

UNIVERSIDADE FEDERAL DO PARANÁ

THIAGO MENEZES SOUSA

A RELAÇÃO ENTRE A DESIGUALDADE DE RENDA E O ENSINO BÁSICO DOS
MUNICÍPIOS BRASILEIROS

CURITIBA

2021

THIAGO MENEZES SOUSA

A RELAÇÃO ENTRE A DESIGUALDADE DE RENDA E O ENSINO BÁSICO DOS
MUNICÍPIOS BRASILEIROS

Trabalho apresentado como requisito parcial
para a conclusão do Curso de Ciências Econô-
micas, setor de Ciências Sociais Aplicadas da
Universidade Federal do Paraná.

Orientador: Prof. Dr. Marcos Minoru Hasegawa

CURITIBA

2021

TERMO DE APROVAÇÃO

THIAGO MENEZES SOUSA

A RELAÇÃO ENTRE A DESIGUALDADE DE RENDA E O ENSINO BÁSICO DOS MUNICÍPIOS BRASILEIROS

Trabalho apresentado como requisito parcial para a conclusão do Curso de Ciências Econômicas, setor de Ciências Sociais Aplicadas da Universidade Federal do Paraná, pela seguinte banca examinadora:

Prof. Dr. Marcos Minoru Hasegawa
Orientador

Prof. Dr. Alexandre Alves Porsse
UFPR

Prof. Dr. Paulo de Andrade Jacinto
UFPR

Curitiba, 13 de Dezembro de 2021.

RESUMO

O objetivo desta dissertação é entender o efeito de indicadores educacionais na desigualdade de renda dos municípios brasileiros, de maneira a confirmar a literatura acadêmica, além de mostrar como a melhora da educação é um caminho que deve ser percorrido caso o Brasil queira diminuir sua desigualdade de renda e se aproximar dos países desenvolvidos. Para tal finalidade, o presente trabalho utiliza três formas principais de análise: Espacial, Regressão Linear Múltipla e Algoritmos de Aprendizado de Máquina. Os resultados demonstram que as variáveis dependem de sua localização no espaço e dos indicadores da vizinhança; municípios com maiores taxas de analfabetismo entre jovens, menores notas no Enem e com maiores taxas de entrada tardia no sistema de ensino são mais desiguais, enquanto municípios com maiores médias de horas de aula por dia para a educação infantil são menos desiguais; por fim, o Algoritmo de Aprendizado de Máquina demonstra que, das variáveis utilizadas, a taxa de analfabetismo entre jovens foi a variável mais importante na predição dos municípios mais desiguais. Assim, os resultados da análise mostram que a relação entre educação e desigualdade de renda é real e significativa.

Palavras-chaves: desigualdade; renda; educação; municípios; análise espacial; regressão linear múltipla; aprendizado de máquina; analfabetismo; horas-aula; Enem; entrada-tardia.

ABSTRACT

The objective of this dissertation is to understand the effect of educational indicators on income inequality of the Brazilian cities, to confirm the literature and to reveal that educational improvement must be done if Brazil wants to decrease its income inequality index gap to developed countries. To do so, the following undergraduate thesis applies three ways of analysis, being Spatial Analysis, Multiple Linear Regression and Machine Learning. The results show that the variables depend on its spatial location and neighbors indicators; cities with the higher rates of illiteracy among young people, smaller Scholastic Aptitude Test scores and higher rates of late admission to school are more unequal, while cities with the higher average number of hours in the school day are less unequal. Finally, the Machine Learning algorithm shows that, regarding the studied variables, the indicator of illiteracy among young people was the most important one when it comes to prediction of the most unequal cities. Therefore, the results of the analysis reveal that the connection between education and income inequality is real and significant.

Key-words: inequality. income. education. cities. spatial analysis. multiple linear regression. machine learning. illiteracy. hours in the school day. late admission.

LISTA DE ILUSTRAÇÕES

FIGURA 1 – COEFICIENTE DE GINI	16
FIGURA 2 – CENTROIDE DOS MUNICÍPIOS BRASILEIROS	19
FIGURA 3 – EXEMPLO DE ALGORITMO SUPERVISIONADO DE APRENDI- ZADO DE MÁQUINA	22
FIGURA 4 – CURVA ROC	23
FIGURA 5 – EXEMPLO DE ÁRVORE DE DECISÃO	24
FIGURA 6 – ÍNDICE DE GINI 1981 - 2019	25
FIGURA 7 – ÍNDICE DE GINI POR ESTADO (2010)	27
FIGURA 8 – ANALFABETISMO ENTRE ADOLESCENTES POR ESTADO (2010)	29
FIGURA 9 – ENTRADA TARDIA NO SISTEMA DE ENSINO POR ESTADO (2019)	31
FIGURA 10 – MÉDIA DE HORAS-AULA POR DIA NA EDUCAÇÃO INFANTIL POR ESTADO (2010)	33
FIGURA 11 – MÉDIA DA NOTA DO ENEM POR ESTADO (2010)	35
FIGURA 12 – EXEMPLO DE DISPERSÃO PARA O ÍNDICE DE MORAN	36
FIGURA 13 – GRÁFICOS DE DISPERSÃO DE MORAN PARA OS INDICADO- RES ANALISADOS	37
FIGURA 14 – DISTRIBUIÇÃO DOS INDICADORES PARA A REGIÃO SUL . .	38
FIGURA 15 – DISTRIBUIÇÃO DOS INDICADORES PARA A REGIÃO SUDESTE	39
FIGURA 16 – DISTRIBUIÇÃO DOS INDICADORES PARA A REGIÃO CENTRO- OESTE	40
FIGURA 17 – DISTRIBUIÇÃO DOS INDICADORES PARA A REGIÃO NORDESTE	41
FIGURA 18 – DISTRIBUIÇÃO DOS INDICADORES PARA A REGIÃO NORTE .	42
FIGURA 19 – CURVA ROC	47
FIGURA 20 – IMPORTÂNCIA DE CADA VARIÁVEL PARA O GRADIENT BOOST	47

LISTA DE TABELAS

TABELA 1 – MUNICÍPIOS MENOS DESIGUAIS	26
TABELA 2 – MUNICÍPIOS MAIS DESIGUAIS	26
TABELA 3 – MUNICÍPIOS COM MENORES INCIDÊNCIAS DE ANALFABE- TISMO ENTRE JOVENS	28
TABELA 4 – MUNICÍPIOS COM MAIORES INCIDÊNCIAS DE ANALFABE- TISMO ENTRE JOVENS	28
TABELA 5 – MUNICÍPIOS COM MENORES INCIDÊNCIAS DE ENTRADA TAR- DIA NA ESCOLA	30
TABELA 6 – MUNICÍPIOS COM MAIORES INCIDÊNCIAS DE ENTRADA TAR- DIA NA ESCOLA	30
TABELA 7 – MUNICÍPIOS COM MENORES MÉDIAS DE HORAS-AULA POR DIA NA EDUCAÇÃO INFANTIL	32
TABELA 8 – MUNICÍPIOS COM MAIORES MÉDIAS DE HORAS-AULA POR DIA NA EDUCAÇÃO INFANTIL	32
TABELA 9 – MUNICÍPIOS COM MAIORES MÉDIAS NA NOTA DO ENEM . .	34
TABELA 10 – MUNICÍPIOS COM MENORES MÉDIAS NA NOTA DO ENEM . .	34
TABELA 11 – RESULTADOS DO ÍNDICE DE MORAN PARA OS INDICADORES ANALISADOS	36
TABELA 12 – DESCRIÇÃO DO BANCO DE DADOS	43
TABELA 13 – PRIMEIRA REGRESSÃO	44
TABELA 14 – SEGUNDA REGRESSÃO	45
TABELA 15 – RESULTADOS DO APRENDIZADO DE MÁQUINA	46

SUMÁRIO

1	INTRODUÇÃO	8
2	REVISÃO DE LITERATURA	10
2.1	Variáveis que Mais Impactam a Educação	10
2.2	O Combate à Desigualdade	11
2.3	Qualidade da Educação	12
2.4	Métodos Utilizados Anteriormente	13
3	METODOLOGIA	15
3.1	Dados Utilizados	15
3.1.1	Desigualdade de Renda	15
3.1.2	Indicadores Educacionais	17
3.1.3	Medição do Aprendizado	17
3.2	Estratégia Econométrica	18
3.2.1	Análise Espacial	18
3.2.2	Análise de Regressão	20
3.2.3	Aprendizado de Máquina	21
3.2.3.1	Métodos Utilizados	23
4	RESULTADOS E DISCUSSÃO	25
4.1	Estatísticas Descritivas das Variáveis Estudadas	25
4.1.1	Coeficiente de Gini	25
4.1.2	Analfabetismo Entre Pessoas de 11 a 14 Anos	27
4.1.3	Entrada Tardia no Sistema de Ensino	29
4.1.4	Horas-Aula por Dia	31
4.1.5	Média da Nota do Enem	33
4.2	Análise Espacial	35
4.2.1	Índice de Moran	35
4.2.2	Região Sul	38
4.2.3	Região Sudeste	39
4.2.4	Região Centro-Oeste	40
4.2.5	Região Nordeste	41
4.2.6	Região Norte	42
4.3	Análise de Regressão	43
4.4	Aprendizado de Máquina	45
5	CONSIDERAÇÕES FINAIS	48
	REFERÊNCIAS	50

1 INTRODUÇÃO

Nos últimos anos, o Brasil tem passado por inúmeros desafios. Recessão grave e histórica, lenta recuperação econômica e pandemia. Todos esses fatores são refletidos no índice de desigualdade. Esse indicador, medido a partir do índice de Gini, mostra que o país alcançou em 2018 a sua pior marca em dez anos: 0,539 (BANCO MUNDIAL, 2019). Dessa forma, o Brasil regrediu uma década nesse indicador, influenciado pela queda da renda dos 50% mais pobres (BARBOSA; SOUZA; SOARES, 2020), o que afeta não só a qualidade de vida das pessoas mais carentes, como também o crescimento econômico do país.

De acordo com o Banco Mundial (2019), o Brasil é o nono país mais desigual do mundo e o mais desigual da América Latina. A partir dos dados da PNUD de 2019, podemos constatar que 28,3% da renda é concentrada pelos 1% mais ricos do país. Enquanto isso, os 40% mais pobres detêm apenas 10,4% da riqueza total (SASSE, 2021).

A partir dos dados observados, portanto, é inegável a presença da desigualdade na sociedade brasileira, influenciando todos os elementos da vida da população, principalmente dos indivíduos mais pobres. Esse problema cria diversos obstáculos na vida das pessoas, dificultando o acesso a serviços que deveriam ser garantidos a todos, como a educação, serviços públicos e voz política. Prejudica, também, o acesso à terra, à moradia de qualidade, ao crédito e ao trabalho formal, aumentando o risco de desemprego e conseqüentemente aumentando a pobreza, fazendo com que o desenvolvimento econômico de toda a sociedade seja prejudicado (FERRANTI et al., 2004).

Diversos autores citam que a mais segura e óbvia saída desse cenário desfavorável é a educação. A eficácia desta abordagem pode ser vista em exemplos externos, pois Estados Unidos, Japão, Coreia do Sul e China, entre outros, demonstram como o investimento e foco na educação podem mudar uma sociedade, pois ela gera diversos benefícios, como o forte impacto em aumento da renda, aumento da produtividade e, além disso, potencializa os efeitos da globalização.

Neste contexto, o presente trabalho tem como objetivo demonstrar na prática qual a relação entre os índices educacionais e os índices de desigualdade de renda dos municípios brasileiros. Além disso, abordar como a desigualdade pode afetar a vida das pessoas e exibir como diversos autores citam a educação como a resposta para quebrar esse ciclo.

Dessa forma, o trabalho a seguir emprega três formas principais de análise:

Espacial, Regressão Linear Múltipla e Algoritmos de Aprendizado de Máquina. Essas abordagens foram úteis para entender a relação dos dados no espaço e para entender a relação entre as variáveis. O trabalho está dividido em introdução, revisão de literatura, metodologia utilizada, resultados e discussão, considerações finais e referências.

2 REVISÃO DE LITERATURA

2.1 VARIÁVEIS QUE MAIS IMPACTAM A EDUCAÇÃO

Diversos pesquisadores lidam com os dados da educação e revelam informações relevantes sobre quais são as variáveis que mais afetam o desempenho escolar. Entre eles está Menezes Filho (2012), que com o trabalho intitulado de "Os Determinantes do Desempenho Escolar do Brasil", descreve que as "variáveis que mais explicam o desempenho escolar são as características familiares e do aluno, tais como educação da mãe, cor, atraso escolar e reprovação prévia, número de livros, presença de computador em casa e trabalho fora de casa". Dessa forma, portanto, é possível dizer que as variáveis do aluno e da família são os aspectos que mais explicam a proficiência de um estudante e, portanto, há um forte impacto da desigualdade de renda no desempenho. Além disso, Menezes Filho (2012) destaca como a escolaridade média das mães de todos os alunos de uma escola é um fator mais contundente do que a escolaridade individual das mães de cada aluno sobre suas notas, indicando que o nível de renda da comunidade ao redor é um fator importantíssimo e deve ser considerado. Ademais, também há a constatação, de que crianças com acesso a alguns itens primários, como acesso a livros, jornais, computador e eletricidade têm notas melhores (MENEZES FILHO, 2012). Ou seja, aqueles jovens e crianças com oportunidades de estudo, presumivelmente terão um desempenho melhor.

Logo, é possível entender que os estudantes dependem, entre outras coisas, do meio em que vivem para atingirem ou não todo o seu potencial, de forma que crianças que vivem em lares pobres têm menor probabilidade de atingirem a plenitude de suas capacidades, ao passo que seus colegas mais favorecidos limitam-se apenas por habilidades natas (BUCKINGHAM; WHELDALL; BEAMAN-WHELDALL, 2013).

Em concordância com este argumento, Langoni (1973) corrobora que não somente o nível educacional individual afeta o nível de renda de uma pessoa, mas também o "background familiar", fazendo com que o efeito da desigualdade seja ainda mais potencializado.

Isto posto, é clara a relação entre a desigualdade e a educação, pois os estudos que tratam da relação entre resultados escolares e o capital humano das famílias demonstram a tendência da reprodução das desigualdades nas gerações futuras (FERRÃO; FERNANDES, 2003), de maneira que países e comunidades mais desiguais produzem menos oportunidades de desenvolvimento humano para seus residentes, o que, em uma espécie de ciclo perverso, causa ainda mais desigualdade.

2.2 O COMBATE À DESIGUALDADE

Segundo Ferranti et al. (2004), a desigualdade social aumenta a pobreza e consequentemente diminui o impacto do desenvolvimento econômico na sociedade, prejudicando não só a renda individual, mas também o crescimento agregado da economia. Ainda existem outras razões para que a desigualdade seja combatida, pois afeta fortemente outros indicadores, como a taxa de poupança e a taxa de mortalidade infantil (BARROS; MENDONÇA, 1995).

