UNIVERSIDADE FEDERAL DO PAMPA

LOURENÇO NATANIEL PINHEIRO PORTELLA

ANÁLISE DE POLÍTICAS PÚBLICAS DE TRANSFERÊNCIA DE RENDA UTILIZANDO TÉCNICAS DE CIÊNCIA DE DADOS

Bagé
2025

LOURENÇO NATANIEL PINHEIRO PORTELLA

ANÁLISE DE POLÍTICAS PÚBLICAS DE TRANSFERÊNCIA DE RENDA UTILIZANDO TÉCNICAS DE CIÊNCIA DE DADOS

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Engenharia de Computação como requisito parcial para a obtenção do grau de Bacharel em Engenharia de Computação.

Orientador: Carlos Michel Betemps
Coorientador: Sandro da Silva Camargo

Bagé
2025

Ficha catalográfica elaborada automaticamente com os dados fornecidos pelo(a) autor(a) através do Módulo de Biblioteca do Sistema GURI (Gestão Unificada de Recursos Institucionais).

P843a Portella, Lourenço Nataniel Pinheiro

ANÁLISE DE POLÍTICAS PÚBLICAS DE TRANSFERÊNCIA DE RENDA UTILIZANDO TÉCNICAS DE CIÊNCIA DE DADOS / Lourenço Nataniel Pinheiro Portella.

131 f.: il.

Orientador: Carlos Michel Betemps
Coorientador: Sandro da Silva Camargo
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal do Pampa, Engenharia de Computação, 2025.

1. Política pública. 2. Ciência de dados. 3. Transferência de renda. 4. Programa Bolsa Família. 5. IDESE. I. Título.

LOURENÇO NATANIEL PINHEIRO PORTELLA

ANÁLISE DE POLÍTICAS PÚBLICAS DE TRANSFERÊNCIA DE RENDA UTILIZANDO TÉCNICAS DE CIÊNCIA DE DADOS

Trabalho de Conclusão de Curso apresentado ao curso de Engenharia de Computação como requisito parcial para a obtenção do grau de Bacharel em Engenharia de Computação.

Dissertação defendida e aprovada em: 10 de dezembro de 2025.

Banca examinadora:

_____

Prof. Dr. Carlos Michel Betemps

Orientador

(UNIPAMPA)

_____

Prof. Dr. Julio Saraçol Domingues Jr.

(UNIPAMPA)

_____

Profª Drª Sandra Dutra Piovesan

(UNIPAMPA)

Assinado eletronicamente por **SANDRA DUTRA PIOVESAN**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 18/12/2025, às 18:56, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **CARLOS MICHEL BETEMPS**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 18/12/2025, às 20:21, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **JULIO SARACOL DOMINGUES JUNIOR**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 19/12/2025, às 11:35, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

A autenticidade deste documento pode ser conferida no site https://sei.unipampa.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1919677** e o código CRC **DBE74462**.

AGRADECIMENTO

Neste momento, dirijo meus agradecimentos a todos que me auxiliaram nesta jornada acadêmica, culminando na apresentação deste trabalho, que representa parte do conhecimento adquirido ao longo destes anos.

Agradeço a todo o corpo docente da UNIPAMPA, que durante minha graduação demonstrou solicitude em compreender as dificuldades de um aluno com deficiência visual, sem adotar atitudes capacitistas. Um especial reconhecimento aos professores do curso de Engenharia de Computação, pela nobre transmissão de seu conhecimento de diversas formas, contribuindo não apenas para minha formação profissional, mas também para o meu desenvolvimento humano e caráter.

Agradeço aos meus familiares, amigos e colegas de trabalho, que durante toda a graduação mantiveram a confiança em meu potencial, acreditando que eu chegaria a este ponto e que poderei alcançar objetivos ainda mais elevados.

Agradeço à área de computação, que possui em sua essência a capacidade de evoluir de forma acelerada, propiciando a ampliação contínua do aprendizado e, consequentemente, o acesso a conhecimentos antes de difícil obtenção.

A todos, expresso minha profunda gratidão pelo apoio indispensável nesta trajetória.

RESUMO

As políticas públicas de transferência de renda de caráter redistributivo no Brasil têm como objetivo romper o ciclo intergeracional da pobreza entre os beneficiários. A grande maioria dos estudos busca verificar os impactos dessas políticas na sociedade, sem, contudo, examinar como as características de desenvolvimento podem influenciar os indicadores associados a tal política. O Programa Bolsa Família (PBF), já consolidado no Brasil, destaca-se por sua abrangência e pelas condicionalidades para seu recebimento, tornando-se um exemplo relevante para a análise de políticas públicas de transferência de renda. Por outro lado, o Índice de Desenvolvimento Socioeconômico do Rio Grande do Sul (IDESE), inspirado no IDH, mensura o desenvolvimento no estado com base em indicadores e índices compostos de educação, renda e saúde, constituindo um parâmetro adequado para investigar como o desenvolvimento regional afeta a implementação dessa política. Na busca de uma análise eficiente, empregaram-se técnicas de Ciência de Dados, campo interdisciplinar que permite criar e interpretar modelos robustos, com foco na obtenção de métricas que relacionem variáveis dependentes e independentes. Neste trabalho, aplicou-se o aprendizado de máquina para a geração de resultados avaliativos: o *K-means* foi utilizado para identificar agrupamentos (*clusters*) de amostras com características semelhantes; subsequentemente, utilizaram-se os métodos de Regressão Linear, Redes Neurais *Multilayer Perceptron* (MLP) e *XGBoost* para validar as possíveis relações entre os atributos selecionados. O coeficiente de determinação R^2 foi adotado como a principal métrica de avaliação dessas relações. Observou-se que o *K-means* possui a capacidade de identificar relações e associar amostras em grupos coesos, os quais foram validados pelos métodos computacionais, com destaque para o *XGBoost*, que obteve um R^2 de 81,39% para a combinação do total de beneficiários com todos os índices calculados pelo IDESE. Ao extrair *insights* das combinações com os melhores valores de R^2, constatou-se que os valores médios de todos os índices do IDESE, bem como os índices que compõem o bloco de educação, tendem a afetar inversamente a quantidade de beneficiários do PBF. Índices acima da média estão associados a uma quantidade de beneficiários 48,15% inferior à média geral, enquanto índices abaixo da média resultam em um número 36,35% superior à média de beneficiários.

Palavras-chave: Política pública. ciência de dados. transferência de renda. Programa Bolsa Família. IDESE.

ABSTRACT

Public policies for income transfer with a redistributive character in Brazil aim to break the intergenerational cycle of poverty among beneficiaries. The vast majority of studies seek to verify the impacts of these policies on society without, however, examining how development characteristics can influence the indicators associated with such a policy. The Programa Bolsa Família (PBF), already consolidated in Brazil, stands out for its scope and the conditionalities for its receipt, becoming a relevant example for the analysis of public income transfer policies. On the other hand, the Índice de Desenvolvimento Socioeconômico do Rio Grande do Sul (IDESE), inspired by the HDI, measures development in the state based on indicators and composite indices of education, income, and health, constituting an adequate parameter to investigate how regional development affects the implementation of this policy. In pursuit of an efficient analysis, Data Science techniques were employed—an interdisciplinary field that allows the creation and interpretation of robust models, focusing on obtaining metrics that relate dependent and independent variables. In this work, machine learning was applied to generate evaluative results: K-means was used to identify clusters of samples with similar characteristics; subsequently, Linear Regression, Multilayer Perceptron (MLP) Neural Networks, and XGBoost methods were used to validate possible relationships between the selected attributes. The coefficient of determination R^2 was adopted as the main evaluation metric for these relationships. It was observed that K-means has the capacity to identify relationships and associate samples into cohesive groups, which were validated by computational methods, with emphasis on XGBoost, which obtained an R^2 of 81.39% for the combination of the total number of beneficiaries with all indices calculated by IDESE. By extracting insights from the combinations with the best R^2 values, it was found that the average values of all IDESE indices, as well as the indices that compose the education block, tend to inversely affect the number of PBF beneficiaries. Indices above the average are associated with a number of beneficiaries 48.15% lower than the overall average, while indices below the average result in a number 36.35% higher than the average number of beneficiaries.

Keywords: public policy, data science, income transfer, *bolsa família* program, IDESE.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Coredes	Conselhos Regionais de Desenvolvimento
DEE	Departamento de Economia e Estatística
EF	Ensino Fundamental
FEE	Fundação de Economia e Estatística
IBGE	Instituto Brasileiro de Geografia e Estatística
IDH	Índice de Desenvolvimento Socioeconômico
IDESE	índice de Desenvolvimento Socioeconômico do Rio Grande do Sul
MDS	Ministério do Desenvolvimento e Assistência Social, Família e Combate à Fome
MLP	Redes neurais Multilayers Perceptrom
MSE	Erro Quadrático Médio
ONU	Organização das Nações Unidas
PBF	Programa Bolsa Família
PIB	Produto Interno Bruto
RS	Rio Grande do Sul
Sagicad	Secretaria de Avaliação, Gestão da Informação e Cadastro Único
SAEB	Sistema de Avaliação da Educação Básica
UNIPAMPA	Universidade Federal do Pampa

SUMÁRIO

1 INTRODUÇÃO

As políticas públicas de transferência de renda no Brasil, implementadas desde a década de 1990, visam mitigar desigualdades socioeconômicas por meio de assistência financeira condicionada a critérios específicos. Dentre as iniciativas de auxílio empregadas no país, o Programa Bolsa Família (PBF), instituído em 2003, destaca-se por sua abrangência e duração. Estudos compilados na literatura científica analisam os impactos diretos e indiretos do programa na educação, saúde e economia, além de investigarem percepções sociais desde sua criação até o presente (CAMPELLO; NERI, 2013; VIANA; KAWAUCHI; BARBOSA, 2018). Contudo, persistem lacunas na avaliação quantitativa da correlação entre os indicadores de desenvolvimento nacionais ou regionais e os números do PBF, especialmente no estado do Rio Grande do Sul (KÜHN; TONETTO, 2017).

O Índice de Desenvolvimento Socioeconômico do Rio Grande do Sul (IDESE), inspirado no Índice de Desenvolvimento Humano (IDH) criado pela Organização das Nações Unidas (ONU) na década de 1990 (FEE ACCURSO; HEUSER, 2003), oferece uma métrica multidimensional que avalia, em módulos distintos, a educação, a renda e a saúde em municípios e regiões gaúchas. Embora estudos longitudinais comparem beneficiários e não beneficiários do PBF (SANTOS, 2021; BONILHA, 2024), a aplicação de técnicas de ciência de dados para correlacionar os índices socioeconômicos apresentados nas dimensões do IDESE com as valores encontrados ao longo do tempo no PBF (como número de beneficiários e valores transferidos) de forma integrada tem sido pouco explorada. Essa lacuna conduz à seguinte questão de pesquisa: ***”De que forma a ciência de dados pode contribuir para analisar a relação entre os dados históricos do PBF e as variações do IDESE ao longo dos anos nos municípios e regiões do estado do Rio Grande do Sul?”***

Propõe-se a utilização de técnicas de ciência de dados (área que engloba diversos campos do conhecimento, como computação e estatística), aliadas ao conhecimento de domínio necessário para aplicá-las ao problema em análise. O processo inicia-se com a obtenção dos dados, seguida da aplicação de métodos computacionais de aprendizado de máquina supervisionado e não supervisionado. Por fim, analisam-se os resultados dos modelos gerados, visando compreender as correlações das mudanças nos valores do IDESE sobre o PBF.

A aplicação dos métodos computacionais inicia com a formação de agrupamentos

por meio do *K-means* (K-médias) (HARRISON, 2019), uma técnica de aprendizado não supervisionado que busca reunir em grupos os registros que apresentem características similares. Estes resultados subsidiam a aplicação de métodos de aprendizado supervisionado, variando em nível de complexidade, com o intuito de obter o melhor valor do coeficiente de determinação (R^2), validando as características do IDESE selecionadas como atributos preditores do PBF. Dentre os diversos métodos computacionais de aprendizado supervisionado existentes, optou-se por três que trouxessem configurações relevantes a este estudo e cujos resultados pudessem ser comparados entre si, a fim de obter informações adicionais.

- **Regressão Linear**: Método fundamental de aprendizado supervisionado, usualmente empregado na identificação de relações lineares simples e diretas, envolvendo uma ou múltiplas variáveis explicativas (WITTEN et al., 2025);

- **Redes Neurais *Multilayer Perceptron* (MLP)**: Método avançado inserido no contexto do aprendizado profundo, que visa identificar relações não lineares e interações complexas entre as variáveis selecionadas, demonstrando capacidade de lidar com um elevado número de características para explicar um determinado valor (ZHANG et al., 2023);

- ***XGBoost***: Técnica que combina algoritmos de aprendizado de máquina considerados simples (modelos fracos) para interpretar com maior acurácia as relações não lineares entre as características, resultando em desempenho superior devido ao seu alto grau de generalização e precisão (GÉRON, 2019).

A seleção desses métodos justifica-se pela necessidade de comparar abordagens clássicas, como a Regressão Linear, com técnicas mais avançadas e robustas, como as Redes Neurais MLP e o *XGBoost*, associadas ao uso do *K-means* para estabelecer a melhor segmentação dos dados. Com os resultados almeja-se subsidiar avaliações críticas sobre os indicadores do PBF, contribuindo para a formulação de políticas públicas mais eficazes e auxiliando os gestores no entendimento do impacto dos índices de desenvolvimento. Adicionalmente, busca-se validar a aplicabilidade de técnicas de ciência de dados em contextos socioeconômicos, nos quais a interpretação dos resultados obtidos pelos modelos é fundamental (HOSSIN et al., 2023).

1.1 Objetivo geral

A definição de um objetivo de pesquisa constitui uma etapa complexa e fundamental para o trabalho (WAZLAWICK, 2009), visto que não deve se limitar à simples reescrita do tema. Nesse sentido, realizou-se um aprofundamento na literatura para a verificação de processos previamente executados. Tais contribuições são detalhadas no Capítulo 4.2 e integradas aos demais capítulos e seções que compõem este trabalho.

Desta forma, o objetivo geral é: ***analisar, por meio de técnicas de ciência de dados, o impacto das variações nos valores dos índices socioeconômicos de educação, renda e saúde, calculados pelo IDESE, sobre os números históricos do Programa Bolsa Família (PBF) nos municípios e regiões do Rio Grande do Sul.***

Acredita-se que o objetivo geral desta pesquisa seja plenamente alcançável, dada a extensa literatura disponível sobre os temas abordados. Além disso, deve-se considerar que diversas outras pesquisas adotaram abordagens inversas, focando na verificação de como o governo influencia a sociedade a partir das modificações nos índices em um determinado período ou após a implantação de uma política pública. Este trabalho, por sua vez, investiga a relação oposta: como a sociedade (representada pelos índices socioeconômicos) influencia os resultados de uma política pública. Tal análise baseia-se na forte correlação existente entre estes dois temas, a qual pode ser visualizada de forma coerente através do uso da ciência de dados.

1.2 Objetivos específicos

Com a finalidade de direcionar a pesquisa em consonância com o objetivo geral proposto, estabelecem-se objetivos específicos que buscam compreender questões mais intrínsecas, listados a seguir:

1° Entender as características apresentadas nos dados do PBF e IDESE, considerando as características e metodologias originais;

2° Validar o uso de ciência de dados para analisar como as políticas públicas interagem com a sociedade;

3° Investigar as relações entre os dados do PBF e as dimensões do IDESE nas regiões e municípios gaúchos com o uso de métricas estatísticas de avaliação obtidas de modelos computacionais.

Tais objetivos conferem maior rigor à construção da metodologia e à realização da análise dos resultados, tornando-a mais focada nos pontos necessários para responder às questões levantadas. Isso ocorre ao integrar a ciência de dados nos processos de construção e análise dos resultados dos métodos computacionais e das métricas estatísticas, considerando as características intrínsecas a cada um dos assuntos alvo do estudo, no caso, o PBF e o IDESE.

1.3 Organização do texto

Esta monografia é estruturada em diferentes capítulos e respectivas seções, com o objetivo de assegurar uma linha de conhecimento e raciocínio para o entendimento das conclusões apresentadas, as quais derivam das respostas às questões de pesquisa geral e específicas.

O Capítulo 2 apresenta a metodologia de pesquisa, descrevendo o processo efetuado para a obtenção, tratamento e validação dos conjuntos de dados e métodos computacionais utilizados neste trabalho. O capítulo inclui uma seção dedicada às ferramentas de *hardware* e *software* utilizadas, detalhando as estratégias empregadas para alcançar os objetivos propostos. Em seguida, no Capítulo 3, expõe-se a fundamentação teórica, com a definição e contextualização dos conceitos-chave relacionados aos temas abordados neste estudo, como técnicas de ciência de dados, políticas públicas e índices socioeconômicos.

No Capítulo 4, descreve-se como foram realizadas a pesquisa e a seleção de estudos correlatos, os quais possuem características semelhantes a este trabalho, subsidiando assim a abordagem e a validade do estudo.

O Capítulo 5 aborda o desenvolvimento e a aplicação da ciência de dados, incluindo a mensuração de métricas básicas sobre os dados, a aplicação dos métodos computacionais e a interpretação dos resultados obtidos a partir dos modelos estatísticos construídos. Por fim, as considerações finais do trabalho são apresentadas no Capítulo 6, onde são sintetizadas as conclusões obtidas, limitações, além das implicações para políticas públicas e possíveis direções para trabalhos futuros.

2 MATERIAIS E MÉTODOS

Neste capítulo são descritos os procedimentos metodológicos de pesquisa deste trabalho e as ferramentas tecnológicas empregadas na obtenção dos resultados, com a intenção de gerar resultados precisos e reprodutíveis. A estrutura adotada divide-se em duas seções principais: Organização Metodológica e Ferramentas que, em conjunto, evidenciam o rigor e a consistência dos métodos empregados, para respaldar as conclusões contidas neste trabalho.

2.1 Organização metodológica

Este trabalho adota uma combinação de metodologias indutiva, estatística e exploratória que, em conjunto, asseguram alto rigor científico ao longo do estudo. A análise da bibliografia será relevante antes e após a obtenção dos resultados, de modo a permitir a indução e generalização de informações para a elaboração de conclusões que possam ser comparadas ou que sirvam para futuros trabalhos (MARCONI; LAKATOS, 2004).

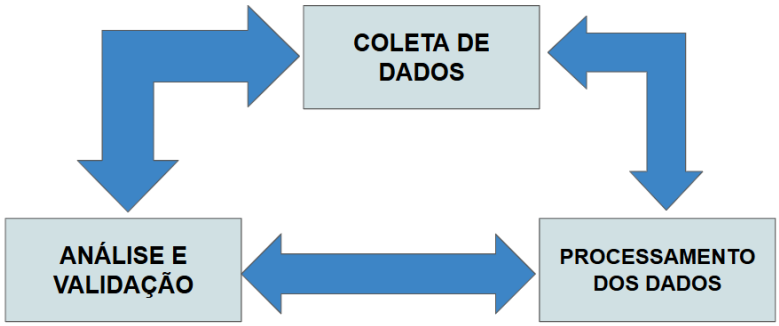
O intuito desta pesquisa científica é explorar e explicar, por meio de dados quantitativos (FREITAS; PRODANOV, 2013), as possíveis relações entre políticas públicas de transferência de renda e índices socioeconômicos. A análise fundamenta-se na utilização da ciência de dados para a correta resposta às questões de pesquisa, e na possibilidade de uso de técnicas computacionais e métricas estatísticas aplicadas aos dados governamentais disponibilizados a respeito do PBF e do IDESE.

A escolha dessas abordagens baseia-se na necessidade de extrair conhecimento a partir de bases de dados históricas complexas, explorando as correlações diretas e indiretas entre os dados selecionados. Para tanto, o Programa Bolsa Família (PBF) foi adotado como política pública de transferência de renda, e é usado neste trabalho como alvo dos resultados, e o índice socioeconômico Índice de Desenvolvimento Socioeconômico do Rio Grande do Sul (IDESE), que fornece as características, considerando apenas o período de 2007 a 2021.

A metodologia adotada foi simplificada para contemplar apenas três fases principais: (*i*) coleta de dados, (*ii*) processamento de dados, e (*iii*) análise e validação. Tais etapas são descritas nas subseções seguintes e na Figura 1, que detalha, de forma visual, o ciclo metodológico referente ao uso dos dados empregados neste trabalho, de modo

a possibilitar uma compreensão clara do processo investigativo e evidenciar as escolhas técnicas e os procedimentos adotados para alcançar os resultados esperados.

Figura 1 – Fluxograma da metodologia



Fonte: Elaborado pelo Autor.

2.1.1 Coleta de dados

Nesta seção são apresentadas as fontes e os procedimentos executados que envolvem a integração e organização de tabelas e atributos, com o objetivo de coletar os dados utilizados neste estudo. Esses conjuntos de dados possuem diversas características distintivas que necessitaram de adaptações para a geração do conjunto unificado e completo. A elaboração do conjunto final de dados exigiu cuidados especiais para assegurar a qualidade e a comparabilidade dos dados, constituindo uma etapa fundamental para a consistência das análises subsequentes.

Após a definição do problema de pesquisa, buscou-se encontrar locais onde estão disponíveis os dados mais relevantes a serem utilizados no trabalho. Com a busca concluída, efetuou-se a coleta dos dados a partir de repositórios governamentais oficiais que seguem as diretrizes contidas na Lei nº 12.527 de 2011 (BRASIL, 2011), que define a liberdade de acesso à informação aos cidadãos no Brasil. Essa pesquisa por informações originais e de fontes confiáveis foi importante para conferir maior solidez ao prosseguimento das demais etapas descritas neste trabalho.

O Programa Bolsa Família (PBF) tem seus dados distribuídos em diversas fontes do governo federal, incluindo dados simples ou extremamente detalhados. No entanto, pela simplicidade de navegação e adaptação a este trabalho, preferiu-se o uso do sistema **VIS DATA** (SAGICAD, 2025), mantido pela Secretaria de Avaliação, Gestão da Informação e Cadastro Único (Sagicad), vinculado ao Ministério do Desenvolvimento

e Assistência Social, Família e Combate à Fome (MDS).

A Sagicad disponibiliza diversos filtros de acesso a dados de programas sociais do governo federal, com a possibilidade de geração de tabelas no formato *.csv* para posterior manipulação. Além disso, permite que o usuário que consulta o sistema adquira os dados gerais referentes ao país, estado, município, ou combinação entre os entes federativos disponíveis, e também indique um determinado período que se deseja observar. Do VIS DATA foram retiradas informações do período iniciando no ano de 2007 e terminando em 20221 (em que o PBF disponibiliza dados até outubro de 2021), então temos as seguintes informações:

- **Código**: Código do município fornecido pelo IBGE;
- **Unidade Territorial**: Nome do município;
- **UF**: Sigla do estado onde se encontra o município;
- **Referência**: Ano/mês dos dados;
- **Famílias PBF (até Out/2021)**: Quantidade de famílias beneficiárias no período de referência;
- **Valor repassado às famílias PBF (até Out/2021)**: Valor total repassado no período de referência.

É importante ressaltar alguns aspectos relativos ao conjunto de dados obtido do PBF. A variável **Famílias PBF (até Out/2021)** contempla exclusivamente as famílias beneficiárias, o que pode gerar certas discrepâncias por representar uma quantidade reduzida em comparação com informações mais detalhadas disponíveis em outras fontes, as quais, contudo, apresentam maior complexidade para extração completa dos valores. Quanto à variável **Valor repassado às famílias PBF (até Out/2021)**, os dados sofreram distorções devido à disponibilização das informações do ano de 2020, que estão constantes apenas nos registros do programa Auxílio Emergencial. A inclusão desses registros acarretaria a mesma discrepância na variável, porém para valores mais elevados, optando-se, portanto, pela manutenção apenas dos dados originais. Por fim, ressalta-se que não houve a incorporação do conjunto de dados que complementaria os meses de outubro, novembro e dezembro de 2021, os quais estão disponíveis nos registros do programa Auxílio Brasil.

Os dados referentes ao Índice de Desenvolvimento Socioeconômico do Rio Grande do Sul (IDESE) foram disponibilizados pelo sistema **IdeseVis** (DEE-RS-DEEDADOS, 2025) sob controle do Departamento de Economia e Estatística do Rio Grande do Sul (DEE-RS). Tais dados possuem informações apenas do período de 2013 a 2021, que se referem às últimas atualizações apresentadas no Capítulo 3. Esses dados foram interpretados como insuficientes para o trabalho; desta forma, procurou-se os dados mais antigos que seguiam a mesma metodologia de cálculo. Esses dados estão arquivados atualmente no sitema **DataRS** que é intitulado **Repositório de dados socioeconômicos do Estado do Rio Grande do Sul** (DEE-RS, 2025). Com a mescla desses dois sistemas, foi possível obter dados metodologicamente equivalentes de 2007 a 2021, dos quais os seguintes atributos foram extraídos:

- **TIPO UNID**: Rótulo para descrever se o índice se refere a uma unidade da federação, divisão de território, ou conselho/entidade regional;
- **COD**: Código do local indicado pelo IBGE ou que se refere ao conselho/entidade regional;
- **NOME**: Nome do ente federativo, divisão do território, ou conselho/entidade regional;
- **CATEGORIA**: Nome do índice/indicador;
- **ANO**: Ano a que se refere o índice/indicador;
- **VALOR**: Valor do índice/indicador.

A escolha do período de análise ser de 2007 a 2021 fundamenta-se na disponibilidade dos dados do IDESE, cuja metodologia de cálculo sofreu alterações significativas fora desse intervalo (KANG et al., 2014), inviabilizando a disponibilização por parte do governo dos dados anteriores. Também relaciona-se a este recorte temporal as alterações efetuadas no PBF em outubro de 2021, que poderiam influenciar os resultados das análises.

Complementarmente, foram integrados aos dados as informações referentes à população estimada de cada município do estado, calculada e disponibilizada pelo DEE-RS (DEE-RS-POPVIS, 2025), e os metadados municipais, que contêm os identificadores numéricos que indicam o município e a divisão territorial em que está localizada a micro/mesorregião do estado do RS, extraídos do sistema de armazenamento

de informações geográficas do Instituto Brasileiro de Geografia e Estatística (IBGE) (IBGE, 2024).

2.1.2 Processamento dos dados

Nesta seção são apresentados os processos de pré-processamento dos dados obtidos a partir dos procedimentos efetuados na Seção 2.1.1. Os processos aqui descritos foram necessários para obter a concatenação de várias tabelas em um único arquivo tabular, de modo a servir como um agregador ou banco de dados de todas as informações que servem de base para as análises e resultados apresentados. A formulação da tabela única foi executada com o auxílio da linguagem de programação Python (FOUNDATION, 2025b) e das bibliotecas Pandas (PANDAS, 2024) e NumPy (NUMPY, 2022), que oferecem amplo suporte para manipulação de estruturas de dados tabulares.

Inicialmente, foi efetuada a concatenação e formatação das tabelas que continham os dados do IDESE. Partiu-se do princípio de que, como os dados estavam organizados quase de forma unidimensional, seria gerada uma coluna para referenciar cada um dos índices calculados. Tais modificações foram possíveis com o uso das estruturas de manipulação do Pandas chamadas *DataFrames*. Essas alterações foram realizadas utilizando os atributos **ANO** e **COD**, que serviram como etiquetas para identificar os municípios e criar uma amostra para cada ano de referência.

Em seguida, foi feita a união das tabelas do IDESE e do PBF. Esse processo foi viabilizado pelo uso do atributo que fazia referência, em cada tabela, ao **código/ID do IBGE** e também ao **ano** da amostra. A finalização da adição de conteúdo à tabela foi realizada com a inserção das informações de população e das informações regionais do IBGE referentes a cada município. Para a correta manipulação, os dados numéricos foram ajustados para valores inteiros ou reais, conforme a necessidade, com o uso da biblioteca *NumPy*. Após esses processos, percebeu-se que era necessário que a tabela fosse convertida para uma codificação que aceitasse as regras ortográficas da língua portuguesa. Dessa maneira, a tabela foi convertida para a codificação *latin-1*.

Por último, para corrigir possíveis equívocos, foram alterados os nomes das colunas (atributos), de modo a tornar mais intuitiva a manipulação futura durante a execução de pesquisas e construção de testes computacionais. Com todas estas ações descritas realizados, foi possível obter os seguintes atributos, que são visíveis na Tabela 1, que apresenta de qual conjunto os dados foram importados, seu nome e o tipo de

informação que contém.

Tabela 1 – Descrição dos atributos contidos na tabela usada no trabalho.

Conjunto	Atributo	Tipo de dado
IBGE	Nome cidade	String
	Id cidade	Inteiro
	Id microrregião	Inteiro
	Nome microrregião	String
	Id mesorregião	Inteiro
	Nome mesorregião	String
Todos	Ano	Inteiro
PBF	Total de beneficiários	Inteiro
	Valor total repassado	Real
IDESE	Idese	Real
	Anos Finais EF	Real
	Anos Iniciais EF	Real
	Ensino Fundamental	Real
	Ensino Médio	Real
	Escolaridade Adulta	Real
	Pré Escola	Real
	Bloco Educação	Real
	Apropriação da Renda	Real
	Geração da Renda	Real
	Bloco Renda	Real
	Mortes por Causas Evitáveis	Real
	Condições Gerais de Saúde	Real
	Óbitos por Causas Mal Definidas	Real
	Longevidade	Real
	Consultas Pré Natal	Real
	Mortalidade de Menores de 5 anos	Real
	Saúde Materno Infantil	Real
	Bloco Saúde	Real
População	População	Inteiro

Fonte: Elaborado pelo Autor

A tabela final foi colocada em um formato que trouxesse maior possibilidade de manipulação com o uso das bibliotecas contidas no Pythone, por isso, foi escolhido o formato de arquivo *.csv*. A tabela final contém 15 anos de registros do IDESE e PBF, referentes aos 497 municípios que pertencem ao estado do Rio Grande do Sul. Obtém-se, assim, um conjunto de dados composto por 29 colunas (atributos) e 7455 linhas (amostras).

