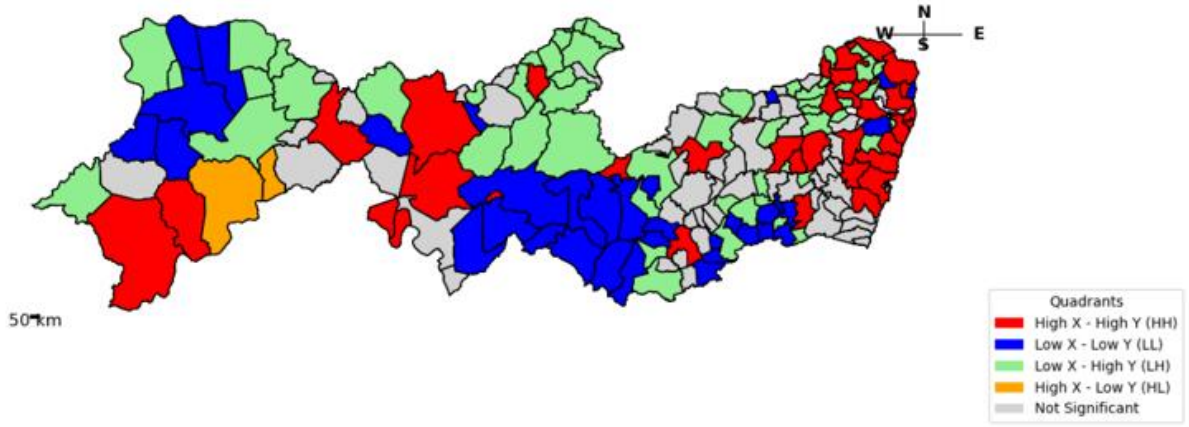
Nesse sentido, pode indicar que ocorreu uma suavização da análise AEDE comparado aos dados reais a comparação dos mapas apontam que houve uma ampliação dos pontos e formação de *clusters* utilizando a predição do IFDM em municípios próximos Tabira mostrando que a maior parte das cidades próximas têm uma formação Alto-Alto indicando um possível efeito transbordamento quando se olha para o IFDM, assim como as mesorregiões do Agreste, Mata Pernambucana e da Região Metropolitana de Recife estão com regiões mais amplas para *clusters* Alto-Alto que antes não apresentavam-se no modelo.

Posteriormente, foi feita outra análise bivariada utilizando o IFDM predito como o *Y* e o PIB per capita como *X* da mesma forma que aplicado com o IFDM real com objetivo de verificar as mudanças da análise com a utilização da variável predita. A Figura 20 apresenta o resultado do LISA bivariado entre o IFDM predito e o PIB per capita dos municípios de Pernambuco em 2021, buscando evidenciar um possível desenvolvimento municipal e relação espacialmente com o desempenho econômico regional. Os *clusters* Alto-Alto indicam municípios que apresentam alto PIB per capita associados a vizinhos com elevado IFDM predito, formando áreas de desenvolvimento integrado. Esses agrupamentos aparecem em boa parte da Região Metropolitana do Recife na Mata Pernambucana e em alguns pontos no Agreste, sugerindo a presença de polos regionais com maior dinamismo socioeconômico corroborando com a teoria de Perroux (1955) sobre cidades ou regiões polos.

Por outro lado, os *clusters* Baixo-Baixo, revelam regiões em que baixos níveis de renda coincidem com baixos níveis de desenvolvimento social entre os municípios vizinhos, predominando no Sertão e em parte do Agreste Central. Esse padrão reforça a persistência de desigualdades estruturais no estado. Além disso, observam-se *outliers* espaciais do tipo Baixo-Alto, caracterizados por municípios com baixo IFDM predito, apesar de estarem cercados por vizinhos com alto PIB per capita, o que sugere limitações na capacidade local de absorver ou converter o dinamismo econômico em melhoria social.