Diversos autores concordam que uma das principais causas da desigualdade é a dificuldade de acesso à educação por todas as camadas da população, pois “a expansão do setor educacional, em geral, acompanha o processo de desenvolvimento econômico” (LANGONI, 1973). Existem diversos benefícios estratégicos de se investir na educação como um agente de mudança da condição social de um município ou país, primeiramente, como Ferranti et al. (2004) expõem: “sua distribuição pode aumentar sem a necessidade de redistribuir tirando de outra pessoa”, ou seja, é uma forma de crescimento muito eficiente. Além disso, uma educação mais igualitária influencia resultados econômicos com mais de um mecanismo, são eles: “aumenta o capital humano da força de trabalho, que por si aumenta a produtividade do trabalho e assim gera um crescimento em direção a um novo ponto de equilíbrio da produção” e, concomitantemente, aumenta a capacidade de inovação da economia de forma a criar novas tecnologias, produtos e processos que promovem o crescimento, além de democratizar o conhecimento às tecnologias pré-existentes (HANUSHEK; WOESSMANN, 2007).

Além disso, Hanushek e Woessmann (2007) discorrem sobre o impacto da melhoria na educação de uma sociedade e o divide como ganhos individuais e ganhos sociais, e argumentam que, embora haja muitas incertezas, há pouca razão para crer que os ganhos individuais sejam maiores do que os ganhos sociais, na medida que existem outros efeitos positivos associados, como a redução do crime, melhoria na saúde, fertilidade, participação cidadã e na produtividade da economia como um todo. Os autores defendem que a renda individual é influenciada muito fortemente pela proficiência cognitiva, argumentando que uma distribuição de proficiência educacional mais igualitária numa sociedade está intimamente ligada à distribuição de renda e, dessa forma, o crescimento econômico é fortemente afetado pelas habilidades dos trabalhadores.

Como exemplo internacional de como a educação pode mudar uma economia, é possível estudar o caso da China, que passou de um PIB de US\$191bi em 1980 e chegou a US\$14,7tri (BANCO MUNDIAL, 2021), números que levaram o país a se tornar a segunda maior economia do mundo em 2010. No entanto, evidentemente, não se deve confundir crescimento econômico com redução da desigualdade. De acordo com Lv et al. (2015), vários movimentos foram feitos pelo governo Chinês para melhoria

da educação. Como ações principais: promoveu a alocação mais eficiente de recursos para universidades; aprovou uma lei de educação compulsória em 1986, impondo nove anos de educação compulsória para todas as crianças chinesas; propôs uma estratégia de “avanço da nação através da educação e ciência” e, finalmente, realizou projetos em 1995 e 1998 que buscaram construir diversas universidades de qualidade no século 21. Dessa forma, a China basicamente eliminou o analfabetismo entre jovens e jovens adultos e, portanto, observou um enorme crescimento econômico nas últimas décadas.

Existem dois aspectos muito importantes a serem considerados quando se discute o efeito da educação na renda e, portanto, na desigualdade. São elas: o grau de desigualdade educacional da força de trabalho e a sensibilidade dos salários ao nível educacional (BARROS; MENDONÇA, 1995). Ao tratar do primeiro item, Barros e Mendonça afirmam que o Brasil é um dos países mais desiguais neste sentido, pois 15% dos trabalhadores não têm escolaridade, enquanto 10% da população conta com ensino superior. Em relação ao segundo item, os autores afirmam que a cada ano adicional de estudo, o trabalhador brasileiro tende a elevar o nível salarial, em média, em 15%. De acordo com eles, “à medida que o sistema educacional se expande, a sensibilidade dos salários à educação tende a reduzir, levando a um declínio na desigualdade salarial”.

Estes dois fatores combinados: alta desigualdade escolarial e alta sensibilidade dos salários à educação, fazem com que seja difícil pensar em uma solução para a desigualdade social do país que não passe por políticas que visem diminuir a desigualdade educacional

O trabalho de Ferreira e Cruz (2010), por sua vez, propõe que existe um nível de riqueza inicial necessário para que uma pessoa busque investir em capital humano. Neste caso, quanto maior a desigualdade, maior a chance deste nível não ser alcançado e, portanto, os indivíduos acabarem trabalhando como mão de obra não qualificada, fazendo com que exista mais desigualdade em relação àquelas pessoas qualificadas, que são mais produtivas e têm maiores salários. Ou seja, a desigualdade de renda inicial pode gerar desigualdade de renda no longo prazo.

2.3 QUALIDADE DA EDUCAÇÃO

Como já declarado, países como os Estados Unidos, Japão, Coreia do Sul e China demonstram como o investimento e o foco na educação podem mudar uma sociedade. No entanto, o simples investimento de dinheiro em educação não resolve todos os problemas. Na verdade, o Brasil é um dos países que mais comprometem parte do PIB em educação, chegando a 6%, enquanto a média dos países da OCDE é 5,5% (OLIVEIRA, 2018). Contudo, mesmo o país investindo parte considerável de sua riqueza no setor, obtém resultados ruins quando comparado às médias internacionais.

Como descrito por Menezes Filho (2012), alguns municípios brasileiros chegam a gastar R\$1.000,00 por aluno para alcançar uma média de 250 pontos no exame de proficiência do 9º ano, enquanto outros municípios alcançam o mesmo desempenho com cerca de R\$200,00, não sendo o montante investido, portanto, o maior determinante da performance escolar.

Um fator que explica o motivo pelo qual o gasto público não é o aspecto mais importante na melhoria da educação é trazido por Hanushek e Woessmann (2007), ao dizerem que, evidentemente, um ano de estudo em um local pobre não é o mesmo do que um ano de estudo em um local rico, definem que as diferenças estão na "eficiência do sistema educacional, qualidade do ensino, infra-estrutura educacional ou do currículo", e estabelecem que ao estimar o efeito da educação em algum fator econômico, deve-se utilizar uma métrica que capte o quanto os alunos aprenderam em sala de aula. Dessa forma, eles indicam que não se deve tratar educação e escolaridade da mesma forma, pois habilidades cognitivas podem ser desenvolvidas em escolas, pela família, pares e pela cultura ao redor da criança, de forma que as características locais, como instituições econômicas fortes, sistema de direitos de propriedade bem definidos, nível de abertura da economia e segurança são como um pré-requisito para o desenvolvimento econômico e que, sem tais atributos, a educação e as habilidades individuais podem não ajudar a sociedade em questão a atingir todo o potencial econômico almejado. Dessa forma, pode-se dizer que gastos municipais com educação são necessários, mas não suficientes e, portanto, não foram levados em consideração no presente trabalho.

2.4 MÉTODOS UTILIZADOS ANTERIORMENTE

Para tratar da relação entre desigualdade e educação nas regiões do Brasil, Langoni (1973) utilizou dados de 1960 e 1970. O autor utiliza dados censitários para medir as características das regiões e, ao realizar uma regressão log-linear, estudou a relação entre a variável dependente: log da renda, e as variáveis independentes: nível de educação, sexo, região e atividade. De acordo com Langoni, as três primeiras justificadas pela teoria do capital humano, e as demais para refletir as diferenças nas estruturas produtivas e a existência de imperfeições de mercado. O conjunto de variáveis possibilitou explicar 51% da variância do log da renda em 1960 e 59% em 1970, ao passo que a contribuição marginal da educação para a variância da renda aumentou 33% no período.

Outros trabalhos importantes neste assunto são os de Romer (1990) e Barro (1991). Ambos apresentam o capital humano como o fator que sustenta o crescimento econômico. Como a forma de medir a qualidade da força de trabalho é principalmente pela educação, o primeiro utiliza anos de estudo, enquanto o segundo utiliza as taxas

de matrículas escolares para tal finalidade.

Já Lv et al. (2015), utilizaram a técnica econométrica de modelo de dados em painel para estimar os impactos de fatores educacionais no crescimento econômico entre 31 províncias chinesas entre os anos de 1996 e 2010.

Por fim, como citado anteriormente, Ferreira e Cruz (2010) sustentam a tese que as pessoas que não têm um certo nível inicial de renda não conseguem investir em capital humano (educação). Dessa maneira, para medir o impacto do nível de educação dos municípios brasileiros na desigualdade utilizam a média de anos de estudo num modelo de regressão não linear de efeito limiar (threshold).

3 METODOLOGIA

3.1 DADOS UTILIZADOS

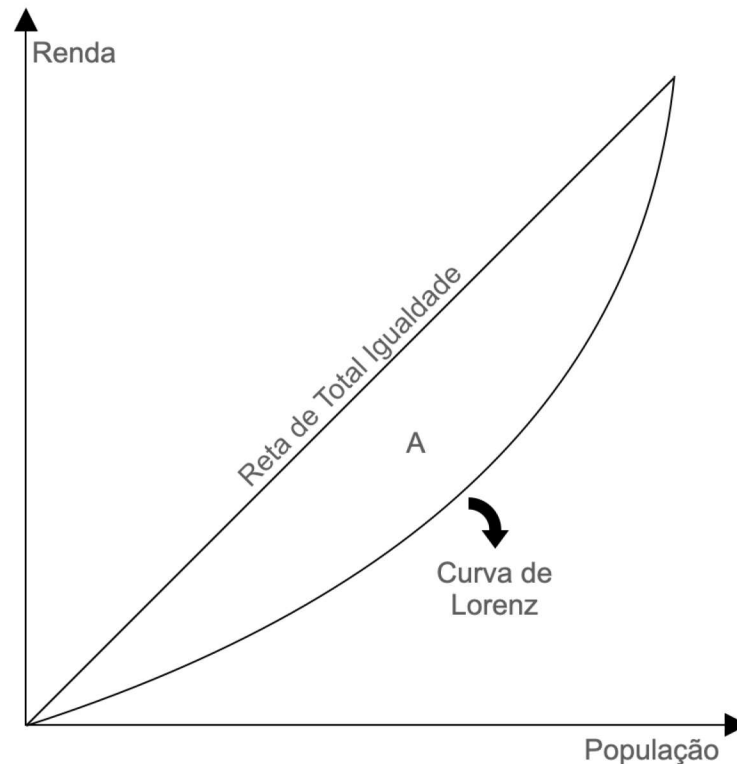
A escolha das variáveis que fariam parte deste estudo foi uma das etapas mais indispensáveis e, ao mesmo tempo, complexa. Assim, é fundamental destacar que as variáveis utilizadas foram definidas com base nos trabalhos apresentados na revisão de literatura e, além disso, foram testadas mais de 60 variáveis que potencialmente poderiam ser incluídas no estudo, no entanto, a partir da análise prévia, foram selecionadas uma variável dependente e quatro variáveis independentes: aquelas que mais se mostraram relevantes para a pesquisa.

3.1.1 Desigualdade de Renda

Este trabalho trata da desigualdade de renda dos municípios brasileiros, portanto, o indicador utilizado para tal finalidade será o Coeficiente de Gini, pois oferece uma maneira consistente para este objetivo (FARRIS, 2010). Publicado pela primeira vez por Corrado Gini em 1912, tornou-se a principal maneira de quantificar o grau de concentração da riqueza em um determinado local. Em termos teóricos, é definido como "um menos duas vezes a área entre a curva de Lorenz e a linha que representa igualdade perfeita" (LERMAN; YITZHAKI, 1984).

O índice de Gini compara a renda de uma população à uma distribuição perfeita em que todos os cidadãos têm a mesma renda. Para computar o Coeficiente de Gini, primeiro se calcula a curva de Lorenz para a população em questão. A curva é uma representação da distribuição de renda ou riqueza em uma sociedade. É útil também, pois ao contrário de muitas outras medidas de desigualdade, permite valores negativos, de maneira que se um indivíduo têm a riqueza negativa, é porque tem dívidas.

FIGURA 1 – COEFICIENTE DE GINI



Fonte: Elaboração Própria (2021).

Legenda: "A" significa a área de concentração, ou seja, o coeficiente de Gini.

De modo alternativo, a fórmula pode ser descrita matematicamente como:

$$G = \sum_{i=1}^n (Y_i + Y_{i-1})(X_i + X_{i-1}) \quad (3.1)$$

onde Y_i é a proporção acumulada da população para i , enquanto X_i é a proporção acumulada da renda para i .

O índice varia de 0 a 1, onde zero significa igualdade perfeita (em que todos os integrantes de um determinado local têm a mesma renda). De forma contrária, um valor igual a um expressa desigualdade máxima. Por exemplo, em uma população fictícia, se um único indivíduo concentra toda a renda disponível, o coeficiente de Gini será igual a um.

Os dados de desigualdade de renda foram obtidos para os 5.568 municípios brasileiros a partir da base de dados do IBGE referente ao Censo Demográfico de 2010.

3.1.2 Indicadores Educacionais

Entre dados de bases censitárias e tabelas do INEP, foram testadas mais de 60 variáveis e quatro foram escolhidas para este trabalho; entre elas a informação sobre a porcentagem da população de 11 a 14 anos de cada município que não sabe ler nem escrever um bilhete simples. Essa informação é importante, pois a quantidade de pessoas sem acesso à alfabetização "reflete a exclusão do passado, faz sombra ao presente e mina possibilidades do futuro" (COSTA, 2019).

Adicionalmente, este trabalho também utilizou a informação de entrada tardia no sistema de ensino. Foram considerados neste critério aqueles alunos que entraram na escola com 8 anos ou mais, ou seja, de forma atrasada. Esta informação está disponível nas bases de dados mais recentes do Sistema Nacional da Educação Básica, de 2019, e reflete a porcentagem de estudantes do nono ano do ensino fundamental, por município, que se encaixa neste grupo.

Também foram considerados na análise os dados de média de horas-aula diárias por município. Foram utilizados os indicadores disponibilizados pelo INEP para o ano de 2010, que contêm a média de horas-aula para a educação infantil, ensino fundamental e médio. O indicador para a educação infantil foi o que se mostrou mais importante na relação com a desigualdade social dos municípios e, por isso, nesta análise, será utilizado o indicador dessa categoria. A relação existente entre horas por dia que crianças e adolescentes passam na escola e indicadores sociais já é conhecida: Jacob e Lefgren (2003) afirmam que quando há oferta de atividades estruturadas e monitoradas para jovens, o índice de crimes juvenis tende a diminuir, enquanto Kruger, Berthelon e Navia (2009) afirmam que, no Chile, mulheres que vivem em municípios com maior disponibilidade de escolas com ensino integral têm menos chance de apresentarem gravidez na adolescência. Além disso, como afirma Menezes Filho (2012), alunos que passam mais tempo em sala de aula têm um desempenho melhor do que o outro grupo. No Brasil, o Plano Nacional da Educação define que pelo menos 50% das escolas públicas deverão ofertar ensino integral até 2024, pois entende-se que há ganhos sociais nesta ação. Portanto, este estudo buscará testar a influência da carga horária média dos municípios na desigualdade de renda do mesmos.

3.1.3 Medição do Aprendizado

A partir do que foi sugerido por Hanushek e Woessmann (2007) em relação à importância de, em estudos econométricos, mensurar a qualidade da educação e não somente a frequência nas escolas, neste trabalho as notas do Exame Nacional do Ensino Médio de 2010 (mesmo ano da medição do último censo demográfico disponível) foram utilizadas como forma de medição do capital humano, devido ao fato de que capturam o quanto os alunos aprenderam.

Os Microdados são disponibilizados pelo INEP e permitiram a exploração dos dados para os mais de 4,6 milhões de indivíduos que realizaram a prova daquele ano. Dessa forma, construiu-se uma nova variável composta pela nota média entre as áreas do conhecimento avaliadas no exame, são elas: linguagens e códigos e suas tecnologias, ciências humanas e suas tecnologias, ciências da natureza e suas tecnologias, matemática e suas tecnologias e redação. Depois disso, os resultados foram agregados para a obtenção da nota média para cada município.