Adicionalmente, durante os testes e a execução dos métodos computacionais, foram criados três atributos derivados da combinação do total de beneficiários, do valor total repassado e da população. Tais atributos são: População Beneficiária, Repasse por Beneficiário e Repasse por População, cujas fórmulas de cálculo são descritas nas

Equações 1, 2 e 3, respectivamente.

$$\text{População beneficiária} = \frac{\text{Total de beneficiários}}{\text{População}} \tag{1}$$

$$\text{Repasse por beneficiário} = \frac{\text{Valor total repassado}}{\text{Total de beneficiários}} \tag{2}$$

$$\text{Repasse por população} = \frac{\text{Valor total repassado}}{\text{População}} \tag{3}$$

Durante a execução dos testes dos métodos computacionais, fez-se necessário empregar a combinação de atributos do IDESE. Esses indicadores foram utilizados para criar casos de teste e agrupar informações dos blocos do IDESE que apresentam forte correlação entre si. As novas variáveis são descritas a seguir, com seu nome acompanhado dos componentes que as integram:

- **blocos**: Bloco Educação, Bloco Renda, Bloco Saúde;
- **blocoEducacaoResumido**: Ensino Fundamental, Ensino Médio, Escolaridade Adulta, Pré Escola;
- **blocoEducacaoTodos**: Anos Finais EF, Anos Iniciais EF, Ensino Fundamental, Ensino Médio, Escolaridade Adulta, Pré Escola;
- **blocoRendaTodos**: Apropriação da Renda, Geração da Renda;
- **blocoSaudeResumido**: Condições Gerais de Saúde, Longevidade, Saúde Materno Infantil;
- **blocoSaudeTodos**: Mortes por Causas Evitáveis, Condições Gerais de Saúde, Óbitos por Causas Mal Definidas, Longevidade, Consultas Pré Natal, Mortalidade de Menores de 5 anos, Saúde Materno Infantil.

2.1.3 Análise e validação

Nesta seção são apresentados os processos metodológicos empregados para a obtenção das análises e resultados preliminares e finais contidos neste trabalho. Tais procedimentos sofrem influência das seções anteriores e, consequentemente, também as influenciam, estabelecendo um ciclo iterativo até a validação final dos resultados.

Como primeira ação, realizou-se uma busca por métricas básicas nos conjuntos de dados do PBF e IDESE. Essa etapa foi conduzida com o auxílio da biblioteca Pandas do Python, que dispõe de ferramentas específicas para o tratamento de dados tabulares. As métricas incluem média, mediana, desvio padrão, entre outras, considerando um período anual ou abrangendo os 15 anos reunidos nos dados. Por meio de atributos específicos, selecionaram-se divisões territoriais, como micro/mesorregiões ou o estado agregado, para consolidar informações de todos os municípios. Tais medidas servem como subsídio para a apresentação numérica dos dados, mediante a construção de texto ou tabelas, e principalmente para a geração de gráficos, os quais podem facilitar a visualização das informações de maneira mais didática ao leitor. Essas ações estão alinhadas com os princípios da ciência de dados, foco deste trabalho (MORETTIN; SINGER, 2022).

A tabela unificada com todos os dados foi primordial para o desenvolvimento de testes e execução de métodos computacionais, os quais geraram condições para análises preliminares e resultados finais, contidos no Capítulo 5. Os dados foram utilizados para a aplicação de técnicas de aprendizado de máquina (*machine learning*) e aprendizado profundo (*deep learning*), sendo selecionados dois tipos de aprendizado para este trabalho: supervisionado e não supervisionado.

Cada técnica de aprendizado de máquina possui configurações e objetivos distintos. O aprendizado não supervisionado é representado pelo método ***K-means***, que visa rotular em agrupamentos as amostras contidas em um conjunto de dados. Para o aprendizado supervisionado, foram escolhidas três técnicas distintas, permitindo a comparação de resultados e desempenho: a **Regressão Linear**, por ser uma técnica simples para verificar a relação entre uma ou mais características com o atributo a ser predito; a técnica conhecida como **Redes Neurais *Multilayer Perceptron*** (MLP), que possibilita verificar relações não lineares presentes nos dados; e a técnica ***XGBoost***, que apresenta alta capacidade de generalização de relações entre as amostras, possibilitando um elevado grau de acerto nos resultados finais. Essas técnicas são descritas em maiores detalhes no Capítulo 3. A seguir, apresenta-se uma breve descrição de cada método computacional de aprendizado de máquina utilizado, contendo seu principal modo de uso e aplicação no trabalho:

- ***k-means***: O método *k-means*, disponibilizado no conjunto de bibliotecas do *Sci-kit Learning* (SCIKIT-LEARN, 2025), foi empregado para agrupar amostras dos dados em um determinado número k de grupos (*clusters*) (HARRISON, 2019), com base nos atributos previamente selecionados e extraídos do conjunto de dados. O

K-means foi utilizado para apresentar grupos com amostras relacionadas entre si, o que possibilitou um melhor emprego dos métodos de aprendizado supervisionado, os quais são alimentados com os dados e, dessa forma, podem entregar resultados mais precisos. Consequentemente, os melhores resultados são separados e as características dos grupos são analisadas em detalhes.

- **Regressão Linear**: O método de Regressão Linear, obtido do conjunto de bibliotecas do *Scikit-learn* (SCIKIT-LEARNING, 2025), teve como intuito a identificação de correlações lineares simples ou múltiplas entre as variáveis previamente selecionadas e os grupos gerados pelo *K-means*. Esta técnica foi escolhida por ser simples, de fácil entendimento e amplamente utilizada historicamente em comparação com métodos mais complexos (HARRISON, 2019).

- **Redes Neurais *Multilayer Perceptron***: Este método é aplicado com o uso de duas bibliotecas, *Keras* (KERAS, 2025) e *Tensorflow* (TENSORFLOW, 2024), necessárias para a construção robusta de sua arquitetura de implementação, que visa emular o cérebro humano (HARRISON, 2019). A MLP foi empregada para explorar relações não lineares entre as amostras de cada grupo separado pelo *K-means*, visando aprofundar a análise e ampliar a compreensão dos dados, de modo a alcançar melhores resultados nas métricas sob análise.

- ***XGBoost***: Este método, pertencente à biblioteca Documentation (2023), tem como principal objetivo a descoberta de relações não lineares entre as amostras selecionadas, com capacidade de generalizar as relações de forma mais aprimorada que a **MLP**, por meio da agregação e aprimoramento de uma ou mais técnicas de aprendizado de máquina consideradas simples (HARRISON, 2019). Neste trabalho, o *XGBoost* foi aplicado para verificar as relações dos dados entre os grupos gerados pelo *K-means*, apresentando uma forma de comparação com os demais métodos.

Para aprimorar os resultados dos métodos de aprendizado de máquina, utilizaram-se algoritmos específicos que auxiliam na obtenção de melhores desempenhos nas fases de treinamento e execução. No caso do aprendizado não supervisionado, representado pelo método *k-means*, empregou-se o cálculo do **coeficiente de silhueta**, o qual permitiu determinar a quantidade ideal de grupos (*clusters*) para o conjunto de dados, evitando a utilização de abordagens puramente empíricas e de tentativa e erro para o valor de k, que representa a quantidade de agrupamentos que o *K-means* deve separar.

Nos métodos supervisionados, utilizou-se a **validação cruzada *k-fold***, técnica que divide os dados fornecidos ao método computacional, de modo que este execute múltiplas vezes com uma diversidade de amostras, permitindo que uma amostra esteja tanto no conjunto de treinamento quanto no de validação do modelo gerado. A validação cruzada *k-fold* tem a finalidade de evitar discrepâncias que podem existir nos dados e que poderiam afetar os resultados, causando análises equivocadas (IZBICKI; SANTOS, 2020).

Para a comparação dos desempenhos da Regressão Linear, MLP e *XGBoost*, utilizou-se a média de duas métricas amplamente empregadas em ciência de dados (MORETTIN; SINGER, 2022): o **coeficiente de determinação**, conhecido como R^2, e o **Erro Quadrático Médio (*Mean Squared Error*) (MSE)**. Ambos os valores foram calculados utilizando dados separados para validação dos modelos gerados em cada um dos métodos de aprendizado supervisionado apresentados.

A aplicação dessas técnicas de aprimoramento de modelos e das métricas de desempenho assegura que os métodos computacionais empregados estejam alinhados aos objetivos deste trabalho, fornecendo um suporte analítico robusto para as conclusões e interpretações propostas na pesquisa.

2.2 Ferramentas

O objetivo desta seção é apresentar de forma clara as ferramentas de hardware, software e de programação empregadas neste trabalho. Tais descrições são de extrema relevância para possibilitar reproduções futuras e para a reflexão sobre aprimoramentos em pesquisas subsequentes.

2.2.1 Ambiente de desenvolvimento e execução

Nesta seção são apresentadas as ferramentas e configurações de sistema utilizadas para o desenvolvimento dos algoritmos de programação, bem como os softwares empregados na execução dos códigos e a máquina responsável pela realização dos métodos computacionais descritos na seção anterior.

Utilizaram-se dois ambientes de desenvolvimento durante a criação de testes e execução dos códigos. Essa opção visou identificar a ferramenta mais adequada aos propósitos do trabalho. Inicialmente, o ***Google Colab*** (COLAB, 2025) foi escolhido

por facilitar a integração entre código, visualização de resultados e documentação ao longo das etapas da pesquisa, sem a necessidade de armazenamento ou execução em máquina física do pesquisador. Posteriormente, ao analisar as características do trabalho, verificou-se que o conjunto de dados final possuía volume reduzido de informações e que o tempo de execução dos métodos computacionais encontrava-se dentro de uma margem aceitável. Assim, optou-se pelo uso do *framework* ***Jupyter Notebook*** (JUPYTER, 2025), que permite a criação e armazenamento de códigos e arquivos diversos diretamente na máquina do pesquisador. A escolha por essa ferramenta visou a obtenção de resultados robustos nas análises subsequentes.

As configurações de hardware e software para criação e execução de todas as técnicas computacionais aqui apresentadas são descritas na Tabela 2, que apresenta as especificações dos componentes físicos da máquina utilizada, além de seu sistema operacional:

Tabela 2 – Especificações técnicas do ambiente de desenvolvimento.

Componente	**Especificação**	**Detalhes / Notas**
Tipo de Ambiente	Local (Máquina Física)	Notebook Acer Aspire A515-57, Utilizado para testes, validação de desempenho e comparações.
Sistema Operacional	Windows 11 (64 bits)	Edição Home Single Language e Versão 24H2
Processador (CPU)	Intel(R) Core(TM) i5-12450H	8 Cores, 12 Threads, Clock de 2.0 GHz a 4.4 GHz
Memória RAM	32 GB DDR4	2667 MTs, Dual-Channel
Placa de Vídeo (GPU)	Intel® UHD Graphics for 12th Gen Intel® Processors	Frequência: 1.20 GHz, Memória compartilhada com CPU: 16 GB
Armazenamento	SSD (NVMe) 256 GB (Hana HFS256GEJ9X110N)	Leitura: 4.696 MB/s, Escrita: 2.710 MB/s

Fonte: Elaborado pelo Autor

Ressalta-se que as informações técnicas apresentadas correspondem às disponibilizadas pelos fabricantes do hardware e dos softwares utilizados. No entanto, o desempenho pode não ser considerado fixo devido a oscilações de frequência ou mau gerenciamento de memória em algum componente específico, o que pode acarretar variações nos resultados que venham a ser reproduzidos futuramente.

2.2.2 Ambiente de programação

Nesta seção apresentam-se as informações e justificativas para a seleção da linguagem de programação e das bibliotecas de algoritmos utilizadas neste trabalho. Essas escolhas foram motivadas principalmente pelas possibilidades de manipulação de dados, criação e execução de algoritmos, e pelo amplo emprego no campo da ciência de dados.

Para a manipulação de dados, criação e execução dos códigos das técnicas computacionais e apresentação de informações, optou-se pela linguagem de programação **Python** (FOUNDATION, 2025b), uma das mais bem-sucedidas atualmente e com grande adesão na área de ciência de dados. O Python foi empregado na implementação de todas as etapas metodológicas do estudo, com destaque para o uso de bibliotecas amplamente reconhecidas na literatura científica e consolidadas na prática acadêmica. Essas bibliotecas desempenharam um papel essencial na execução das análises propostas e são detalhadas a seguir:

- **Pandas e NumPy**: Utilizadas para a manipulação e organização de diferentes tipos de dados, fornecendo ferramentas robustas para transformação, estruturação e análise de grandes volumes de informações, sejam dados únicos ou tabulares (PANDAS, 2024; NUMPY, 2022).

- **Matplotlib, Seaborn e Geopandas**: Empregadas para a visualização gráfica dos dados e resultados, permitindo a criação de representações visuais claras e informativas (MATIPLOTLIB, 2025; SEABORN, 2025; TEAM, 2025).

- **Scikit-learn**: Por se tratar de uma biblioteca vasta, foi extremamente útil em diversas etapas do trabalho, sendo aplicada na implementação de métodos computacionais de aprendizado de máquina, como Regressão Linear e o algoritmo *k-means*. Além disso, contribuiu com técnicas algorítmicas para aprimoramento dos métodos, como a aplicação de validação cruzada *K-fold* e do coeficiente de silhueta, e possibilitou a geração automática de métricas como R^2 e *MSE* (SCIKIT-LEARNING, 2025).

- **TensorFlow e Keras**: Empregadas na construção e no treinamento de Redes Neurais Artificiais do tipo *Multilayer Perceptron* (MLP), proporcionando flexibilidade na configuração da arquitetura das redes e eficiência no processamento dos dados (TENSORFLOW, 2024; KERAS, 2025).

- ***XGBoost***: Foi aplicada para facilitar a criação e execução do método computacional *XGBoost* (DOCUMENTATION, 2023), possibilitando a configuração de parâmetros e a integração com as demais bibliotecas utilizadas.

- **Requests e JSON**: Bibliotecas destinadas à requisição de informações no formato JSON, armazenadas em ambientes externos à aplicação (COMMUNITY, 2025; FOUNDATION, 2025a).

- **Time**: Biblioteca específica para extração de tempos de execução, utilizada para verificar diferenças na performance dos métodos computacionais.

- **Warnings**: Utilizada para suprimir mensagens de alerta exibidas por bibliotecas durante a execução de códigos específicos.

Apresentam-se acima vasta quantidade das ferramentas de programação utilizadas. No entanto, deve-se ressaltar que o Python possui bibliotecas integradas, cuja indicação de uso não é necessária, dispensando, portanto, descrição, justificativa e apresentação.

3 FUNDAMENTAÇÃO TEÓRICA

O presente capítulo tem por objetivo apresentar a fundamentação teórica que sustenta este trabalho, organizando-se em eixos temáticos distintos, que envolvem aspectos de políticas públicas e ciência de dados como temas principais. Na Seção 3.1 são discutidos os conceitos de políticas públicas, mais especificamente os referentes à transferência de renda, com ênfase na natureza redistributiva dessas ações e no ciclo de vida das políticas públicas, culminando em um estudo detalhado do Programa Bolsa Família (PBF), analisando suas condicionalidades e os principais desafios de monitoramento e avaliação. A Seção 3.2 explora os índices socioeconômicos como ferramentas analíticas essenciais para a mensuração de impactos em programas sociais. Inicialmente, define-se o que é um índice, com o intuito de discutir sua construção e utilização em políticas públicas, tendo em vista a análise do Índice de Desenvolvimento Socioeconômico do Rio Grande do Sul (IDESE) exposto na subseção 3.2.1, onde são detalhadas a sua origem, a estrutura metodológica e as categorias de interpretação dos resultados.

Em seguida, são tratados os temas referentes à ciência de dados, onde a Seção 3.3 apresenta os fundamentos da ciência de dados, destacando sua natureza interdisciplinar e o fluxo típico de um projeto de análise de dados. A subseção 3.3.1 introduz as principais técnicas de aprendizado de máquina que serão empregadas neste estudo, discriminando os métodos não supervisionados, *K-means*, e supervisionados, Regressão Linear, Redes Neurais MLP e *XGBoost*. Nas Subseções 3.3.1.5 e 3.3.1.6 detalham-se os procedimentos de aprimoramento aplicados durante o uso de modelos, como validação cruzada *k-fold* e coeficiente de silhueta, voltados a garantir robustez e qualidade nas predições e agrupamentos. Por fim, a Subseção 3.3.2 apresenta as métricas estatísticas que serão utilizadas para avaliar o comportamento dos dados e a eficiência dos modelos, incluindo média, mediana, desvio padrão, resíduos, coeficiente de determinação (R^2) e Erro Quadrático Médio (MSE).

Dessa forma, este capítulo estabelece o arcabouço teórico necessário para compreender tanto o contexto institucional das políticas de transferência de renda quanto as ferramentas quantitativas e computacionais empregadas na análise utilizando ciência de dados.

3.1 Políticas pública de transferência de renda

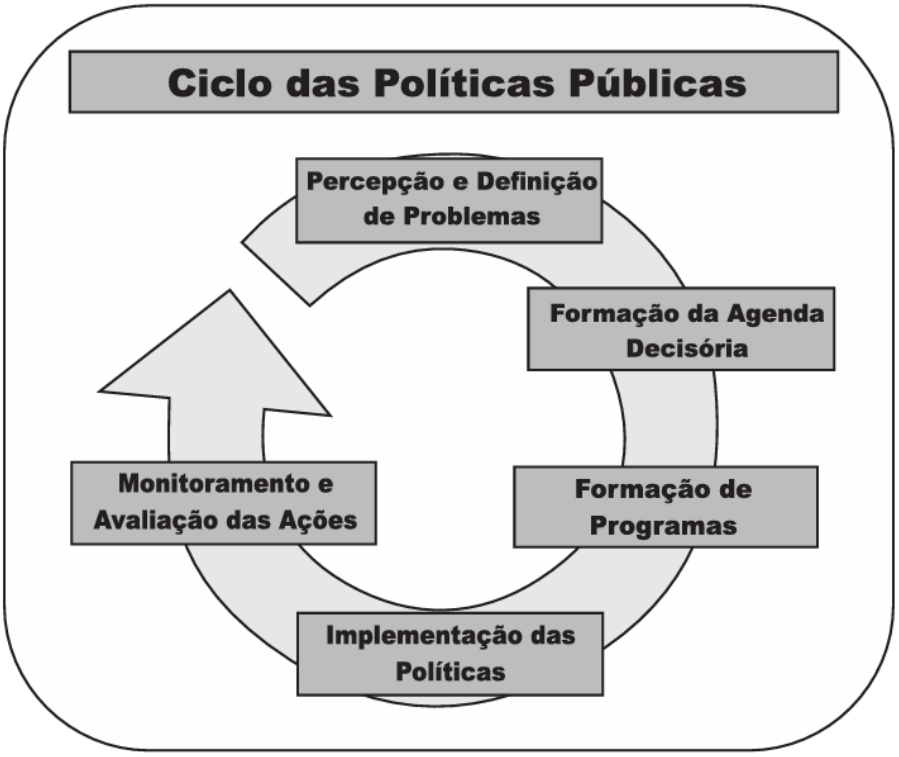
O termo **política pública** possui definições variadas dependendo da área de aplicação, sendo amplamente discutido nas ciências políticas, na sociologia, na economia e em áreas correlatas. Sob uma perspectiva social, políticas públicas são entendidas como conjuntos de ações destinadas a ampliar o acesso a serviços essenciais, seja para parcelas específicas da população ou para sua totalidade, além de oferecer subsídios para pesquisas sobre seus impactos diretos e indiretos. Essas políticas, em sua maioria, são formuladas por instâncias governamentais, mas também podem adotar modelos híbridos, envolvendo parcerias com a iniciativa privada, com o objetivo de promover o bem-estar social durante todo o ciclo de sua implementação (SOUZA, 2006; MASTRODI; IFANGER, 2019).

Neste trabalho, o foco recai sobre políticas públicas implementadas por governos, que frequentemente seguem diretrizes definidas como direitos sociais constitucionais de caráter universal. Tais políticas podem ser direcionadas a objetivos específicos ou mais amplos. Exemplos incluem ações voltadas para a redução da desigualdade social, o combate à pobreza e o acesso universal a serviços de saúde e educação (MELAZZO, 2010). A complexidade conceitual na definição de políticas públicas também se reflete nas etapas do ciclo de vida dessas políticas. Diversos autores propõem classificações distintas para essas etapas. Neste trabalho, adota-se o modelo de Raeder (2014), que sintetiza a literatura em cinco fases principais:

1. Percepção e definição de problemas;
2. Formação da agenda decisória;
3. Formulação de programas e projetos;
4. Implementação das políticas delineadas;
5. Monitoramento e avaliação das ações planejadas.

Este trabalho concentra-se na fase de monitoramento e avaliação das ações planejadas, que permeia todo o ciclo de vida da política pública, fornecendo retroalimentação para ajustes e mensuração de eficácia, esse ciclo é mostrado na Figura 2. Segundo Worthen et al. (2004), essa etapa visa atribuir valor aos objetivos pensados, identificando se os resultados obtidos foram produtivos.

Figura 2 – Etapas do ciclo de vida das políticas públicas.



Fonte: (RAEDER, 2014).

Segundo a classificação de Lowi (1964), que identificou uma tipologia para as políticas públicas e que é amplamente adotada na gestão pública, pode-se categorizar as políticas em:

- **Regulatórias**: Estabelecem normas e controles;
- **Distributivas**: Alocam recursos sem redistribuição explícita;
- **Redistributivas**: Transferem recursos entre grupos sociais;
- **Constitutivas**: Definem regras gerais de processo.

As políticas públicas de transferência de renda, foco deste trabalho, enquadram-se no tipo redistributiva, caracterizado pela realocação de recursos de um grupo para outro, frequentemente este tipo de política pode gerar conflitos devido à natureza conhecida como soma zero (SECCHI, 2014).

3.1.1 Programa bolsa família

Uma das ações adotadas no Brasil que obteve maior sucesso na implementação de uma política pública redistributiva de renda é o Programa Bolsa Família (PBF) (MDS, 2023). Este programa pode ser mais precisamente definido como um programa de transferência direta e condicionada de renda. O PBF foi criado em 2003 pela Medida Provisória nº 132/2003 (BRASIL, 2003), que unificou diversos outros programas anteriormente estabelecidos com a mesma finalidade de auxiliar parcelas vulneráveis da população. No entanto, o PBF introduziu a exigência de condicionalidades que os beneficiários deveriam cumprir para adquirir e também continuar a receber o recurso.

O objetivo central do PBF é quebrar o ciclo intergeracional da pobreza das famílias beneficiárias, para isso as condicionalidades de recebimento dos valores são usados como suporte para se obter uma melhoria dos níveis de educação, saúde e bem-estar social.

O acompanhamento do cumprimento dessas regras é realizado por meio do sistema Cadastro Único (CADÚnico) que é operado de forma decentralizada pela união, estados e municípios, que serve como base para verificar a progressão dos beneficiários durante sua permanência no programa. O CADÚnico também é utilizado para determinar se o beneficiário deve ser removido do programa por descumprimento de alguma das regras estabelecidas em lei.

No ano de 2004, com a Lei Nº.10.836/2004 (BRASIL, 2004), foram formadas as bases para o funcionamento do programa até o momento em que o PBF sofreu alterações fundamentais por meio da Lei Nº.14.284/2021 (BRASIL, 2021) e posteriormente pela Lei N°.14.601/2023 (BRASIL, 2023).

As condicionalidades contidas no Artigo 3° da lei de 2003, apresentado a seguir, são alvo de discussões continuas sobre a sua eficácia na sociedade.

> "A concessão dos benefícios dependerá do cumprimento, no que couber, de condicionalidades relativas ao exame pré-natal, ao acompanhamento nutricional, ao acompanhamento de saúde, à frequência escolar de 85% (oitenta e cinco por cento) em estabelecimento de ensino regular, sem prejuízo de outras previstas em regulamento."(BRASIL, 2004, art. 3º)

Os trabalhos dedicados à análise das condicionalidades e demais impactos do PBF nos grupos que recebem ou não o benefício devem ser bem estruturados para que possam ser utilizados como suporte pelos governos, a fim de verificar se houve geração de valor relevante para os objetivos desejados em períodos que antecedem, sucedem e também durante a aplicação do programa social (RAMOS; LIMA, 2014).

O estudo de Martins e Rückert (2022) demonstra, por meio de investigação empírica, que não há discrepâncias significativas no desempenho escolar entre discentes beneficiários e não beneficiários em contextos socioeducacionais restritos. A pesquisa sugere que variáveis estruturais do sistema público de ensino, como a qualidade da infraestrutura escolar e a capacitação docente, exercem influência mais determinante no rendimento discente do que o mero cumprimento da frequência escolar obrigatória. Esses achados apontam para a necessidade de políticas complementares que integrem melhorias sistêmicas no ecossistema educacional.

No âmbito da saúde, a investigação de Damião et al. (2021) identifica um fenômeno de relaxamento progressivo na aplicação das condicionalidades por parte dos profissionais da área. A análise evidencia que, embora o acompanhamento estatístico permaneça formalizado, ocorre uma burocratização crescente dos processos de coleta de dados, com redução do rigor analítico nas consultas pediátricas. Essa tendência gera distorções na avaliação territorial das políticas públicas, podendo comprometer a precisão de indicadores socioassistenciais e, consequentemente, a alocação estratégica de recursos. Tais evidências reforçam a importância de mecanismos de auditoria contínua e capacitação técnica para preservar a integridade operacional do programa.

3.2 Índices socioeconômicos

Em primeiro lugar, é necessário associar o termo indicador como a base para a criação de um índice, pois um indicador é uma ferramenta que sintetiza dados quantitativos ou qualitativos, representando realidades complexas de maneira objetiva. A construção de um indicador é realizada a partir da combinação estruturada de variáveis brutas, como dados provenientes de pesquisas, censos ou registros administrativos. Isoladamente, um indicador possui um significado limitado, mas sua integração com outros indicadores pode fornecer subsídios para a criação de um índice, que permitirá traduzir fenômenos sociais e econômicos em métricas comparáveis (JANNUZZI, 2009).

Por meio de cálculos estatísticos, como razões, médias ou valores compostos, os índices tornam-se essenciais para orientar decisões governamentais e da iniciativa privada. Esses dados são aplicados em todo o ciclo de vida de uma política pública, incluindo a etapa de avaliação, auxiliando na compreensão de como os resultados da política pública são influenciados pelos índices selecionados (MPOG; SPI, 2012).

A Figura 3 ilustra o processo padrão de criação de um indicador, que inclui

etapas como coleta de dados brutos, normalização dos dados, geração e interpretação dos resultados. Esse fluxo destaca a importância da transparência metodológica para evitar distorções na análise e garantir que os indicadores sejam utilizados de maneira eficaz na formulação de políticas públicas.

Figura 3 – Fluxo de criação de um indicador.

Fonte: Elaborado pelo Autor, baseado em Jannuzzi (2009).

A interação entre as dimensões econômica e social gera os índices socioeconômicos, que revelam como fatores econômicos influenciam, e são influenciados, pelas dinâmicas sociais. A construção desses índices varia conforme o contexto analítico; por exemplo, na dimensão social, eles quantificam aspectos intangíveis do bem-estar coletivo, como acesso à educação, condições de saúde ou percepção de segurança (JANNUZZI, 2009). Já os índices econômicos concentram-se em variáveis mensuráveis relacionadas à produção, distribuição e consumo de recursos, incluindo métricas macroeconômicas, como o Produto Interno Bruto (PIB) e a inflação, além de elementos microeconômicos, como a renda familiar e o custo de vida (JANNUZZI, 2014).

O avanço além de paradigmas reducionistas, que mediam o progresso social apenas pela riqueza material, levou à criação de índices mais abrangentes. Nesse cenário, o Índice de Desenvolvimento Humano (IDH), desenvolvido pelo Programa das Nações Unidas para o Desenvolvimento (PNUD) em 1990 (STANTON, 2007), tornou-se uma referência global, dando origem no Brasil a vários outros índices socioeconômicos ao longo do tempo, como FIRJAM, IDHM e outros índices regionais, como o Índice de Desenvolvimento Socioeconômico (IDESE), sob estudo neste trabalho.

Em relação ao IDH, seu diferencial reside na combinação de três dimensões essenciais: educação, medida por anos de escolaridade; saúde, medida pela expectativa de vida; e renda, medida pela renda nacional bruta per capita. Essa abordagem permite avaliar como países ou regiões transformam recursos econômicos em qualidade de vida, oferecendo uma visão mais equilibrada do desenvolvimento humano (STANTON, 2007).

Apesar de sua ampla adoção em políticas públicas, o IDH não está livre de

críticas. Estudos apontam que sua metodologia simplificada pode ignorar desigualdades internas e fatores ambientais da sociedade em estudo (NAYAK, 2015). Mesmo assim, sua capacidade de sintetizar dados complexos em um único índice mantém sua relevância para comparações internacionais (SAAB et al., 2021). As atualizações metodológicas ao longo dos anos, como ajustes para disparidades regionais, mostram que o IDH continua evoluindo, servindo como base para o desenvolvimento de novos índices que complementam sua análise.

3.2.1 Índice de desenvolvimento socioeconômico do Rio Grande do Sul (IDESE)

O Índice de Desenvolvimento Socioeconômico do Rio Grande do Sul (IDESE), criado em 2003 pela extinta Fundação de Economia e Estatística (FEE/RS) (FEE ACCURSO; HEUSER, 2003), é um índice sintético inspirado no IDH, exposto na Seção 3.2, mas com adaptações metodológicas para análise regionalizada. Sua principal inovação reside na periodicidade anual e na granularidade geográfica, que permite comparar municípios, microrregiões, mesorregiões e Conselhos Regionais de Desenvolvimento (Coredes) do estado do Rio Grande do Sul, a partir de três eixos principais: educação, renda e saúde (DEE-RS, 2024).

Em 2013, após uma década de aplicação, o IDESE passou por revisão metodológica para corrigir limitações identificadas em sua estrutura original (KANG et al., 2014). A versão atualizada, cuja arquitetura é ilustrada na Figura 4, organiza-se em um valor principal, 3 blocos temáticos, 9 sub-blocos compostos de índices ou indicadores, e mais 6 indicadores que compõem os índices. Os valores são expressos na faixa de 0 a 1, e os valores iniciais são obtidos de indicadores que passam por uma normalização de valores, a qual pode ser generalizada pela Equação 4, garantindo comparabilidade entre indicadores de escalas distintas (BERNARDINI et al., 2017):

$$I_{x_j,t} = \frac{y_{x_j,t} - LI_x}{LS_x - LI_x} \tag{4}$$

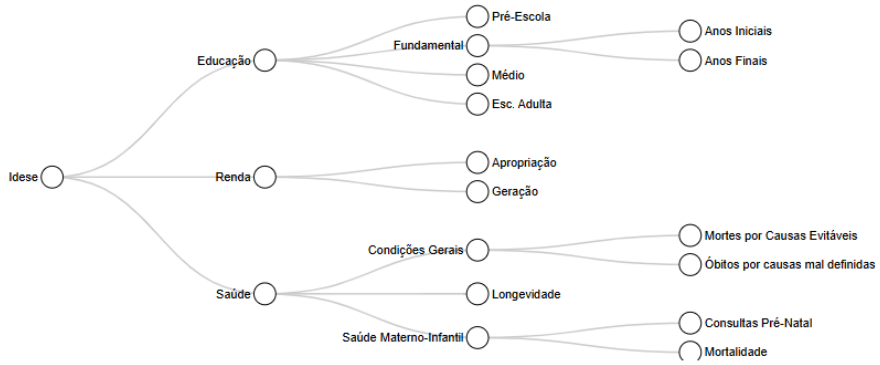
$I_{x_j,t}$: Índice normalizado do indicador x na unidade geográfica j no ano t

$y_{x_j,t}$: Valor bruto do indicador x

LI_x, LS_x: Limites inferior e superior pré-definidos para o indicador x

Cada bloco temático e o valor final do IDESE resultam da média aritmética

Figura 4 – Estrutura metodológica do IDESE.