Já os *outliers* Alto-Baixo representam municípios que apresentam alto PIB per capita mesmo sendo vizinhos de áreas com baixo IFDM predito, indicando possíveis boas práticas de gestão pública, capacidades institucionais diferenciadas ou estruturas econômicas locais específicas que sustentam melhores resultados. Por fim, os municípios em cinza não apresentaram autocorrelação espacial significativa, indicando que, nessas áreas, a relação entre PIB per capita e IFDM predito dos vizinhos não segue um padrão espacial definido. Esse conjunto de padrões revela uma dinâmica regional heterogênea e reforça a importância de análises espaciais para orientar políticas públicas diferenciadas no território pernambucano bem como a integração do IFDM e de outros indicadores de desenvolvimento como o IDHM com a proxy de crescimento econômico (PIB per capita).

Figura 20: Análise de dependência espacial local (LISA) bivariada do IFDM Predito e PIB per capita dos municípios de Pernambuco - 2021



Fonte: Elaboração própria com base nos resultados.

Posto isto, há uma figura com mais áreas significativamente relevantes do ponto de vista estatístico no mapa com o IFDM predito vemos regiões de baixo desenvolvimento predominantemente no Sertão de Pernambuco que pode indicar que essas áreas devem ser levadas em consideração como alvos de políticas públicas de desenvolvimento, no entanto para melhores hipóteses é preciso se profundar nos estudos. Além disso, podemos interpretar a Região Metropolitana de Recife como apontam os estudos como cidade polo corroborada com a predição do RF a exceção da cidade de São Lorenço da mata que nas estatísticas apresentadas já apresenta um dos IFDMs mais baixos da Região e no mapa se mostra como um *cluster* Baixo-Baixo apontando possível subdesenvolvimento. Ademais, a cidade de Afrânio se mostra um *cluster* Baixo-Alto validando a análise dos dados reais, contudo seu entorno próximo tem cidades com desenvolvimento positivo, as quais não se apresentavam na leitura anterior, que pode indicar que Afrânio não está conseguindo incorporar esse desenvolvimento.

5 CONSIDERAÇÕES FINAIS

Este estudo teve como objetivo analisar e prever o Índice FIRJAN de Desenvolvimento Municipal (IFDM) dos municípios de Pernambuco em 2021, empregando o modelo *Random Forest* associado a técnicas de Análise Exploratória de Dados Espaciais (AEDE). Para isso, utilizaram-se dados do Atlas do Desenvolvimento Humano e do portal da Firjan referentes ao período de 2013 a 2021, de modo a capturar a trajetória recente do desenvolvimento socioeconômico do estado e justificar a aplicação de um método preditivo robusto, capaz de lidar com múltiplas variáveis e identificar padrões espaciais. A escolha dessa abordagem se fundamenta-se na necessidade de aprimorar o entendimento sobre o comportamento regional do crescimento econômico e apoiar estudos e políticas públicas voltadas ao desenvolvimento de Pernambuco, pois é um dos estados com menor índice de Desenvolvimento quando se comparado com os demais.

Os Resultados do modelo preditivo apontam que é o modelo mais abordado na literatura recente de predição de desenvolvimento e se mostra eficaz no caso do estado de Pernambuco corroborando com a literatura. Além disso, é possível testar o modelo robusto e passar com uma predição de 80 por cento em média sem ter um superajuste e a incorporação de ruído já que o modelo generalizou bem para os dados de teste. Posto isto, a melhora na predição da AEDE corrobora na hipótese que é incorporada na própria literatura do modelo de ser robusto a ruídos. Dessa forma, é possível verificar que as correlações espaciais que já se mostram significativas no AEDE padrão sem integração com o *Random Forest* identificando possíveis *clusters* espaciais próximas a Região metropolitana de Recife corroborando com a literatura de economia urbana que é intensificada mostrando mais *clusters* próxima a região quando integrada com *Random Forest.* Os mesmos resultados no mapa são vistos para o Sertão de Pernambuco Agreste e Mata Pernambucana.

Nesse sentido, ainda assim, reconhece-se que fatores não incluídos na análise, como características do mercado de trabalho, oferta de ensino técnico e privado, e outros indicadores socioeconômicos correlatos ao PIB per capita, podem contribuir para aprofundar futuras investigações. Além disso, avaliar a eficácia do modelo à medida que novos dados forem disponibilizados após 2021 pode fortalecer o desenvolvimento do estudo, assim como aplicá-lo a um município específico, testando o desempenho do modelo em escala local.