3.2 ESTRATÉGIA ECONOMETRICA

3.2.1 Análise Espacial

A busca constante de novas e melhores formas de se entender a imensa quantidade de dados produzidos no século XXI se estende por todas as áreas, incluindo a Economia. Mesmo não sendo atividade principal da disciplina, atualmente o campo da análise espacial têm se tornado cada vez mais relevante. Em vista disso, a econometria espacial têm como objetivo analisar atividades e indicadores econômicos em diferentes regiões, tornando-se, assim, uma maneira importante de se entender dados com características de georreferenciamento (FISCHER; GETIS, 2010).

Existem duas maneiras principais de aplicar técnicas de análise espacial. São elas: criação de mapas que permitem a visualização de como as variáveis se relacionam no espaço, e a aplicação de estatísticas espaciais para dar suporte à análise visual. Este trabalho aplica as duas técnicas como forma de entender a relação entre desigualdade de renda e indicadores educacionais, de forma a identificar padrões espaciais.

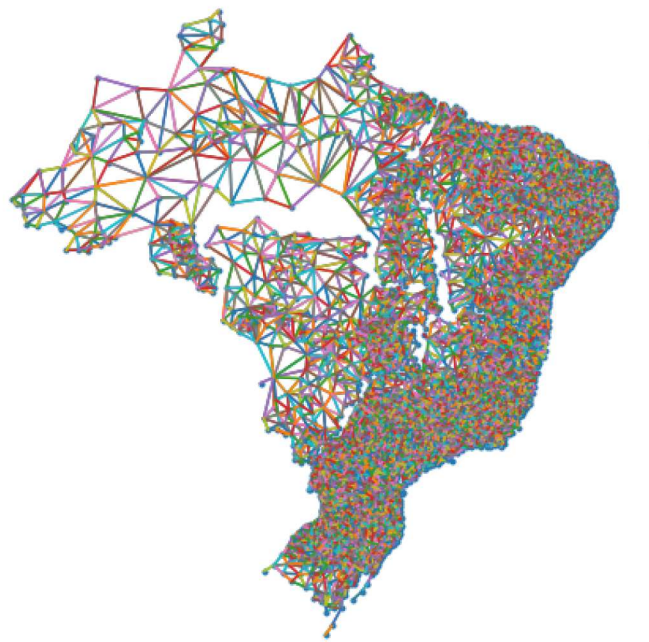
Em relação à aplicação de estatísticas espaciais, neste trabalho foi analisada a correlação entre os municípios, calculada a partir do índice de Moran Global, um dos métodos estatísticos mais difundidos para esta finalidade; este coeficiente identifica a relação existente entre um determinado polígono e a sua vizinhança, medindo a dependência espacial entre valores da mesma variável em diferentes locais do espaço por intermédio da covariância entre eles (LUZARDO; CASTAÑEDA FILHO; RUBIM, 2017).

Esta análise é útil para entender padrões espaciais entre as regiões, dando suporte estatístico às análises visuais de mapas coropléticos. Portanto, em outras palavras, o índice de Moran indica a autocorrelação espacial entre os municípios vizinhos, indicando o quanto uma variável é dependente dos valores dessa mesma variável nas localidades próximas. Dessa forma, para capturar a relação espacial entre os municípios, foi necessária a utilização de uma matriz de estrutura espacial. Os valores do índice de Moran podem variar de menos um a um, onde valores negativos indicam

correlação inversa, zero indica relação nula, e valores positivos indicam correlação positiva.

O mapa abaixo é um exemplo do que foi necessário elaborar para que fosse possível calcular a distância entre os municípios e, assim, calcular o índice de Moran. A localização dos municípios foi obtida a partir de uma base de dados pública do *website* "Exploratory", que contém as coordenadas das retas que formam os limites de cada município, assim, torna-se conhecido o local de cada cidade no espaço, tal como o formato geométrico de cada uma, tornando possível, por fim, calcular os centroides de cada município e a distância entre eles. Os cálculos e mapas, por sua vez, foram realizados na linguagem de programação *python*, utilizando a biblioteca "ESDA" e "Folium", respectivamente.

FIGURA 2 – CENTROIDE DOS MUNICÍPIOS BRASILEIROS



Fonte: Elaboração Própria (2021).

Legenda: Exemplo de técnica utilizada neste trabalho. Os pontos no mapa representam os centroides de cada município brasileiro e são utilizados para calcular a distância entre cada local.

Matematicamente, o coeficiente de Moran é descrito da seguinte maneira:

$$I_w = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} \quad i \neq j \quad (3.2)$$

onde " n " é o número de unidades espaciais (no caso deste trabalho, o número de municípios na análise); " y " é o valor da variável de interesse considerado na área " i " (no caso deste trabalho, o índice de Gini de cada município); $\bar{y}$ é o valor médio da variável de interesse na região e " W_{ij} " é a matriz de pesos espaciais da vizinhança normalizada (LUZARDO; CASTAÑEDA FILHO; RUBIM, 2017).

A matriz de pesos espaciais " W_{ij} " utilizada neste trabalho é do tipo "Rainha", que identifica relações de adjacência de forma binária, distinguindo se um polígono (limite municipal) partilha alguma borda com outro polígono ou não.

A utilização deste método provou-se proveitosa para entender o comportamento das variáveis de Gini e de proficiência estudantil na análise a seguir.

3.2.2 Análise de Regressão

Todo o processo de manipulação de dados, como a seleção das informações relevantes, tratamento de *outliers*, agregação dos dados por municípios, junção das bases de dados e demais cuidados estatísticos foram realizados através da linguagem de programação *python*. O dinamismo e poder de processamento que esta ferramenta oferece tornou possível a análise de milhões de registros.

Para que fosse possível transformar esses dados em informações relevantes, o principal instrumento de análise do presente trabalho é a regressão linear múltipla utilizando o método de mínimos quadrados ordinários. A análise de regressão com dados de corte transversal é o instrumento mais aplicado na análise empírica em economia e outras ciências sociais, sendo assim muito útil para testar hipóteses econômicas (WOOLDRIDGE, 2016), e tem como objetivo explicar a relação linear entre uma variável dependente y como função de n variáveis independentes, fazendo possível entender o relacionamento entre elas.

Em geral, pode-se descrever matematicamente um modelo com n variáveis independentes da seguinte forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_n x_n + \mu \quad (3.3)$$

onde β_0 é o intercepto, β_1 avalia a mudança em y em relação a x_1 mantendo os outros fatores fixos, β_2 avalia a mudança de y relação a x_2 mantendo os outros fatores fixos e assim por diante.

3.2.3 Aprendizado de Máquina

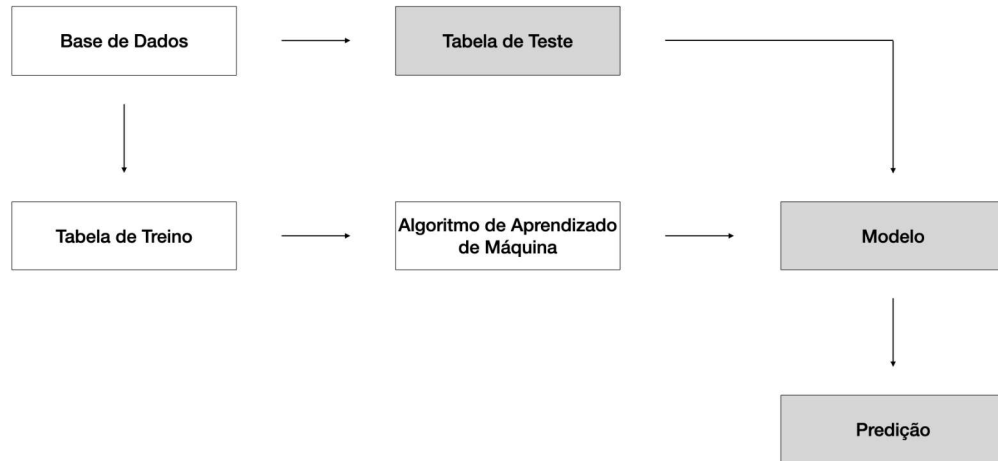
De acordo com Athey (2018), algoritmos de aprendizado de máquina terão um impacto dramático na área da Economia em pouco tempo, portanto, adicionalmente à análise de regressão linear múltipla e da análise espacial, buscou-se também, neste trabalho, a utilização de técnicas de aprendizado de máquina para corroborar a existência da relação entre os indicadores de desigualdade de renda e de educação.

Existem muitos benefícios dessa abordagem mais moderna, incluindo a capacidade de analisar muitos dados e de captar relações que outros métodos não são capazes. Contudo, diferentemente das análises anteriores, que focavam diretamente na inferência causal entre as variáveis já citadas, a abordagem e utilização de *machine learning* fornece algum tipo de predição baseada nos dados, com o objetivo de otimizar o ajuste do modelo em uma amostra de teste. Neste caso, explorou-se a capacidade de um algoritmo de descobrir, a partir das variáveis de educação, quais são os municípios com os piores desempenhos no índice de Gini. Ainda de acordo com Athey (2018), o ideal é combinar as análises econométricas mais tradicionais com os métodos de aprendizado de máquina, pois apesar deste último ser muito eficiente em descobrir padrões, não é focada em causalidade.

Não é o objetivo deste trabalho definir e abordar os detalhes do que são e como funcionam algoritmos de *machine learning*, pois é um campo muito amplo e trabalhos podem ser realizados com este exclusivo propósito. No entanto, para definir resumidamente, algoritmos de aprendizado de máquina podem ser separados em duas grandes áreas: a) algoritmos não supervisionados, que normalmente são usados para classificação de grupos e são comumente mais utilizados para fotos, vídeos e textos, e b) algoritmos supervisionados, que normalmente utilizam variáveis X para predizerem uma saída Y (ATHEY, 2018).

Algoritmos supervisionados, método que foi utilizado nesse trabalho, utilizam uma técnica de separação da base de dados entre tabelas de treino e de teste, onde o algoritmo aprende como se comportam os dados e aplica esse aprendizado na tabela de teste. Neste caso, utilizamos as mesmas variáveis preditoras utilizadas anteriormente para cada município: a) nota do Enem; b) porcentagem de analfabetismo entre jovens de 11 a 14 anos; c) porcentagem de estudantes que entraram tardiamente na escola e d) média de horas-aula por dia, enquanto a variável objetivo, aquela que o algoritmo procura predizer, é uma variável binária que identifica os municípios que estão no terceiro e quarto quartil em termos de desigualdade de renda, ou seja, a metade mais desigual do país.

FIGURA 3 – EXEMPLO DE ALGORITMO SUPERVISIONADO DE APRENDIZADO DE MÁQUINA



Fonte: (JONES, 2018).

Neste trabalho, foram utilizados diversos métodos de *machine learning* para que fosse possível a comparação da eficiência entre os modelos. São eles: a) Regressão Logística, b) Árvore de Decisão, c) *Random Forest* e d) *Gradient Boost*. Para a medição da eficiência da predição, foi utilizado o método de Característica de Operação do Receptor, ou Curva ROC, de acordo com Fawcett (2006), esta é uma técnica para visualizar, organizar e selecionar os modelos baseados nas suas performances. Este método é muito utilizado para este propósito, pois a classificação simples da acurácia normalmente é uma métrica demasiadamente fraca para a medição da performance.

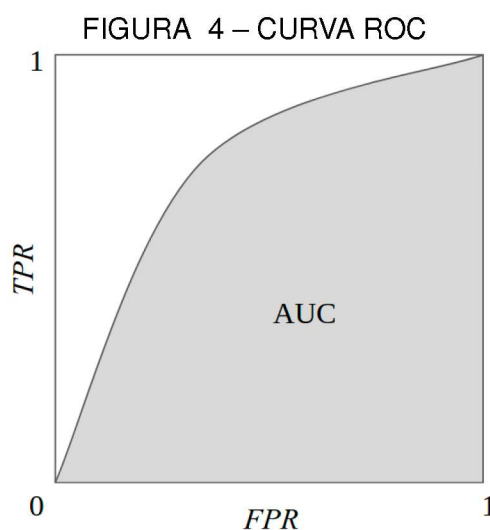
Geralmente, na medição simples da acurácia, a performance de um modelo é medida em quantas vezes ele acerta a predição em relação ao número total de registros. No entanto, este não é o melhor método para se avaliar a performance de um modelo, pois não leva em conta a capacidade do algoritmo de avaliar verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. O método da Curva ROC resolve este problema devido ao fato de considerar todas as possibilidades. O cálculo se dá por:

$$TPR = \frac{\sum \text{Verdadeiros Positivos}}{\sum \text{Positivos}} \quad (3.4)$$

$$FPR = \frac{\sum \text{Falsos Positivos}}{\sum \text{Negativos}} \quad (3.5)$$

Dessa forma, é possível elaborar um gráfico onde o eixo x representa a taxa de verdadeiros positivos (*TPR*) e o eixo y representa a taxa de falsos positivos (*FPR*).

A área abaixo da curva é chamada de AUC (*Area Under the Curve*) e representa a capacidade de um modelo de fazer previsões corretas.



Fonte: (PAULINO, 2018)

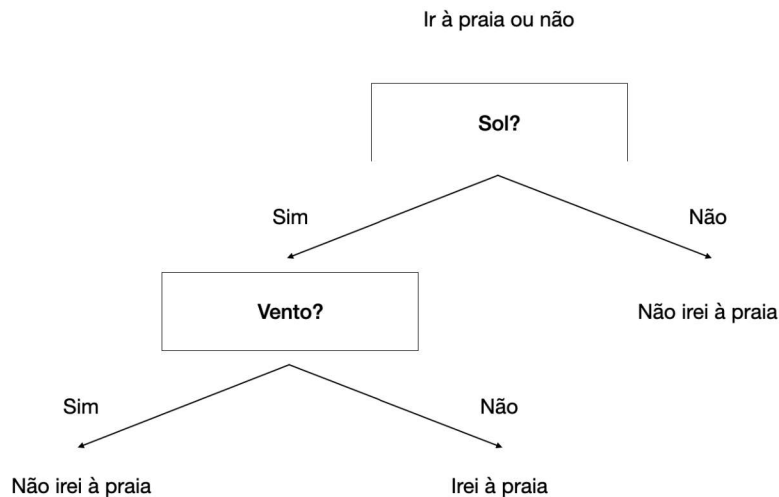
Assim, um modelo em que todas as previsões estão erradas tem um resultado de AUC igual a zero, enquanto um modelo que acerta todas as previsões tem AUC igual a um.

3.2.3.1 Métodos Utilizados

Em relação aos métodos de aprendizado de máquina, como mencionado anteriormente, foram utilizados quatro, são eles: a) Regressão Logística, b) Árvore de Decisão, c) *Random Forest* e d) *Gradient Boost*. Primeiramente, a Regressão Logística é um recurso que permite estimar a probabilidade associada à ocorrência de determinado evento (neste caso, do município ter um desempenho ruim no índice de Gini), em face de um conjunto de variáveis explanatórias (variáveis educacionais). O que distingue um modelo de regressão logística de um modelo de regressão linear múltipla é que, na regressão logística, além da relação entre os dados não ser linear, a variável resposta é binária (HOSMER JR.; LEMESHOW; STURDIVANT, 2013). Uma vez que essas diferenças são levadas em consideração, os dois métodos são muito parecidos em seus princípios estatísticos, no entanto, atendendo a propósitos diferentes.