Fonte: (DEE-RS-ESTRUTURA, 2025).

simples de seus sub-blocos, em que os valores dos índices são obtidos conforme detalhado por BERNARDINI et al. (2017), resumidos a seguir:

- **Bloco Educação**: Combina quatro sub-blocos que avaliam desde a universalização da educação infantil e do ensino médio até o desempenho no Sistema de Avaliação da Educação Básica (SAEB). A inclusão de métricas de qualidade educacional visa superar a mera mensuração de acesso.

- **Bloco Renda**: Divide-se em duas dimensões: geração de renda, por meio do cálculo da renda per capita, e apropriação, obtida pelo cálculo da proporção de ocupados com renda acima de dois salários mínimos. Essa dualidade busca equilibrar análises coletivas e individuais, corrigindo vieses históricos de indicadores puramente monetários.

- **Bloco Saúde**: Integra três valores com indicadores objetivos para minimizar subjetividades, por meio do cálculo da mortalidade infantil, cobertura de atenção básica e proporção de óbitos por causas mal definidas. A seleção prioriza variáveis sensíveis a políticas públicas.

Após a mensuração, cada valor pode ser categorizado e analisado em três faixas: valores inferiores a 0,500 indicam baixo desenvolvimento; valores entre 0,500 e 0,799 são classificados como desenvolvimento médio; e valores iguais ou superiores a 0,800 denotam alto desenvolvimento (DEE-RS, 2020).

Com o passar do tempo, agregou-se grande valor às dimensões e ao valor total do IDESE, gerando estudos que analisaram esses dados de forma ampla ou direcionada. Dias (2021) analisou os impactos dos valores do IDESE nos municípios onde foi implantada a

Universidade Federal do Pampa (UNIPAMPA) e constatou que houve melhora nos valores gerais do IDESE na década após a implantação, em comparação com a década anterior.

Uma análise ampla realizada por Lied (2024) observou a existência de discrepâncias críticas entre as características dos municípios e regiões do estado do Rio Grande do Sul, e recomenda um foco das gestões estadual, regional e municipal. Um exemplo de discrepância são municípios que possuem altos índices de desenvolvimento em algum bloco específico do IDESE, mas não necessariamente alcançam êxito nos outros blocos, o que pode impedir a obtenção de um valor geral elevado do IDESE.

3.3 Ciência de Dados

A Ciência de Dados (*Data Science*) constitui um campo interdisciplinar de conhecimentos que engloba estatística, computação e ciências humanas, essa convergência é ilustrada na Figura 5, que representa um diagrama de Venn, com a ciência de dados no centro e os diversos conhecimentos sendo sobrepostos. Estes campos de conhecimento em conjunto são necessários para que se realize o tratamento e a extração de conhecimentos significativos a partir de grandes conjuntos de dados, visando apresentar tais informações de maneira clara e eficiente (PIERSON, 2019; GRUS, 2021).

Figura 5 – Diagrama de Venn para ciência de dados.

Fonte: Baseado em Conway (2013).

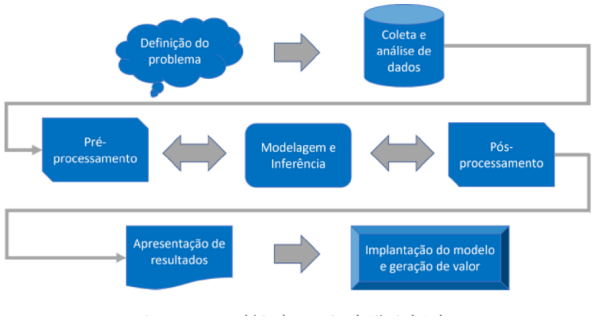
Como visto, o cientista de dados deve se aprofundar em diversas áreas de

conhecimento para obter recursos de análise. Na estatística, o conhecimento é útil para a coleta, organização e análise das informações obtidas a partir dos dados iniciais e posteriores resultados. A computação desempenha um papel essencial na gestão de bancos de dados e no uso de linguagens de programação para o desenvolvimento e aplicação de algoritmos de aprendizado de máquina (*Machine Learning*) e aprendizado profundo (*Deep Learning*), ambos subáreas da Inteligência Artificial. Por sua vez, as ciências humanas são indispensáveis na definição do problema a ser abordado, na escolha das técnicas adequadas e na interpretação dos resultados de forma clara e objetiva (MORETTIN; SINGER, 2022).

A ciência de dados expandiu-se significativamente nas últimas décadas, principalmente devido ao crescimento do fenômeno conhecido como *Big Data*. Este crescimento tem sido impulsionado pelo aumento exponencial no volume, na variedade e na velocidade de geração de dados, assim como pelos avanços no poder de processamento dos computadores (KALINOWSKI et al., 2023). Embora um cientista de dados possa realizar a coleta, tratamento e armazenamento dos dados, esta é uma especialização da engenharia de dados, que atualmente necessita executar tais ações para o bom uso do grande volume de dados (PIERSON, 2019). Por considerar que o volume de dados analisado neste trabalho não é excessivo, estes processos de engenharia de dados foram executados e são apresentados no Capítulo 2.

Como se observa, a área de ciência de dados vai muito além da aplicação mecânica de algoritmos. É exigida uma abordagem sistêmica que abrange desde a formulação precisa do problema de pesquisa até a validação dos dados coletados, com a inclusão de pré-processamento dos dados, seleção criteriosa das técnicas analíticas e, por fim, a interpretação crítica dos resultados. A Figura 6 ilustra essa estrutura metodológica, enfatizando a importância de cada etapa no desenvolvimento de projetos em ciência de dados.

Figura 6 – Fluxo de projeto de ciência de dados



Fonte: (ESCOVEDO; KOSHIYAMA, 2020).

No decorrer desta seção, o foco será em apresentar conceitos usados neste trabalho, para duas das três grandes disciplinas internas à ciência de dados: a computação, voltada à criação e aplicação de métodos computacionais de aprendizado de máquina para interpretação e extração de informações dos dados brutos, e a estatística, essencial para a obtenção de informações relevantes de avaliação dos dados iniciais e de desempenho dos resultados obtidos por meio dos métodos computacionais. As ciências humanas, por sua vez, têm sua aplicação em todo o âmbito do trabalho, visando a formulação de questões de pesquisa e a obtenção de respostas para tais perguntas.

3.3.1 Aprendizado de máquina

A etapa de criação e aplicação de modelos estatísticos e métodos computacionais deve ser realizada após uma adequada organização dos dados coletados e refinados, conforme apresentado no Capitulo 2. Essa etapa computacional é essencial para identificar padrões que estejam presentes nos dados de forma explícita ou implícita (MORETTIN; SINGER, 2022).

As técnicas computacionais que serão exploradas neste trabalho incluem aprendizado de máquina supervisionado e não supervisionado. Ambas desempenham um papel fundamental na análise e interpretação dos dados, permitindo que se encontrem correlações relevantes e agrupamentos significativos (GÉRON, 2019; ESCOVEDO; KOSHIYAMA, 2020).

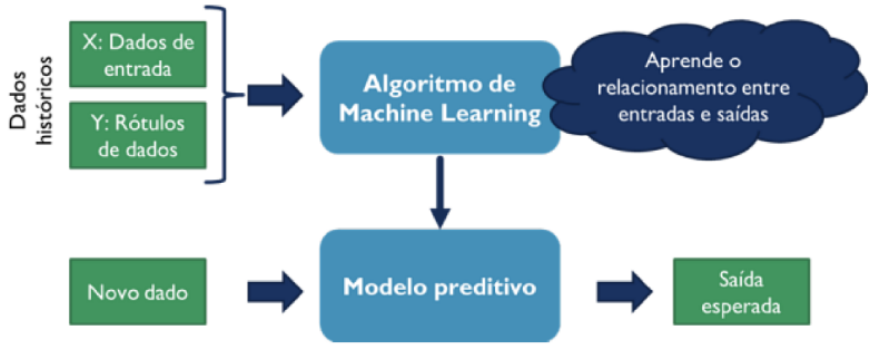
- **Aprendizado supervisionado**: A abordagem se inicia com um conjunto de dados, que será dividido entre conjuntos de treinamento e validação. Esses conjuntos contêm, respectivamente, as entradas, definidas como características ou atributos, e os rótulos que contêm as respostas esperadas. Utilizando esses conjuntos, os modelos são aplicados de forma a relacionar as entradas com as saídas.

- **Aprendizado não supervisionado**: Nesta abordagem os dados disponibilizados não possuem rótulos de saída. O objetivo principal é interpretar e agrupar os dados com base em suas características intrínsecas, identificando padrões e relações relevantes que possam emergir do conjunto analisado.

As Figuras 7 e 8 são utilizadas para observar as diferenças da construção e aplicação de cada tipo de aprendizado de máquina apresentado neste trabalho. Essas

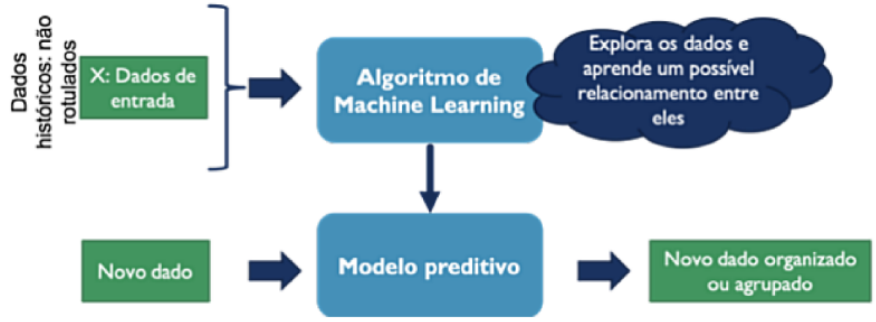
aplicações desempenham um papel central no processo que abrange desde a agregação dos conjuntos de dados até a obtenção de resultados, sendo essenciais para compreender a dinâmica e as contribuições do aprendizado de máquina no fluxo de ciência de dados (ESCOVEDO; KOSHIYAMA, 2020).

Figura 7 – Aprendizado de máquina supervisionado no fluxo de ciência de dados.



Fonte: (KALINOWSKI et al., 2023)

Figura 8 – Aprendizado de máquina não supervisionado no fluxo de ciência de dados.



Fonte: (KALINOWSKI et al., 2023)

Para que haja integração entre as diferentes técnicas de aprendizado de máquina, neste trabalho, optou-se por selecionar, em cada categoria, um ou mais algoritmos de grande relevância e uso cotidiano, além de utilizar técnicas que auxiliam na melhoria de alguma etapa da execução do método computacional.

Para o aprendizado não supervisionado, optou-se pelo uso do algoritmo *k-means*, que divide os dados de entrada em k agrupamentos (*clusters*), buscando identificar os grupos que melhor correspondem às características originalmente apresentadas. Utiliza-se a técnica do cálculo do coeficiente de silhueta para obter o melhor número de grupos que representam os dados de entrada.

Já os algoritmos de aprendizado supervisionado escolhidos foram a Regressão Linear, Redes Neurais *Multilayer Perceptron* (MLP) e *XGBoost*. A Regressão Linear

utiliza técnicas simples e diretas para identificar relações lineares nos dados, enquanto as Redes Neurais MLP e o *XGBoost* são capazes de modelar relações não-lineares complexas. Utiliza-se a técnica de validação cruzada *K-fold* na divisão dos dados de treinamento e validação, a fim de obter uma melhor generalização nas saídas e, com isso, maior robustez dos modelos.

3.3.1.1 K-means

O aprendizado de máquina não supervisionado objetiva encontrar relações entre os dados sem que haja uma saída previamente definida. Uma das técnicas empregadas nessa área é a identificação de agrupamentos (*clusters*), que utiliza um algoritmo capaz de dividir o conjunto de dados em grupos com características semelhantes. Essa separação é realizada com base em uma métrica de distância entre as amostras, a qual pode ser constante ou variável no espaço euclidiano. Além disso, a segmentação é feita com base em interpretações automáticas extraídas dos atributos ou variáveis presentes no conjunto de dados (HARRISON, 2019; IZBICKI; SANTOS, 2020).

O *K-means* (k-médias) é um algoritmo de agrupamento (*clustering*) que particiona um conjunto de dados em k grupos, onde k pode ser definido de forma direta pelo usuário, por tentativa e erro, ou determinado por cálculos executados por algoritmos interpretadores dos resultados das métricas usadas para dividir os grupos do algoritmo (como o coeficiente de silhueta apresentado na Seção 3.3.1.6). A divisão é normalmente efetuada com base na distância dos pontos de um determinado k grupo até a média deste grupo, com o valor da média denominada centroide, de modo que cada amostra é associada a um centroide calculado a partir das suas características, conforme demonstrado na Equação 5. Por fim, se executa em cada ciclo de teste a verificação para confirmar que a classificação das amostras em um dos diversos k grupos é a adequada (MORETTIN; SINGER, 2022).

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_p - y_p)^2} \tag{5}$$

A Equação 5 representa a distância Euclidiana entre os pontos $\mathbf{x}$ e $\mathbf{y}$, onde x_i e y_i correspondem às coordenadas dos pontos no espaço de dimensão p.

Existem diversas classes de algoritmos criados para se executar o *K-means* de forma bem sucedida, com a implementação mais usada sendo aquela apresentada no trabalho de Lloyd (1982), que segue os seguintes passos de forma resumida (IZBICKI;

SANTOS, 2020):

1. Calcula-se a variação interna de cada grupo por meio da soma dos quadrados das distâncias euclidianas entre as amostras do grupo e seu centroide principal. Esse cálculo é apresentado na Equação 6;

$$W(G_k) = \sum_{x_i \in G_k} \|x_i - \mu_k\|^2, \tag{6}$$

   Em que x_i representa um ponto no grupo G_k e μ_k é o centroide correspondente.

2. Define-se a quantidade K de grupos e, em seguida, escolhem-se K pontos iniciais que possivelmente pertencem a esses grupos;

3. Consideram-se esses pontos como os centroides iniciais;

4. Insere-se cada ponto restante com base na proximidade ao centroide dos grupos já formados;

5. Após cada inserção, recalcula-se o centroide principal;

6. Minimiza-se iterativamente a soma total dos quadrados dentro de cada grupo, conforme apresentado na Equação 7, até que os centroides se estabilizem.

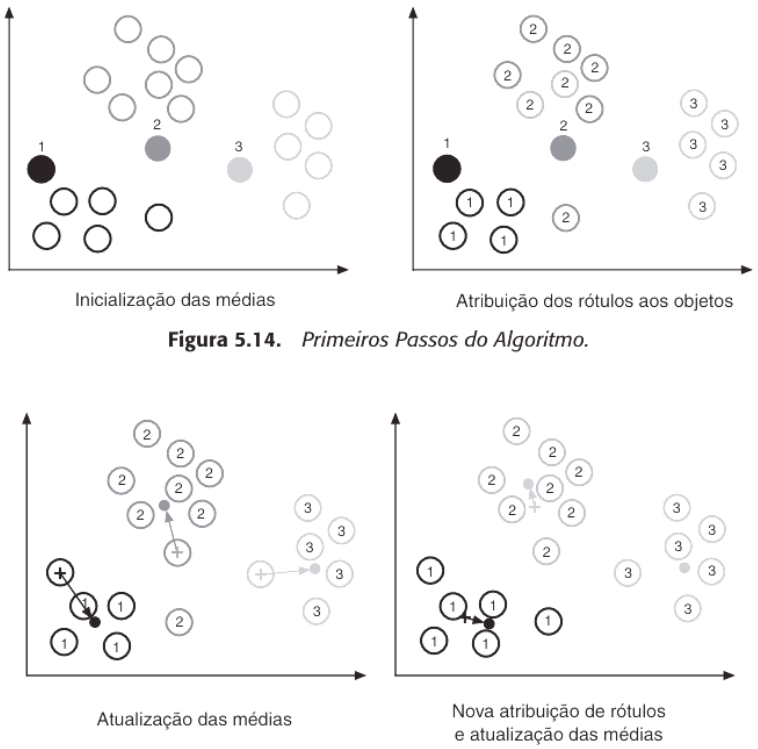
$$\sum_{k=1}^{K} W(G_k) = \sum_{k=1}^{K} \sum_{x_i \in G_k} \|x_i - \mu_k\|^2 \tag{7}$$

É importante ressaltar que a normalização dos dados para mantê-los dentro de uma faixa de valores adequada é fundamental. Caso haja discrepâncias exageradas, os cálculos dos centroides podem ser comprometidos, levando a agrupamentos imprecisos.

A apresentação dos resultados da execução do algoritmo *K-Means* é comumente realizada por meio de gráficos de dispersão, os quais destacam os grupos (*clusters*) formados a partir do valor k escolhido. A Figura 9 ilustra a aplicação do *K-means* em um conjunto de 20 amostras e $k = 3$.

Inicialmente, os centroides são posicionados em locais aleatórios. De forma iterativa, o algoritmo *K-means* busca um centro comum (média) para cada grupo, e as amostras são agrupadas conforme a distância que possuem desses centroides.

Figura 9 – Execução do *K-means*.



Fonte: (GOLDSCHMIDT; PASSOS, 2005)

3.3.1.2 Regressão Linear

A Regressão Linear é amplamente utilizada na estatística para examinar a relação entre uma ou mais variáveis quantitativas. Devido à sua natureza comparativa e quantitativa, é comum seu uso na ciência de dados para determinar modelos ou equações que demonstrem como determinadas entradas influenciam as saídas (MORETTIN; SINGER, 2022; WITTEN et al., 2025).

Essa abordagem permite identificar padrões e relações matemáticas que são fundamentais para prever resultados com base em dados históricos ou observacionais. Por ser uma técnica simples e eficiente, a Regressão Linear serve como a base para métodos mais complexos aplicados no aprendizado de máquina, especialmente na construção de algoritmos preditivos.

A Regressão Linear simples é amplamente utilizada para prever um valor baseado em outro. A Equação 8 apresenta o formato básico dessa abordagem, onde o termo independente representado por x é multiplicado por um peso ou coeficiente, ajustado para aproximar-se do valor correto. O parâmetro α, por sua vez, representa o valor básico ou a interseção, utilizado para definir o valor de x quando y possuir valor zero ou o valor

mínimo do conjunto de dados (HARRISON, 2019).

$$y = \alpha + \beta x \tag{8}$$

O formato básico pode ser expandido para a Equação 9, que incorpora múltiplos atributos x, cada um associado ao seu respectivo coeficiente β. Essa abordagem permite uma análise mais robusta e detalhada das relações entre os diferentes atributos e a variável dependente.

$$y = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}, \quad i = 1, \ldots, n, \tag{9}$$

- Onde i representa a amostra analisada, e cada x corresponde ao atributo ou característica a ela associado, definindo ao final o valor de y.

Com base na Equação 8 como modelo básico para o cálculo dos valores resposta na Regressão Linear, devemos compreender que nem sempre esses valores se apresentam de forma linear. Quando a relação não é estritamente linear, a obtenção direta dos coeficientes α e β pode apresentar discrepâncias que precisam ser tratadas. Dessa forma, aplicamos equações que geram os valores resposta estimados $\hat{y}$, os quais são comparados com os valores que seriam observados caso a linearidade fosse plenamente atendida.

Para isso, calcula-se a soma dos quadrados dos erros conforme a Equação 10 (MORETTIN; SINGER, 2022):

$$Q(\alpha, \beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2, \tag{10}$$

Onde e_i representa o erro associado à observação i.

Em seguida, minimizamos essa soma dos quadrados utilizando o método dos mínimos quadrados para obter os estimadores de α e β (Equações 11 e 12):

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{11}$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\,\overline{x}, \tag{12}$$

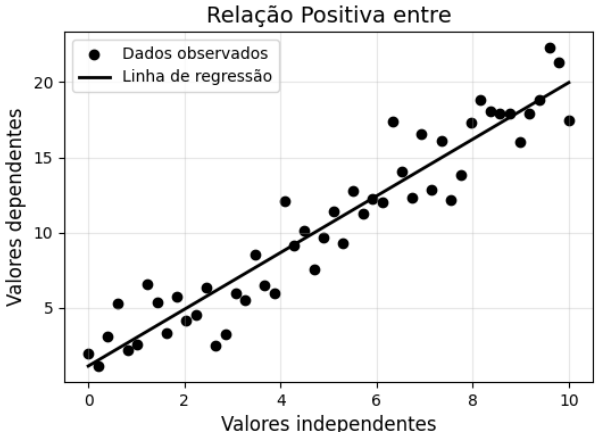
Em que $\overline{x}$ e $\overline{y}$ representam as médias amostrais das variáveis x e y, respectivamente, definidas pelas Equações 13 e 14:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{13}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{14}$$

Com os cálculos efetuados, podemos obter uma equação que modela as relações, a qual pode ser utilizada para descrever a interação entre as variáveis. Essas relações podem ser representadas por valores quantitativos diretos ou expostas por meio de gráficos, como o apresentado na Figura 10.

Figura 10 – Gráfico de dispersão Regressão Linear



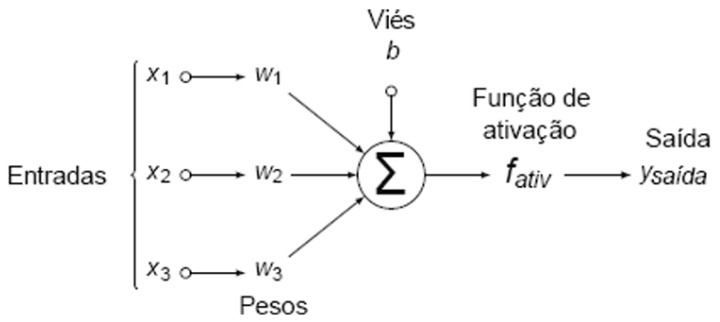
Fonte: Elaborado pelo Autor.

3.3.1.3 Redes Neurais multilayer perceptron

O desenvolvimento do Perceptron, por Rosenblatt (1958), é considerado uma das primeiras fases da inteligência artificial no campo do aprendizado supervisionado, tendo por objetivo emular o funcionamento de um neurônio cerebral (MORETTIN; SINGER, 2022). O algoritmo é altamente funcional, sendo aplicado na resolução de problemas de classificação e regressão.

O Perceptron tem seu algoritmo básico dividido em partes bem definidas (MORETTIN; SINGER, 2022): primeiramente, são recebidas as características ou entradas x. Para cada entrada, é atribuído um valor denominado peso w. Após a atribuição dos pesos, calcula-se a soma ponderada das entradas com seus respectivos pesos (Equação 15). O resultado, z, é então aplicado em uma função de ativação que pode incluir um valor de viés fixo, por exemplo a função de ativação *rectified linear unit* (ReLU), descrita na

Figura 11 – Funcionamento de um Perceptron.



Fonte: (MORETTIN; SINGER, 2022)

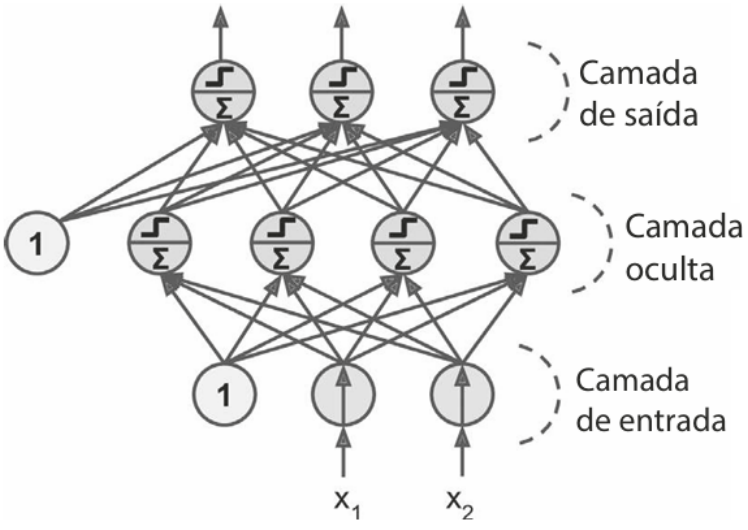
Equação 16, que retorna a saída y. Esse fluxo básico é apresentado na Figura 11.

$$z = w_1.x_1 + w_2.x_2 + \cdots + w_n.x_n \tag{15}$$

$$\text{ReLU}(z) = \begin{cases} 0 & \text{se } z \leq 0 \\ z & \text{se } z > 0 \end{cases} \tag{16}$$

A união de múltiplas camadas de Perceptrons (neurônios), com conexões diretas entre camadas consecutivas, é denominada rede neural *Multilayer Perceptron* (MLP). Essa arquitetura é composta por uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Cada camada contém um número variado de neurônios. Essa arquitetura é representada na Figura 12.

Figura 12 – Representação de uma Rede Neural *Multilayer Perceptron* (MLP).



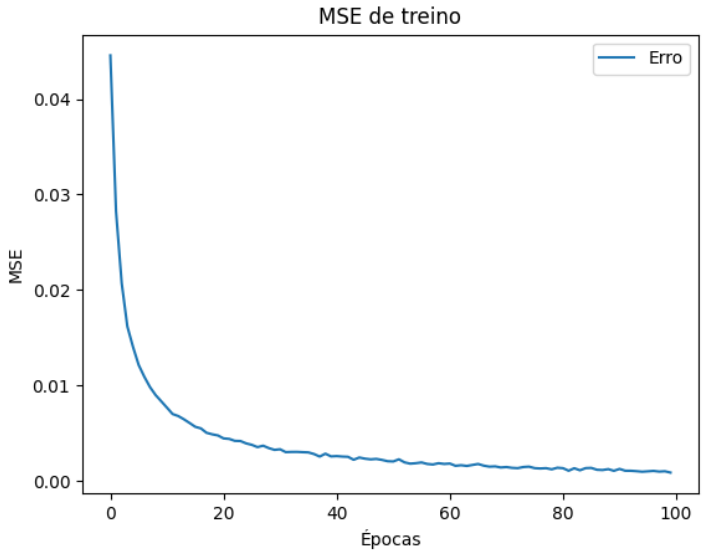
Fonte: (GÉRON, 2019)

Com o aumento da complexidade dos dados, tornou-se cada vez mais necessário aprofundar e ampliar o número de camadas, originando um novo campo na computação denominado aprendizado de máquina profundo (*Deep Learning*). É relevante ressaltar que, na aplicação de uma MLP, podem ser empregadas diversas camadas com quantidades variadas de neurônios. Entretanto, se o objetivo for predizer valores, a última camada deve obrigatoriamente conter um único neurônio de saída (ZHANG et al., 2023; MORETTIN; SINGER, 2022).

Com o objetivo de melhorar os resultados de uma MLP, é necessário que durante a etapa de treinamento seja aplicado o algoritmo de retropropagação (*backpropagation*) descrito por Rumelhart, Hinton e Williams (1986). Ele é empregado com o objetivo de aprimorar os valores dos pesos w associados às saídas de cada uma das camadas ocultas. A retropropagação funciona com cada neurônio da camada oculta atual conectado aos neurônios da camada seguinte, assim sucessivamente até a camada de saída.

Com a obtenção de um valor de saída y em um determinado instante do treinamento (conhecido como época), é aplicado o cálculo de uma função de erro, para que seja possível compará-lo com o valor real da amostra. A função mais popular para essa etapa é o *Mean Square Error* (MSE) (Erro Quadrático Médio), detalhada na Seção 3.3.2.2. O MSE pode ser apresentado em um gráfico para evidenciar a evolução dos valores e, consequentemente, indicar se o aprendizado do modelo está sendo bem executado, conforme ilustrado na Figura 13.

Figura 13 – Exemplo de MSE na execução de treinamento de uma MLP.



Fonte: Elaborado pelo Autor

Logo após a aplicação do MSE, o algoritmo armazena os pesos, para que sejam utilizados na próxima rodada de treinamento, de forma a aproximar o resultado do valor real associado à amostra de treinamento, obtendo resultados mais precisos (GÉRON, 2019; GRUS, 2021).

3.3.1.4 XGBoost

O *XGBoost* (*Extreme Gradient Boosting*), desenvolvido por Chen e Guestrin (2016), é uma biblioteca de código desenvolvida e distribuída para diversas linguagens de programação (DOCUMENTATION, 2023). Essa biblioteca aprimora as ideias implementadas nos algoritmos de *Gradient Boosting* (FRIEDMAN, 2001), que utilizam modelos computacionais fracos que, combinados, formam um modelo mais robusto (GÉRON, 2019; BRUCE; BRUCE, 2019). No âmbito dos algoritmos de *Boosting*, esse método vem se destacando desde a sua criação, sendo amplamente utilizado para resolver problemas de aprendizado de máquina supervisionado nos campos de classificação e regressão, contribuindo significativamente para projetos de ciência de dados (Kaggle, 2022).

O *XGBoost* baseia-se na combinação e iteração de algoritmos de Árvores de Decisão (QUINLAN, 1986), organizando-os em paralelo e aplicando o algoritmo de *Gradient Boosting*. Nesse processo, os dados são divididos em pequenos conjuntos e distribuídos entre as árvores, permitindo a avaliação da qualidade dos resultados parciais, a fim de fornecer melhores decisões para os níveis subsequentes das árvores de decisão (BRUCE; BRUCE, 2019). O *XGBoost* depende de parâmetros de ajuste com o objetivo de reduzir o viés e o sobreajuste (*overfitting*), possibilitando que as árvores de decisão se interliguem e gerem resultados mais precisos.

Por ser um pacote robusto, o *XGBoost* permite a combinação de diversos parâmetros, também conhecidos como hiperparâmetros, que auxiliam no ajuste da fase de treinamento. Isso visa melhorar tanto a velocidade de execução quanto a qualidade dos resultados obtidos nos dados de validação. O algoritmo básico de *Boosting*, que também pode ser utilizado no *XGBoost*, pode ser descrito conforme apresentado em Izbicki e Santos (2020):

1. Definimos $g(x) \equiv 0$ e $r_i = y_i \; \forall i = 1, \ldots, n$.

2. Para $b = 1, \ldots, B$:

   (a) Ajustamos uma árvore com p folhas para $(x_1, r_1), \ldots, (x_n, r_n)$. Seja $g^b(x)$ sua

respectiva função de predição.