REFERÊNCIAS

ALMEIDA, E. **Econometria espacial aplicada**. Campinas, SP: Alínea, 2012.

ALVES, D. F. **Ensaios em economia regional e urbana**. 2024. Recife: Universidade Federal de Pernambuco, Centro de Ciências Sociais Aplicadas, Departamento de Economia, Programa de Pós-Graduação em Economia (PIMES), 2024.

ALVES, D. F. **Estrutura produtiva e desigualdade intermunicipal de renda no Brasil: uma abordagem regional**. 2020. Dissertação (Mestrado em Economia) – Universidade Federal do Rio Grande do Norte, Centro de Ciências Sociais Aplicadas, Natal, 2020.

ARUMNISAA, R. I.; WIJAYANTO, A. W. Comparison of ensemble learning method: Random Forest, Support Vector Machine, AdaBoost for classification Human Development Index (HDI). **Sistemasi: Jurnal Sistem Informasi**, v. 12, n. 1, p. 206–218, 31 jan. 2023.

ATHEY, S. The impact of machine learning on economics. In: AGRAWAL, A.; GANS, J.; GOLDFARB, A. (org.). **The economics of artificial intelligence: an agenda**. Chicago: University of Chicago Press, [s.d.].

BARROS, L. A. B. C.; BERGMANN, D. R.; CASTRO, F. H.; SILVEIRA, A. D. M. **Endogeneidade em regressões com dados em painel: um guia metodológico para pesquisa em finanças corporativas**. *Revista Brasileira de Gestão de Negócios*, v. 22, n. 0, 2020.

BAVARESCO, M. Z.; WEBBER, C. G. IA explicável para reduzir a assimetria de informação no consumo: uma análise comparativa de ferramentas e implicações educacionais. **Redin**, v. 13, n. 2, p. 59-77, 2024.

CALIFE, Danilo Bruno. **Com maior PIB, mas sem desenvolvimento**? Indústria automobilística e desenvolvimento econômico: o caso de Goiana, Pernambuco. 2021. Dissertação (Mestrado em Economia) – Programa de Pós-Graduação em Economia, Universidade Federal de Pernambuco, Recife, 2021.

CLIFF, A. D.; ORD, J. K. **Spatial processes: models & applications**. London: Pion, 1981.

COLOMBO, Jefferson A.; LAZZARI, Martinho R. **Timing, duração e magnitude da recessão econômica de 2014-2016 nos estados brasileiros.** *Estudos Avançados*, v. 31, n. 89, jan./abr. 2017.

CUNHA, J. P. Z. **Um estudo comparativo das técnicas de validação cruzada aplicadas a modelos mistos**. 2019. Dissertação (Mestrado em Ciências – Estatística) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2019.

DALCHIAVON, E. C. **Desenvolvimento econômico dos municípios paranaenses: análise a partir do Índice FIRJAN de Desenvolvimento Municipal (IFDM) para o período de 2005 a 2013**. 2017. 150 f. Dissertação (Mestrado em Gestão e Desenvolvimento Regional) – Universidade Estadual do Oeste do Paraná, Francisco Beltrão, 2017.

DURÃES, R. G.; SALIS, T. T.; COELHO, F. G. F.; BRAGA, A. P. Análise de explicabilidade de um modelo de aprendizado de máquina para aplicações industriais. In: **CONGRESSO BRASILEIRO DE AUTOMÁTICA – CBA**, 24., 2022, João Pessoa. *Anais*. João Pessoa: Sociedade Brasileira de Automática, 2022.

EFRON, B. Bootstrap methods: another look at the jackknife. **Annals of Statistics**, v. 7, n. 1, p. 1–26, 1979.

FEDERAÇÃO DAS INDÚSTRIAS DO ESTADO DO RIO DE JANEIRO (FIRJAN). **Índice FIRJAN de Desenvolvimento Municipal (IFDM)**. Rio de Janeiro: FIRJAN, 2023. Disponível em: https://www.firjan.com.br/ifdm. Acesso em: 21 jul. 2025.

GRUS, J. **Data science do zero**. 2. ed. São Paulo: O'Reilly, 2021.