O método de Árvore de Decisão, por sua vez, utiliza os dados para criar um modelo em forma de árvore e, dessa forma, pode prever a chance de um evento acontecer ou não. Nesta "árvore", cada nó representa um teste, e cada teste gera uma nova subárvore (MONARD; BARANAUSKAS, 2003), de maneira que à medida que o algoritmo realiza os testes nas variáveis preditoras, calcula a probabilidade do evento da variável resposta acontecer ou não. A Figura 5 representa um simples exemplo fictício de como funciona uma árvore de decisão.

FIGURA 5 – EXEMPLO DE ÁRVORE DE DECISÃO



Fonte: (TECH, 2021)

Já o método de *Random Forest* cria uma massiva quantidade de Árvores de Decisões, onde o resultado é a previsão da maioria das árvores. Por exemplo, nos dados deste trabalho, este método cria diversas árvores onde cada uma calcula a probabilidade de determinado município ter um desempenho ruim no índice de Gini ou não, dessa forma, caso a maior parte delas preveja que a cidade tem um desempenho ruim, o modelo selecionará essa resposta (BREIMAN, 2001).

Por fim, neste trabalho também foi utilizado o método de *Gradient Boost*. De forma descomplicada, este tipo de algoritmo faz uma combinação de modelos simples para criar um modelo robusto que lida melhor com grandes quantidades de dados e relações complexas. O modelo começa fazendo previsões com uma regressão ou uma árvore de decisão, depois, aplica um novo modelo que foca exclusivamente em casos em que a previsão não foi boa. Este processo é repetido diversas vezes, em que cada repetição consiste na tentativa de utilização de um novo modelo que seja melhor do que o anterior, minimizando o erro (HOARE, 2017). Essas diversas combinações geralmente fazem com que este seja um dos métodos mais eficazes na previsão de eventos.

4 RESULTADOS E DISCUSSÃO

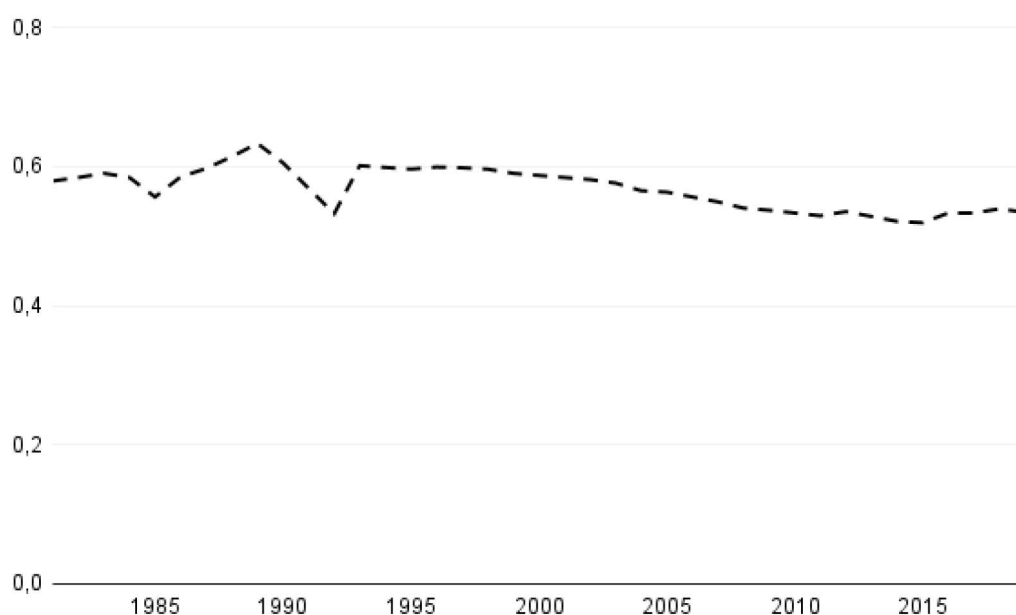
4.1 ESTATÍSTICAS DESCRITIVAS DAS VARIÁVEIS ESTUDADAS

Aqui se apresentam os resultados e informações extraídos dos dados. Importante ressaltar que todos os cálculos e estimações apresentados neste trabalho foram realizados no nível municipal, não sendo objetivo deste estudo a análise dos resultados entre os estados ou macrorregião. No entanto, para facilitar a apresentação, os mapas aqui exibidos serão por estado.

4.1.1 Coeficiente de Gini

Como relatado anteriormente, após décadas de avanço na construção de uma economia com menor concentração de renda, o índice de Gini brasileiro atingiu em 2018 sua pior marca em uma década: 0,539.

FIGURA 6 – ÍNDICE DE GINI 1981 - 2019



Fonte: (BANCO MUNDIAL, 2019).

A partir da base de dados do censo de 2010, é possível verificar quais são os melhores e os piores municípios brasileiros em termos de desigualdade de renda. As tabelas consideram apenas cidades com mais de 50 mil habitantes.

TABELA 1 – MUNICÍPIOS MENOS DESIGUAIS

Posição	UF	Cidade	GINI
1	MG	Nova Serrana	0,35
2	PR	Sarandi	0,35
3	RS	Parobe	0,37
4	SC	Gaspar	0,37
5	SC	Biguaçu	0,39
6	MG	Ribeirão das Neves	0,39
7	SP	Várzea Paulista	0,39
8	SC	Indaial	0,39
9	SP	Francisco Morato	0,40
10	SP	Votorantim	0,40
11	SC	Brusque	0,40
12	SC	Camboriú	0,40
13	SC	Içara	0,40
14	SC	Palhoça	0,40
15	SP	Hortolândia	0,41

FONTE: IBGE (2010)

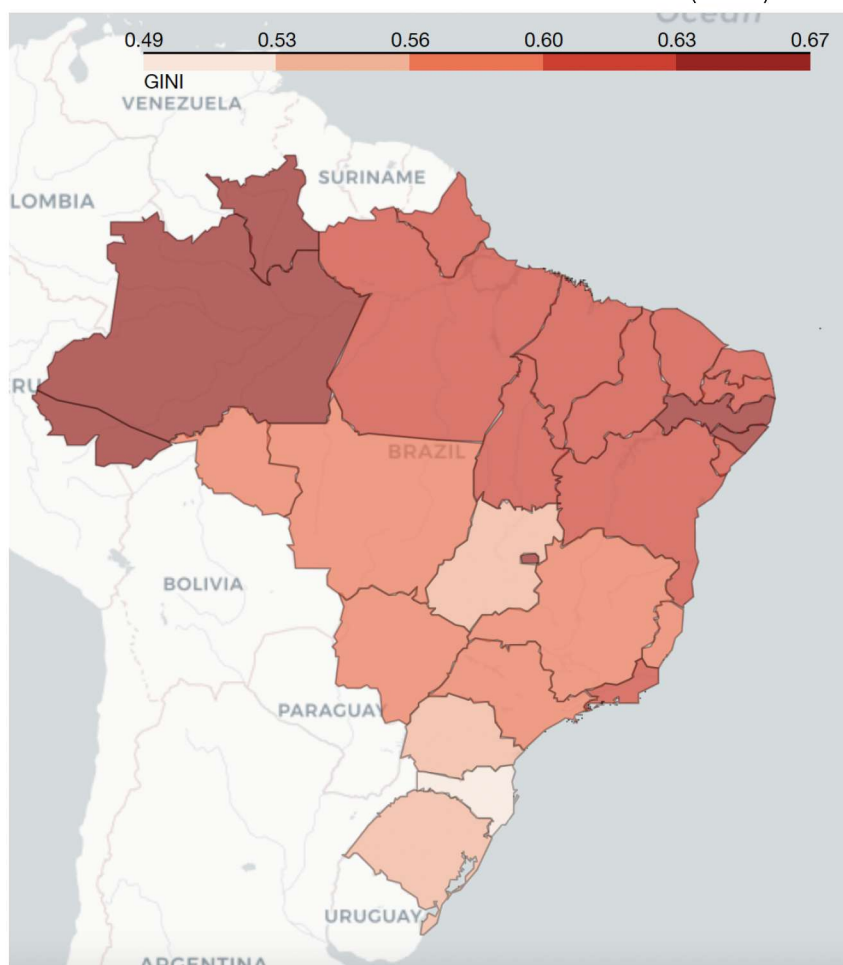
TABELA 2 – MUNICÍPIOS MAIS DESIGUAIS

Posição	UF	Cidade	GINI
1	PB	Cabedelo	0,70
2	PE	Recife	0,68
3	MG	Nova Lima	0,68
4	SP	Santana de Parnaíba	0,67
5	BA	Tucano	0,65
6	AM	Tabatinga	0,65
7	AC	Cruzeiro do Sul	0,64
8	AM	Coari	0,64
9	AM	Maués	0,64
10	PA	Oriximiná	0,64
11	PA	Portel	0,64
12	PA	Moju	0,63
13	AL	Maceió	0,63
14	BA	Lauro de Freitas	0,63
15	BA	Salvador	0,63

FONTE: IBGE (2010)

Em relação aos estados brasileiros, considerando dados de 2010, Santa Catarina é o estado com a menor concentração de renda (0,4942), enquanto o estado de Amazonas é o mais desigual (0,6664). A média do coeficiente de Gini para os municípios brasileiros é de 0,49. Esse número é muito alto quando comparado aos países que compõem a OCDE (Organização para a Cooperação e Desenvolvimento Econômico). A média desses países é de 0,32 (FERNANDES, 2015). Ou seja, os países da OCDE, na média, são consideravelmente menos desiguais, inclusive, do que o estado menos desigual do Brasil. Para melhor visualização, foi produzido um mapa com a distribuição do índice de Gini para os estados.

FIGURA 7 – ÍNDICE DE GINI POR ESTADO (2010)



Fonte: Elaboração Própria com dados do DATASUS (2010).

4.1.2 Analfabetismo Entre Pessoas de 11 a 14 Anos

Também disponível através das bases censitárias, o indicador da porcentagem da população de 11 a 14 anos de cada município que não sabe ler nem escrever um bilhete simples foi muito importante para a análise. Abaixo, estão listadas as piores e melhores cidades neste quesito. As tabelas consideram apenas aquelas com mais de 50 mil habitantes.

TABELA 3 – MUNICÍPIOS COM MENORES INCIDÊNCIAS DE ANALFABETISMO ENTRE JOVENS

Posição	UF	Cidade	Analfabetismo (%)
1	SC	Indaial	0,39
2	MG	Itaúna	0,49
3	MG	Poços de Caldas	0,60
4	RS	Farroupilha	0,62
5	PR	União da Vitória	0,63
6	MG	São João Del Rei	0,65
7	PR	Cianorte	0,65
8	MG	Divinópolis	0,66
9	SP	Assis	0,66
10	PR	Arapongas	0,69
11	PR	Rolândia	0,69
12	RS	Campo Bom	0,72
13	SP	Vinhedo	0,73
14	SC	Balneário Camboriú	0,73
15	PR	Cambé	0,73

FONTE: IBGE (2010)

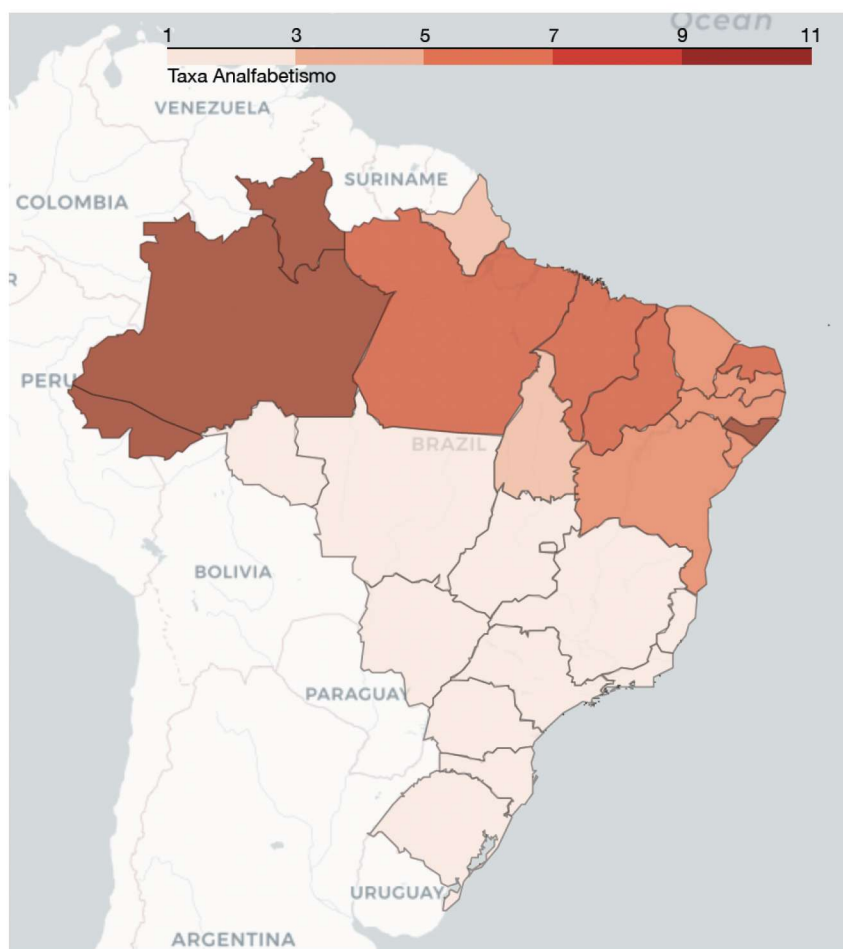
TABELA 4 – MUNICÍPIOS COM MAIORES INCIDÊNCIAS DE ANALFABETISMO ENTRE JOVENS

Posição	UF	Cidade	Analfabetismo (%)
1	PA	Portel	20,3
2	PA	Acará	16,3
3	PA	Breves	15,9
4	PE	Buíque	15,1
5	BA	Araci	15,1
6	CE	Icó	13,8
7	PE	Ouricuri	13,0
8	MA	Tutóia	12,8
9	PA	Viseu	12,6
10	PA	Igarapé-Mirim	12,36
11	PA	Moju	12,04
12	AL	União dos Palmares	11,7
13	PA	Itupiranga	11,3
14	BA	Monte Santo	11,2
15	MA	Buritcupu	10,7

FONTE: IBGE 2010)

A média deste indicador, levando em consideração todos os municípios, é de 3,69%. Em termos de média dos municípios de cada estado, Santa Catarina é o estado que apresenta o menor valor, com 1,15% e o pior é o Amazonas, com 11,1%. Para melhor visualização, foi produzido um mapa com a distribuição deste indicador para os estados.

FIGURA 8 – ANALFABETISMO ENTRE ADOLESCENTES POR ESTADO (2010)



Fonte: Elaboração Própria com dados do IBGE 2010.

4.1.3 Entrada Tardia no Sistema de Ensino

A informação da taxa de alunos do nono ano do ensino fundamental que entraram tardiamente no sistema de ensino mostrou-se importantíssima para a relação dos indicadores educacionais com o índice de desigualdade de renda. Encaixam-se nesse grupo aquelas crianças que entraram na escola com oito anos de idade ou mais. Abaixo, estão listadas as piores e melhores cidades neste quesito. As tabelas consideram apenas aquelas com mais de 50 mil habitantes.