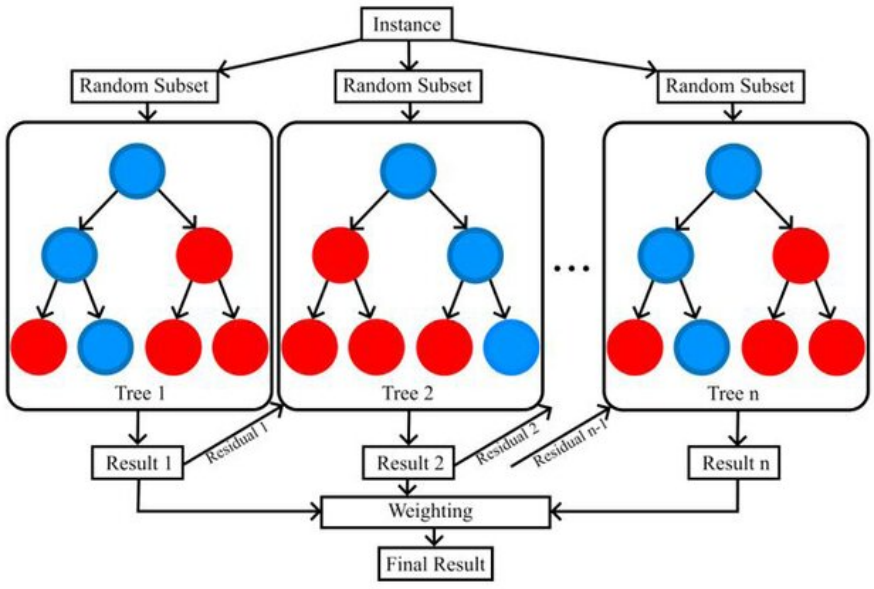
(b) Atualizamos g e os resíduos: $g(x) \leftarrow g(x) + \lambda g^b(x)$ e $r_i \leftarrow Y_i - g(x)$.

3. Retornamos o modelo final $g(x)$.

Sendo $g(x)$ denominada função estimadora, λ representa a taxa de aprendizado, p corresponde ao número de folhas das árvores de decisão e B à quantidade de árvores. Os valores desses parâmetros devem ser ajustados de acordo com o problema por meio de testes automáticos ou manuais para se obter os melhores resultados. Os valores mais frequentes são: λ é definido com um valor pequeno entre $0,001$ e $0,1$, B varia de 100 a valores acima de 1000, e o parâmetro p pode variar de 2 a 6.

A Figura 14 apresenta a estrutura do treinamento do *XGBoost* (ÖZTORNACI; ATA; KARTAL, 2024), que consiste em um conjunto de árvores de decisão construídas de forma sequencial a partir de subconjuntos aleatórios dos dados. Cada árvore é responsável por minimizar os erros residuais das anteriores, e seus resultados são combinados por meio de uma ponderação que gera a saída final do modelo. Esse processo iterativo permite ao *XGBoost* alcançar alta performance e generalização em tarefas de classificação e regressão.

Figura 14 – Representação do algoritmo *XGBoost*.



Fonte: (ÖZTORNACI; ATA; KARTAL, 2024).

O *XGBoost* não é a única biblioteca que implementa o algoritmo de *Gradient Boosting*, competindo com diversas outras que poderiam ser utilizadas neste estudo. No

entanto, optou-se por esta biblioteca por ter sido a primeira a se popularizar, além de apresentar um desempenho que equilibra o tempo de treinamento com resultados finais satisfatórios (BENTÉJAC; CSÖRGŐ; MARTÍNEZ-MUÑOZ, 2021).

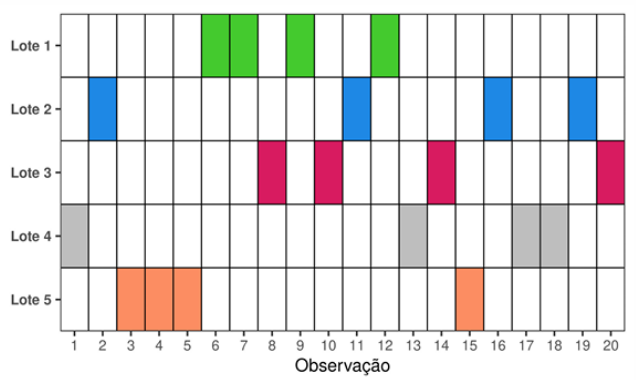
3.3.1.5 Validação cruzada k-fold

Embora seja possível utilizar o conjunto completo de dados para validação do modelo final, a abordagem mais comum é a validação cruzada, cuja ideia principal consiste em dividir aleatoriamente o conjunto de dados em subconjuntos de treinamento e validação. Dessa forma, o treinamento do modelo final é conduzido sem a influência dos dados de validação. Isso constitui uma forma de minimizar a ocorrência de subajuste (*underfitting*) ou sobreajuste (*overfitting*) com os dados de treinamento (WITTEN et al., 2025; IZBICKI; SANTOS, 2020).

Para minimizar o impacto das oscilações decorrentes de uma divisão inadequada dos dados de treinamento e validação, pode-se optar pela validação cruzada *k-fold* (k-dobras). Esse método oferece uma maior robustez aos modelos computacionais construídos, pois constrói o modelo diversas vezes, utilizando a cada iteração uma parte do conjunto de dados para treinamento e outra para validação. Ao realizar k divisões, a amostra do conjunto de dados participara tanto do processo de treinamento em $n-1$ vezes e também fará parte da validação em uma das iterações, garantindo uma interdependência que aprimora os resultados (IZBICKI; SANTOS, 2020; MORETTIN; SINGER, 2022).

A validação cruzada *k-fold* pode ser aplicada com diferentes números de iterações (divisões), sendo 2, 5 e 10 os mais utilizados (FUSHIKI, 2011). A Figura 15 exemplifica, em um conjunto de 20 amostras, a execução da divisão dos dados usados em treinamento e validação em cada etapa de validação cruzada *k-fold*.

Figura 15 – Exemplo de execução da validação cruzada *k-fold* com $k = 5$.



Fonte: (IZBICKI; SANTOS, 2020)

3.3.1.6 Coeficiente de silhueta

Conforme já explicado na Seção 3.3.1.1, o *K-means* depende da definição de um valor k, que irá dividir os dados em *clusters* (agrupamentos) que representem os grupos da melhor forma possível. O método convencional de escolha de k costuma ser meramente empírico, o que pode inviabilizar uma divisão adequada para o conjunto de dados analisado (BRUCE; BRUCE, 2019).

O coeficiente de silhueta surge como uma boa opção para identificar o melhor valor de k, pois analisa a coesão dos *clusters* tanto internamente quanto em relação uns aos outros. O valor do coeficiente varia de -1 a 1, sendo que valores próximos de -1 indicam uma representação insatisfatória dos agrupamentos (*clusters* mal definidos), enquanto o valor 1 representa uma coesão ideal para os agrupamentos inferidos pelo *K-means* (HARRISON, 2019). O cálculo do coeficiente de silhueta leva em consideração todo o conjunto de dados: após a determinação dos centroides, verifica-se a proximidade entre os dados dentro de cada grupo e entre grupos distintos.

O silhueta se mostra uma ótima opção para encontrar o melhor valor para k, visto que sua definição, apresentada nas Equações 17, 18, 19 e 20, não se limita a verificar as amostras dentro de um grupo, mas também avalia se os grupos estão realmente bem separados entre si (WANG et al., 2017).

$$sil(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{17}$$

$$a(i) = \frac{\sum_{p=1}^{n} w_{p,h}\, d_E(X_i, X_p)}{n_h - 1} \tag{18}$$

$$b(i) = \min_{l \neq h} \left(\frac{\sum_{p=1}^{n} w_{p,l}\, d_E(X_i, X_p)}{n_l} \right) \tag{19}$$

$$Sil = \frac{1}{n} \sum_{i=1}^{n} sil(i) \tag{20}$$

Onde:

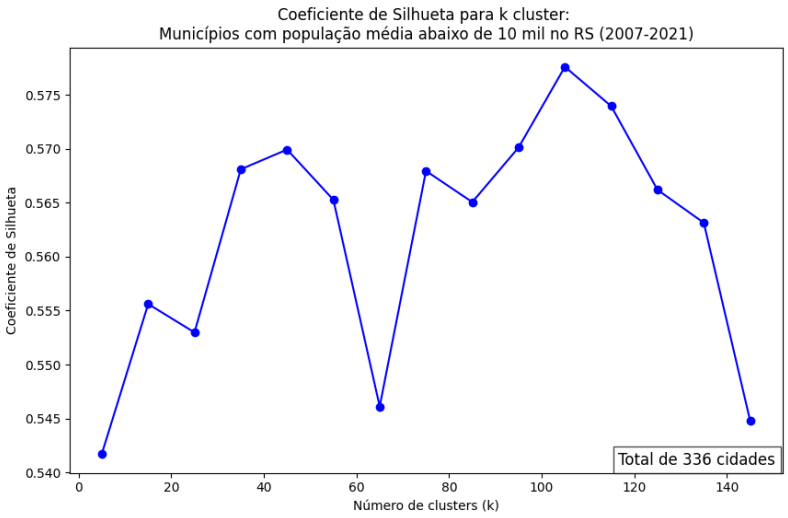
- $a(i)$ (Equação 18) é a distância média de X_i a todos os outros pontos em seu próprio *cluster* h;
- $b(i)$ (Equação 19) é a menor distância média de X_i aos pontos do *cluster* vizinho mais próximo;

- $sil(i)$ (Equação 17): Medida individual de qualidade do agrupamento, variando entre -1 (ponto mal agrupado) e 1 (ponto bem agrupado);

- Sil (Equação 20) é o coeficiente de silhueta global do agrupamento, calculado como a média de $sil(i)$ para todos os pontos, variando de -1 (agrupamento ruim) a 1 (agrupamento ideal).

O algoritmo para encontrar o valor de silhueta é executado para valores variados de k, permitindo a concatenação em um vetor para comparação. Assim, busca-se identificar o valor de k cujo coeficiente se aproxima mais de 1, para que o maior valor seja selecionado como o melhor número de k grupos divididos e também para observar se outros valores podem ser considerados utilizáveis. A Figura 16 mostra os resultados da execução do coeficiente de silhueta para diversos valores de k.

Figura 16 – Exemplo de gráfico do coeficiente de silhueta para um valor de k na aplicação do *K-means*.



Fonte: Elaborado pelo Autor

3.3.2 Métricas estatísticas

Este trabalho utiliza métricas estatísticas de análise que se aplicam ao conjunto de dados, a fim de verificar como estão distribuídas uma ou mais características. Também são empregadas métricas de desempenho, que permitem comparar os resultados obtidos dos métodos computacionais selecionados, para fornecer subsídio interpretativo necessário ao Capítulo 5. Além de valores quantitativos, serão utilizados gráficos que

demonstram a evolução de determinadas métricas analíticas ou que apresentam relações intrínsecas geradas após a execução dos métodos computacionais (BRUCE; BRUCE, 2019; MORETTIN; SINGER, 2022).

3.3.2.1 Métricas de análise

Entre os valores básicos de análise a serem obtidos estão a **média**, **mediana** e **desvio padrão**, que podem oferecer uma visão abrangente do comportamento dos dados ao longo do tempo. Essas métricas são fundamentais para uma análise preliminar estratificada ao utilizar métodos de ciência de dados (BRUCE; BRUCE, 2019; MORETTIN; SINGER, 2022).

As métricas de análise selecionadas são apresentadas e equacionadas a seguir, com o objetivo de fornecer maiores recursos de entendimento durante a leitura dos demais capítulos:

- **Média**: a média simples é calculada como o somatório de todas as amostras x_i dividido pelo número total n de amostras, conforme a Equação 21. A média é importante para identificar o valor representativo do conjunto de dados.

$$\text{Média}(\bar{x}) = \frac{\sum_{i=1}^{n} x_i}{n} \tag{21}$$

- **Mediana**: a mediana busca identificar o valor central de um conjunto de dados ordenado, eliminando o impacto de valores extremos. O valor central pode ser obtido de maneira distinta se o total de amostras n for ímpar ou par, conforme apresentado na Equação 22. A mediana é utilizada para comparar diretamente diferentes populações de dados que possam apresentar características similares.

$$\text{Mediana} = \begin{cases} x_{\frac{n+1}{2}}, & \text{se } n \text{ é ímpar,} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \text{se } n \text{ é par.} \end{cases} \tag{22}$$

- **Desvio padrão**: o desvio padrão mede a dispersão dos valores do conjunto de dados em relação à média. Quanto mais próximo de zero, menos dispersos estarão os dados. Seu cálculo é feito conforme a Equação 23, que considera o somatório do quadrado da diferença entre cada amostra x_i e a média $\bar{x}$, dividido pelo número total

de amostras n, seguido da extração da raiz quadrada.

$$\text{Desvio Padrão} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \tag{23}$$

3.3.2.2 Métricas de desempenho

Para os métodos computacionais analisados neste estudo, serão utilizadas métricas que expressem de forma clara e abrangente o quão bem ajustado está o modelo resultante de predição. Após a etapa de treinamento dos modelos de regressão, o principal valor obtido são os **resíduos** ou erros, que correspondem à diferença entre o valor real y_i e o valor predito $\hat{y}_i$, conforme a Equação 24. Esse valor pode variar tanto positivamente quanto negativamente.

$$\hat{e}_i = y_i - \hat{y}_i \tag{24}$$

A partir dos resíduos, pode-se calcular uma das métricas mais importantes para validar o potencial de um modelo de regressão em explicar a relação entre as variáveis independentes usadas na sua criação e a variável dependente, que é o **coeficiente de determinação** (R^2). Esse coeficiente é calculado pela Equação 25 e o seu valor pode variar de 0 a 1 e, se for preciso, representado em termos percentuais, sendo que: valores mais próximos de 1 indicam um modelo confiável e ajustado, já, valores próximos de 0, indicam um modelo que explica pouco ou não tendo melhor desempenho que apenas o uso da média. O valor de R^2 pode ser negativo, isso ocorre quando o modelo se comporta pior que apenas utilizar a média dos dados como o valor encontrado para uma dada entrada.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{25}$$

A métrica conhecida como **Erro Quadrático Médio (MSE)** é amplamente utilizada para comparar modelos estatísticos de regressão, pois avalia com precisão os resíduos. A Equação 26 define o MSE como a média dos quadrados dos resíduos, ou seja, o somatório dos quadrados das diferenças entre os valores reais e os preditos, dividido pelo número total de amostras. Quanto menor o valor do MSE, melhor será o ajuste do modelo, indicando que os dados usados nos atributos explicam bem o resultado previsto.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{26}$$

4 TRABALHOS CORRELATOS

Este capítulo foi construído para apresentar os métodos de pesquisa bibliográfica adotados para a consulta de trabalhos relacionados aos temas abordados neste estudo, que buscam estabelecer uma relação entre políticas públicas, especificamente as de transferência de renda, e índices socioeconômicos. Tal investigação visa verificar em que medida modificações em uma dessas áreas podem gerar consequências relevantes na outra, com o requisito primordial de que a bibliografia consultada inclua, em alguma etapa, o emprego de uma ou mais disciplinas que estão englobadas na ciência de dados.

Segundo Abbasi, Chiang e Xu (2023), os temas de ciência de dados e políticas públicas vêm sendo amplamente explorados nas últimas décadas, aproximando-se em termos de volume de estudos publicados. Dessa forma, a ciência de dados passa a convergir e, consequentemente, a auxiliar em todo o ciclo de vida de uma política pública. Um exemplo dessa convergência são os projetos de Ghani (2018), que empregam técnicas de ciência de dados em diversas áreas da sociedade para predizer e classificar ações governamentais de bem-estar social.

Nas seções subsequentes, apresentam-se os critérios de pesquisa e seleção, um breve resumo de alguns trabalhos relevantes e, por fim, uma discussão acerca de como este trabalho se diferencia após a realização da investigação bibliográfica.

4.1 Pesquisa e seleção

Esta seção descreve o processo de levantamento bibliográfico realizado com o objetivo de identificar estudos relacionados ao tema deste trabalho. A pesquisa concentrou-se em fontes acadêmicas, tais como artigos científicos, monografias, dissertações, teses e publicações técnicas, com ênfase na seleção final de investigações que empregam técnicas de ciência de dados para auxiliar no ciclo de vida das políticas públicas de transferência de renda em conjunto com índices socioeconômicos, sem ignorar possíveis contribuições de trabalhos cujo foco seja apenas um dos temas principais.

A investigação bibliográfica não se concentrou em fontes que agregassem um volume excessivo de informações, como livros ou compilações de produções acadêmicas, uma vez que tais publicações costumam ser excessivamente amplas e, consequentemente, não se relacionam diretamente com este estudo. Essas obras podem surgir como auxiliares

na construção da fundamentação teórica e metodológica. Optou-se, portanto, por analisar publicações que abordem um problema de pesquisa bem definido e que seja relevante para este trabalho.

O processo de pesquisa seguiu uma abordagem padronizada, iniciando com uma busca ampla e, em seguida, aplicando critérios de filtragem para identificar estudos confiáveis e relevantes no contexto do tema em desenvolvimento. Em um primeiro momento, definiram-se as palavras-chave diretamente relacionadas aos assuntos abordados, tanto em termos amplos quanto específicos. A partir disso, obteve-se uma agregação de termos que orientaram a busca bibliográfica, conforme apresentado na Tabela 3.

Tabela 3 – Palavras-chave de busca.

Palavras-chave (Português)	**Palavras-chave (Inglês)**	**Justificativa**
Ciência de dados	*Data science*	Por ser um dos focos principais deste estudo, este termo foi considerado obrigatório na busca para que apareçam trabalhos relacionados.
Política pública / políticas públicas	*public policy / public policies*	Termo fundamental, pois possibilita a amplitude de resultados.
Transferência de renda / redistribuição de renda	*income transfer / income redistribution / social assistance*	Termos que acompanham os tipos de políticas públicas de interesse neste estudo, em especial a política redistributiva analisada.
Índice Socioeconômico / IDESE / IDH	*Socioeconomic Index / IDESE / HDI*	Indicadores utilizados para avaliar o desenvolvimento social e econômico da sociedade.

Fonte: Elaborado pelo Autor.

O mecanismo de busca utilizado foi o **Google *Scholar* (Google Acadêmico)** (SCHOLAR, 2025), ferramenta que agrega diferentes tipos de fontes e repositórios acadêmicos, oferecendo diversas publicações por meio de uma única pesquisa, além de proporcionar amplas possibilidades de filtragem para seleção e eliminação de trabalhos retornados após a realização de uma investigação. Para auxiliar no refinamento dos resultados, utilizou-se a combinação de palavras-chave com operadores lógicos Booleanos *AND/OR* (E/OU) criando-se, assim, uma ***string* de busca** que possibilita maior capacidade de filtragem e retorno de estudos mais correlacionados com esta pesquisa.

Inicialmente, a *string* de busca foi elaborada em português. Para ampliar a quantidade de trabalhos correlatos encontrados e possibilitar análise posterior, efetuou-se a tradução dos termos necessários para o inglês. Ademais, com o objetivo de restringir

as buscas, aplicou-se um conjunto de critérios eliminatórios, listados abaixo, os quais, de forma sequencial e cumulativa, integram critérios de inclusão e exclusão de trabalhos selecionados para serem resumidos na próxima seção. Esta lista faz uso de ferramentas disponíveis no Google *Scholar* e da análise empírica do pesquisador sobre o estudo selecionado. Tais regras servem para selecionar os trabalhos com maior relevância para serem analisados e sumarizados na seção subsequente.

1° **Critério**: O trabalho não foi publicado no período de 2020 a 2025 (ano de finalização desta monografia);

2° **Critério**: O trabalho é considerado uma obra agregadora de trabalhos (livro, coletânea de artigos, etc) ou é apenas um resultado repetitivo;

3° **Critério**: O trabalho utiliza técnicas de ciência de dados apenas para dar suporte superficial ao tema de pesquisa;

4° **Critério**: O trabalho não converge técnicas de ciência de dados para analisar políticas públicas e índices socioeconômicos;

5° **Critério (final/correlato)**: O trabalho utiliza ciência de dados para analisar políticas públicas de transferência de renda ou índices socioeconômicos e como eles afetam um ao outro.

A partir da lista de critérios de seleção, realiza-se uma análise detalhada do trabalho para verificar se este transcende os critérios 2 e 3. Este procedimento é efetuado mediante a observação do título, resumo e capítulo ou seção de conclusões, para que, finalmente, seja examinado o trabalho em sua totalidade e sejam definidas as semelhanças com esta pesquisa. Dessa forma, atinge-se o critério 4, e com isto o estudo é considerado um trabalho correlato, prosseguindo-se para a análise aprofundada na próxima etapa.

Com os critérios de eliminação definidos, foram realizadas buscas utilizando as *strings* de busca, cujos valores e os resultados em termos de quantidade de trabalhos retornados após a aplicação de cada critério de exclusão são apresentados na Tabela 4.

As *strings* foram concebidas de modo que, em cada bloco separado pelo operador booleano *AND* (E), seja apresentada uma ideia ou tema abordado, além de realizar a combinação de termos entre parênteses com o operador booleano *OR* (OU), o que amplia o escopo da investigação, permitindo a filtragem por trabalhos que abordem variações conceituais relevantes. Ressalta-se que a seleção desses termos pode ter resultado na

Tabela 4 – *Strings* de busca e seus resultados.

***String* de busca**	**Resultado**	**Trabalhos excluídos**				**Correlatos**
		1	**2**	**3**	**4**	
"ciência de dados" AND ("política pública" OR "políticas públicas") AND ("transferência de renda" OR "redistribuição de renda") AND ("índice socioeconômico" OR "idese" OR "idh")	35	8	11	8	3	5
"data science" AND ("public policy" OR "public policies") AND ("income transfer" OR "income redistribution") AND ("socioeconomic index" OR "IDESE" OR "HDI")	10	0	3	3	3	1
Total de resultados	45	8	14	11	6	6

Fonte: Elaborado pelo autor.

não identificação de estudos com terminologias ou combinações lexicais distintas e, consequentemente, na sua não citação e análise.

Os trabalhos retornados após a aplicação do critério de eliminação 4 e, por consequência, considerados de alguma forma correlatos, são examinados na seção subsequente. Os temas identificados apresentam correlação direta com este estudo, uma vez que serão investigados os impactos nas esferas econômica, de saúde e de educação, bem como a influência desses fatores sobre uma política pública de transferência de renda.

4.2 Análise de trabalhos

Esta seção foi elaborada para realizar a descrição e a análise dos trabalhos que possuem correlação total ou parcial, seguindo os critérios impostos na seção anterior. Inicialmente, apresentam-se os trabalhos selecionados que possuem uma correlação direta; logo após, comenta-se sobre os trabalhos com informações relevantes, mas que foram eliminados pelos critérios de exclusão por possuírem baixa correlação com o escopo deste trabalho.

Ao revisar a Tabela 4, informa-se a quantidade de 6 trabalhos considerados diretamente correlatos, os quais focam em temas de economia, saúde e bem-estar social. Estes temas interagem diretamente com as dinâmicas que os governos desejam alterar

quando é implantada uma política pública de transferência de renda. Segue-se agora para uma descrição resumida dos trabalhos considerados correlatos, nos quais se encontram informações de como foi feito o uso de técnicas de Ciência de Dados para a aplicação desse campo em políticas públicas de transferência de renda e índices socioeconômicos.

Maia, Noguti e Ara (2020) aplicam técnicas de Ciência de Dados para analisar fatores associados ao que foi definido como **taxa de utilização do Programa Bolsa Família**. Este estudo fez uso de 1,2 bilhão de registros dos pagamentos aos beneficiários, usando como local de análise os municípios do estado da Bahia. O trabalho emprega o método computacional conhecido como **Máquinas de Vetores de Suporte para Regressão** (SVR), para identificar variáveis socioeconômicas com maior poder explicativo para o uso do PBF, destacando indicadores como formalização do trabalho, infraestrutura básica domiciliar e composição da população economicamente ativa. Essa abordagem demonstra como métodos avançados de Ciência de Dados podem apoiar a compreensão e a gestão de políticas públicas de transferência de renda, reforçando a importância de indicadores sociais na avaliação de vulnerabilidade municipal.

Pardita et al. (2024) realizam uma análise da dinâmica da pobreza na Indonésia, por meio de diferentes métodos de regressão, avaliando o impacto do PIB, da distribuição de renda e do IDH sobre os níveis de pobreza provinciais entre 2018 e 2022. Os autores demonstram que o crescimento econômico contribui para a redução da pobreza, enquanto a desigualdade de renda atua como fator agravante. Embora o IDH apresente relação negativa com a pobreza, seu efeito não se mostrou estatisticamente significativo. O estudo destaca a importância de políticas públicas como tributação progressiva, programas de transferência direta de renda e investimentos educacionais, além de evidenciar a utilidade de métricas socioeconômicas integradas como suporte analítico para a avaliação de políticas sociais, uma abordagem alinhada ao uso de índices compostos no presente trabalho.

Feitosa Júnior (2023) emprega redes neurais para investigar diversos dados, entre eles o investimento em políticas públicas de assistência social, com destaque para o Programa Bolsa Família (PBF), a fim de encontrar correlações e entender a dinâmica epidemiológica da hanseníase no estado do Pará. O estudo demonstrou que o gasto *per capita* com o PBF e outros índices socioeconômicos envolvendo a educação estão entre as variáveis mais relevantes que interagem e impactam os indicadores de saúde. Este trabalho fornece uma base metodológica sólida ao exemplificar o uso de modelos avançados de Inteligência Artificial para quantificar a relação entre as políticas de

transferência de renda e os determinantes sociais.

Simonato (2023) investiga os impactos na economia brasileira do Auxílio Emergencial, uma importante política de transferência de renda criada em decorrência da pandemia de COVID-19. O estudo fez uso de ferramentas avançadas de Ciência de Dados e simulação econômica. A pesquisa quantifica os efeitos da política em variáveis regionais, setoriais e, fundamentalmente, por faixas de rendimento familiar, evidenciando como a transferência de renda interagiu com os índices socioeconômicos para gerar diferentes impactos no mercado de trabalho e no consumo.

Gomes et al. (2025) utilizam aprendizado de máquina supervisionado para investigar a insegurança alimentar de mulheres adultas. O trabalho correlaciona o consumo alimentar com índices socioeconômicos de vulnerabilidade de saúde, demonstrando o potencial da modelagem computacional para o diagnóstico e a classificação de grupos em risco, subsidiando a avaliação e o aprimoramento de políticas públicas de assistência e transferência de renda.

Oliveira Filho (2025) discute a relevância do *Big Data* como insumo fundamental para a formulação de políticas públicas eficazes na erradicação da pobreza, tomando como base o estudo comparativo entre o estado do Ceará e experiências e aplicadas na China. O autor argumenta que a utilização estratégica da análise massiva de dados permite um diagnóstico mais acurado da pobreza multidimensional e a criação de intervenções mais precisas e integradas, correlacionando as variáveis socioeconômicas e os resultados das políticas por meio de simulações de cenários. Esta abordagem demonstra o papel crucial da Ciência de Dados na transformação de indicadores socioeconômicos em evidências.

Durante a aplicação dos critérios de exclusão de trabalhos correlatos, houve trabalhos com grande relevância em outras áreas, mas que não dialogavam em profundidade com o tema de Ciência de Dados ou faziam uso de políticas públicas de transferência de renda de forma extremamente básica. Citam-se a seguir alguns trabalhos envolvendo temas diversos que não foram considerados como trabalhos correlatos por não terem atendido aos critérios 3 e 4.

Em relação à construção de políticas públicas voltadas à melhoria das questões de desigualdade econômica, Silva (2022) e Januario (2024) investigam como fatores diversos podem afetar os níveis de pobreza em regiões do Brasil. Para isso, utilizam-se várias variáveis, incluindo fatores sociais e índices calculados por órgãos governamentais, buscando identificar relações entre tais fatores. Dessa forma, o uso de Ciência de Dados é empregado para propor melhorias a segmentos específicos da população e

para identificar quais fatores possuem maior impacto; contudo, estes trabalhos abordam o tema de transferência de renda de forma vaga e sem ênfase. Do ponto de vista microeconômico, Santos (2020) observou como os municípios são afetados quando os valores de aposentadorias e seguridade social são aplicados localmente ou transferidos para outras regiões.

Nas políticas públicas relacionadas à educação, Queiroga et al. (2022) apresentam um panorama para avaliar a qualidade do ensino em escolas do Uruguai, utilizando dados históricos e aprendizado de máquina, com o objetivo de evitar altas taxas de retenção ou evasão escolar. Em outra abordagem, Santos (2022) buscou correlacionar fatores de criminalidade com a evasão escolar, analisando tanto os dados quanto o ambiente em que as escolas estão inseridas, bem como a incidência de crimes na região. O estudo demonstrou que, por meio de técnicas diversas de Ciência de Dados e de um volume adequado de dados, é possível identificar variáveis explicativas relevantes para a compreensão dos fenômenos escolares. Já Brusaca (2025) verifica como a Ciência de Dados pode auxiliar em fatores educacionais do serviço público.

Aguiar (2021) investigou quais fatores de risco estão presentes entre mulheres que vivem em situação de insegurança alimentar, utilizando técnicas estatísticas para evidenciar o peso de tais variáveis nessa condição. Ademais, os trabalhos de Silva (2023) e Paiva et al. (2022) têm como objetivo empregar técnicas de Ciência de Dados para auxiliar na detecção de doenças, as quais podem ser ou não transmitidas de forma massiva, através da análise de dados históricos, a fim de identificar correlações entre as informações.

Para verificar o nível de correlação temática dos trabalhos expostos anteriormente com esta pesquisa, elaborou-se a Tabela 5. Nela são sumarizados os principais temas abordados em cada estudo, apresentando o campo da ciência de dados aplicado, a política pública analisada, o tipo de índice utilizado e o objetivo da análise.

Com essas informações, é possível constatar a relevância deste trabalho, que visa examinar e consolidar diversas abordagens presentes na literatura. A pesquisa expande a aplicação de técnicas de ciência de dados, compreendidas como métodos computacionais e análise de dados, além de empregar índices socioeconômicos para investigar como a sociedade influencia as decisões governamentais, alterando os valores históricos de uma política pública de transferência de renda.

Tabela 5 – Revisão de literatura sobre aplicação de ciência de dados em políticas públicas

Estudo	**Ciência de dados**	**Política Pública**	**Índices**	**Objetivo da Análise**
Maia, Noguti e Ara (2020)	Máquina aleatória de regressão e análise de *Big Data*	Programa Bolsa Família	Índices socioeconômicos variados	Índices sobre a política
Pardita et al. (2024)	Regressão de dados em painel	Combate a pobreza	IDH, PIB do país e PIB per capita	Índices sobre a política
Oliveira Filho (2025)	Análise aprofundada de dados com *Big Data*	Combate a pobreza	Índices socioeconômicos variados	Índices sobre a política
Feitosa Júnior (2023)	Redes neurais artificiais	Política pública de combate à hanseníase	Indicadores de saúde e índices sociais variados	politica sobre os índices
Simonato (2023)	Criação de modelos estatísticos e análise de dados	Auxílio emergencial	Índices relacionados a consumo e trabalho	política sobre os índices
Gomes et al. (2025)	Métodos computacionais e análise de dados	Política de combate à insegurança alimentar	Índices diversos	Índices sobre a política
Este trabalho	Análise de dados e métodos computacionais	Programa Bolsa Família	IDESE (índices diversos)	Índices sobre a política

Fonte: Elaborado pelo Autor

5 CIÊNCIA DE DADOS APLICADA NA ANÁLISE DO IMPACTO DO IDESE SOBRE O PBF

Este capítulo tem como intuito apresentar como a interdisciplinaridade da ciência de dados pode ser empregada na análise e na extração de resultados relevantes para a sociedade, com foco em compreender valores de métricas básicas e no uso de técnicas computacionais complexas para identificar as relações de impacto dos índices do IDESE sobre os valores do PBF.