HAN, J.; KAMBER, M.; PEI, J. **Data mining: concepts and techniques**. 3. ed. Burlington, MA: Elsevier, 2012.

HARRISON, M. **Machine learning: guia de referência rápida**. São Paulo: O'Reilly, 2020.

KAUR, M. et al. Supervised machine-learning predictive analytics for national quality of life scoring. **Applied Sciences**, v. 9, n. 8, p. 1613, 18 abr. 2019.

MCBRIDE, L.; NICHOLS, A. Improved poverty targeting through machine learning: an application to the USAID Poverty Assessment Tools. **Unpublished manuscript**, 2015. Disponível em: http://www.econthatmatters.com/wp-content/uploads/2015/01/improvedtargeting_21jan2015.pdf.

KRUGER, R. V.; BOURSCHIEDT, D. M. Mercado de trabalho e o Índice FIRJAN de Desenvolvimento Municipal: padrões espaciais dos municípios do estado do Paraná. **Estudios Económicos**, v. 38, n. 77, p. 99-117, 2021.

LOPES, P. C. B.; PEREIRA, L. A. G. Análise espacial do Índice de Desenvolvimento Humano Municipal (IDHM) no Brasil. **Anais do ENANPEGE**, 2021. Disponível em: https://www.editorarealize.com.br/editora/anais/enanpege/2021/TRABALHO_COMPLETO_EV154_MD1_SA129_ID65420092021213600.pdf.

MARCONATO, M.; COELHO, M. H. O IDHM dos municípios brasileiros sob a perspectiva da análise exploratória de dados espaciais. 2019. Disponível em: https://revistas2.uepg.br/index.php/economia/article/view/14507.

OLIVEIRA, M. A. **Aplicação do machine learning na previsão de índices econômicos: o IDHM com o modelo Random Forest de 2012 à 2021**. 2023. Trabalho de Conclusão de Curso (Bacharelado em Ciências Econômicas) – Universidade Federal de Mato Grosso do Sul, Campo Grande, 2023.

OZDEN, E.; GULERYUZ, D. Optimized machine learning algorithms for investigating the relationship between economic development and human capital. **Computational Economics**, 24 set. 2021.

PERROUX, François. *Note sur la notion de "pôle de croissance"*. **Économie Appliquée**, v. 8, n. 1–2, p. 307–320, 1955.

PNUD; IPEA; FJP. **Atlas do desenvolvimento humano nas regiões metropolitanas brasileiras: Índice de Desenvolvimento Humano Municipal (IDHM)**. Brasília: PNUD; Ipea; FJP, 2014. Disponível em: https://portalantigo.ipea.gov.br/agencia/images/stories/PDFs/livros/livros/atlasdodesenvolvimentohumanorms_o-indice-de-desenv.pdf. Acesso em: 21 jul. 2025.

REIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, jan. 2001. Disponível em: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf. Acesso em: 21 jul. 2025.

ROMERO, J. A. R. Análise espacial da pobreza municipal no estado de Minas Gerais – 1991-2000. In: **XIV ENCONTRO NACIONAL DE ESTUDOS POPULACIONAIS – ABEP**, Caxambu, 18–22 set. 2006.

RODRIGUES, K. C. T. T. et al. Uma análise espacial da imigração no Brasil. **Economia e Desenvolvimento**, v. 27, n. 1, 2015. Disponível em: https://bit.ly/3bzISuq.

SANTANA, J. H. C. **Random Forest: uma abordagem poderosa para aprendizado de máquina**. 2022. Trabalho de Conclusão de Curso – Faculdade de Educação Tecnológica do Estado do Rio de Janeiro, Paracambi, 2022.

SANTOS, C. B. **Previsão do Índice de Desenvolvimento Humano e da expectativa de vida na América Latina por meio de técnicas de mineração de dados**. 2016. Tese (Doutorado em Engenharia de Produção) – Universidade Tecnológica Federal do Paraná, Ponta Grossa, 2016.