TABELA 5 – MUNICÍPIOS COM MENORES INCIDÊNCIAS DE ENTRADA TARDIA NA ESCOLA

Posição	UF	Cidade	Entrada Tardia (%)
1	PA	Santana do Araguaia	0
2	RS	Canguçu	0,27
3	CE	Acopiara	0,37
4	BA	Guanambi	0,40
5	SP	Mogi Mirim	0,42
6	PR	Irati	0,58
7	RJ	Valença	0,59
8	SP	Itapira	0,59
9	RN	Ceará-Mirim	0,62
10	SP	Ourinhos	0,62
11	MG	São João Del Rei	0,63
12	PE	Salgueiro	0,65
13	PI	Floriano	0,68
14	SP	Leme	0,68
15	BA	Paulo Afonso	0,69

FONTE: SAEB (2019)

TABELA 6 – MUNICÍPIOS COM MAIORES INCIDÊNCIAS DE ENTRADA TARDIA NA ESCOLA

Posição	UF	Cidade	Entrada Tardia (%)
1	SC	Lages	6,92
2	PA	Belém	6,49
3	RJ	Campos dos Goytacazes	6,22
4	PB	Bayeux	5,74
5	RJ	Niterói	5,62
6	MT	Cuiabá	5,38
7	PA	Breves	5,20
8	AP	Santana	5,17
9	RJ	São Gonçalo	5,14
10	PA	Novo Repartimento	5,14
11	RJ	São João de Meriti	5,03
12	MA	Caxias	4,98
13	PB	Santa Rita	4,97
14	AM	Tabatinga	4,89
15	CE	Maracanau	4,82

FONTE: SAEB (2019)

A média deste indicador, levando em consideração todos os alunos que responderam a prova do SAEB, é de 2,36%. Em termos de estados, o Espírito Santo é o estado que apresenta o menor valor, com 1,33% e o pior é o Amapá, com 3,96%. Para melhor visualização, foi produzido um mapa com a distribuição deste indicador para os estados.

TABELA 7 – MUNICÍPIOS COM MENORES MÉDIAS DE HORAS-AULA POR DIA NA EDUCAÇÃO INFANTIL

Posição	UF	Cidade	Horas
1	MA	Santa Inês	3,1
2	MA	Santa Luzia	3,3
3	PI	Floriano	3,6
4	BA	Jaguaquara	3,6
5	MA	Chapadinha	3,6
6	MA	Coroata	3,6
7	PA	Itupiranga	3,6
8	MA	Paço do Lumiar	3,6
9	PA	Viseu	3,6
10	MA	Zé Doca	3,7
11	PA	Moju	3,7
12	MA	Pinheiro	3,7
13	PA	Ipixuna do Pará	3,8
14	AM	Parintins	3,8
15	PA	Breu do Branco	3,8

FONTE: INEP (2010)

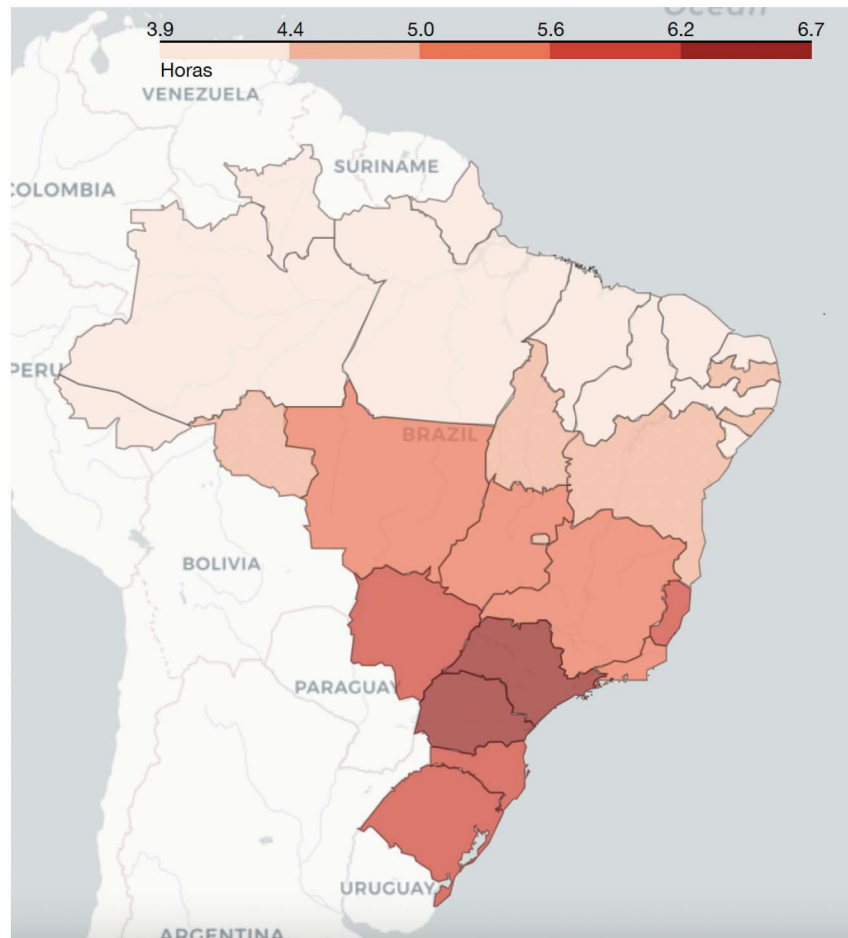
TABELA 8 – MUNICÍPIOS COM MAIORES MÉDIAS DE HORAS-AULA POR DIA NA EDUCAÇÃO INFANTIL

Posição	UF	Cidade	Horas
1	RS	Campo Bom	11,5
2	RS	Cachoeirinha	10,9
3	SC	Rio do Sul	10,6
4	RS	Lajeado	10,6
5	GO	Caldas Novas	10,4
6	RS	Venâncio Aires	10,2
7	RS	Canoas	10,1
8	RS	Alvorada	10,0
9	SP	Votuporanga	10,0
10	SC	Blumenau	9,9
11	RS	Santa Cruz do Sul	9,7
12	SC	Navegantes	9,6
13	SP	Praia Grande	9,6
14	SC	Criciúma	9,5
15	SP	Biguaçu	9,4

FONTE: INEP (2010)

A média deste indicador, levando em consideração todos os municípios do Brasil, é de 5,27 horas. Em termos de estados, o Paraná é o estado que apresenta o maior valor, com 6,73 horas e o pior é o Piauí, com 3,88 horas. Para comparação, a média de horas por dia que um aluno passa na escola nos Estados Unidos é de 6,64 (U.S. DEPARTMENT OF EDUCATION, 2008), portanto, o melhor estado brasileiro apresenta um número consideravelmente bom, no entanto, todos os outros estados estão abaixo da média norte-americana. Para melhor visualização, foi produzido um mapa com a distribuição deste indicador para os estados.

FIGURA 10 – MÉDIA DE HORAS-AULA POR DIA NA EDUCAÇÃO INFANTIL POR ESTADO (2010)



Fonte: Elaboração Própria com dados do INEP 2010.

4.1.5 Média da Nota do Enem

Este indicador foi utilizado como uma aproximação para a medição do capital humano devido ao fato de capturar, de fato, o quanto os alunos de cada cidade aprenderam durante a vida escolar. Abaixo, estão listadas as piores e melhores cidades neste quesito. As tabelas consideram apenas aquelas com mais de 50 mil habitantes.

TABELA 9 – MUNICÍPIOS COM MAIORES MÉDIAS NA NOTA DO ENEM

Posição	UF	Cidade	Nota
1	RJ	Niterói	580,86
2	SC	Florianópolis	577,65
3	RJ	Nova Friburgo	574,37
4	SP	Valinhos	573,24
5	SP	Vinhedo	572,99
6	MG	Itajubá	572,68
7	SP	São Caetano do Sul	572,14
8	MG	Juiz de Fora	571,95
9	MG	Araxá	571,57
10	SP	Araraquara	570,78
11	SP	Botucatu	568,87
12	MG	Patos de Minas	567,63
13	SP	São João da Boa Vista	567,24
14	MG	Formiga	565,76
15	MG	Barbacena	565,71

FONTE: INEP (2010)

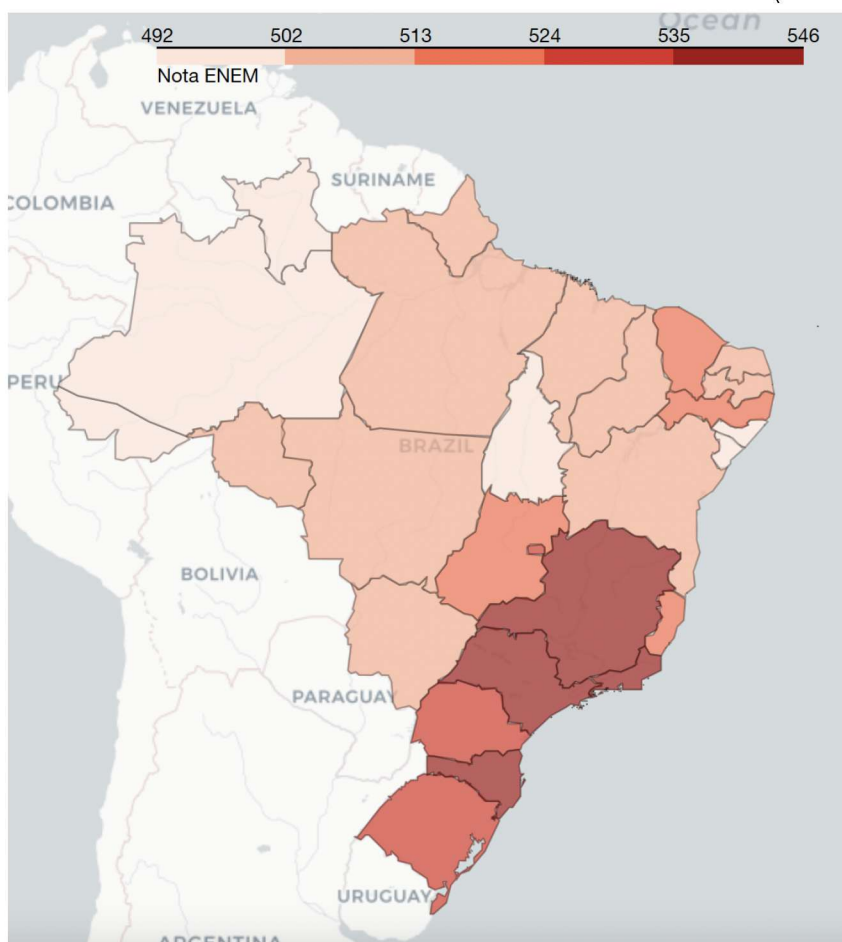
TABELA 10 – MUNICÍPIOS COM MENORES MÉDIAS NA NOTA DO ENEM

Posição	UF	Cidade	Nota
1	MA	Grajaú	451,54
2	AM	Coari	453,70
3	CE	Trairi	464,12
4	MA	Barreirinhas	466,79
5	PA	Monte Alegre	466,87
6	CE	Granja	469,06
7	AM	Itacoatiara	469,81
8	AM	Tefé	470,15
9	PA	Alenquer	470,62
10	MA	Barra do Corda	473,09
11	PA	Ipixuna do Pará	473,09
12	MA	Tutóia	473,53
13	AM	Manacapuru	474,42
14	CE	Itapipoca	474,97
15	MA	Itapecuru Mirim	474,65

FONTE: INEP (2010)

A média deste indicador, levando em consideração todos os municípios do Brasil, é de 504,58. Em termos de estados, o Rio de Janeiro é o estado que apresenta o maior valor, com 546,30; o pior é o Acre, com 491,53. Para melhor visualização, foi produzido um mapa com a distribuição deste indicador para os estados.

FIGURA 11 – MÉDIA DA NOTA DO ENEM POR ESTADO (2010)



Fonte: Elaboração Própria com dados do INEP 2010.

4.2 ANÁLISE ESPACIAL

4.2.1 Índice de Moran

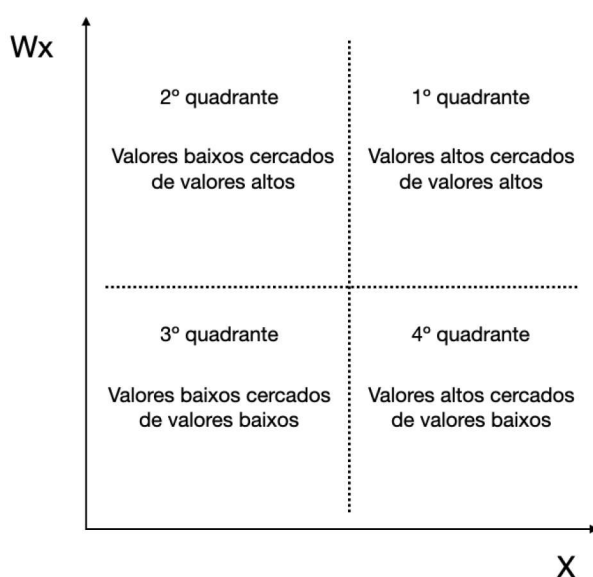
Na primeira parte da análise espacial, o índice de Moran foi utilizado para verificar empiricamente a correlação dos indicadores levando em consideração a proximidade dos municípios. Esta análise se mostrou útil para entender os padrões espaciais entre as variáveis e mostrar o quanto elas são dependentes da região em que estão inseridas. A técnica foi executada para todos os 5.568 municípios, englobando os cinco indicadores presentes neste trabalho: a) Coeficiente de Gini, b) Analfabetismo Entre Pessoas de 11 a 14 Anos, c) Entrada Tardia no Sistema de Ensino, d) Horas-Aula por Dia e e) Média da Nota do Enem.

Caso os dados apresentem a tendência de se agruparem (valores altos próximos de valores altos ou valores baixos próximos a valores baixos), o índice de Moran será positivo. Quando valores altos "repelirem" outros valores altos e apresentarem tendência de estar próximos de valores baixos, o indicador será negativo. Importante notar que, por ser uma ferramenta de inferência estatística, o resultado da análise deve

ser sempre analisado à luz do contexto da hipótese nula. Neste caso, a hipótese nula é de que o atributo sendo analisado é distribuído randomicamente (não existe correlação espacial) (ARCGIS PRO, 2021).

Adicionalmente, foi utilizado o *Moran Scatterplot*, um gráfico de dispersão dividido em quadrantes que indica a distribuição dos municípios nos grupos da análise, como demonstrado pelo exemplo abaixo:

FIGURA 12 – EXEMPLO DE DISPERSÃO PARA O ÍNDICE DE MORAN



Fonte: Elaboração Própria (2021).