Inicia-se com a Seção 5.1, onde de forma individualizada é realizada uma análise aprofundada dos conjuntos de dados do PBF e do IDESE, visando a extração de informações que servem de base para as análises e comparações empregadas nas seções seguintes. Em seguida, na Seção 5.2 mostra-se quais os parâmetros e respectivos valores configurados nos métodos computacionais e algoritmos auxiliares.

A Seção 5.3 se divide primeiramente em mostrar a execução do método *K-means*, que agrupou os registros com base em diversas combinações de atributos a fim de encontrar agrupamentos coerentes. Ao final, os métodos computacionais de aprendizado supervisionado são acionados para comparar entre si o desempenho por meio de métricas previamente definidas e também a eficácia das divisões efetuadas pelo *K-means*.

Na Seção 5.4 são apresentados, de forma aprofundada, os melhores ou mais relevantes resultados, que fornecem importantes *insights* sobre as execuções dos métodos computacionais. Tais informações oferecem maior embasamento às considerações finais expostas no Capítulo 6.

Para acompanhamento ou consulta de todos os dados gerados e apresentados neste trabalho, disponibiliza-se um repositório versionado que inclui tabelas (conjuntos de dados), *notebooks* de código para os métodos computacionais, geração de métricas estatísticas e resultados em formatos textuais ou figuras. Essas informações foram disponibilizadas em Portella (2025), onde está compilado todos os processos efetuados para que se alcançasse todas os valores aqui dispostos.

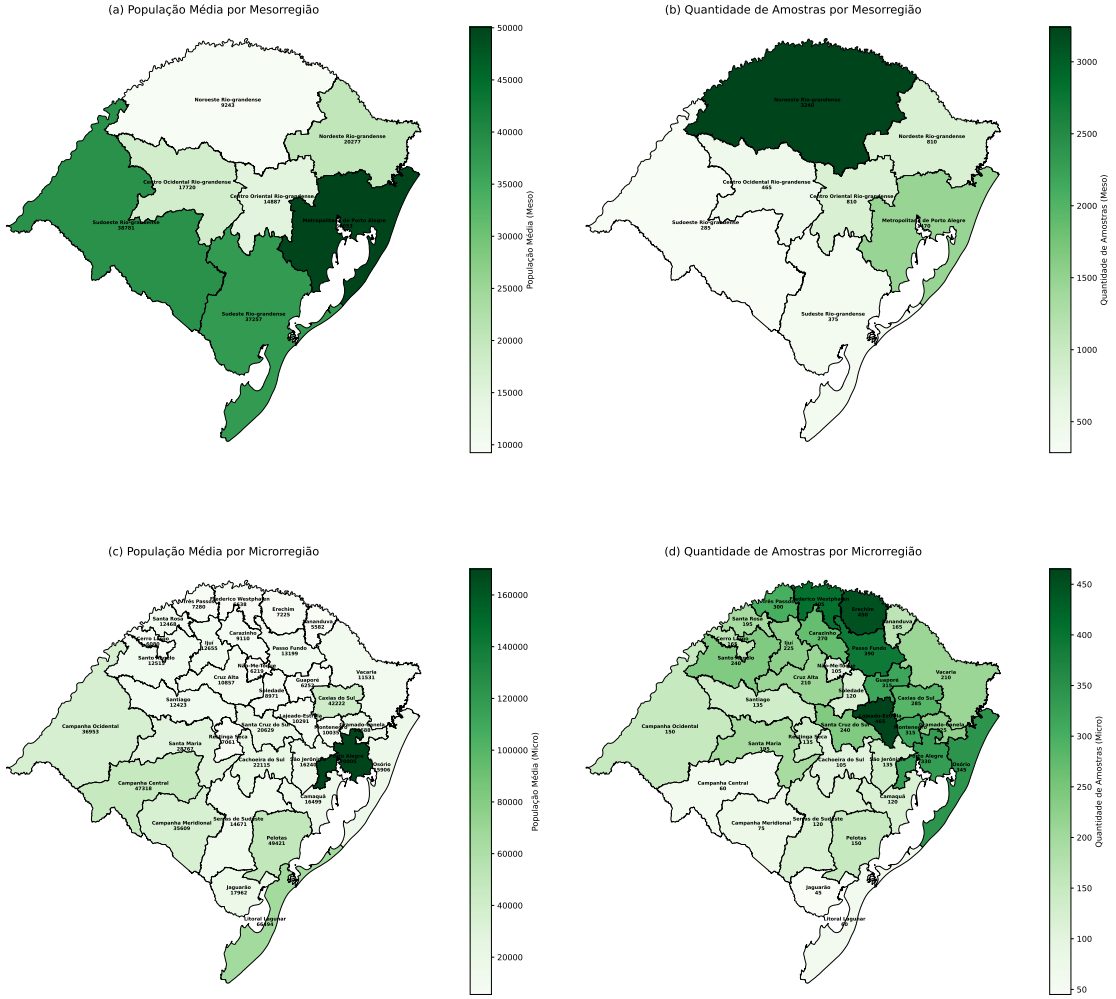
5.1 Estatísticas básicas dos conjuntos de dados

Nesta seção são apresentadas as métricas estatísticas básicas obtidas a partir do conjunto de dados utilizado neste trabalho, o qual compreende informações referentes

ao PBF e ao IDESE. Foram conduzidos processos de análise estatística descritiva considerando as amostras dos 15 anos de cada um dos 497 municípios, que fazem parte das 35 microrregiões e 7 mesorregiões do estado, conforme delimitações do IBGE (2017).

Para uma ideia de como estão distribuídos os registros (amostras) do conjunto de dados, e também para se descobrir a média da população, em relação às micro e mesorregiões, elaborou-se a Figura 17, em que o mapa do estado do Rio Grande do Sul é utilizado para a apresentação direta dessas informações. Este tipo de visualização serve para descrever como as regionalidades internas do estado podem afetar os resultados extraídos nas subseções seguintes.

Figura 17 – Mapa do estado dividido em micro e mesorregiões para demonstrar a média populacional e total de amostras constantes no conjunto de dados.



Fonte: Elaborado pelo Autor

A seguir são apresentadas informações resumidas. Para maiores referências, sugere-se o acesso ao Apêndice A, onde estão localizadas todas as informações das métricas básicas totais e anuais dos conjuntos de dados, em que a Seção A.1 expõe os

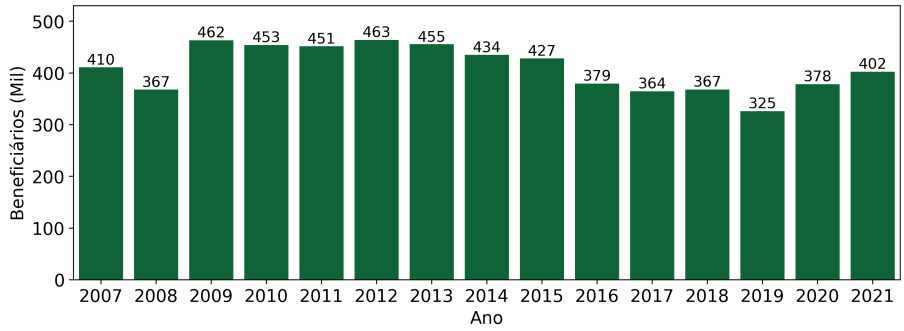
dados para o PBF e na Seção A.2 são mostrados os dados para cada um dos índices que pertencem ao IDESE.

5.1.1 Métricas PBF

Nesta seção, buscando trazer maior embasamento sobre o conjunto de dados do Programa Bolsa Família, apresentam-se os dados agregados do estado, municípios, micro e mesorregiões gaúchas, com foco nas cinco variáveis originais ou criadas para este estudo (descritas na Subseção 2.1.2), mas com maior foco no total de famílias beneficiárias e valor total repassado. Inicia-se com um comentário sobre as quantidades totais e anuais, com posterior aprofundamento nos dados, para identificar quais informações possuem maior relevância e detectar possíveis discrepâncias.

O PBF, após sua implantação, apresentou diversas modificações, tanto nos critérios de elegibilidade das famílias quanto no volume de recursos transferidos. Para demonstrar essas alterações, apresenta-se, na Figura 18, um gráfico de barras contendo a evolução, durante o período de 2007 a 2021, da quantidade total de famílias beneficiárias no estado do Rio Grande do Sul. Destaca-se o ano de 2012 com o maior volume, atingindo mais de 463 mil beneficiários, equivalendo a aproximadamente 5,3% da população, e o ano de 2019, com aproximadamente 325 mil beneficiários, como o menor valor agregado.

Figura 18 – Gráfico da evolução da quantidade total de famílias beneficiárias do PBF entre 2007 a 2021 no RS.

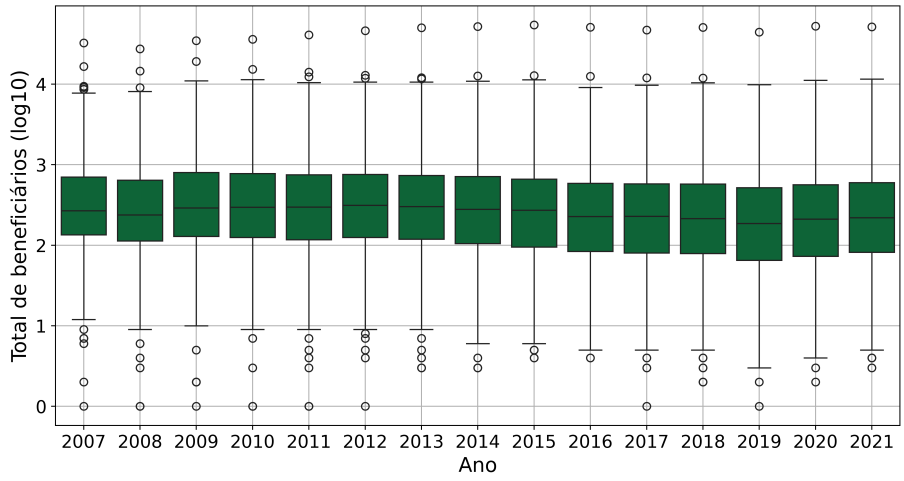


Fonte: Elaborado pelo Autor

Ao aprofundar a análise para verificar como estão distribuídas as famílias beneficiárias entre as amostras de cada ano dos 497 municípios, observa-se que existem grandes discrepâncias e alta concentração em determinados valores. Isso pode ser visualizado no gráfico *boxplot* da Figura 19, que contém os dados normalizados pela função log10, para melhor apresentação. Com este gráfico pode verificar-se que o valor

máximo de mais de 54 mil famílias beneficiárias foi obtido em 2015 na cidade de Porto Alegre. Entre 2007 e 2012, não houve registro de famílias beneficiárias em Pinto Bandeira (devido a nesta época o município não existir (BANDEIRA, 2025)), enquanto a maior taxa de população beneficiária, de 15,71%, refere-se a Benjamin Constant do Sul em 2021, valor superior à média do estado no período de 2007 a 2021, que foi de 4,45%.

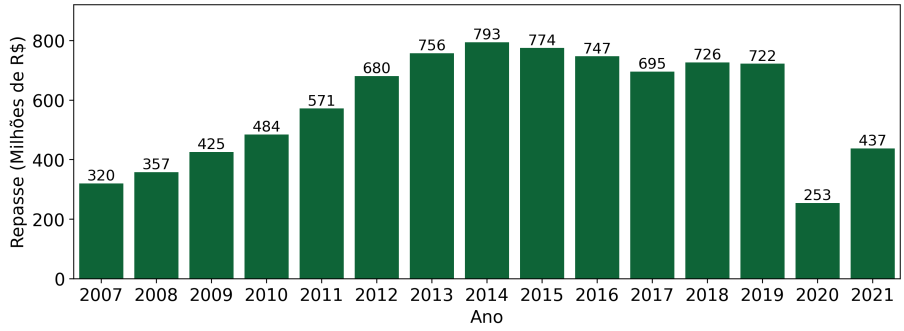
Figura 19 – Gráfico da evolução da quantidade de famílias beneficiárias do PBF entre 2007 a 2021 nos municípios do RS.



Fonte: Elaborado pelo Autor

Para os dados referentes ao valor repassado às famílias beneficiárias, observa-se, na Figura 20, um crescimento consistente no volume transferido entre 2007 e 2014, seguido por uma expressiva queda em 2019, decorrente de alterações implementadas durante a pandemia de COVID-19, quando foi instituído pelo governo federal o programa Auxílio Emergencial, que absorveu os valores anteriormente destinados às famílias beneficiárias do PBF.

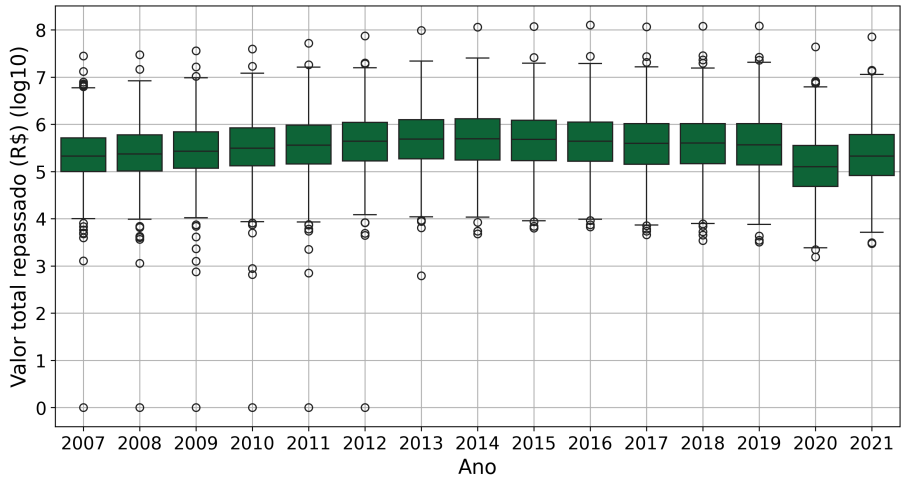
Figura 20 – Gráfico da evolução do volume total repassado às famílias beneficiárias do PBF entre 2007 a 2021 no RS.



Fonte: Elaborado pelo Autor

Ao examinar os registros referentes a cada um dos municípios gaúchos, apresenta-se a Figura 21, na qual se verifica a existência de discrepâncias relevantes, com a maior parte dos registros concentrando-se em valores anual repassado próximos a R$ 870 mil, mas com valores máximos alcançados pela cidade de Porto Alegre, que chegou em 2021 a R$ 126 milhões. Um dado importante a ser enfatizado é que, no período de 2007 a 2021, a média de valor repassado às famílias foi de R$ 1.388,44, com o maior valor de repasse aos beneficiários registrado em Fagundes Varela no ano de 2018, no montante de R$ 7.545,50.

Figura 21 – Gráfico da evolução do volume total de recursos repassados às famílias beneficiárias do PBF entre 2007 a 2021 aos municípios do RS.



Fonte: Elaborado pelo Autor

Para apresentar as principais métricas extraídas dos atributos do PBF utilizados neste trabalho, elaborou-se a Tabela 6, que resume a evolução das estatísticas entre 2007 e 2021, indicando o volume total, média total anual, desvio padrão, média, mediana, quartis inferior (25%/Q1) e superior (75%/Q3), e valores mínimo e máximo. Ressalta-se que os valores totais e média total anual não são apresentados para os atributos que são percentuais ou derivados de outros valores.

Tabela 6 – Estatísticas de beneficiários e valores repassados

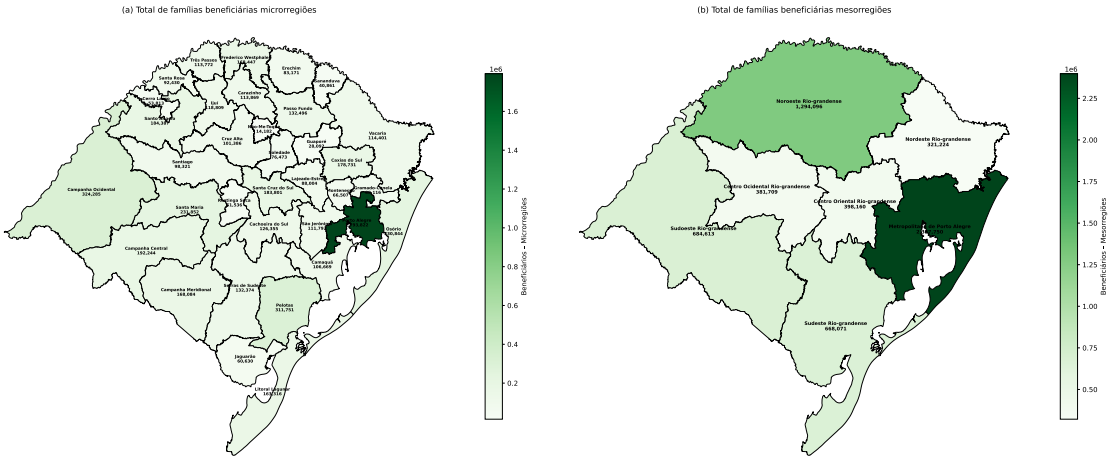
Atributo	**Total**	**Média anual**	**Média**	**Desv. pad.**	**Q1**	**Mediana**	**Q3**	**Mín.**	**Máx.**
Famílias beneficiárias	$6{,}15 \times 10^6$	409,7k	824,4	2.506,7	99	253	667	1	54.272
Valor repassado (R$)	$8{,}75 \times 10^3$	583,1k	1.173	4.467	0,12	0,34	0,87	620	126,7k
População beneficiária (%)	-	-	4,47%	3%	2,03%	4,05%	6,4%	0,036%	15,71%
Repasse por fam. benef. (R$)	-	-	1.388,4	532,0	944,1	1.373,1	1.769,7	206,7	7.545,5
Repasse por população (R$)	-	-	62,19	51,0	23,65	50,04	87,50	0,23	429,9

Fonte: Elaborado pelo Autor

Como análise complementar, criou-se a Figura 22, na qual é apresentada a soma

total de famílias beneficiárias no estado durante o período de análise, destacando-se na Figura 22(a) as microrregiões e na Figura 22(b) as mesorregiões segundo divisão do IBGE. Percebe-se que as regiões com maior volume populacional concentram os maiores totais de beneficiários, enquanto se destaca a microrregião de São Jerônimo, onde se localizam municípios com baixas populações, que resulta em apenas 1,82% do total de famílias beneficiárias no RS.

Figura 22 – Total de famílias beneficiárias do PBF por micro e mesorregiões do RS.



Fonte: Elaborado pelo Autor

Observa-se que, durante o período de 2007 a 2021, o RS experimentou algumas oscilações nos valores referentes às famílias beneficiárias, mantendo uma distribuição com pequenos incrementos ao longo dos anos. Já os valores repassados demonstram como as mudanças nos parâmetros do PBF afetaram a renda dos beneficiários e municípios, com destaque para as regiões mais populosas que, consequentemente, concentram os maiores volumes de recursos.

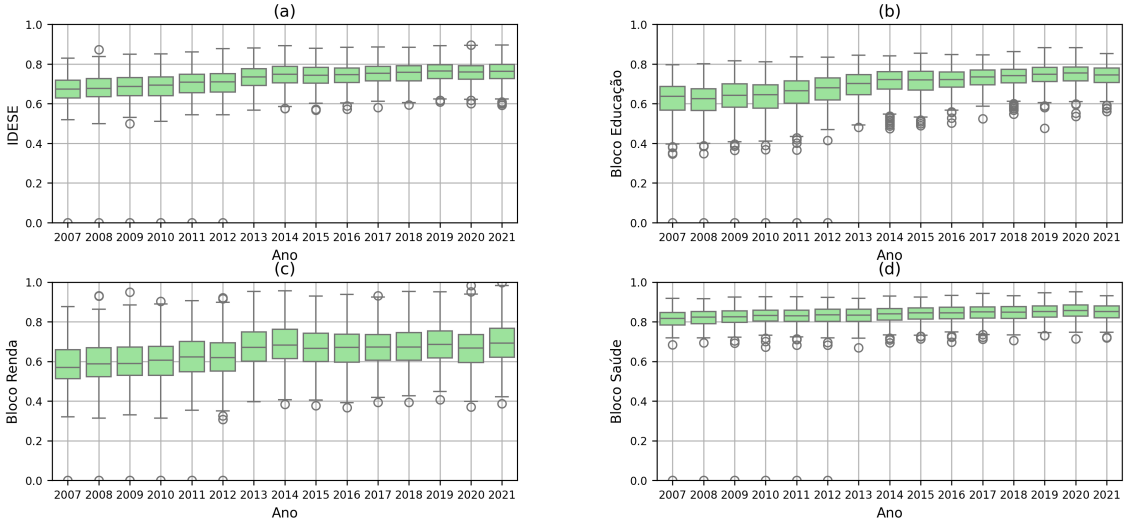
5.1.2 Métricas IDESE

Nesta seção são apresentadas as métricas estatísticas básicas obtidas sobre o conjunto de dados do Índice de Desenvolvimento Socioeconômico do Rio Grande do Sul (IDESE). Será analisado o índice principal, os três blocos temáticos: educação, renda e saúde, e serão abordados sub-blocos específicos que apresentem valores relevantes para este estudo e análises futuras.

Para ilustrar a distribuição dos índices no período de 2007 a 2021, foi gerada a Figura 23, a qual exibe quatro *boxplots* com todas as amostras do conjunto de dados

utilizado neste estudo. A figura está dividida da seguinte forma: (a) valor IDESE geral; (b) bloco Educação; (c) bloco Renda; e (d) bloco Saúde. Cada *boxplot* demonstra como os valores se distribuem entre os municípios do estado em cada ano.

Figura 23 – Distribuição do IDESE e seus blocos temáticos principais no RS entre 2007 a 2021.



Fonte: Elaborado pelo Autor

Ao observar a Figura 23, verifica-se que o valor médio do IDESE geral, apresentou aumento constante, mantendo-se na faixa classificada como média pela metodologia do índice (apresentada na Subseção 3.2.1). A mesma evolução foi alcançada pelos blocos Educação e Renda. Destaca-se, positivamente, o bloco Saúde, por apresentar valor médio sempre acima de 0,800, classificado como valor alto.

A Tabela 7 sintetiza a distribuição dessas métricas ao longo de todo o período analisado. Com a ressalva de que os valores mínimos foram obtidos mediante a remoção dos registros zerados de Pinto Bandeira entre 2007 e 2012, uma vez que tais valores poderiam afetar a análise subsequente. Destaca-se que os valores apresentados na Tabela 7 referem-se exclusivamente ao conjunto de dados utilizado neste estudo e podem divergir dos dados oficialmente divulgados por DEE-RS (2024).

Tabela 7 – Métricas do IDESE e blocos de educação, renda e saúde no RS entre 2007 a 2021.

Índice	Média	Desvio Padrão	0,25%	Mediana	0,75%	Mínimo	Máximo
IDESE	0.725	0.064	0.682	0.730	0.774	0.498	0.896
Bloco Educação	0.691	0.083	0.642	0.705	0.753	0.346	0.883
Bloco Renda	0.648	0.107	0.572	0.648	0.724	0.308	0.997
Bloco Saúde	0.836	0.042	0.807	0.839	0.868	0.668	0.951

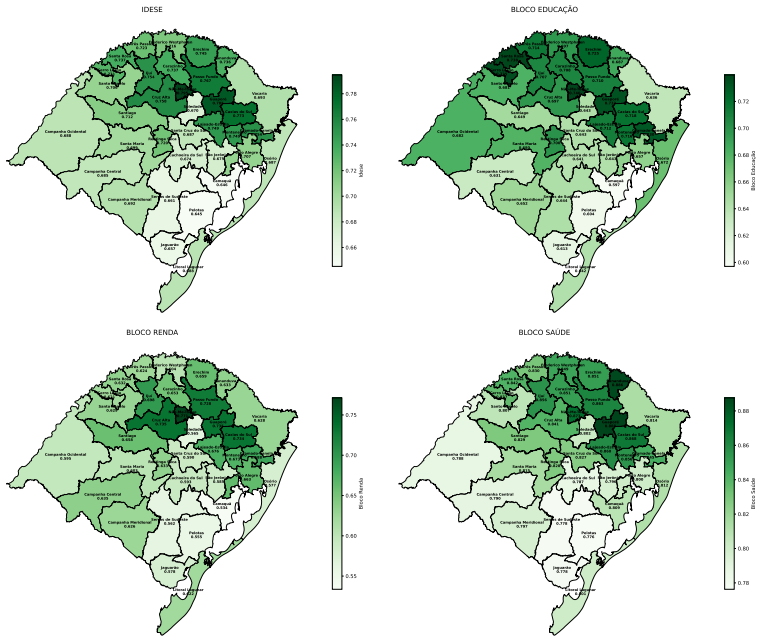
Fonte: Elaborado pelo Autor

Os valores máximos do IDESE e dos blocos Educação e Renda foram observados em 2020, nos municípios de Carlos Barbosa, Picada Café e Santo Expedito do Sul, respectivamente. O maior valor do bloco Saúde ocorreu em Água Santa, em 2021. Quanto aos menores valores, destaca-se o valor do IDESE geral de 0,498 registrado em Jaquirana em 2019, o bloco educação apresentou o valor de 0,347 em Charrua em 2007, já o bloco renda teve o valor de 0,308 em Benjamin Constant do Sul em 2012, e o bloco saúde obteve o valor de 0,668 em Pedro Osório em 2013.

Ao examinar as médias dos demais índices que compõem os blocos temáticos, obtidas no período de 2007 a 2021, observa-se que as cinco melhores médias pertencem ao bloco Saúde, com destaque para os índices de Mortalidade de Menores de 5 Anos, Óbitos por Causas Mal Definidas e Longevidade, com valores de 0,898, 0,903 e 0,927, respectivamente. Por outro lado, os índices do bloco Renda apresentam médias inferiores a 0,665 em Apropriação da Renda, sendo os três piores valores os de Geração da Renda (0,631), Mortes por Causas Evitáveis (0,619) e Escolaridade Adulta (0,454).

A Figura 24 apresenta, em um mapa, as médias dos quatro índices principais IDESE geral e os blocos educação, renda e saúde, para cada uma das 35 microrregiões do estado. Esse gráfico facilita a visualização da distribuição dos indicadores em regiões que agregam municípios, permitindo compreender como essas localidades estão se desenvolvendo.

Figura 24 – Valores médios do IDESE e seus blocos temáticos em cada uma das microrregiões do RS entre 2007 a 2021.



Fonte: Elaborado pelo Autor

Com essas informações, é possível compreender que os melhores índices estão concentrados na região norte do estado e na área próxima à capital. Também se pode interpretar que os índices de renda são muito discrepantes no estado, em contraste com os índices de saúde, sempre elevados e com pouca variabilidade. Ademais, é positivo o aumento anual nos valores do IDESE geral e do bloco educação.

5.2 Configurações de teste e aplicação

Está seção interage com este trabalho para apresentar as configurações necessárias para que os testes e as aplicações dos métodos computacionais sejam executados de forma correta. A seção segue com a descrição da edição do conjunto de dados, bem como a apresentação e definição dos parâmetros pertencentes a cada um dos métodos utilizados neste estudo.

A primeira configuração consistiu no ajuste do conjunto de dados. Conforme observado nas Subseções 5.1.1 e 5.1.2, os conjuntos de dados do PBF e IDESE apresentam valores discrepantes (*outliers*). Tais valores podem causar inconsistência na aplicação dos métodos computacionais, além de possivelmente gerar distorções nas análises dos resultados. Por este motivo, optou-se por remover todos os registros nulos, os quais compreendem valores anteriores à nova emancipação do município de Pinto Bandeira, ocorrida em 2013 (BANDEIRA, 2025); este procedimento removeu 7 linhas (amostras) sem relevância.

Logo após, realizou-se a remoção de valores extremos, identificados no conjunto de dados do PBF através do atributo **total de beneficiários**. Estes valores referem-se à capital do estado, Porto Alegre, que apresenta uma discrepância extrema de famílias beneficiárias em comparação aos outros municípios. Desta forma, os 15 registros acima de 20 mil beneficiários foram removidos. Ao final, o conjunto de dados ficou formado por 29 colunas (atributos), envolvendo atributos descritivos e de análise, além do total de 7433 amostras.

Para os métodos computacionais, a seleção dos parâmetros ocorreu após a aplicação de testes baseados na literatura e também por empirismo (tentativa e erro). Definiu-se um parâmetro comum a vários métodos e algoritmos: o *random state* foi fixado no valor igual a 42, este valor se apresenta como universalmente aceito para reprodutibilidade. Cada método computacional possui características próprias de parametrização, em que: as Redes Neurais (MLP) exigem o ajuste fino de diversas

características para a construção da melhor arquitetura; o *XGBoost* permite uma quantidade variada de parâmetros, possuindo alguns cruciais para seu funcionamento; e o *K-means*, que demanda configurações específicas na construção dos grupos. No entanto, a Regressão Linear não necessita de configurações prévias para a criação do modelo.

Para apresentar as configurações dos métodos e permitir comparações em trabalhos futuros, elaborou-se a Tabela 8, que detalha os parâmetros e valores para cada método utilizado.

Tabela 8 – Parâmetros e configurações dos métodos utilizados

Método	**Parâmetro**	**Valor**	**Descrição**
K-means	n_clusters	Indefinido	Define a quantidade de grupos que o K-means deve construir (definido via silhueta).
	random_state	42	Semente para seleção aleatória das amostras.
Regressão Linear	Não possui hiperparâmetros de ajuste.		
MLP	Camadas e neurônios	[16,8,4,2,1]	Quantidade de neurônios em cada camada da rede.
	Função de ativação	relu	Função de ativação usada durante o treinamento (exceto na saída).
	Otimizador	RMSprop	Otimizador utilizado (biblioteca Keras).
	Learning rate	0.001	Taxa de aprendizado durante o treinamento.
	Função de perda	MSE	Métrica de erro usada para ajuste do modelo.
XGBoost	n_estimators	1000	Quantidade de árvores de decisão criadas.
	max_depth	5	Profundidade máxima da árvore.
	learning_rate	0.1	Taxa de aprendizado (contribuição de cada árvore).
	objective	reg:squarederror	Função objetivo para problemas de regressão.
	random_state	42	Semente para seleção aleatória das amostras.

Fonte: Elaborado pelo Autor.

Para a confirmação dos valores apresentados na Tabela 8, foram efetuados testes de desempenho utilizando um conjunto específico de variáveis dependentes e independentes, além do conjunto de dados completo, editado conforme citado no incio desta seção. Buscou-se, ao fim de cada teste, a melhoria do coeficiente de determinação R^2.