SANTOS, T. M. **Métodos preditivos computacionalmente eficientes baseados em floresta aleatória**. 2024. Tese (Doutorado em Probabilidade e Estatística) – Universidade de São Paulo, São Paulo, 2024. Disponível em: https://www.teses.usp.br/teses/disponiveis/45/45133/tde-01052024-164427/publico/tese_Tiago_Mendonca_dos_Santos.pdf.

SANTOS, M. V. **Projeção de demanda elétrica com algoritmos de aprendizado de máquina**. 2022. Trabalho de Conclusão de Curso (Graduação em Economia) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2022.

SOBEL, T. F.; MUNIZ, A. L. P.; COSTA, E. F. Divisão regional do desenvolvimento humano em Pernambuco: uma aplicação da análise de cluster. **Teoria e Evidência Econômica**, v. 15, n. 33, p. 37-62, jul./dez. 2009.

SHERMAN, L. et al. Global high-resolution estimates of the United Nations Human Development Index using satellite imagery and machine-learning. 1 mar. 2023.

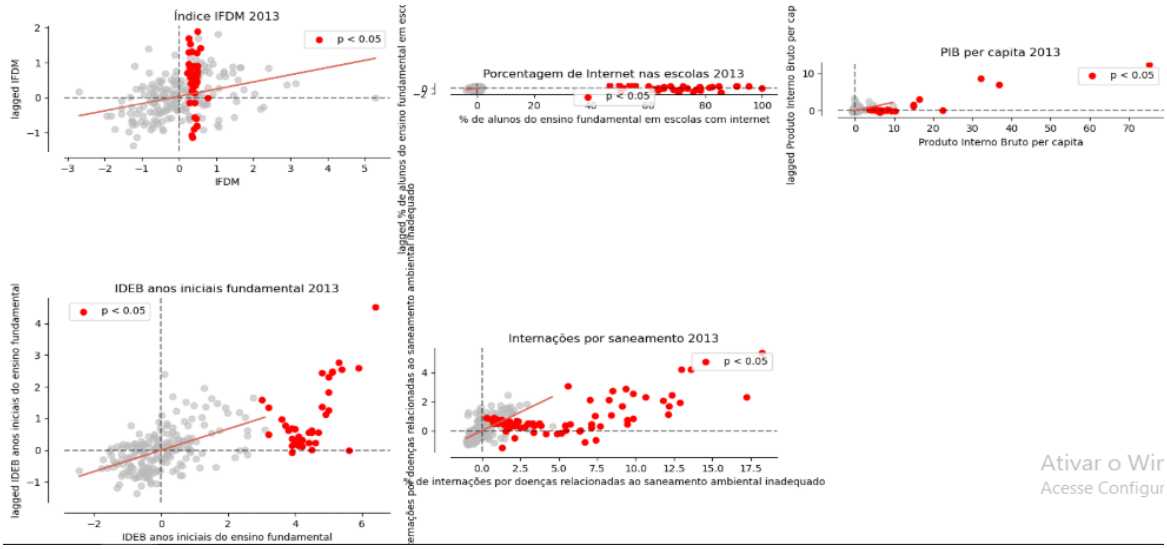
TOBAIGY, F.; ALAMOUDI, M.; BAFAIL, O. Human Development Index: determining and ranking the significant factors. **International Journal of Engineering Research & Technology (IJERT)**, 2023. Disponível em: https://www.ijert.org/human-development-index-determining-and-ranking-the-significant-factors. Acesso em: 21 jul. 2025.