Os resultados dos cálculos do índice de Moran, para os cinco indicadores analisados, foram:

TABELA 11 – RESULTADOS DO ÍNDICE DE MORAN PARA OS INDICADORES ANALISADOS

Variável	Moran	valor-p	z-score
a) Coeficiente de Gini	0,4931	0,0000	60,78
b) Analfabetismo	0,7274	0,0000	89,53
c) Entrada Tardia	0,1298	0,0000	16,00
d) Horas-Aula diárias	0,4948	0,0000	60,91
e) Média do Enem	0,6191	0,0000	76,22

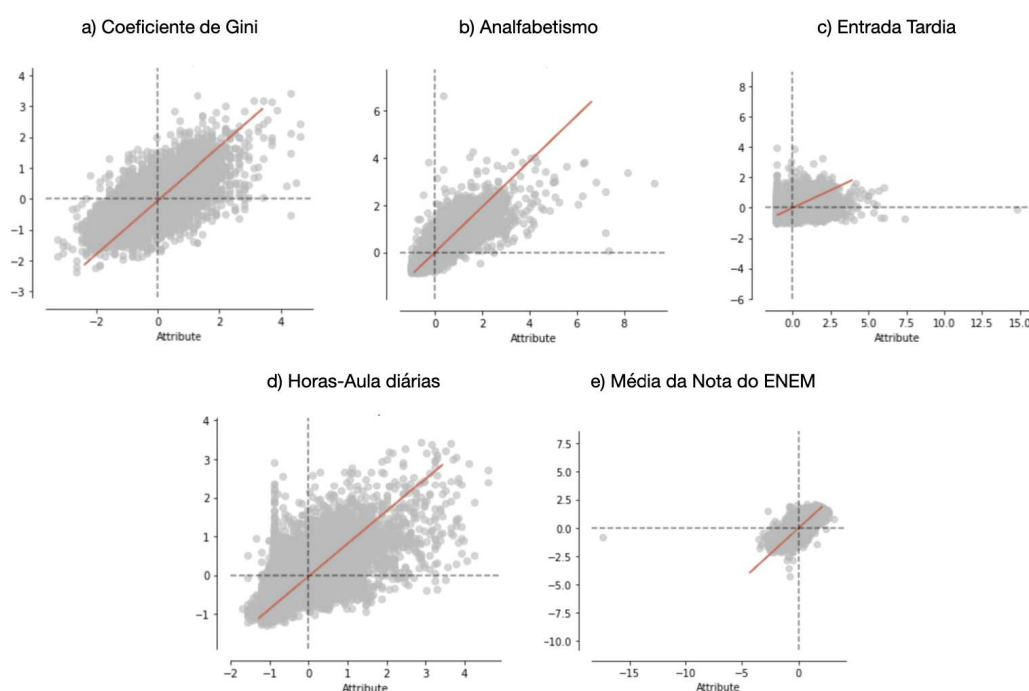
FONTE: Elaboração Própria (2021)

Como observado, para todas as variáveis, o P-valor é menor do que 0,01 e Z-score é maior do que zero; pode-se dizer, portanto, que ao nível de 1% de significância há fortes indícios contra a hipótese nula, dessa forma, todos os resultados são estatisticamente significantes. Ou seja, as distribuições espaciais são mais agrupadas do que o esperado se os dados fossem randômicos (não correlacionados).

A partir disso, observa-se que o indicador mais correlacionado espacialmente é o indicador de analfabetismo entre jovens: a correlação de 0,7274 demonstra uma relevante dependência dos municípios em relação à região em que estão inseridos, ou seja, cada município, em relação à porcentagem da população de 11 a 14 anos que é analfabeta, tende a estar ao lado de um município que apresenta valores próximos.

A média da nota do Enem também se mostrou bastante correlacionada, com um valor de 0,6191. Seguida pelos indicadores de horas-aula diárias (0,4948) e do coeficiente de Gini (0,4931). Todos estes resultados demonstram uma dependência regional considerável. Por último, o indicador de entrada-tardia, com uma correlação de 0,1298, mostrou-se o indicador menos correlacionado espacialmente.

FIGURA 13 – GRÁFICOS DE DISPERSÃO DE MORAN PARA OS INDICADORES ANALISADOS

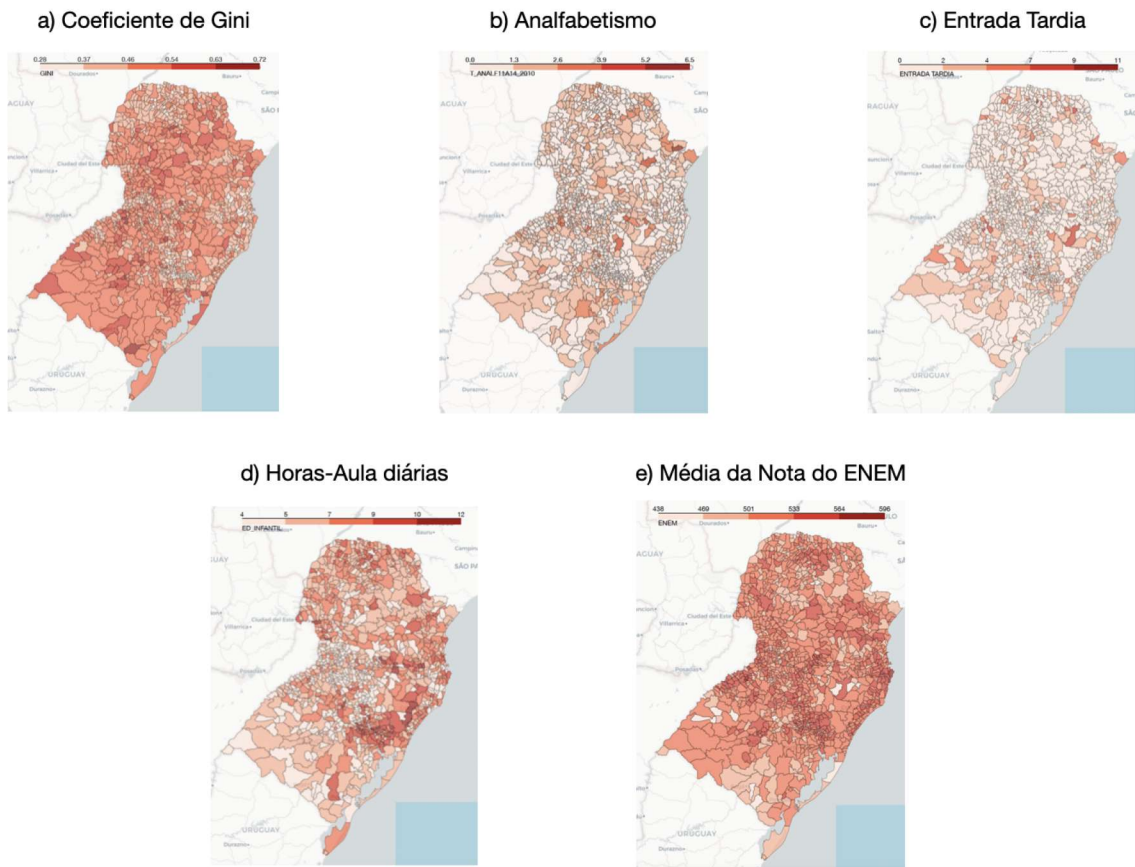


Fonte: Elaboração Própria (2021).

Os gráficos de dispersão, por sua vez, auxiliam na exploração de como é o comportamento da correlação de cada variável. Os gráficos acima mostram que, para a variável do coeficiente de Gini, há um certo equilíbrio, pois da mesma forma que municípios com altos índices de desigualdade tendem a estar cercados de municípios também mais desiguais, municípios menos desiguais tendem a estar próximos de municípios com um índice menor. Já para as variáveis de analfabetismo e de horas diárias de estudo, a correlação está mais concentrada naquelas cidades com altos indicadores, que tendem a estar próximas de cidades que também apresentam valores maiores.

4.2.2 Região Sul

FIGURA 14 – DISTRIBUIÇÃO DOS INDICADORES PARA A REGIÃO SUL

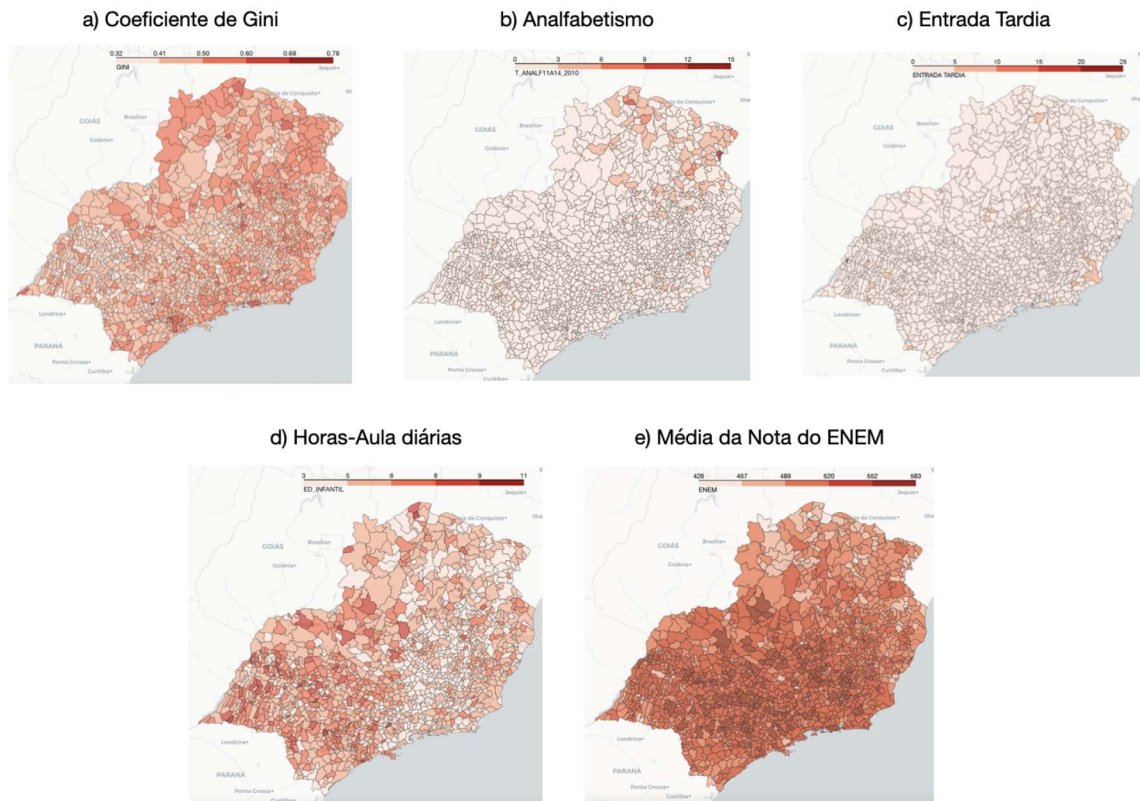


Fonte: Elaboração Própria (2021).

A partir das distribuições dos indicadores para a Região Sul, é possível observar que os locais com as maiores taxas de desigualdade de renda são as regiões do centro e sul do Rio Grande do Sul; serra catarinense; e centro-sul e campos gerais do Paraná. Ao relacionar com os indicadores educacionais, verifica-se uma clara associação com os indicadores de horas-aula diárias e média da nota do Enem, de maneira que as regiões citadas como as piores em termos de desigualdade de renda são as mesmas que apresentam os piores números, comparativamente, em termos de horas-aula diárias. Em relação ao indicador de média da nota do Enem, o sul do Rio Grande do sul e o centro-sul e campos gerais do Paraná são os que apresentam as piores médias.

4.2.3 Região Sudeste

FIGURA 15 – DISTRIBUIÇÃO DOS INDICADORES PARA A REGIÃO SUDESTE



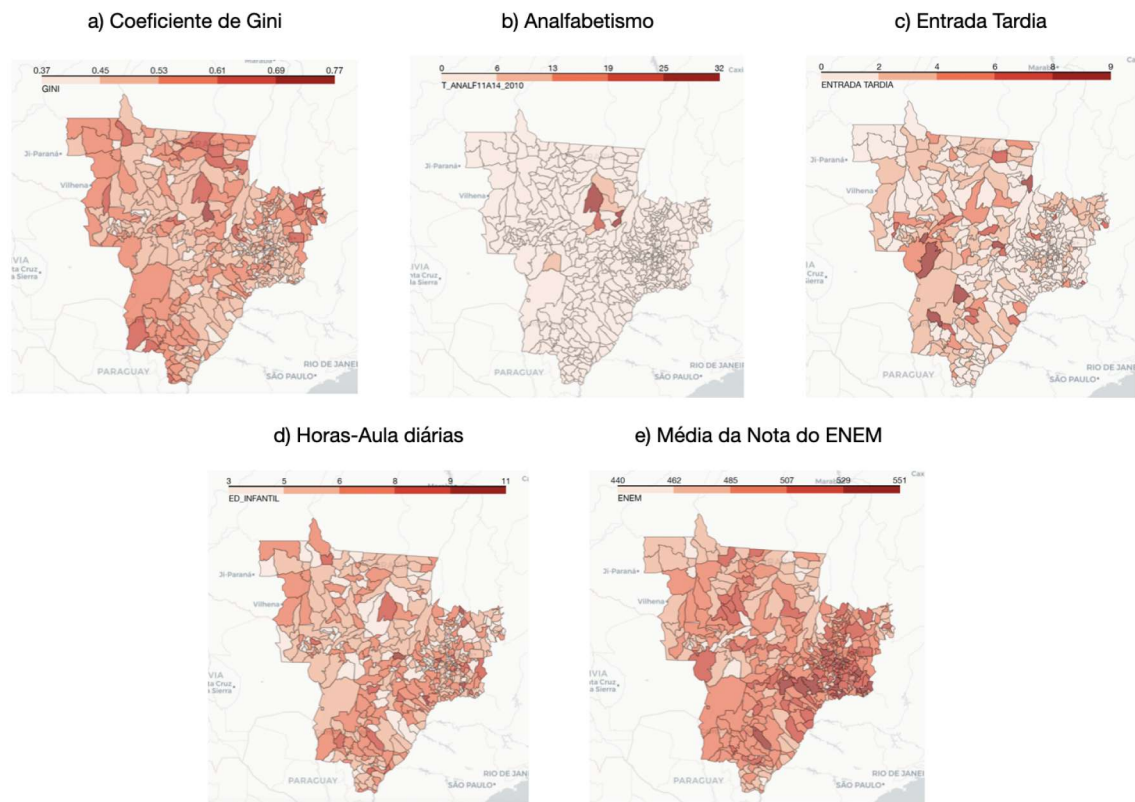
Fonte: Elaboração Própria (2021).

A partir dos gráficos da Região Sudeste, é possível observar que a região do interior de São Paulo é relativamente menos desigual do que as outras localidades, incluindo a faixa litorânea do estado, Rio de Janeiro, Espírito Santo e Minas Gerais. O norte de Minas Gerais, é o local que apresenta a maior concentração de municípios mais desiguais. Ao relacionar com os indicadores educacionais, observa-se que o norte de Minas Gerais também é a região com a maior incidência de analfabetismo entre jovens e o local com a menor média nas notas do Enem.

Ademais, destaca-se que o estado de São Paulo, além de apresentar os menores índices de desigualdade, também apresenta uma maior média de horas de aula por dia para a educação infantil, além de ter praticamente todos os municípios com uma média elevada na nota do Enem.