Como parâmetros de teste, escolheu-se o atributo *total de beneficiários* como variável dependente (alvo), por se tratar de um atributo primordial e completo. Como variável independente, utilizou-se o conjunto de atributos **blocoSaudeTodos** (cuja definição reside no Capítulo 2), por conter um equilíbrio adequado de características.

Com estes atributos selecionados, realizaram-se testes utilizando o conjunto de dados em cada um dos métodos de aprendizado de máquina supervisionado, a fim de alternar configurações, otimizar o R^2 final e consequentemente selecionar os melhores parâmetros. Estes testes fundamentaram as configurações fixas para a aplicação apresentada nas seções seguintes.

Ressalta-se que o *K-means* não necessitou de mudanças manuais para a criação do modelo, uma vez que a definição do número de *clusters* foi guiada pelo cálculo do coeficiente de silhueta. Já para os métodos de Redes Neurais (MLP) e *XGBoost*, foram realizados os seguintes testes para a melhoria de suas arquiteturas:

- **Redes Neurais (MLP)**: Realizou-se uma sequência de testes para adaptar os melhores parâmetros para a criação da arquitetura:

    1. **Otimizador**: Foram testados diversos otimizadores disponíveis na biblioteca *Keras*. O *RMSprop* (KERAS, 2025) apresentou o melhor desempenho, com taxa de aprendizado (*learning rate*) de 0,001;

    2. **Funções**: Escolheu-se a *relu* como função de ativação, por se comportar melhor no tipo de problema proposto. A função de perda selecionada foi o Erro Quadrático Médio (*MSE*), comum na construção de arquiteturas MLP para regressão;

    3. **Arquitetura de camadas**: Para definir a quantidade ideal de camadas e neurônios, utilizou-se a técnica de potências de 2^n, onde n é decomposto a cada nova camada. Testaram-se arquiteturas com n variando de 2 a 8. Ao final, definiu-se $n = 4$, resultando na arquitetura $[16, 8, 4, 2, 1]$, onde a última camada possui obrigatoriamente um (1) neurônio, pois este trabalho se tratar de um para problemas de regressão.

- ***XGBoost***: Este método possui inúmeros hiperparâmetros. Seguindo a literatura (IZBICKI; SANTOS, 2020), sendo alguns com mais destaque para teste e, obrigatoriamente, para adaptações. Abaixo, apresentam-se as configurações de usadas como modelo inicial:

    1. *n_estimators*=100;

    2. *max_depth*=3;

    3. *learning_rate*=0.1;

4. *objective*='reg:squarederror';
5. *random_state*=42).

Segue, abaixo, breve descrição de cada um dos parâmetros e justificativa da sua alteração, seguindo das explicações dos teste efetuados, para chegar em valores mais adequados dos parâmetros iniciais informados acima:

- **Número de árvores**: O parâmetro *n_estimators* define a quantidade de árvores para o aprendizado conjunto. Sua definição incorreta pode causar sobreajuste (*overfitting*). Foram testados os valores 100, 500, 800 e 1000, com a seleção do valor de *n_estimators* = 1000;
- **Profundidade das árvores**: Define, de forma simples, a complexidade das perguntas que cada árvore pode responder. Valores baixos podem gerar subajuste (*underfiting*), enquanto valores altos podem causar sobreajuste. Testaram-se valores de *max_depth* variando entre 3 a 6, e ao final chegou-se a conclusão de que 5 seria o mais ideal;
- **Taxa de aprendizagem**: Define a contribuição de cada árvore no resultado final. O hiperparâmetro foi testado com valores de 0.001, 0.05, 0.01 e 0.1, com *learning_rate* sendo fixado em 0.1;

Além das configurações nos métodos principais, foram ajustados parâmetros nos algoritmos auxiliares. Destaca-se que a validação cruzada *K-fold* teve o número de divisões (*n_splits*) definido como 5, valor adequado para problemas de regressão. Para o cálculo do coeficiente de silhueta, não foram necessárias modificações, visto que sua função é apenas validar os agrupamentos gerados pelo *K-means*.

Com as configurações supracitadas, acredita-se que a replicação dos testes e execuções seja viável. Caso sejam necessárias novas pesquisas, alterações nestes parâmetros podem trazer resultados relevantes para este domínio de estudo.

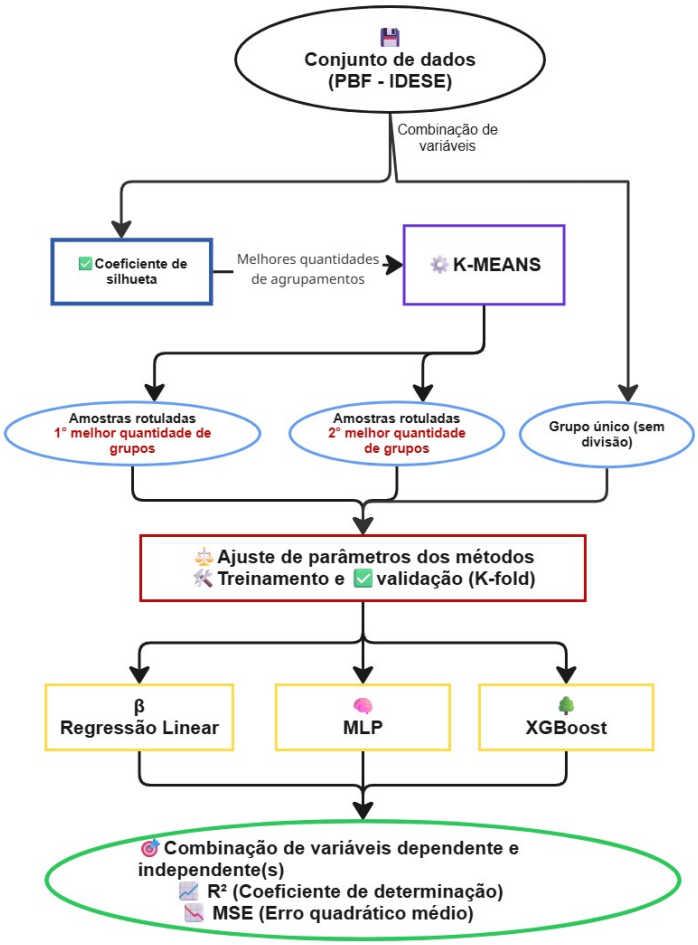
5.3 Aplicação de métodos computacionais

Nesta seção são descritos os processos efetuados para a aplicação dos métodos computacionais de aprendizado de máquina supervisionado e não supervisionado, para que sejam gerados os resultados deste trabalho. Inicia-se com o uso do *K-means* para dividir as combinações de atributos dependentes e independentes em grupos coesos e,

com isso, seguir para a aplicação dos métodos supervisionados, a fim de que se obtenham os melhores valores de R^2 e, assim, validar os agrupamentos gerados anteriormente.

Para consulta, encontram-se disponíveis no Apêndice B, alguns resultados relevantes e informações sobre a execução dos métodos computacionais. Na Seção B.1 são apresentadas as informações referentes ao tempo de construção e execução de cada modelo. Subsequentemente, nas Seções B.2 e B.3, são expostos os dados estatísticos dos grupos que compõem os 15 melhores resultados do coeficiente de determinação (R^2), obtidos com a aplicação dos métodos computacionais de aprendizado supervisionado, seguindo os valores encontrados como as melhores divisões definidas pelo coeficiente de silhueta aplicado aos conjuntos de dados processados pelo *K-means*.

Figura 25 – Fluxograma do processo de validação



Fonte: Elaborado pelo Autor.

A Figura 25 visa apresenta o fluxo de ciência de dados de forma resumida, que é usado nessa seção. Em suma, inicia-se com um conjunto de dados contendo

atributos (variáveis) dependentes e independentes previamente selecionados. Em seguida, selecionam-se os melhores valores para a quantidade de agrupamentos na execução do *K-means*. Posteriormente, com as amostras já rotuladas, ocorre o processo de treinamento e execução dos métodos computacionais (Regressão Linear, Redes Neurais MLP e *XGBoost*). Ao final, são gerados modelos que retornam as métricas de avaliação.

As execuções descritas nas subseções seguintes visaram a busca por resultados preliminares, que serviram de grande valia para a geração de análises mais aprofundadas. Estas análises nos informam quais características dos índices socioeconômicos do IDESE mais impactam o PBF. Tais interpretações podem ser vistas na Seção 5.4.

5.3.1 Agrupamento de amostras com K-means

Nesta seção é descrito o procedimento de divisão dos grupos por meio do método *K-means* aplicado às variáveis dependentes e independentes previamente estabelecidas. O objetivo consiste em obter agrupamentos mais nítidos que possam fornecer informações mais precisas para avaliar a existência de impacto dos valores do IDESE sobre o PBF.

Para implementar a técnica de aprendizado de máquina não supervisionado e segmentar o conjunto de dados em subconjuntos homogêneos, buscou-se melhorar os resultados do coeficiente de determinação (R^2) e Erro Quadratico Médio *(MSE)*, além de estabelecer a correlação entre as variáveis dependentes e independentes. Para isso, empregou-se o algoritmo *K-means*, que realiza esse processo de maneira automatizada.

A implementação do *K-means* inicia com a escolha, no conjunto de dados, da variável dependente (oriunda do PBF) e de uma ou mais variáveis independentes (provenientes do IDESE). Essa escolha baseia-se na Tabela 9, que lista os dados tratados como dependentes e independentes, cujas descrições são elaboradas na Seção 2.1.2.

Tabela 9 – Variáveis dependentes e independentes do estudo

Variável Dependente	**Variável Independente**
Total de beneficiários	idese
Valor total repassado	blocos
População beneficiária	blocoEducacao
Repasse por beneficiário	blocoEducacaoResumido
Repasse por população	blocoEducacaoTodos
	blocoRenda
	blocoRendaTodos
	blocoSaude
	blocoSaudeResumido
	blocoSaudeTodos
	independenteTodos

Fonte: Elaborado pelo Autor

Com a definição dos atributos, passou-se à combinação sucessiva de variáveis dependentes e independentes. Foram executados 11 testes para cada variável dependente, totalizando 55 combinações. Cada combinação foi submetida ao cálculo do coeficiente de silhueta, métrica que indica a quantidade ideal de grupos (*clusters*) para o *K-means*. O coeficiente de silhueta foi calculado através de um laço de repetição para um intervalo de k (número de grupos) variando de 2 a 10. Selecionaram-se os dois melhores resultados de k para a divisão dos grupos pelo *K-means*, permitindo a posterior análise métrica com os métodos de aprendizado supervisionado .

Após a aplicação do coeficiente de silhueta, os valores correspondentes ao **melhor** k e ao **segundo melhor** k foram armazenados para uso na construção do modelo *K-means*. Dessa forma, o método identificou a relação entre as amostras, agrupando-as conforme sua similaridade. Observou-se que o melhor valor de k variou entre 2 e 4, enquanto o segundo melhor valor situou-se entre 2 e 6.

Para complementar a análise e adicionar mais informações aos resultados, optou-se também por realizar testes sem a aplicação do coeficiente de silhueta e sem a divisão em grupos pelo *K-means*. Nestes casos, denominados **grupo único**, utilizou-se o conjunto de dados completo. Estes testes funcionaram como controle, permitindo observar a capacidade de generalização dos modelos diante de um volume maior de informações heterogêneas, que não compartilham necessariamente as mesmas condições.

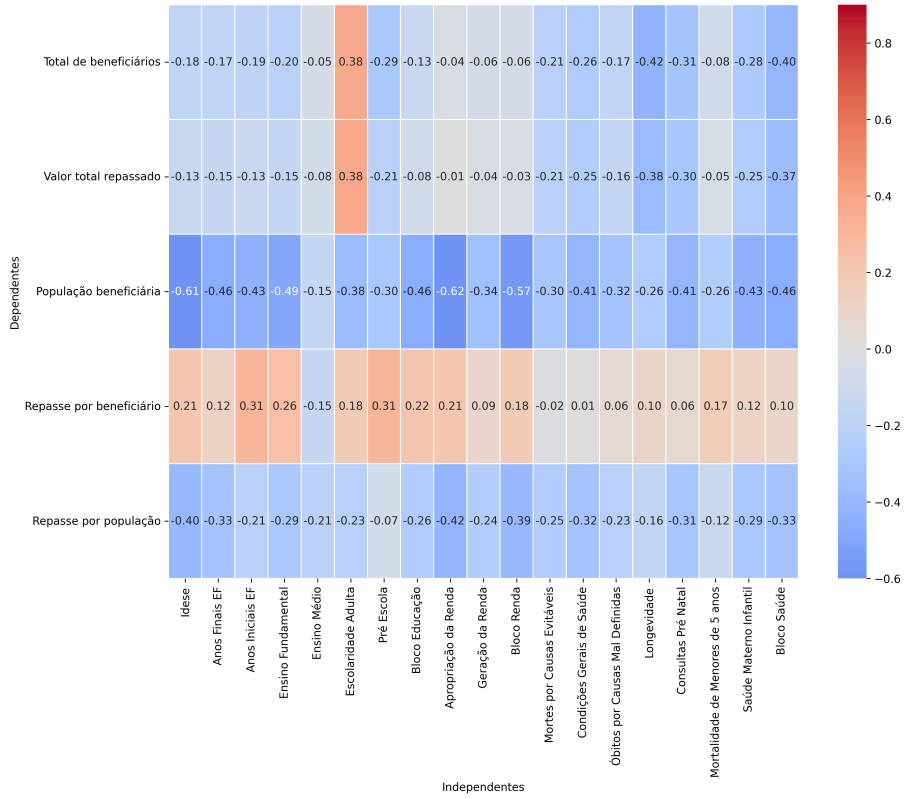
5.3.2 Encontro de relações

Nesta seção são apresentados os resultados da aplicação dos métodos computacionais de aprendizado supervisionado sobre os conjuntos de dados construídos a partir da divisão realizada pelo *K-means*, conforme descrito na Seção 5.3.1. A seção inicia com uma análise preliminar das relações entre os atributos dependentes e independentes, para, em seguida, aprofundar-se nos melhores resultados obtidos com a aplicação dos modelos.

Para observar as relações entre as variáveis dependentes, representadas pelos atributos do PBF, e as independentes, representadas pelos índices do IDESE, foi criada uma matriz de confusão (matriz de calor), exposta na Figura 26, que mostra, por meio do cálculo da raiz quadrada do R^2, a relação existente entre cada par de atributos, tendo cada um dos atributos retirado de um dos dois conjuntos de dados. Pode-se observar que, na sua grande maioria, as correlações são baixas ou negativas. Se destaca positivamente a

Escolaridade Adulta, que apresentou 41% de correlação com os principais atributos (Total de beneficiários e Valor total repassado) do PBF, mas destaca-se ainda mais positivamente o atributo criado para o PBF, Repasse por beneficiários, que possui correlação com a maior parte dos atributos do IDESE. Esses dados podem fornecer subsídios para a geração de *insights* na interpretação dos resultados.

Figura 26 – Matriz de calor/correlação



Fonte: Elaborado pelo Autor

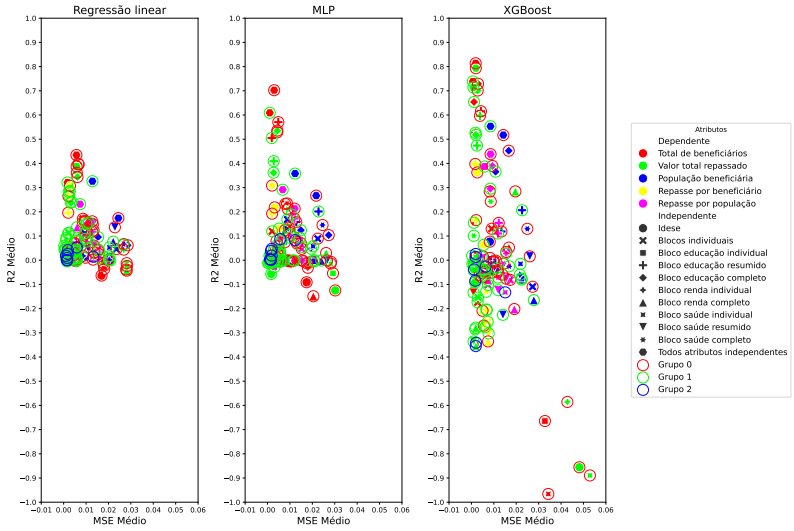
Após a aplicação do *K-means* para separar os grupos de forma otimizada, seguindo as duas divisões sugeridas pelo coeficiente de silhueta para os melhores valores de k grupos, conforme descrito na Seção 5.3.1 e Apêndice B, partiu-se para a próxima etapa. Esta consistiu na execução dos métodos computacionais de Regressão Linear, Redes Neurais (MLP) e *XGBoost*, a fim de encontrar os valores de R^2 e MSE que indicassem quais atributos dependentes e independentes apresentavam a melhor relação entre si.

Com cada combinação e seus respectivos grupos definidos, iniciou-se a aplicação dos três métodos de aprendizado supervisionado. No primeiro momento, realizou-se o treinamento utilizando os dados divididos em 5 etapas através da validação cruzada *K-fold*. Este método dividiu os dados selecionados entre treinamento e validação, garantindo que toda amostra fosse utilizada em uma das etapas. Ao final de cada execução

do *K-fold*, obtiveram-se os valores de R^2 e MSE, extraindo-se posteriormente a média desses resultados.

A sequência de execuções foi aplicada em todas as combinações. A partir disso, foram gerados gráficos onde se representaram todas as combinações e seus grupos através de formas geométricas e cores para melhor visualização. Cada gráfico contém os resultados para os três métodos, indicando os melhores desempenhos na parte superior esquerda. Na Figura 27 são apresentados os resultados para a divisão usando o **melhor** *k*, correspondente aos grupos que agregam amostras com maior similaridade entre si.

Figura 27 – Resultados melhor *k*



Fonte: Elaborado pelo Autor

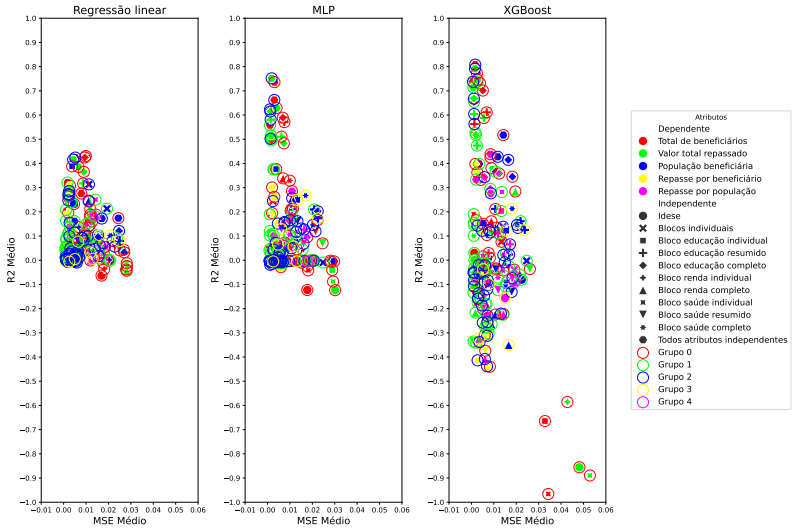
Tabela 10 – 10 melhores resultados obtidos da divisão *K-means* melhor *k*

Dependente	Independente	Modelo	Grupo	R² médio	R² desvio	MSE médio	MSE desvio	Tempo (s)
Total de beneficiários	Todos atributos	XGBoost	0	0.8139	0.0354	0.0019	0.0006	3.00
Valor total repassado	Todos atributos	XGBoost	0	0.7935	0.0340	0.0020	0.0004	3.08
Total de beneficiários	Todos atributos	XGBoost	1	0.7388	0.0628	0.0006	0.0001	2.96
Total de beneficiários	Bloco educação completo	XGBoost	0	0.7291	0.0346	0.0030	0.0005	1.64
Valor total repassado	Todos atributos	XGBoost	1	0.7140	0.0433	0.0009	0.0003	3.13
Total de beneficiários	Todos atributos	MLP	0	0.7025	0.0505	0.0030	0.0004	39.83
Valor total repassado	Bloco educação completo	XGBoost	0	0.7014	0.0590	0.0028	0.0005	1.54
Total de beneficiários	Bloco educação completo	XGBoost	1	0.6539	0.0311	0.0012	0.0002	2.01
Total de beneficiários	Bloco educação resumido	XGBoost	0	0.6155	0.0533	0.0043	0.0012	1.45
Total de beneficiários	Todos atributos	MLP	1	0.6090	0.0990	0.0010	0.0003	46.81

Fonte: Elaborado pelo Autor

Os resultados do melhor k revelam que o maior R^2 médio foi obtido pelo *XGBoost*, atingindo 0,8139 (81,39%) na combinação **Total de beneficiários e Todos atributos independentes**; esta combinação resultou da definição de melhor k igual a 2. É perceptível o poder de generalização e expansão do entendimento dos dados pelo *XGBoost* em comparação com a Regressão Linear, que se manteve em uma estreita faixa de resultados (com destaque para a mesma combinação de atributos gerando um R^2 de 0,4345 ou 43,45%), e com a MLP, que obteve bons resultados, alcançando seu melhor valor de R^2 igual a 0,7024 (70,24%). Na Tabela 10, são apresentados os 10 melhores resultados encontrados para a divisão feita pelo melhor k e o tempo em segundos que levou para o método ser executado.

Figura 28 – Resultados dos modelos com divisão *K-means* do segundo melhor k



Fonte: Elaborado pelo Autor

A Figura 28 apresenta os resultados para o **segundo melhor** k. Neste cenário, observa-se um maior número de combinações, dada a maior quantidade de divisões dos dados realizada pelo *K-means*, possibilitando resultados positivos em maior escala. O melhor desempenho foi novamente do modelo *XGBoost*, com R^2 de 0,8034 (80,34%) na combinação de **Total de beneficiários e Todos atributos independentes**. A MLP obteve 0,7516 (75,16%) na mesma combinação, enquanto a Regressão Linear chegou a um R^2 de 0,4305 (43,05%) na combinação **Total de beneficiários e Bloco educação resumido**, ressaltando-se seu MSE médio de 0,0098. Na Tabela 11, são apresentados os 10 melhores resultados da divisão feita pelo valor do segundo melhor k, destacando-se o *XGBoost*.

Para obter resultados que servissem como prova de conceito (teste de controle), foram efetuados testes sem a divisão do *K-means*; esses resultados são apresentados na

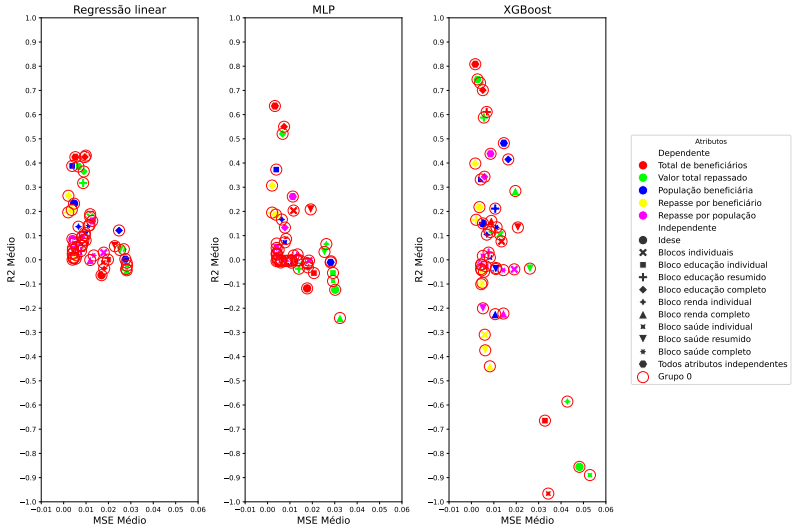
Tabela 11 – 10 melhores resultados obtidos da divisão *K-means* do segundo melhor *K*

Dependente	Independente	Modelo	Grupo	R^2 médio	R^2 desvio	MSE médio	MSE desvio	Tempo (s)
Total de beneficiários	Todos atributos	XGBoost	2	0.8083	0.0453	0.0017	0.0002	2.51
Valor total repassado	Todos atributos	XGBoost	2	0.7915	0.0831	0.0016	0.0010	1.83
Total de beneficiários	Todos atributos	XGBoost	0	0.7728	0.0723	0.0024	0.0007	1.98
Valor total repassado	Todos atributos	MLP	2	0.7517	0.0566	0.0019	0.0009	22.74
Valor total repassado	Todos atributos	XGBoost	0	0.7456	0.0303	0.0027	0.0006	2.64
Total de beneficiários	Todos atributos	XGBoost	1	0.7388	0.0628	0.0006	0.0001	2.87
Total de beneficiários	Bloco educação completo	XGBoost	2	0.7369	0.0526	0.0008	0.0002	1.12
Total de beneficiários	Todos atributos	MLP	0	0.7358	0.0859	0.0031	0.0018	18.42
Valor total repassado	Bloco educação completo	XGBoost	0	0.7314	0.0789	0.0039	0.0014	1.22
Valor total repassado	Todos atributos	XGBoost	1	0.7140	0.0433	0.0009	0.0003	3.16

Fonte: Elaborado pelo Autor

Figura 29. Pode-se verificar que as mesmas tendências observadas anteriormente nos valores de divisão de melhor e segundo melhor *k* se mantêm sem alterações significativas nos resultados dos modelos. O *XGBoost* obteve, na combinação **Total de beneficiários e Todos atributos independentes**, um R^2 de 0,8083 (80,83%); a MLP alcançou 0,6355 (63,55%); e a Regressão Linear teve seu melhor resultado na combinação **Total de beneficiários e Bloco educação resumido**, com valor de 0,4305 (43,05%). Na Tabela 12 são apresentados os 10 melhores resultados para os dados sem divisão de grupos.

Figura 29 – Resultados grupo único



Fonte: Elaborado pelo Autor

Em relação aos demais atributos do PBF criados a partir das variáveis

Tabela 12 – 10 melhores resultados sem divisão de grupos do *K-means*

Dependente	Independente	Modelo	Grupo	R² médio	R² desvio	MSE médio	MSE desvio	Tempo (s)
Total de beneficiários	Todos atributos	XGBoost	0	0.8083	0.0453	0.0017	0.0002	2.63
Valor total repassado	Todos atributos	XGBoost	0	0.7456	0.0303	0.0027	0.0006	2.58
Valor total repassado	Bloco educação completo	XGBoost	0	0.7314	0.0789	0.0039	0.0014	1.29
Total de beneficiários	Bloco educação completo	XGBoost	0	0.7011	0.0544	0.0051	0.0017	1.35
Total de beneficiários	Todos atributos	MLP	0	0.6355	0.0657	0.0033	0.0005	34.65
Total de beneficiários	Bloco educação resumido	XGBoost	0	0.6110	0.0282	0.0068	0.0020	1.15
Valor total repassado	Bloco educação resumido	XGBoost	0	0.5878	0.0789	0.0056	0.0026	1.12
Total de beneficiários	Bloco educação completo	MLP	0	0.5500	0.0628	0.0074	0.0013	24.87
Valor total repassado	Bloco educação completo	MLP	0	0.5205	0.0611	0.0067	0.0012	25.21
População beneficiária	Todos atributos	XGBoost	0	0.4821	0.0434	0.0145	0.0020	2.35

Fonte: Elaborado pelo Autor

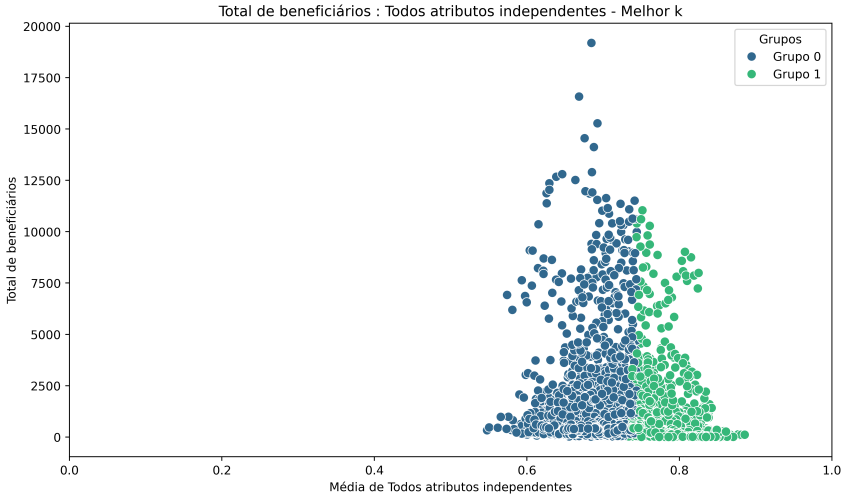
principais, obtiveram-se resultados com menor relevância. Destaca-se um coeficiente de determinação R^2 de 55,34% no grupo 1 da combinação **População beneficiária e Todos atributos independentes**, obtido mediante *XGBoost* com o valor do melhor k. Para os demais casos, os valores de R^2 permaneceram inferiores a 45% nas execuções realizadas com MLP e *XGBoost*.

Como conclusão, após a realização das execuções e análise dos resultados dos modelos, observa-se que as melhores combinações envolvem os atributos do PBF **Total de beneficiários** e **Valor total repassado**, onde a união com um agregado de atributos gera resultados superiores em comparação à combinação com apenas uma característica independente. Em relação aos modelos, obtiveram-se resultados melhores e com maior explicabilidade através do *XGBoost*, apresentando grande discrepância em relação às Redes Neurais (MLP) e à Regressão Linear. Quanto ao tempo de processamento, a Regressão Linear é o modelo mais veloz, porém o *XGBoost* apresenta um tempo extremamente rápido considerando a qualidade de seus resultados, contrastando com o tempo de execução da MLP, que registrou as maiores durações.

5.4 Análise aprofundada dos resultados

Nesta seção aprofundam-se as análises dos resultados apresentados na Subseção 5.3.2, visando validar e confirmar as divisões realizadas pelo método *K-means*, bem como

Figura 30 – Melhor divisão para o *K-means*, usando total de famílias beneficiárias do PBF e média de todos os valores do IDESE



Fonte: Elaborado pelo Autor

os elevados coeficientes de determinação (R^2) em algumas combinações retornadas pelos métodos computacionais de aprendizado supervisionado. Tal análise tem por objetivo verificar o impacto dos valores do IDESE sobre o PBF.

Inicia-se pela análise específica dos resultados da divisão denominada melhor k, valor determinado pelo coeficiente de silhueta aplicado ao *K-means*. Conforme observado na Subseção 5.3.2, os melhores valores de R^2 envolveram os dois principais atributos do PBF, quantidade de famílias beneficiárias e valor total repassado, combinados a todas as características do IDESE. Dessa forma, busca-se, nos cinco melhores resultados, extrair as informações internas que levaram o *K-means* a este agrupamento validado pelos métodos computacionais *XGBoost* e MLP. Por fim, geram-se *insights* relevantes para a compreensão do comportamento dos valores do PBF a partir de determinados índices do IDESE.