APÉNDICE A: Assuntos complementares

Informações

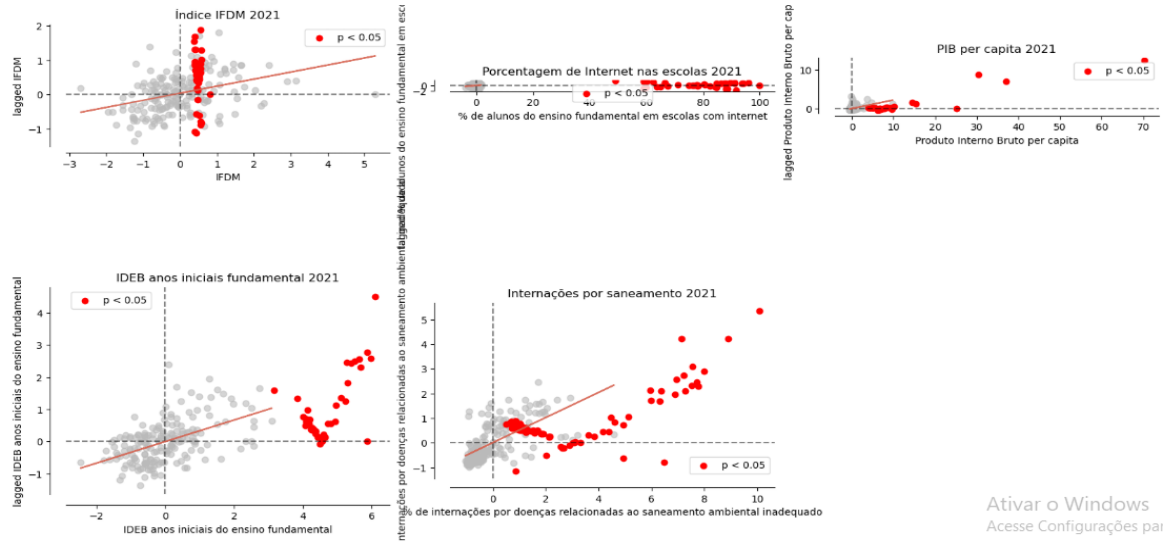
A informações desse apêndice demonstram de forma visual os municípios significativos localmente dentro do gráfico de dispersão global do I de Moran. Apresenta-se dentro de uma Diagrama de dispersão apresentando os quadrantes para as variáveis analisadas nos anos de 2013 e 2021 e posteriormente é demonstrado quais são os municípios e que tipo de variação baseado por quadrante e a sua significância estatística.

Figura 21: Gráfico de Dispersão de Probabilidade global com Índice locais de Moran 2013



Fonte: Elaboração própria com base nos resultados.

Figura 22: Gráfico de Dispersão de Probabilidade global com Índice locais de Moran 2021



Fonte: Elaboração própria com base nos resultados.

Tabela 11: Índice de Moran local ano 2021 PIB per capita municípios Pernambuco.

Territorialidades	**Produto Interno Bruto per capita**	**Quadrante**	**p-valor**
Cabo de Santo Agostinho	30.3375	High-High (HH)	0.002
Sirinhaém	8.4125	High-High (HH)	0.010
Escada	7.9725	High-High (HH)	0.001
Surubim	7.7850	Low-High (LH)	0.001
São Benedito do Sul	3.9550	Low-Low (LL)	0.006
Quipapá	4.9475	Low-Low (LL)	0.004
Cupira	6.6875	Low-Low (LL)	0.001
Triunfo	5.5125	Low-Low (LL)	0.009
Jaqueira	5.0250	Low-Low (LL)	0.002
Itaíba	4.9150	Low-Low (LL)	0.002
Panelas	4.3375	Low-Low (LL)	0.005
Lagoa dos Gatos	4.1200	Low-Low (LL)	0.008
Fernando de Noronha	25.0475	Não significativo	0.001

Fonte: Elaboração própria com base nos resultados.

Tabela 11: Índice de Moran local ano 2021 IFDM municípios Pernambuco.

Territorialidades	**IFDM**	**Quadrante**	**p-valor**
Escada	0.4999	High-High (HH)	0.005
Buenos Aires	0.4728	High-Low (HL)	0.009
Palmares	0.5224	Low-High (LH)	0.009
Serra Talhada	0.5486	Low-High (LH)	0.007
Pedra	0.4152	Low-Low (LL)	0.007
Parnamirim	0.4339	Low-Low (LL)	0.007
Itaíba	0.4009	Low-Low (LL)	0.002
Ouricuri	0.4317	Low-Low (LL)	0.007
Manari	0.4071	Low-Low (LL)	0.005
Tupanatinga	0.3933	Low-Low (LL)	0.006
Fernando de Noronha	0.8182	Não significativo	0.001

Fonte: Firjan e Atlas do desenvolvimento humano, 2021, elaboração própria.

Figura 23: Mapa de Pernambuco e Seus municípios

Fonte: Tribunal judiciário de Pernambuco