4.2.4 Região Centro-Oeste

FIGURA 16 – DISTRIBUIÇÃO DOS INDICADORES PARA A REGIÃO CENTRO-OESTE

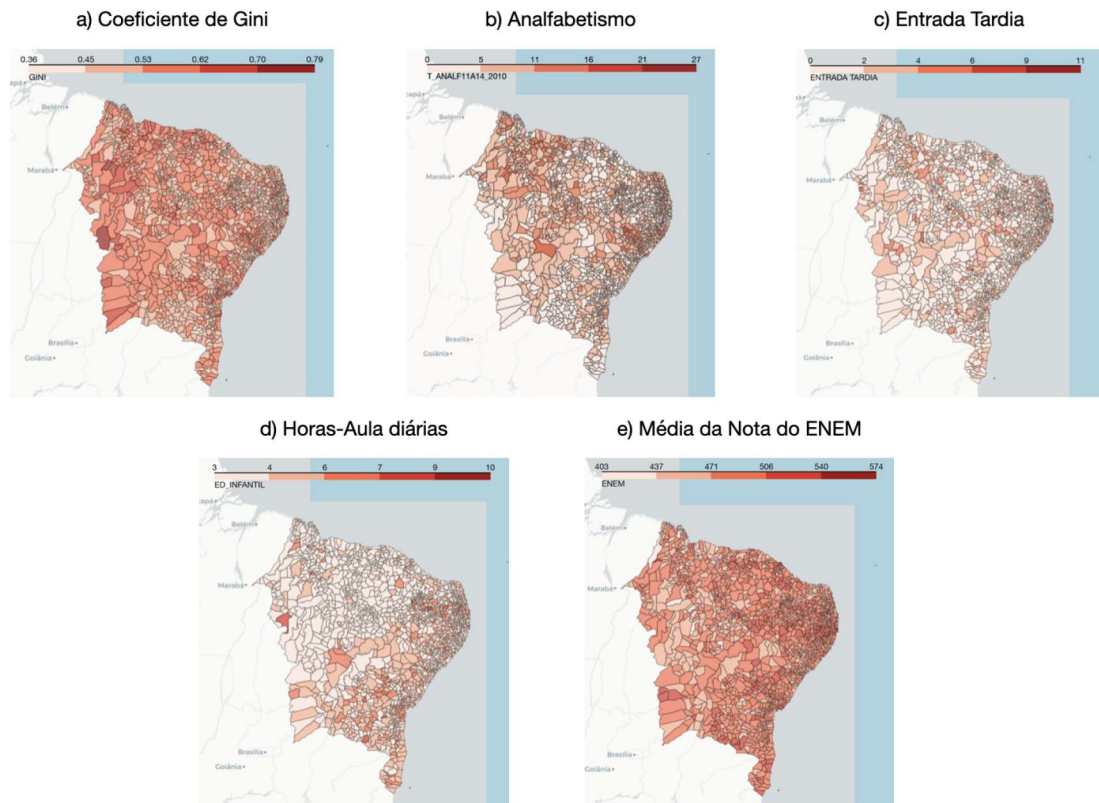


Fonte: Elaboração Própria (2021).

Os mapas da Região Centro-Oeste mostram como os municípios que fazem divisa com Paraguai, Bolívia, e Regiões Norte e Nordeste são os mais desiguais em relação aos municípios mais centrais e àqueles que fazem divisa com a Região Sudeste. Ao comparar com os indicadores educacionais, aparentam ser as mesmas regiões com números maiores de alunos com entrada tardia no sistema de ensino, e as mesmas regiões com as menores médias na nota do Enem.

4.2.5 Região Nordeste

FIGURA 17 – DISTRIBUIÇÃO DOS INDICADORES PARA A REGIÃO NORDESTE

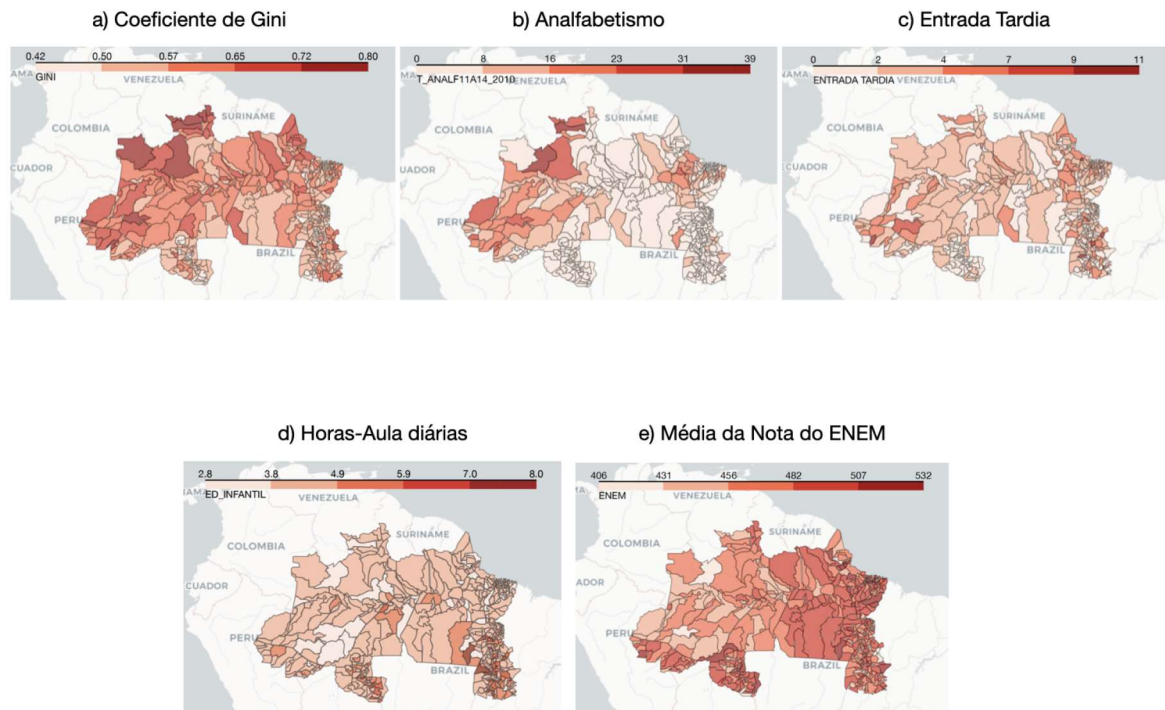


Fonte: Elaboração Própria (2021).

Os municípios da Região Nordeste mais ao Oeste, que fazem divisa com as Regiões Norte e Centro-Oeste, apresentam mais desigualdade de renda do que os municípios mais próximos do litoral e da Região Sudeste. Essas também parecem ser as regiões que apresentam os maiores valores para o nível de horas de aula por dia e as melhores médias na nota do Enem.

4.2.6 Região Norte

FIGURA 18 – DISTRIBUIÇÃO DOS INDICADORES PARA A REGIÃO NORTE



Fonte: Elaboração Própria (2021).

A Região Norte talvez seja a que apresenta padrões mais claros nas distribuições dos indicadores de cada município. Os municípios do Acre, extremo oeste do estado do Amazonas e Roraima são as regiões que apresentam os maiores índices de desigualdade de renda da região e, da mesma forma, são as regiões com os maiores índices de analfabetismo entre jovens, menores valores para o nível de horas de aula por dia e as piores médias na nota do Enem.

4.3 ANÁLISE DE REGRESSÃO

Para confirmar a análise visual e para que seja possível entender a relação das variáveis educacionais com a desigualdade de renda mais adequadamente, neste trabalho foi utilizado o método da regressão linear múltipla utilizando o método de mínimos quadrados ordinários para os dados dos municípios brasileiros.

Foram realizados testes para que se atendessem a validade da inferência estatística, como a quantificação da multicolinearidade através do teste VIF, *Variance Inflation Factor*. Este teste consiste em verificar o quanto a variância de um coeficiente de regressão estimado aumenta se as suas preditoras estiverem correlacionadas. Como há o pressuposto da independência entre as variáveis preditoras, esta avaliação se mostra primordial.

Além disso, foi realizada a verificação do pressuposto da Homocedasticidade através do teste de *Breusch–Pagan*, que verifica se os resíduos apresentam variação constante. Caso essa premissa não seja atendida, verifica-se a presença de Heterocedasticidade, o que faz com que seja mais complicado determinar o verdadeiro desvio padrão dos erros, comprometendo o resultado dos intervalos de confiança.

Por fim, foi realizado o teste para a verificação da suposição de normalidade dos resíduos, através do método de *Jarque-Bera*. Caso este requisito não seja atendido, os resultados de erros-padrão e intervalos de confiança são atrapalhados.

Para que todos esses pressupostos fossem atendidos e se pudesse confirmar a validade da regressão, foi necessária a realização da transformação dos dados em *dummies*, variáveis binárias que indicam a ocorrência de um evento ou não. Além disso, nem todas as variáveis puderam ser utilizadas na mesma regressão, fazendo-se a necessidade, portanto, da realização de dois modelos de regressão, a serem analisados independentemente. Assim, todos os requisitos estatísticos foram validados e a análise se mostrou adequada.

Em síntese, as variáveis utilizadas foram:

TABELA 12 – DESCRIÇÃO DO BANCO DE DADOS

Variável	Descrição	Fonte
Y_i	Medida de desigualdade de renda. Varia de 0 a 1.	IBGE
<i>FLAG_ANALFABETISMO</i>	"1" para municípios acima da mediana em analfabetismo.	IBGE
<i>FLAG_ENTRADA</i>	"1" para municípios acima da mediana em entrada tardia.	SAEB
<i>FLAG_HORAS</i>	"1" para municípios acima da mediana em horas-aula.	INEP
<i>FLAG_ENEM</i>	"1" para municípios no primeiro quartil em média de nota.	INEP

FONTE: Elaboração Própria (2021)

As tabelas abaixo exibem os resultados das estimações via MQO.

TABELA 13 – PRIMEIRA REGRESSÃO

	Coeficiente
<i>Constante</i>	0,4646*** (0,001)
<i>FLAG_ANALFABETISMO</i>	0,0454*** (0,002)
<i>FLAG_ENEM</i>	0,0223*** (0,002)
<i>Adjusted R²</i>	0,197

FONTE: Elaboração Própria (2021).

Nota: *10% de significância, ** 5% de significância,

*** 1% de significância.

O resultado pode ser escrito com a seguinte expressão:

$$Y_i = 0,4646 + 0,0454FLAG_ANALFABETISMO + 0,0223FLAG_ENEM \quad (4.1)$$

Na primeira regressão (tabela 13), que utiliza os dados de analfabetismo e da nota do Enem, o R^2 é 0,197, indicando que essas duas variáveis explicam 19,7% da variação da desigualdade de renda dos municípios. No entanto, como alerta King (1986), em estudos sociais normalmente não há razão para usar este indicador, de maneira que pode ser muito enganoso, pois o objetivo é estabelecer uma relação confiável, mesmo que, muitas vezes, pequena.

Dessa maneira, ao analisar os coeficientes, é possível afirmar que, tudo o mais mantido constante, em municípios que o indicador de analfabetismo se encontra acima da mediana, o coeficiente de Gini é 0,0454 maior do que para as outras localidades. Da mesma maneira, se o município estiver na lista dos 25% piores municípios em termos de nota do Enem, apresentará Gini 0,0223 maior. As duas variáveis se mostraram estatisticamente significantes. Estes resultados, portanto, estão em linha com o esperado: quanto maior o índice de analfabetismo entre jovens de uma cidade, maior será sua desigualdade de renda; da mesma forma, quanto pior a nota do Enem de um município, maior a desigualdade de renda.

TABELA 14 – SEGUNDA REGRESSÃO

	Coeficiente
Constante	0,4985*** (0,001)
FLAG_HORAS	-0,0337*** (0,002)
FLAG_ENTRADA	0,0235*** (0,002)
Adjusted R ²	0,106

FONTE: Elaboração Própria (2021).

Nota: *10% de significância, ** 5% de significância, *** 1% de significância.

O resultado pode ser escrito com a seguinte expressão:

$$Y_i = 0,4985 + -0,0337FLAG_HORAS + 0,0235FLAG_ENTRADA \quad (4.2)$$

Ao analisar a segunda regressão (tabela 14), que utiliza os dados de média de horas de aula por dia e proporção de entrada tardia no sistema de ensino, é possível observar que o R^2 indica que essas duas variáveis, tudo o mais mantido constante, explicam 10,6% da desigualdade de renda dos municípios brasileiros. Ao observar os coeficientes, nota-se que municípios que estão acima da mediana em termos de médias de horas de aula por dia têm, em média, um índice de Gini 0,0337 menor, confirmando a expectativa. Em relação ao indicador de entrada tardia de crianças no sistema de ensino, cidades com o indicador acima da mediana têm o indicador de desigualdade 0,0235 acima das que não se encaixam nesse grupo. As duas variáveis se mostraram estatisticamente significantes. Da mesma forma que na primeira regressão, os resultados se mostram em linha com o esperado: quanto maior a média de horas de aula por dia, menor a desigualdade de renda; além disso, quanto maior a proporção de crianças que entram tardiamente no sistema de ensino, maior a desigualdade de renda.

4.4 APRENDIZADO DE MÁQUINA

De maneira adicional e para que fique clara a relação entre os dados de desigualdade de renda e de educação entre os municípios, abaixo se apresentam os resultados dos algoritmos de aprendizado de máquina. De maneira resumida, estes modelos aprendem com parte dos dados de desigualdade e educação, de modo que "entendem" de forma profunda como funciona a relação entre eles, possibilitando previsões muito acuradas e a compreensão dos dados.

TABELA 15 – RESULTADOS DO APRENDIZADO DE MÁQUINA

MÉTODO	ROC
<i>Gradient Boost</i>	75,7%
<i>Random Forest</i>	73,4%
Regressão Logística	73,2%
Árvore de Decisão	68,5%
Predição Aleatória	50,0%

FONTE: Elaboração Própria (2021).

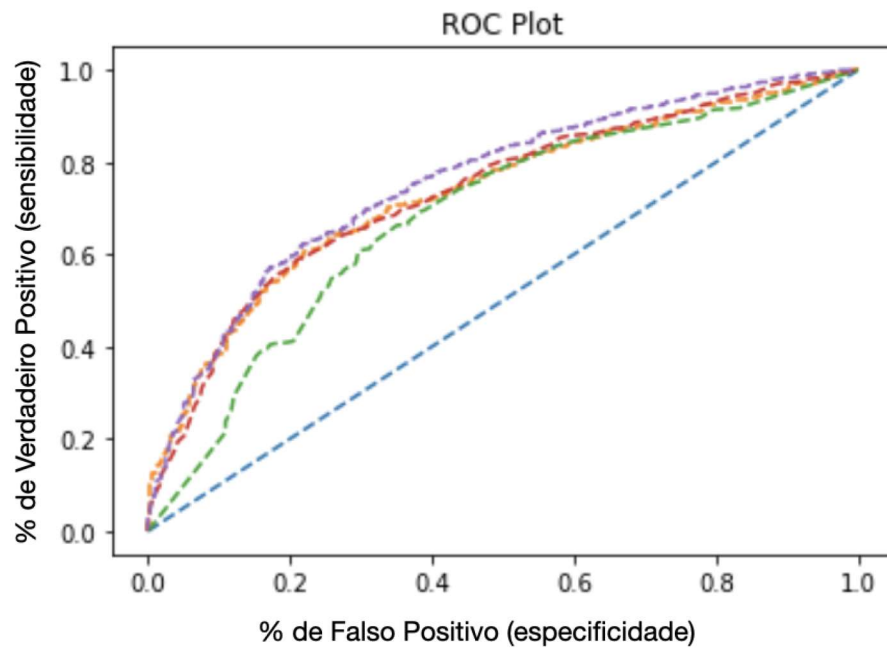
A partir da tabela 15, é possível observar a performance de cada modelo de *machine learning* comparado com a referência (predição aleatória) de 50%. Todos os algoritmos foram melhores do que o *benchmark*, indicando que, de fato, os dados de educação de cada município se relacionam com dados de desigualdade de renda de forma suficiente para que seja possível fazer predições significativas.