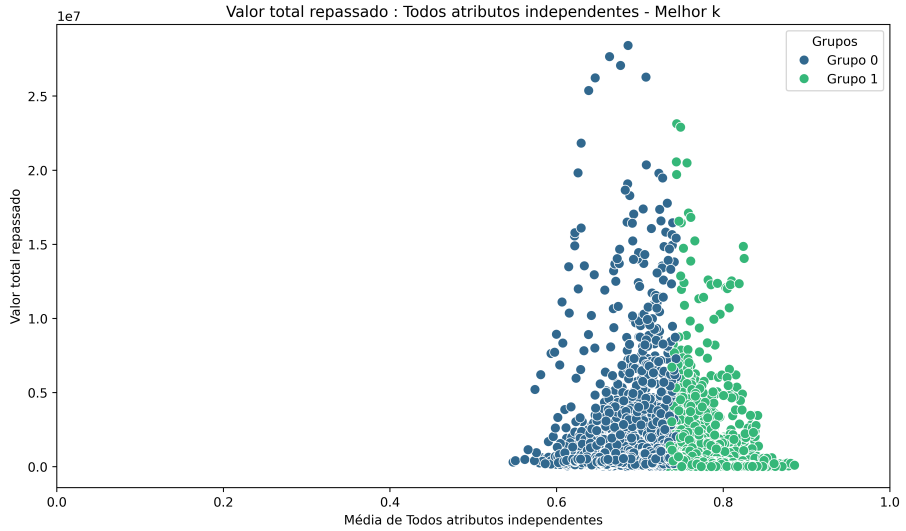
Para a geração de *insights* e melhor visualização dos dados referentes à relação entre o total de famílias beneficiárias do PBF e os valores do IDESE, se fez necessário o cálculo da média dos índices do IDESE que faziam parte de cada amostra nos grupos, consolidando-os em um único valor. Posteriormente, procedeu-se à comparação desses resultados com os valores médios do conjunto de dados apresentados no Apêndice A.

A Figura 30 ilustra a disposição final dos agrupamentos amostrais da junção entre **Total de famílias beneficiárias e a média de todos os índices do IDESE**. Com base na extração das médias dos valores de cada grupo, foi efetuada a comparação com os valores médios obtidos do conjunto completo do PBF e IDESE, dessa forma permitindo

as conclusões detalhadas que se seguem.

Inicialmente, observa-se que o *K-means* efetuou a divisão em dois grupos: o grupo 0, com 3208 amostras, e o grupo 1, com 4226 amostras. Os índices do IDESE no grupo 1 apresentaram média superior à do conjunto de dados global, com destaque para o IDESE geral, cuja média foi de $0,7693$, em comparação ao valor $0,7253$ da média global. Esses índices elevados projetaram uma quantidade de famílias beneficiárias na faixa de $428,2$, valor $48,1\%$ inferior à média do conjunto $(825,03)$. Em contrapartida, o grupo 0 apresentou comportamento inverso, com médias dos índices do IDESE inferiores à média global ($0,6668$ contra $0,7253$). Consequentemente, tais valores implicaram uma quantidade de famílias beneficiárias superior à média neste grupo, atingindo um valor de $1142,27$, o que representa um aumento de $38,45\%$.

Figura 31 – Melhor divisão para o *K-means*, usando valor total repassado do PBF e média de todos os valores do IDESE



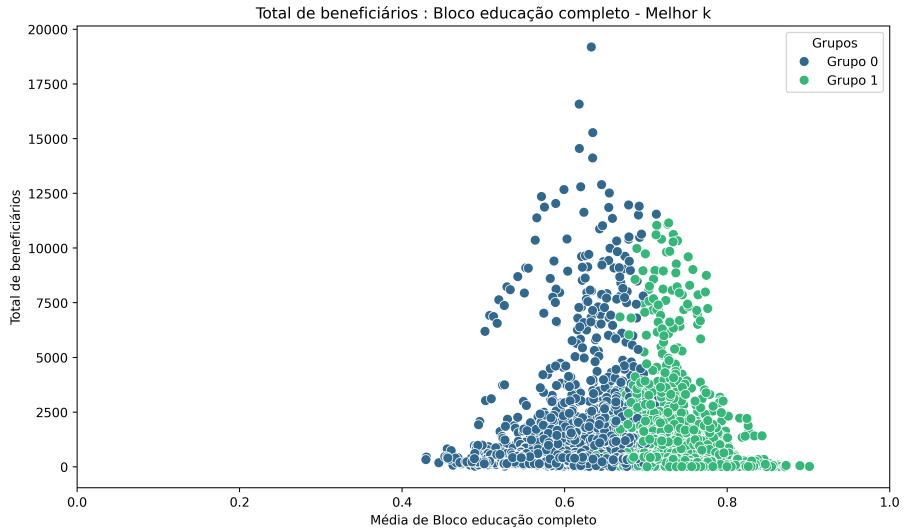
Fonte: Elaborado pelo Autor

Quanto à divisão resultante da combinação entre o **valor total repassado e todas as características do IDESE**, chega-se a conclusões análogas às apresentadas anteriormente: índices reduzidos do IDESE tendem a demandar maiores repasses, ao passo que valores acima da média estão associados a repasses menores. Tais constatações podem ser observadas na Figura 31, que ilustra os agrupamentos realizados pelo *K-means*.

Adicionalmente, examinam-se os índices do IDESE que compõem o bloco educação. Estes também apresentaram valores de R^2 elevados quando combinados com o total de famílias beneficiárias e o valor repassado do PBF. A relação dos agrupamentos gerados para os beneficiários é ilustrada na Figura 32.

Em suma, os mesmos *insights* descritos anteriormente foram observados nesta combinação: índices do IDESE abaixo da média relacionam-se a um maior número de beneficiários. O grupo 0, composto por 2495 amostras, obteve uma média de $1124,55$ beneficiários ($36,3\%$ acima da média global). Já o grupo 1, constituído por 4939 amostras, apresentou média de $540,23$ famílias beneficiárias ($34,52\%$ abaixo da média).

Figura 32 – Melhor divisão para o *K-means*, usando total de famílias beneficiárias do PBF e média dos valores do bloco educação do IDESE



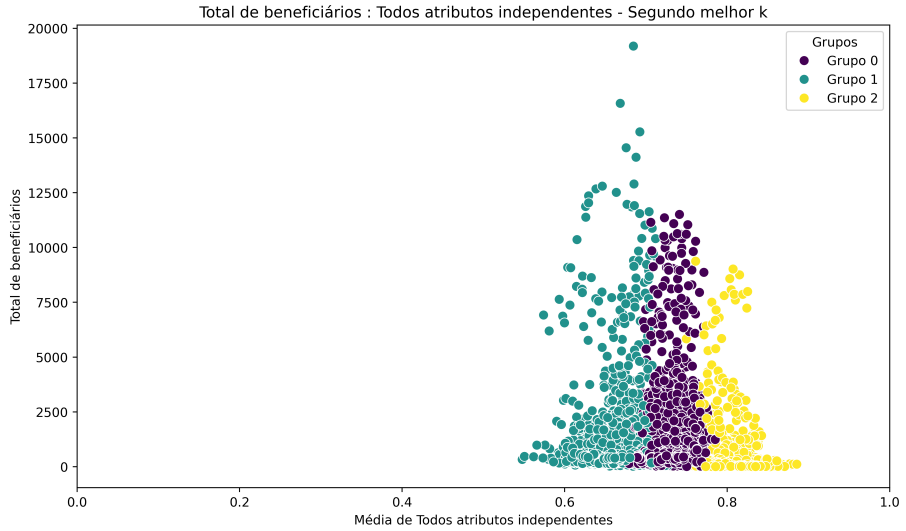
Fonte: Elaborado pelo Autor

Prossegue-se a análise considerando a divisão dos dados em mais grupos pelo *K-means*, baseada no coeficiente de silhueta que definiu o segundo melhor número de k. Observou-se, nesta configuração, a recorrência de valores elevados de R^2 nas variáveis dependentes do PBF, total de beneficiários e valor repassado, combinadas com todos os atributos do IDESE e com aqueles agregados no bloco educação.

Conforme ilustrado na Figura 33, o segundo melhor k para a combinação entre o total de famílias beneficiárias e todos os atributos do IDESE gerou três grupos com características distintas. O grupo 2 apresenta índices médios do IDESE superiores à média global; comparativamente, o valor médio de $0,7874$ para o IDESE geral resultou em uma média de $304,39$ beneficiários ($63,01\%$ inferior à média).

O grupo 0, composto por 3073 amostras, situa-se próximo à tendência central, possuindo valores condizentes com as médias dos atributos dependentes e independentes. Por fim, o grupo 1, com 1613 amostras, exibe valores de IDESE significativamente inferiores à média, com o IDESE geral em $0,6365$. Tal cenário acarreta uma média de $1259,75$ beneficiários nas amostras deste grupo ($52,69\%$ superior à média).

Figura 33 – Segunda melhor divisão para o *K-means*, usando total de famílias beneficiárias do PBF e média de todos os valores do bloco educação do IDESE



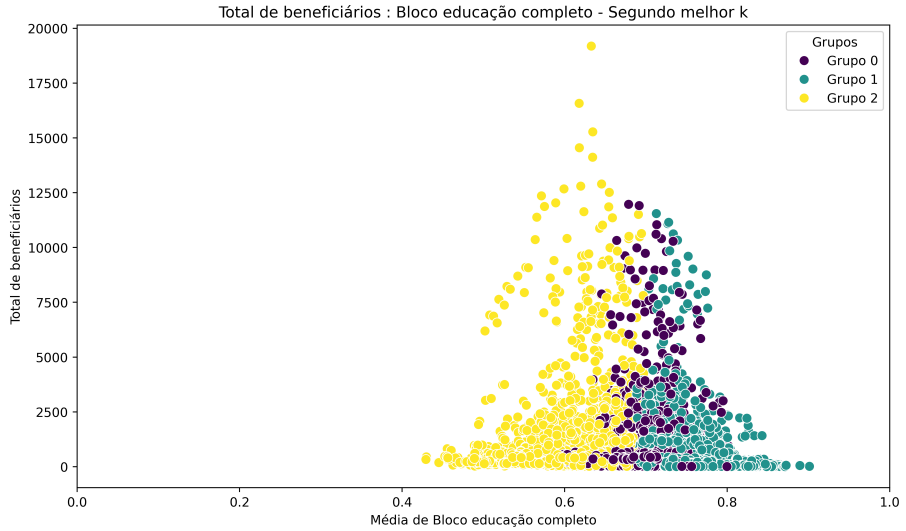
Fonte: Elaborado pelo Autor

Ao analisar as amostras dos três grupos resultantes da combinação entre a quantidade de famílias beneficiárias e os valores do bloco educação do IDESE, conforme ilustrado na Figura 34, observa-se uma proximidade entre os grupos 0 e 1. No entanto, o grupo 1, composto por 2724 amostras, apresenta índices acima da média, resultando em uma média de $477,83$ beneficiários ($42,08\%$ inferior à média global).

O grupo 0, por sua vez, situa-se levemente acima da média nos indicadores, apresentando um total médio de $641,38$ beneficiários ($22,26\%$ abaixo da média). Por fim, o grupo 2, com 2026 amostras, exibe valores de educação extremamente baixos, com destaque para o índice de pré-escola, que obteve média de $0,5830$, em contrapartida à média global de $0,8317$. Tais valores resultaram em uma média de $1209,72$ beneficiários ($46,62\%$ acima da média).

Os resultados obtidos e detalhados nesta seção, fundamentados na validação dos agrupamentos gerados pelo algoritmo *K-means* e na análise dos coeficientes de determinação (R^2), fornecem uma base empírica sólida para o entendimento da relação entre as políticas públicas e os indicadores socioeconômicos. A identificação de padrões distintos, nos quais índices do IDESE inferiores demonstram forte correlação com uma maior demanda e dependência do PBF, não apenas valida a metodologia computacional empregada, mas também destaca a capacidade dos modelos de aprendizado supervisionado em capturar nuances regionais. Portanto, este conjunto de evidências quantitativas atua como um subsídio essencial para as discussões que seguem,

Figura 34 – Segunda melhor divisão do *K-means* para total de famílias beneficiárias do PBF e todos os valores do bloco educação IDESE no RS



Fonte: Elaborado pelo Autor

estruturando as interpretações finais e as conclusões que serão elaboradas no próximo capítulo.

6 CONSIDERAÇÕES FINAIS

Este capítulo apresenta de forma reduzida e concisa os tópicos abordados, com o objetivo principal de integrar todos os capítulos e suas seções que foram formados a partir de um conjunto de perguntas e respostas. Objetiva-se com este capítulo demonstrar de que maneira as questões de pesquisa contribuíram para a progressão deste trabalho, suas limitações e seu potencial para subsidiar estudos futuros que abordem os temas em questão.

Este trabalho teve como propósito responder à seguinte questão: “De que forma a ciência de dados pode contribuir para analisar a relação entre os dados históricos do PBF e as variações do IDESE ao longo dos anos nos municípios e regiões do estado do Rio Grande do Sul?”. Essa questão foi abordada por meio da análise, utilizando técnicas de Ciência de Dados, que aprofundou a compreensão e demonstrou o impacto direto das variações nos valores calculados pelo IDESE, sobre os números históricos do PBF nos municípios e regiões do gaúchas, com tais variações foram analisadas no período de 2007 a 2021.

A questão de pesquisa, e o objetivo de respondê-la, motivaram a busca por trabalhos anteriores e correlatos, que pudessem fornecer embasamento técnico e teórico às abordagens metodológicas aqui empregadas. Com foco na interdisciplinaridade da ciência de dados, foi possível obter respostas para as principais demandas deste trabalho.

Nas seções subsequentes serão discutidos aspectos específicos, onde na Seção 6.1 são apresentados os objetivos e as respostas elaboradas neste trabalho. A Seção 6.2 expõe as limitações identificadas no estudo e como estas podem ser contornadas. Por fim, a Seção 6.3 indica as principais diretrizes para trabalhos futuros nos temas e no contexto desta pesquisa.

6.1 Síntese das conclusões

Nesta seção, são apresentadas as conclusões fundamentadas no objetivo geral e em seus desdobramentos nos objetivos específicos, citados na Seção 1.2. Tais objetivos nortearam o desenvolvimento deste trabalho, sendo fundamentais para a definição da metodologia, a execução da pesquisa e a interpretação dos resultados.

Um aspecto intrínseco relevante ao processo foi a coleta de dados governamentais, disponibilizados em diversas fontes. Contudo, o pré-processamento dessas informações

exigiu tratamento rigoroso para a consolidação de uma base de dados unificada apta às análises. Adicionalmente, a implementação computacional demandou ajustes contínuos, configurando-se como um processo iterativo adaptado às necessidades surgidas ao longo da pesquisa.

Evidenciou-se a viabilidade da aplicação de técnicas de ciência de dados na análise de políticas públicas de transferência de renda. A interdisciplinaridade da área possibilitou a coleta e o processamento dos dados, a aplicação de modelos estatísticos e computacionais, bem como a interpretação dos resultados obtidos. Corroborando a literatura, este trabalho reforça a importância da utilização de métodos computacionais robustos por parte dos governos para mensurar e avaliar os impactos sociais de suas políticas.

Conclui-se que os métodos computacionais selecionados foram eficazes. O algoritmo *K-means* apresentou bom desempenho na segmentação das amostras, revelando padrões relevantes sobre a interação entre o IDESE e o PBF. Além disso, métodos mais recentes, como o *XGBoost*, demonstraram superioridade na capacidade de generalização e modelagem dos dados segmentados pelo *K-means*, superando a Regressão Linear e as Redes Neurais MLP em todas as combinações de variáveis testadas.

Em uma análise aprofundada, constatou-se que os valores do IDESE permitem segmentar o total de famílias beneficiárias em dois grupos com características opostas. Para o Grupo 0, o *XGBoost* obteve um coeficiente de determinação (R^2) de $81,39\%$ e um Erro Quadrático Médio (MSE) de $0,0019$. A análise indica que as amostras desse grupo possuem índices do IDESE elevados, correspondendo a uma quantidade média de famílias beneficiárias $48,1\%$ inferior à do conjunto global. Em contrapartida, o outro grupo apresentou índices do IDESE abaixo da média, associados a um aumento de $38,45\%$ no número de beneficiários.

No que tange aos índices que compõem o bloco temático de educação do IDESE, o *XGBoost* registrou um R^2 de $72,91\%$ e um MSE de $0,0030$. Observou-se que índices reduzidos estão correlacionados a um aumento de $36,3\%$ no número de beneficiários, enquanto índices acima da média resultam em uma redução de $34,52\%$.

Ao examinar os resultados obtidos das combinações do IDESE referentes aos índices que compõem cada bloco temático, destaca-se a relevância da relação entre o bloco de Educação e os valores do PBF. Tal constatação é significativa, uma vez que este bloco é constituído por 6 índices, em contrapartida ao bloco de Saúde, que possui 7. Embora a maior quantidade de variáveis no bloco de Saúde pudesse sugerir um

potencial superior para o estabelecimento de correlações, tal hipótese não se confirmou nos resultados observados.

Em síntese, as conclusões destacam a eficácia das abordagens de ciência de dados na análise da relação entre políticas públicas e indicadores socioeconômicos, reforçando a relevância de métodos computacionais para a interpretação de fenômenos complexos e o subsídio a decisões baseadas em evidências.

6.2 Limitações da pesquisa

Este trabalho foi realizado com grande empenho e trouxe resultados positivos que podem servir de base para trabalhos futuros. No entanto, alguns aspectos importantes devem ser levantados para que haja transparência sobre possíveis problemas que possam representar impedimentos para futuras análises.

O trabalho fez uso de dados que não estão em sua completude, uma vez que os índices do IDESE sofreram grandes modificações em seu cálculo, devido a ajustes necessários para acompanhar os avanços nas áreas avaliadas. Isso limitou a continuidade dos dados para alcançar uma base com um período mais extenso, tornando as análises menos abrangentes. Acrescenta-se a isso as modificações no Programa Bolsa Família, em que mudanças bruscas tornaram suas bases de dados desagregadas e com alterações significativas, dificultando a comparação com anos anteriores e, por esse motivo, impossibilitando sua incorporação neste trabalho.

Em relação ao uso dos métodos computacionais, pode-se destacar que os resultados podem variar conforme a quantidade de amostras e dados provenientes de outras fontes. Ademais, as configurações de alguns métodos podem ser alteradas, uma vez que novas tecnologias podem incorporar outros valores ou parâmetros adicionais durante a criação do modelo final correspondente a cada método.

É importante ressaltar que falhas na pesquisa bibliográfica podem ter ocorrido, devido ao alto número de temas abordados neste trabalho. O refinamento e a especificidade desejados podem ter excluído trabalhos ou repositórios relevantes para esta pesquisa.

O aprofundamento nos resultados poderia ter sido mais elaborado caso houvesse tempo hábil, e a exploração de outras métricas de análise poderia ter sido aplicada para complementar os resultados. Dessa forma, buscou-se neste trabalho refinar e focar em métricas básicas já consolidadas para o tipo de problema investigado.

6.3 Sugestões para trabalhos futuros

Expostos os pontos positivos que conferiram a relevância necessária a este estudo, aliados às limitações encontradas durante o processo de sua realização, faz-se necessária a elaboração de recomendações para futuras pesquisas que desejem aproveitar e aprofundar os conceitos aqui explorados.

Indica-se a utilização de bases de dados mais completas, como a integração de índices socioeconômicos distintos, voltados a temas não abordados nos cálculos do IDESE. O uso de bases diversificadas, incluindo índices regionais, nacionais e internacionais, pode auxiliar na compreensão de como a sociedade influencia a modificação de políticas públicas.

Recomenda-se a incorporação de políticas públicas de outros tipos ou de maior abrangência, que disponham de dados bem discriminados e abertos, assim como o PBF. Ademais, caso seja viável, agregar todas as políticas públicas de transferência de renda instituídas nos estados e no Brasil, a fim de avaliar o quanto essas políticas são afetadas por mudanças nos índices ou indicadores socioeconômicos.

Sugere-se a utilização de métodos computacionais específicos que possam conferir maior interpretabilidade aos dados explorados, permitindo gerar resultados mais robustos que os aqui apresentados. Pretende-se indicar o emprego de mais variáveis interpretativas na avaliação dos resultados gerados pelos modelos, bem como a possível modificação de parâmetros, com o intuito de superar as limitações identificadas.

Caso a ampliação do escopo não seja uma opção, indica-se o aprofundamento temático, utilizando apenas um ou ambos os temas do PBF e IDESE, para explorar uma combinação específica de informações e uma técnica particular de ciência de dados. Essas sugestões visam auxiliar na criação de trabalhos futuros com significativo valor científico para governos e para a sociedade que depende desses temas.

REFERÊNCIAS

ABBASI, A.; CHIANG, R. H.; XU, J. J. Data science for social good. **arXiv preprint arXiv:2311.14683**, 2023.

AGUIAR, I. W. O. d. **Fatores associados à insegurança alimentar domiciliar em uma coorte de mulheres residentes em áreas vulneráveis a arboviroses de Fortaleza-CE**. 108 p. Dissertação (Dissertação (Mestrado em Saúde Pública)) — Universidade Federal do Ceará, Fortaleza, CE, Brasil, 2021. Orientador: Bernard Carl Kendall; Coorientadora: Ligia Regina Franco Sansigolo Kerr. Disponível em: http://www.repositorio.ufc.br/handle/riufc/56271.

BANDEIRA, P. M. P. **História de Pinto Bandeira**. 2025. Disponível em: <<https://www.pintobandeira.rs.gov.br/secao.php?id=2>>. Acessado em: 12 nov. 2025.

BENTÉJAC, C.; CSÖRGŐ, A.; MARTÍNEZ-MUÑOZ, G. A comparative analysis of gradient boosting algorithms. **Artificial Intelligence Review**, Springer, v. 54, n. 3, p. 1937–1967, 2021.

BERNARDINI, R. C. et al. **Aspectos metodológicos do Índice de Desenvolvimento Socioeconômico (Idese)**. Porto Alegre: FEE, 2017.

BONILHA, J. M. M. **Transferência de Renda Condicionada e o Mercado de Trabalho Formal: Uma Análise dos Municípios Brasileiros**. [S.l.], 2024.

BRASIL. **Medida Provisória no 132, de 24 de outubro de 2003**: Dispõe sobre o programa bolsa família e dá outras providências. [S.l.]: Presidência da República, 2003. Disponível em: <https://www.planalto.gov.br/ccivil_03/mpv/antigas_2003/132.htm>. Acesso em: 08 maí. 2025.

BRASIL. **Lei nº 10.836, de 9 de janeiro de 2004**: Cria o programa bolsa família e dá outras providências. [S.l.]: Presidência da República, 2004. Disponível em: <<https://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/l10.836.htm>>. Acesso em: 10 jun 2025.

BRASIL. **Lei nº 12.527, de 18 de novembro de 2011**: Regula o acesso a informações previsto no inciso xxxiii do art. 5º, no inciso ii do § 3º do art. 37 e no § 2º do art. 216 da constituição federal. [S.l.]: Presidência da República, 2011. Disponível em: <<https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>>. Acesso em: 25 out 2025.

BRASIL. **Lei nº 14.284, de 29 de dezembro de 2021**: Altera a lei nº 10.836, de 9 de janeiro de 2004, que cria o programa bolsa família, para estabelecer o auxílio brasil; e dá outras providências. [S.l.]: Presidência da República, 2021. Disponível em: <<https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14284.htm>>. Acesso em: 10 jun 2025.

BRASIL. **Lei nº 14.601, de 12 de junho de 2023**: Restabelece o programa bolsa família e o cadastro Único para programas sociais do governo federal; revoga a lei nº 14.284, de 29 de dezembro de 2021; e dá outras providências. [S.l.]: Presidência da República, 2023. Disponível em: <<https://www.planalto.gov.br/ccivil_03/_Ato2023-2026/2023/Lei/L14601.htm>>. Acesso em: 15 jul 2025.

BRUCE, A.; BRUCE, P. **Estatística prática para cientistas de dados**. [S.l.]: Alta Books, 2019.

BRUSACA, L. L. M. D. S. A efetividade das políticas públicas de inovação no estado do maranhão. Universidade Federal do Tocantins, 2025.

CAMPELLO, T.; NERI, M. C. Programa bolsa família: uma década de inclusão e cidadania. Instituto de Pesquisa Econômica Aplicada (Ipea), 2013.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 785–794.

COLAB, G. **Google Colab Home**. 2025. Disponível em: <<https://colab.research.google.com/>>. Acessado em: 01 fev. 2025.

COMMUNITY, R. **Requests: HTTP for Humans**. 2025. Disponível em: <<https://requests.readthedocs.io/en/latest/>>. Acessado em: 12 nov. 2025.

CONWAY, D. **The Data Science Venn Diagram**. 2013. Disponível em: <<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>>. Acesso em: 15 out. 2025.

DAMIÃO, J. d. J. et al. Condicionalidades de saúde no programa bolsa família e a vigilância alimentar e nutricional: narrativas de profissionais da atenção primária à saúde. **Cadernos de Saúde Pública**, SciELO Public Health, v. 37, p. e00249120, 2021.

DEE-RS. **DATA.RS: Portal de Dados Abertos do Rio Grande do Sul. Departamento de Economia e Estatística - RS.** 2025. Disponível em: <<https://data.rs.gov.br/>>. Acessado em: 30 out. 2025.

DEE-RS, D. de Economia e E. **Índice de Desenvolvimento Socioeconômico (IDESE)**. 2024. Disponível em: <https://dee.rs.gov.br/idese>. Acessado em: 06 de mai. de 2025.

DEE-RS, D. de Economia e Estatística do R. **Apresentação do Índice de Desenvolvimento Socioeconômico (IDESE)**. 2020. Disponível em: <https://dee.rs.gov.br/upload/arquivos/202012/23120114-idese-apresentacao-dee-final.pdf>. Acessado em: 10 de mai. de 2025.

DEE-RS-DEEDADOS. **DEEDADOS: Base de Dados da DEE/FEE. Departamento de Economia e Estatística - RS.** 2025. Disponível em: <<http://feedados.fee.tche.br/feedados/>>. Acessado em: 01 mar. 2025.

DEE-RS-ESTRUTURA. **Estrutura hierárquica do Índice de Desenvolvimento Socioeconômico (IDESE). Departamento de Economia e Estatística - RS.** 2025. Disponível em: <<https://idesevis.dee.rs.gov.br/>>. Acessado em: 01 mar. 2025.

DEE-RS-POPVIS. **POPVIS: Plataforma de Visualização de Dados Populacionais. Departamento de Economia e Estatística - RS.** 2025. Disponível em: <<https://popvis.dee.rs.gov.br/>>. Acessado em: 01 mar. 2025.

DIAS, G. R. Análise comparativa do índice de desenvolvimento socioeconômico nos municípios de abrangência da unipampa. Universidade Federal do Pampa, 2021.

DOCUMENTATION, X. Xgboost documentation. **URL: https://xgboost. readthedocs. io/en/stable/index. html (date of access: 25.05. 2024)**, 2023.

ESCOVEDO, T.; KOSHIYAMA, A. **Introdução a Data Science: Algoritmos de Machine Learning e métodos de análise**. [S.l.]: Casa do Código, 2020.

FEE ACCURSO, J. d. S. Fundação de Economia e E.; HEUSER, S. E. **Índice de Desenvolvimento Socioeconômico do RS (IDESE) — 1991-00**. [S.l.]: FEE, 2003. 31 p. (Documentos FEE, 58). 31 p.: tab. ISSN 1676-1375. ISBN 85-7173-024-5.

FEITOSA JúNIOR, D. J. S. **Políticas públicas e fatores socioeconômicos na dinâmica da hanseníase no Estado do Pará**. Tese (Doutorado) — Universidade de São Paulo, 2023.

FOUNDATION, P. S. **json — JSON encoder and decoder**. 2025. Disponível em: <<https://docs.python.org/3/library/json.html>>. Acessado em: 12 nov. 2025.

FOUNDATION, P. S. **Python Language Reference**. 2025. Disponível em: <<https://www.python.org/>>. Acessado em: 10 nov. 2025.

FREITAS, E. C. d.; PRODANOV, C. C. Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico. **Novo Hamburgo: Feevale**, 2013.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, JSTOR, p. 1189–1232, 2001.

FUSHIKI, T. Estimation of prediction error by using k-fold cross-validation. **Statistics and Computing**, Springer, v. 21, p. 137–146, 2011.

GÉRON, A. **Mãos à obra: aprendizado de máquina com Scikit-Learn & TensorFlow**. [S.l.]: Alta Books, 2019.

GHANI, R. Data science for social good and public policy: examples, opportunities, and challenges. In: **The 41st international ACM SIGIR conference on research & development in information retrieval**. [S.l.: s.n.], 2018. p. 3–3.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining: um guia prático**. [S.l.]: Gulf Professional Publishing, 2005.

GOMES, N. I. G. et al. Análise do consumo alimentar e a situação de insegurança alimentar de mulheres adultas usuárias do sus de joão pessoa, paraíba. Universidade Federal da Paraíba, 2025.

GRUS, J. **Data science do zero: noções fundamentais com Python**. [S.l.]: Alta Books, 2021.

HARRISON, M. **Machine Learning–Guia de referência rápida: trabalhando com dados estruturados em Python**. [S.l.]: Novatec Editora, 2019.

HOSSIN, M. A. et al. Big data-driven public policy decisions: Transformation toward smart governance. **Sage Open**, SAGE Publications Sage CA: Los Angeles, CA, v. 13, n. 4, p. 21582440231215123, 2023.

IBGE. **43 Regiões Geográficas - Rio Grande do Sul**. 2017. Disponível em: <<https://geoftp.ibge.gov.br/organizacao_do_territorio/divisao_regional/divisao_regional_do_brasil/divisao_regional_do_brasil_em_regioes_geograficas_2017/mapas/43_regioes_geograficas_rio_grande_do_sul.pdf>>. Acessado em: 12 nov. 2025.

IBGE. **API de Localidades do IBGE. Instituto Brasileiro de Geografia e Estatística. Brasil.** 2024. Disponível em: <<https://servicodados.ibge.gov.br/api/docs/localidades>>. Acessado em: 01 de mar. 2025.

IZBICKI, R.; SANTOS, T. M. dos. **Aprendizado de máquina: uma abordagem estatística**. [S.l.]: Rafael Izbicki, 2020.

JANNUZZI, P. d. M. Indicadores sociais no brasil: conceitos, fontes de dados e aplicações. In: **Indicadores sociais no Brasil: conceitos, fontes de dados e aplicações**. [S.l.: s.n.], 2009. p. 141–141.

JANNUZZI, P. d. M. **Indicadores Socioeconômicos na Gestão Pública**. Florianópolis : Departamento de Ciências da Administração / UFSC;[Brasília] : CAPES : UAB, 2014. Acessado em: 07 de mai. de 2025. Disponível em: https://educapes.capes.gov.br/handle/capes/145395.