O melhor modelo, em termos de desempenho da Característica de Operação do Receptor, foi o *Gradient Boost*. Este modelo consegue distinguir com 75,7% de precisão quais são os municípios que estão acima da mediana em termos do índice de Gini, ou seja, a metade mais desigual do Brasil. Como exposto anteriormente, o algoritmo de *Gradient Boost* faz uma combinação de modelos simples para criar um modelo robusto que lida melhor com grandes quantidades de dados e relações complexas, mesclando diferentes métodos e aplicando a melhor técnica para cada grupo de dados repetitivamente.

De acordo com Hosmer Jr., Lemeshow e Sturdivant (2013), valores de área abaixo da curva ROC entre 70% e 80% apresentam uma discriminação aceitável para um modelo de *machine learning*. Isto indica que a relação entre educação e desigualdade de renda é forte o suficiente para que se obtenha um modelo com desempenho satisfatório.

Abaixo, na figura 19, é possível observar o gráfico da Curva ROC, que avalia o desempenho dos algoritmos. No eixo X apresentam-se os falsos positivos, enquanto no eixo Y, os verdadeiros positivos. De maneira que é possível traçar uma reta (linha azul) que representa um modelo de predição aleatória, que teria uma performance de 50%. Quanto melhor o modelo, mais verdadeiros positivos terá e mais perto do canto superior esquerdo estará. Assim, a performance do modelo é medida a partir do cálculo da área abaixo da curva. Na imagem é possível observar que todos os modelos foram melhores do que a predição aleatória.

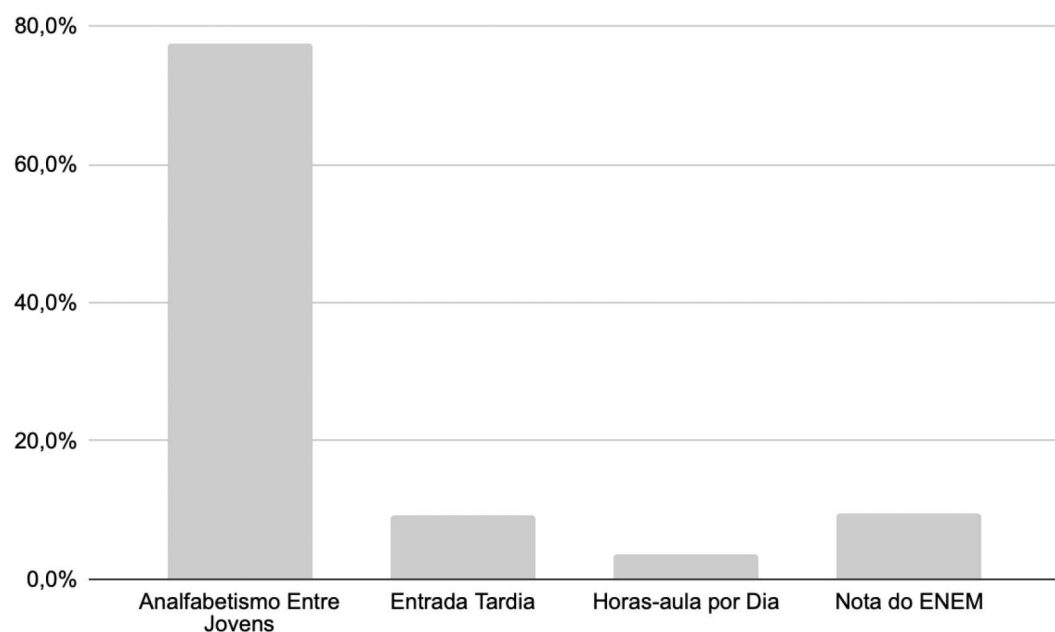
FIGURA 19 – CURVA ROC



Fonte: Elaboração Própria (2021).

Ademais, além de medir a acurácia do modelo, também é importante conhecer qual a contribuição de cada variável para o modelo de *Gradient Boost* efetuar a discriminação entre os melhores e piores municípios em termos de desigualdade de renda. Abaixo, na figura 20, é possível observar que a variável que representa a porcentagem de pessoas de 11 a 14 anos que são analfabetas é a mais significativa, de maneira que quase 80% dos ganhos do modelo se devem a ela.

FIGURA 20 – IMPORTÂNCIA DE CADA VARIÁVEL PARA O GRADIENT BOOST



Fonte: Elaboração Própria (2021).

5 CONSIDERAÇÕES FINAIS

Portanto, os resultados das análises do presente trabalho confirmam a expectativa, resultados e diagnósticos de diversos autores, que citam a educação como a principal saída para a desigualdade. As evidências sugerem esta relação em exemplos de países externos e, como verificado nesta dissertação, também é observada para os 5.568 municípios brasileiros. A educação melhora a renda, a produtividade e aumenta os efeitos da globalização, sendo, à vista disso, ponto central para qualquer plano de futuro de um país.

O objetivo desta dissertação foi entender o efeito de indicadores educacionais na desigualdade de renda do ano de 2010 (última data disponível), de maneira a confirmar os estudos anteriores e literatura acadêmica, além de mostrar como a melhora da educação é um caminho que deve ser percorrido caso o Brasil queira diminuir sua desigualdade de renda e se aproximar dos países desenvolvidos. Os resultados da análise mostram que a relação entre esses dois elementos foi real e significativa.

Esta relação foi analisada através de três perspectivas: a) análise espacial, com a produção de mapas para que se pudesse estabelecer uma relação visual do comportamento das variáveis nas distintas regiões e análise dos Índices de Moran para que fosse possível entender melhor o comportamento das variáveis no espaço, b) análise de regressão, para que se pudesse estatisticamente comprovar a relação entre as variáveis, obedecendo os princípios de autocorrelação, heterocedasticidade e normalidade dos resíduos; e c) análise via algoritmo de aprendizado de máquina, de forma a descobrir se a relação entre educação e desigualdade de renda seria forte o suficiente para que houvesse um algoritmo com boa performance.

A Análise Espacial com o índice de Moran proporcionou uma melhor compreensão do comportamento das variáveis nos municípios em relação à região em que estão inseridos, de maneira que determinou que as variáveis de Gini, analfabetismo, horas-aula por dia e média do Enem são relativamente dependentes da região em que estão inseridas, indicando que são afetadas por múltiplos fatores.

A Análise de Regressão, por sua vez, demonstrou como municípios com maiores taxas de analfabetismo entre jovens, menores notas no Enem e com maiores taxas de entrada tardia no sistema de ensino são mais desiguais, enquanto municípios com maiores médias de horas de aula por dia para a educação infantil são menos desiguais.

Por fim, a análise via algoritmo de Aprendizado de Máquina demonstrou como

é possível, com uma boa acurácia, determinar o nível de desigualdade de um município somente à partir das variáveis educacionais dos mesmos, indicando que essa relação é robusta. Da mesma maneira, o algoritmo com a melhor performance indicou que a variável mais importante para essa predição foi o indicador de analfabetismo entre jovens. Os resultados encontrados neste trabalho reforçam a importância da educação como meio transformador da realidade corrente.

Sendo assim, fica evidente a importância de reforçar o processo de educação via maiores investimentos, mas mais do que isso, é necessário dar condições de estudo para crianças e jovens, melhorar a qualidade e aumentar as horas de aula por dia.

Como sugestão, futuros pesquisadores podem estudar essa relação ao longo tempo, caso existam dados disponíveis e comparáveis. Ademais, cabe ressaltar que a educação é um fator importantíssimo para a desigualdade de renda de uma região, mas não é a única. Por se tratar de um tema complexo e com diversos ângulos e abordagens distintas, trabalho futuros podem abordar outras causas utilizando os dados dos municípios brasileiros, de maneira a enriquecer a literatura acadêmica sobre o tema e dar suporte à tomada de decisão de políticas públicas baseada em dados.

REFERÊNCIAS

- ARCGIS PRO. **How Spatial Autocorrelation (Global Moran's I) works**. [S.l.], 2021. Disponível em: <https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/h-how-spatial-autocorrelation-moran-s-i-spatial-st.htm>. Citado na p. 36.
- ATHEY, S. **The Impact of Machine Learning on Economics**. [S.l.], 2018. P. 507–547. Citado na p. 21.
- BANCO MUNDIAL. **GDP (current US\$)**. [S.l.], 2021. Disponível em: <https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=CN/>. Citado na p. 11.
- _____. **Gini index (World Bank estimate) - Brazil**. [S.l.], 2019. Disponível em: https://data.worldbank.org/indicator/SI.POV.GINI?most_recent_value_desc=true. Citado nas pp. 8, 25.
- BARBOSA, R.; SOUZA, P. F. de; SOARES, S. **Desigualdade de renda no Brasil de 2012 a 2019**. [S.l.], 2020. Disponível em: <http://dados.iesp.uerj.br/desigualdade-brasil/>. Citado na p. 8.
- BARRO, R. J. **Economic Growth in a Cross Section of Countries**. [S.l.], 1991. P. 407–443. Citado na p. 13.
- BARROS, R. P. de; MENDONÇA, R. S. P. de. **Os Determinantes da Desigualdade no Brasil**. [S.l.], 1995. Citado nas pp. 11, 12.
- BREIMAN, L. **Random Forests**. [S.l.], 2001. P. 5–32. Citado na p. 24.
- BUCKINGHAM, J.; WHELDALL, K.; BEAMAN-WHELDALL, R. **Why poor children are more likely to become poor readers: The school years**. [S.l.], 2013. P. 190–213. Citado na p. 10.
- COSTA, G. **Analfabetismo resiste no Brasil e no mundo do século 21**. [S.l.], 2019. Disponível em: <https://agenciabrasil.ebc.com.br/educacao/noticia/2019-09/analfabetismo-resiste-no-brasil-e-no-mundo-do-seculo-21>. Citado na p. 17.

DATASUS. **Índice de Gini da renda domiciliar per capita segundo Região, Unidade da Federação e Região Metropolitana.** [S.l.], 2010. Disponível em: <http://tabnet.datasus.gov.br/cgi/ibge/censo/cnv/giniuf.def>. Citado na p. 27.

FARRIS, F. A. **The Gini Index and Measures of Inequality.** [S.l.], 2010. P. 851–864. Citado na p. 15.

FAWCETT, T. **An introduction to ROC analysis.** [S.l.], 2006. P. 861–874. Citado na p. 22.

FERNANDES, D. **OCDE: desigualdade aumenta em países ricos, mas cai na América Latina.** [S.l.], 2015. Disponível em: https://www.bbc.com/portuguese/noticias/2015/05/150520_desigualdade_estudo_ocde_df_rb. Citado na p. 26.

FERRANTI, D. de; PERRY, G. E.; FERREIRA, F. H. G.; WALTON, M. **Inequality in Latin America - Breaking With History?** [S.l.], 2004. Citado nas pp. 8, 11.

FERRÃO, M. E.; FERNANDES, C. **O Efeito-Escola e a Mudança - Dá Para Mudar? Evidências da Investigação Brasileira.** [S.l.], 2003. Citado na p. 10.

FERREIRA, R. T.; CRUZ, M. S. da. **Efeito da Educação, da Renda do Trabalho, das Transferências e das Condições Iniciais na Evolução da Desigualdade de Renda nos Municípios Brasileiros no Período de 1991 a 2000.** [S.l.], 2010. Citado nas pp. 12, 14.

FISCHER, M. M.; GETIS, A. **Handbook of Applied Spatial Analysis.** [S.l.], 2010. Citado na p. 18.

HANUSHEK, E. A.; WOESSMANN, L. **The Role of Education Quality in Economic Growth.** [S.l.], 2007. Citado nas pp. 11, 13, 17.

HOARE, J. **Gradient Boosting Explained – The Coolest Kid on The Machine Learning Block.** [S.l.], 2017. Disponível em: <https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/>. Citado na p. 24.

HOSMER JR., D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied Logistic Regression.** 3ª Edição. [S.l.], 2013. Citado nas pp. 23, 46.

JACOB, B. A.; LEFGREN, L. **Are Idle Hands the Devil's Workshop? Incapacitation, Concentration and Juvenile Crime.** [S.l.], 2003. Citado na p. 17.

JONES, T. **Supervised learning models.** [S.l.], 2018. Disponível em: <https://developer.ibm.com/articles/cc-supervised-learning-models/>. Citado na p. 22.

KING, G. **How Not to Lie With Statistics: Avoiding Common Mistakes in Quantitative Political Science.** [S.l.], 1986. P. 666–687. Citado na p. 44.

KRUGER, D.; BERTHELON, M. E.; NAVIA, R. **Adolescent Motherhood and Secondary Schooling in Chile.** [S.l.], 2009. Citado na p. 17.

LANGONI, C. G. **Distribuição da Renda e Desenvolvimento Econômico do Brasil: Uma Reafirmação.** [S.l.], 1973. Citado nas pp. 10, 11, 13.

LERMAN, R. I.; YITZHAKI, S. **A Note on The Calculation And Interpretation of the Gini Index.** [S.l.], 1984. P. 363–368. Citado na p. 15.

LUZARDO, A. J. R.; CASTAÑEDA FILHO, R. M.; RUBIM, I. B. **Análise Espacial Exploratória Com o Emprego do Índice de Moran.** [S.l.], 2017. P. 161–179. Citado nas pp. 18, 20.

LV, K.; YU, A.; GONG, S.; WU, M.; XU, X. **Impacts of educational factors on economic growth in regions of China: a spatial econometric approach.** [S.l.], 2015. P. 827–847. Citado nas pp. 11, 14.

MENEZES FILHO, N. **Os Determinantes do Desempenho Escolar no Brasil.** [S.l.], 2012. Citado nas pp. 10, 13, 17.

MONARD, M. C.; BARANAUSKAS, J. A. **Indução de Regras e Árvore de Decisão.** [S.l.], 2003. P. 115–139. Citado na p. 23.

OLIVEIRA, K. **Brasil gasta 6% do PIB em educação, mas desempenho escolar é ruim.** [S.l.], 2018. Disponível em: <https://agenciabrasil.ebc.com.br/educacao/noticia/2018-07/brasil-gasta-6-do-pib-em-educacao-mas-desempenho-escolar-e-ruim/>. Citado na p. 12.

PAULINO, A. **Área Abaixo da Curva ROC**. [S.l.], 2018. Disponível em: <https://medium.com/ensina-ai/%C3%A1rea-abaixo-da-curva-roc-15d2ae93a577>. Citado na p. 23.

ROMER, P. M. **Endogenous Technological Change**. [S.l.], 1990. P. 71–102. Citado na p. 13.

SASSE, C. **Recordista em desigualdade, país estuda alternativas para ajudar os mais pobres**. [S.l.], 2021. Disponível em: <https://www12.senado.leg.br/noticias/infomaterias/2021/03/recordista-em-desigualdade-pais-estuda-alternativas-para-ajudar-os-mais-pobres/>. Citado na p. 8.

TECH, D. **Como funciona o algoritmo Árvore de Decisão**. [S.l.], 2021. Disponível em: <https://didatica.tech/como-funciona-o-algoritmo-arvore-de-decisao/>. Citado na p. 24.

U.S. DEPARTMENT OF EDUCATION. **Average number of hours in the school day and average number of days in the school year for public schools, by state: 2007–08**. [S.l.], 2008. Disponível em: https://nces.ed.gov/surveys/sass/tables/sass0708_035_s1s.asp. Citado na p. 32.

WOOLDRIDGE, J. **Introdução à Econometria: Uma Abordagem Moderna**. 6ª Edição. [S.l.], 2016. Citado na p. 20.