JANUARIO, V. D. F. P. **Perfil da pobreza nos estados brasileiros através do valor de Shapley: análise do período pré e pós pandemia**. 2024. Trabalho de Conclusão de Curso – Curso de Administração, Universidade Federal do Mato Grosso do Sul, Chapadão do Sul - MS, 2024. Disponível em: <https://repositorio.ufms.br/retrieve/8de935cb-6f69-4ecf-a99c-cdaff93b4c2b/19643.pdf>. Acesso em: 29 de Junho de 2025.

JUPYTER, P. **Project Jupyter: Open-source software, open standards, and services for interactive computing across dozens of programming languages**. 2025. Disponível em: <<https://jupyter.org/>>. Acessado em: 10 nov. 2024.

Kaggle. **Kaggle Machine Learning & Data Science Survey 2022**. 2022. Disponível em: <<https://www.kaggle.com/kaggle-survey-2022>>. Acesso em: 30 out. 2025.

KALINOWSKI, M. et al. **Engenharia de Software para Ciência de Dados: Um guia de boas práticas com ênfase na construção de sistemas de Machine Learning em Python**. [S.l.]: Casa do Código, 2023.

KANG, T. H. et al. O novo índice de desenvolvimento socioeconômico (novo idese): aspectos metodológicos. **Textos para discussão FEE**, n. 127, 2014.

KERAS. **Keras Página Inicial**. 2025. Disponível em: <<https://keras.io/>>. Acessado em: 10 abr. 2025.

KÜHN, D. D.; TONETTO, E. d. S. O programa bolsa família e os indicadores sociais de combate à pobreza no rio grande do sul: um olhar multidimensional. **Desenvolvimento em questão: revista do programa de pós-graduação em desenvolvimento. Ijuí. Ano 15, n. 39 (abr./jun. 2017), p.[86]-111**, 2017.

LIED, Ú. M. C. Análise comparativa entre os desempenhos dos municípios gaúchos no índice de desenvolvimento socioeconômico do estado–idese. Universidade Federal de Santa Maria, 2024.

LLOYD, S. Least squares quantization in pcm. **IEEE transactions on information theory**, IEEE, v. 28, n. 2, p. 129–137, 1982.

LOWI, T. J. American business, public policy, case-studies, and political theory. **World politics**, Cambridge University Press, v. 16, n. 4, p. 677–715, 1964.

MAIA, M.; NOGUTI, M. Y.; ARA, A. Predição da taxa de utilização do programa bolsa família na bahia: um estudo via máquinas aleatórias. **Bahia Análise & Dados**, v. 30, n. 2, p. 53–74, 2020.

MARCONI, M. d. A.; LAKATOS, E. M. **Metodologia científica**. [S.l.]: Atlas São Paulo, 2004. v. 4.

MARTINS, B. A.; RÜCKERT, F. Q. O programa bolsa família e a condicionalidade educacional: uma análise do desempenho escolar de estudantes em situação de pobreza. **Revista Brasileira de Educação**, ANPED, v. 27, 2022.

MASTRODI, J.; IFANGER, F. C. de A. Sobre o conceito de políticas públicas. **Revista de direito brasileira**, v. 24, n. 9, p. 03–16, 2019.

MATIPLOTLIB. **Matiplotlib Documentation**. 2025. Disponível em: <<https://matplotlib.org/stable/api/pyplot_summary.html>>. Acessado em: 10 abr. 2025.

MDS, M. d. D. S. e. C. F. **Revista Comemorativa: 20 anos do Bolsa Família**. [S.l.]: Ministério do Desenvolvimento Social e Combate à Fome, 2023. Disponível em: <<https://www.mds.gov.br/webarquivos/MDS/2_Acoes_e_Programas/Bolsa_Familia/Eventos/Revista_Comemoracao_20_anos_BF.pdf>>. Acessado em: 08 de mai. de 2025.

MELAZZO, E. S. Problematizando o conceito de políticas públicas: Desafios à análise e à prática do planejamento e da gestão. **Revista Tópos**, v. 4, n. 2, p. 9–32, 2010.

MORETTIN, P. A.; SINGER, J. d. M. ***Estatística e ciência de dados***. 1. ed.. ed. Rio de Janeiro: LTC, 2022.

MPOG, M. d. P. O. e. G.; SPI, S. d. P. I. E. **Indicadores: orientações básicas aplicadas à gestão pública**. 2012. Disponível em: <<http://bibliotecadigital.economia.gov.br/handle/777/46>>. Acessado em: 28 de Mai. de 2025.

NAYAK, P. Methodological developments in human development literature. **International Journal of Applied Management Research**, v. 2, n. 2, p. 1–22, 2015.

NUMPY. **NumPy Documentation**. 2022. Disponível em: <<https://numpy.org/doc/>>. Acessado em: 10 abr. 2025.

OLIVEIRA FILHO, J. S. P. d. O "big data" como insumo para a formulação de políticas públicas à erradicação da pobreza no estado do ceará: um estudo comparativo entre a realidade cearense e a experiência chinesa. 2025.

ÖZTORNACI, B.; ATA, B.; KARTAL, S. Analysing household food consumption in turkey using machine learning techniques. **Agris on-line Papers in Economics and Informatics**, v. 16, n. 2, 2024.

PAIVA, J. P. S. de et al. Time trend, social vulnerability, and identification of risk areas for tuberculosis in brazil: an ecological study. **Plos one**, Public Library of Science San Francisco, CA USA, v. 17, n. 1, p. e0247894, 2022.

PANDAS. **Pandas Documentation**. 2024. Disponível em: <<https://pandas.pydata.org/docs/>>. Acessado em: 01 mar. 2025.

PARDITA, D. P. Y. et al. Understanding poverty dynamics in indonesia: The role economic growth, income distribution, and human development. **Jurnal Riset Ilmu Ekonomi**, v. 4, n. 2, p. 156–167, 2024.

PIERSON, L. **Data science para leigos**. [S.l.]: Alta Books Editora, 2019.

PORTELLA, L. N. P. **Repositório do Trabalho de Conclusão de Curso**. 2025. Disponível em: <<https://github.com/loportella/pbf-idese-tcc-ii>>. Acessado em: 20 nov. 2025.

QUEIROGA, E. M. et al. Early prediction of at-risk students in secondary education: A countrywide k-12 learning analytics initiative in uruguay. **Information**, MDPI, v. 13, n. 9, p. 401, 2022.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, Springer, v. 1, n. 1, p. 81–106, 1986.

RAEDER, S. T. O. Ciclo de políticas: uma abordagem integradora dos modelos para análise de políticas públicas. **Perspectivas em Políticas Públicas**, v. 7, n. 13, p. 121–146, 2014.

RAMOS, M. P.; LIMA, L. L. Avaliação de impacto de políticas públicas: desafios e perspectivas a partir do programa bolsa família. **Pesquisa em desenvolvimento rural: aportes teóricos e proposições metodológicas. Porto Alegre: Editora da UFRGS, 2014. p. 77-91**, 2014.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, American Psychological Association, v. 65, n. 6, p. 386, 1958.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **nature**, Nature Publishing Group UK London, v. 323, n. 6088, p. 533–536, 1986.

SAAB, F. et al. Políticas públicas e desenvolvimento humano: fatores que impactam o idh em municípios brasileiros. **RACE-Revista de Administração, Contabilidade e Economia**, v. 20, n. 2, p. 209–230, 2021.

SAGICAD, S. de Avaliação Gestão da Informação e Cadastro Único. **VIS DATA 3 beta - Explorador de Dados: plataforma de visualização de dados governamentais. Secretaria de Avaliação, Gestão da Informação e Cadastro Único. Ministério do Desenvolvimento e Assistência Social Família e Combate à Fome (MDS).** 2025. Disponível em: <<https://aplicacoes.cidadania.gov.br/vis/data3/data-explorer.php>>. Acessado em: 01 mar. 2025.

SANTOS, G. C. A. **Impactos da Criminalidade sobre evasão escolar: um estudo na cidade do Rio de Janeiro**. 2022. Trabalho de Conclusão de Curso – Curso de Economia, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro - RJ. Disponível em: <https://www.econ.puc-rio.br/uploads/adm/trabalhos/files/Graciele_Claudine_Aguiar_Santos_Mono_22.1.pdf>. Acesso em: 29 de Junho de 2025.

SANTOS, G. T. d. **Análise de dados amostrais longitudinais da pesquisa de avaliação de impacto do Bolsa Família**. 2021. Disponível em: https://repositorio.ufjf.br/jspui/handle/ufjf/13434.

SANTOS, S. M. d. **Data science aplicado a dados abertos do Governo Federal: estudos de caso sobre a economia dos municípios brasileiros**. 71 p. Dissertação (Dissertação (Mestrado em Engenharia Elétrica)) — Universidade Federal do Pará, Belém, PA, Brasil, 2020. Orientador: Marcelino Silva da Silva. Disponível em: https://repositorio.ufpa.br/jspui/handle/2011/17216.

SCHOLAR, G. **Google Scholar About**. 2025. Disponível em: <<https://scholar.google.com/intl/en/scholar/about.html>>. Acessado em: 01 de jun. de 2025.

SCIKIT-LEARN. **Sklearn.cluster.KMeans**. 2025. Disponível em: <<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>>. Acessado em: 10 abr. 2025.

SCIKIT-LEARNING. **Scikit-learning Documentation**. 2025. Disponível em: <<https://scikit-learn.org/stable/l>>. Acessado em: 10 abr. 2025.

SEABORN. **Seaborn Documentation**. 2025. Disponível em: <<https://seaborn.pydata.org/>>. Acessado em: 10 de Mai. 2025.

SECCHI, L. **Políticas públicas: conceitos, esquemas de análise, casos práticos**. [S.l.]: Cengage Learning, 2014.

SILVA, A. C. d. **Previsão da pobreza do Estado do Ceará**. 77 p. Dissertação (Dissertação (Mestrado em Economia Rural)) — Universidade Federal do Ceará, Fortaleza, CE, Brasil, 2022. Orientador: Jair Andrade de Araujo; Coorientador: Guaracyane Lima Campêlo. Disponível em: http://www.repositorio.ufc.br/handle/riufc/70919.

SILVA, L. F. M. d. Dissertação (Mestrado Acadêmico em Administração), **Análise preditiva baseada em inteligência artificial: um caminho para a transformação do modelo de vigilância das doenças crônicas não transmissíveis**. Santana do Livramento, RS, Brasil: [s.n.], 2023. Orientador: Rafael Camargo Ferraz. Disponível em: https://repositorio.unipampa.edu.br/jspui/handle/riu/8666.

SIMONATO, T. C. **Impactos na economia brasileira do Auxílio Emergencial durante a pandemia do Covid-19: efeitos regionais, setoriais, familiares e no mercado de trabalho**. Tese (Tese (Doutorado em Economia)) — Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil, 2023. Orientador: Edson Paulo Domingues; Coorientadora: Aline Souza Magalhães. Disponível em: http://hdl.handle.net/1843/60068.

SOUZA, C. Políticas públicas: uma revisão da literatura. **Sociologias**, SciELO Brasil, p. 20–45, 2006.

STANTON, E. A. The human development index: A history. 2007.

TEAM, G. D. **GeoPandas Documentation**. 2025. Disponível em: <<https://geopandas.org/en/stable/docs.html>>. Acessado em: 12 nov. 2025.

TENSORFLOW. **Tensorflow Documentation**. 2024. Disponível em: <<https://www.tensorflow.org/api_docs/python/tf/all_symbols>>. Acessado em: 10 abr. 2025.

VIANA, I. A. V. O.; KAWAUCHI, M. O.; BARBOSA, T. V. O. Bolsa família 15 anos (2003-2018). Escola Nacional de Administração Pública (Enap), 2018.

WANG, F. et al. An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity. In: SPRINGER. **Machine Learning and Data Mining in Pattern Recognition: 13th International Conference, MLDM 2017, New York, NY, USA, July 15-20, 2017, Proceedings 13**. [S.l.], 2017. p. 291–305.

WAZLAWICK, R. S. **Metodologia de pesquisa para ciência da computação**. [S.l.]: Elsevier Rio de Janeiro, 2009. v. 2.

WITTEN, I. H. et al. Data mining: Practical machine learning tools and techniques. In: . [S.l.]: Elsevier, 2025.

WORTHEN, B. R. et al. Avaliação de programas: concepções e práticas. In: **Avaliação de programas: concepções e práticas**. [S.l.: s.n.], 2004. p. 730–730.

ZHANG, A. et al. **Dive into Deep Learning**. [S.l.]: Cambridge University Press, 2023. <https://D2L.ai>.

APÊNDICE A – MÉTRICAS PBF E IDESE

Este apêndice tem como finalidade apresentar as métricas estatísticas fundamentais identificadas para os conjuntos de dados do PBF e do IDESE. São expostas informações detalhadas sobre cada uma das variáveis (atributos) pertencentes a esses conjuntos, incluindo tabelas com métricas para os dados agregados e anuais referentes ao período de 2007 a 2021.

Os dados do PBF e do IDESE foram filtrados para remover registros com valores nulos ou zerados, classificados como extremos. Tal escolha visou mitigar possíveis inconsistências e evitar o uso de informações incorretas nas análises. Os códigos de programação desenvolvidos para a extração e o tratamento desses dados encontram-se disponíveis no repositório remoto deste trabalho Portella (2025).

A.1 Dados estatísticos do PBF

- **Total de Beneficiários**

Total	Média Anual	Média	Desvio	Q0	Mediana	Q1	Mín	Máx
6.145.623	409.708,2	825,03	2.507,56	99,0	254,0	668,0	1,0	54.272

Mínimos:	
Cidade - Ano	**Beneficiários**
Boa Vista do Sul 2017, Monte Belo do Sul 2008, Monte Belo do Sul 2010, Monte Belo do Sul 2011, São Vendelino 2019	1

Máximos:		
Cidade	**Ano**	**Beneficiários**
Porto Alegre	2015	54.272

Métricas Anuais:								
Ano	**Total**	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	410.540	827,70	2.080,68	135,75	269,5	703,25	2	32.534
2008	367.631	741,19	1.808,57	113,00	238,5	642,00	1	27.479
2009	462.966	933,40	2.356,44	129,75	290,5	800,00	2	34.682
2010	453.761	914,84	2.290,54	125,75	298,0	776,75	1	36.130
2011	451.438	910,16	2.419,22	118,50	298,5	750,50	1	40.912
2012	463.519	934,51	2.596,76	125,75	316,0	758,25	4	46.223
2013	455.421	916,34	2.727,38	119,00	303,0	736,00	3	50.196
2014	434.715	874,68	2.751,06	105,00	280,0	713,00	3	52.060
2015	427.939	861,04	2.843,63	95,00	273,0	664,00	4	54.272
2016	379.234	763,05	2.632,11	84,00	228,0	587,00	4	51.039
2017	364.325	733,05	2.457,02	80,00	229,0	578,00	1	46.994
2018	367.805	740,05	2.622,69	79,00	214,0	575,00	2	50.617
2019	325.960	655,86	2.310,14	65,00	186,0	518,00	1	44.406
2020	378.103	760,77	2.741,43	73,00	211,0	566,00	2	52.597
2021	402.266	809,39	2.754,43	82,00	220,0	600,00	3	51.478

- **Valor Total Repassado (R$)**

Total	Média Anual	Média	Desvio	Q0	Mediana	Q1	Mín	Máx
8,75E+09	5,83E+08	1,17E+06	4,47E+06	121.220	340.379	871.155	620	1,27E+08

Mínimos:		
Cidade	**Ano**	**Valor (R$)**
Pinto Bandeira	2013	620,00

Máximos:								
Cidade			Ano			Valor (R$)		
Porto Alegre			2016			126.671.933		
Métricas Anuais:								
Ano	**Total**	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	320,2 mi	645.499	1,72E+06	99.600	211.452	513.196	1.280	28,1 mi
2008	357,8 mi	721.412	1,88E+06	104.678	236.478	596.418	1.134	29,9 mi
2009	425,4 mi	857.731	2,26E+06	117.230	270.371	690.718	752	36,1 mi
2010	484,2 mi	976.164	2,49E+06	132.105	313.992	836.223	660	39,2 mi
2011	571,8 mi	1,15E+06	3,10E+06	144.724	360.720	959.138	708	51,9 mi
2012	680,5 mi	1,37E+06	4,06E+06	167.702	439.751	1,11E+06	4.408	74,1 mi
2013	756,8 mi	1,52E+06	5,02E+06	184.132	486.852	1,25E+06	620	97,1 mi
2014	793,7 mi	1,60E+06	5,78E+06	175.330	495.185	1,31E+06	4.759	115,1 mi
2015	774,7 mi	1,56E+06	5,85E+06	168.955	480.532	1,21E+06	6.342	117,1 mi
2016	747,1 mi	1,50E+06	6,20E+06	164.498	434.417	1,12E+06	6.725	126,7 mi
2017	695,1 mi	1,40E+06	5,76E+06	142.179	393.064	1,04E+06	4.569	117,0 mi
2018	726,2 mi	1,46E+06	5,97E+06	146.066	402.790	1,04E+06	3.447	119,9 mi
2019	722,1 mi	1,45E+06	6,05E+06	137.969	368.344	1,03E+06	3.216	121,4 mi
2020	254,0 mi	511.062	2,17E+06	48.273	125.394	357.245	1.550	43,7 mi
2021	437,4 mi	880.016	3,59E+06	82.656	212.324	602.375	2.958	70,9 mi

- **População Beneficiária (%)**

Total	Média Anual	Média	Desvio	Q0	Mediana	Q1	Mín	Máx
333,59	22,24	0,0448	0,0300	0,0204	0,0406	0,0641	0,0004	0,1571
Mínimos:								
Cidade			Ano			Pop. (%)		
Boa Vista do Sul			2017			0,0004		
Máximos:								
Cidade			Ano			Pop. (%)		
Benjamin Constant do Sul			2021			0,1571		
Métricas Anuais:								
Ano	**Total**	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	23,916	0,0482	0,0266	0,0261	0,0463	0,0694	0,0007	0,1186
2008	21,872	0,0441	0,0259	0,0223	0,0410	0,0635	0,0004	0,1141
2009	25,749	0,0519	0,0313	0,0261	0,0485	0,0739	0,0007	0,1471
2010	26,046	0,0525	0,0322	0,0264	0,0487	0,0767	0,0004	0,1434
2011	25,703	0,0518	0,0327	0,0250	0,0483	0,0775	0,0004	0,1385
2012	26,394	0,0532	0,0330	0,0261	0,0487	0,0777	0,0014	0,1492
2013	25,356	0,0510	0,0325	0,0241	0,0462	0,0747	0,0010	0,1462
2014	23,894	0,0481	0,0317	0,0223	0,0438	0,0703	0,0012	0,1468
2015	22,635	0,0455	0,0300	0,0199	0,0413	0,0680	0,0015	0,1361
2016	19,907	0,0401	0,0272	0,0169	0,0359	0,0585	0,0019	0,1315
2017	19,250	0,0387	0,0276	0,0169	0,0333	0,0539	0,0004	0,1420
2018	18,468	0,0372	0,0271	0,0159	0,0330	0,0513	0,0008	0,1462
2019	16,445	0,0331	0,0254	0,0136	0,0281	0,0473	0,0005	0,1408
2020	18,345	0,0369	0,0271	0,0156	0,0324	0,0514	0,0009	0,1485
2021	19,612	0,0395	0,0280	0,0167	0,0348	0,0560	0,0013	0,1571

- **Repasse por Beneficiário (R$)**

Total	Média Anual	Média	Desvio	Q0	Mediana	Q1	Mín	Máx
1,03E+07	689.500	1.388,44	532,01	944,06	1.373,14	1.769,66	206,67	7.545,50
Mínimos:								
Cidade			Ano			Repasse (R$)		
Pinto Bandeira			2013			206,67		
Máximos:								
Cidade			Ano			Repasse (R$)		
Fagundes Varela			2018			7.545,50		

Métricas Anuais:								
Ano	**Total**	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	370.606	747,19	108,30	686,26	750,67	818,96	398,88	1.199,94
2008	459.304	926,02	123,67	855,22	939,56	1.005,97	405,78	1.335,33
2009	456.001	919,36	139,55	849,09	921,23	998,52	376,00	1.839,15
2010	506.215	1.020,59	136,16	952,45	1.040,81	1.100,37	293,33	1.433,11
2011	604.391	1.218,53	135,35	1.154,77	1.234,96	1.298,60	608,42	1.587,89
2012	692.558	1.396,29	152,83	1.317,15	1.408,10	1.488,34	618,50	2.035,29
2013	808.434	1.626,63	252,98	1.472,40	1.582,43	1.745,15	206,67	2.939,33
2014	873.436	1.757,42	272,84	1.600,91	1.723,30	1.894,02	679,86	3.886,67
2015	885.726	1.782,14	285,84	1.613,17	1.739,55	1.908,47	927,76	3.637,29
2016	951.887	1.915,27	310,65	1.724,29	1.883,46	2.062,42	826,00	3.715,18
2017	899.338	1.809,53	334,54	1.615,07	1.762,26	1.960,86	907,33	4.569,00
2018	953.868	1.919,25	422,97	1.695,77	1.853,59	2.091,97	861,75	7.545,50
2019	1.042.251	2.097,08	376,84	1.838,40	2.056,75	2.309,77	1.170,23	4.280,00
2020	318.180	640,20	152,04	529,78	622,60	720,76	231,53	1.410,67
2021	520.305	1.046,89	306,53	832,59	1.011,84	1.207,63	360,75	2.738,00

- **Repasse por População (R$)**

Total	**Média Anual**	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
463.660	30.911	62,24	50,98	23,71	50,12	87,53	0,23	429,61
Mínimos:								
Cidade			**Ano**			**Repasse (R$)**		
Pinto Bandeira			2013			0,23		
Máximos:								
Cidade			**Ano**			**Repasse (R$)**		
Redentora			2019			429,61		
Métricas Anuais:								
Ano	**Total**	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	18.361	37,02	22,27	19,06	34,40	52,38	0,46	101,94
2008	20.754	41,84	26,14	20,26	37,53	60,08	0,41	114,19
2009	24.166	48,72	30,81	23,95	44,05	70,11	0,28	134,40
2010	27.646	55,74	36,66	25,91	49,40	80,93	0,24	156,88
2011	31.972	64,46	41,88	29,82	58,97	94,54	0,26	182,12
2012	37.573	75,75	48,57	36,02	68,30	110,76	1,54	230,73
2013	41.559	83,62	56,30	38,15	74,33	121,79	0,23	303,33
2014	42.511	85,54	60,53	37,86	74,49	123,38	1,61	376,00
2015	40.971	82,44	59,58	35,34	69,76	117,45	2,32	391,17
2016	38.692	77,85	58,42	31,49	68,25	110,12	2,49	381,36
2017	35.826	72,08	57,87	27,85	61,37	102,11	1,64	405,70
2018	36.295	73,03	59,98	28,52	62,71	100,40	1,79	426,70
2019	35.092	70,61	60,26	26,48	58,35	95,95	1,27	429,61
2020	11.729	23,60	20,43	10,00	19,78	30,63	0,63	180,13
2021	20.514	41,28	36,47	17,22	34,40	53,08	0,83	320,65

A.2 Dados estatísticos do IDESE

- **Média dos índices do IDESE**

Índice	Valor
IDESE	0,7253
Bloco Educação	0,6914
Bloco Renda	0,6482
Bloco Saúde	0,8363
Anos Finais EF	0,6780
Anos Iniciais EF	0,7608
Ensino Fundamental	0,7194
Ensino Médio	0,7607
Escolaridade Adulta	0,4538
Pré Escola	0,8317
Apropriação da Renda	0,6650
Geração da Renda	0,6314
Mortes por Causas Evitáveis	0,6187
Condições Gerais de Saúde	0,7609
Óbitos por Causas Mal Definidas	0,9032
Longevidade	0,8982
Consultas Pré Natal	0,7726
Mortalidade de Menores de 5 anos	0,9269
Saúde Materno Infantil	0,8498

- **IDESE Geral**

Média	Desvio	Q0	Mediana	Q1	Mín	Máx
0,725285	0,06438	0,68181	0,730506	0,773712	0,498572	0,89599

Mínimos:		
Cidade	**Ano**	**IDESE**
Jaquirana	2009	0,498572

Máximos:		
Cidade	**Ano**	**IDESE**
Carlos Barbosa	2020	0,89599

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,672798	0,061909	0,628372	0,672940	0,718887	0,518758	0,829596
2008	0,678056	0,061487	0,634828	0,677054	0,726052	0,499881	0,872022
2009	0,686136	0,061449	0,639929	0,687736	0,731190	0,498572	0,849440
2010	0,689920	0,062476	0,640766	0,693576	0,735045	0,510146	0,851099
2011	0,702479	0,062116	0,655058	0,707513	0,748656	0,544406	0,861406
2012	0,708920	0,060487	0,659034	0,710413	0,752298	0,544277	0,876996
2013	0,732661	0,058656	0,691462	0,735514	0,776737	0,567477	0,881585
2014	0,744597	0,056199	0,705013	0,747949	0,787520	0,575721	0,892221
2015	0,740376	0,054659	0,703923	0,743300	0,783019	0,566789	0,878516
2016	0,743729	0,052065	0,705893	0,745624	0,780277	0,571577	0,883502
2017	0,749025	0,050406	0,714295	0,752441	0,788019	0,580549	0,885159
2018	0,753144	0,051332	0,715241	0,757296	0,791443	0,593893	0,884854
2019	0,760662	0,049523	0,726162	0,764374	0,796300	0,609463	0,892477
2020	0,755989	0,050524	0,722597	0,759332	0,791758	0,600243	0,895990
2021	0,760354	0,051669	0,727213	0,763488	0,797061	0,591968	0,895419

- **Bloco Educação**

Média	Desvio	Q0	Mediana	Q1	Mín	Máx
0,69139	0,083275	0,642433	0,705181	0,752917	0,34689	0,883096

Mínimos:		
Cidade	**Ano**	**Educação**
Charrua	2007	0,34689

Máximos:		
Cidade	**Ano**	**Educação**
Picada Café	2020	0,883096

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,623132	0,087121	0,567871	0,637290	0,687492	0,346890	0,796051
2008	0,618272	0,083526	0,564794	0,625049	0,675428	0,347957	0,800935
2009	0,635829	0,084288	0,582042	0,641681	0,700095	0,365372	0,815962
2010	0,634831	0,083786	0,577195	0,646128	0,695605	0,367747	0,811478
2011	0,654244	0,081039	0,603229	0,665904	0,715322	0,366216	0,835309
2012	0,670663	0,075207	0,619198	0,680010	0,729585	0,414148	0,833802
2013	0,691266	0,072929	0,644469	0,702239	0,747119	0,481303	0,844984
2014	0,711623	0,068167	0,672708	0,721504	0,761679	0,473632	0,841585
2015	0,712170	0,065235	0,669036	0,719570	0,763430	0,489832	0,854847
2016	0,718066	0,056783	0,682545	0,722286	0,760411	0,501941	0,847601
2017	0,729832	0,054794	0,694479	0,734333	0,769965	0,523622	0,846332
2018	0,737001	0,052998	0,704652	0,741112	0,773663	0,547738	0,862233
2019	0,745198	0,053349	0,711834	0,748500	0,782948	0,475230	0,881936
2020	0,747767	0,052783	0,714665	0,753851	0,784589	0,535354	0,883096
2021	0,740333	0,052007	0,705987	0,744716	0,778873	0,560878	0,853110

- **Bloco Renda**

Média	Desvio	Q0	Mediana	Q1	Mín	Máx
0,648167	0,107146	0,571995	0,648535	0,724004	0,307709	0,997402

Mínimos:		
Cidade	**Ano**	**Renda**
Benjamin Constant do Sul	2012	0,307709

Máximos:		
Cidade	**Ano**	**Renda**
Água Santa	2021	0,997402

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,580124	0,101428	0,513573	0,570102	0,658703	0,321235	0,876009
2008	0,594449	0,103041	0,522437	0,587985	0,669168	0,314266	0,930745
2009	0,598870	0,102227	0,530117	0,590196	0,672422	0,329930	0,949004
2010	0,605870	0,102809	0,529446	0,607207	0,676149	0,313495	0,903270
2011	0,623813	0,102548	0,548269	0,622406	0,700513	0,353167	0,905582
2012	0,623555	0,104981	0,552031	0,619551	0,693835	0,307709	0,921771
2013	0,675204	0,101817	0,601221	0,671060	0,748309	0,397046	0,952901
2014	0,686216	0,100311	0,614526	0,682969	0,761484	0,383444	0,956213
2015	0,668931	0,098503	0,598761	0,665009	0,741374	0,377371	0,930189
2016	0,670994	0,097404	0,595906	0,669881	0,737334	0,367608	0,937164
2017	0,671422	0,094925	0,606235	0,672625	0,735307	0,393605	0,930633
2018	0,676129	0,099009	0,605905	0,671948	0,746028	0,394372	0,953618
2019	0,687109	0,094590	0,618232	0,686008	0,753835	0,406941	0,951835
2020	0,666240	0,096752	0,594769	0,667033	0,735737	0,371122	0,982044
2021	0,693045	0,105129	0,620205	0,692804	0,766188	0,387034	0,997402

- **Bloco Saúde**

Média	Desvio	Q0	Mediana	Q1	Mín	Máx
0,836298	0,041996	0,807438	0,83879	0,868139	0,668209	0,950625

Mínimos:		
Cidade	**Ano**	**Saúde**
Pedro Osório	2013	0,668209

Máximos:		
Cidade	**Ano**	**Saúde**
Santo Expedito do Sul	2020	0,950625

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,815138	0,042936	0,783709	0,816976	0,845924	0,684210	0,918332
2008	0,821448	0,041147	0,790336	0,823186	0,851217	0,694674	0,916622
2009	0,823708	0,040791	0,797576	0,825065	0,855883	0,691445	0,925283
2010	0,829058	0,040719	0,803590	0,831267	0,859200	0,672395	0,925810
2011	0,829379	0,041452	0,801245	0,830559	0,858658	0,682614	0,926797
2012	0,832543	0,041940	0,801558	0,834837	0,862973	0,682768	0,922169
2013	0,831514	0,041238	0,802576	0,833205	0,863159	0,668209	0,917734
2014	0,835952	0,039784	0,808681	0,839392	0,866281	0,693251	0,929016
2015	0,840028	0,039339	0,813824	0,844237	0,869927	0,714490	0,924011
2016	0,842128	0,040184	0,815428	0,845700	0,872768	0,697206	0,932771
2017	0,845821	0,039355	0,818721	0,849450	0,874498	0,711510	0,943270
2018	0,846302	0,040284	0,816221	0,847828	0,876941	0,705206	0,930519
2019	0,849680	0,039264	0,823408	0,850763	0,880542	0,730441	0,945643
2020	0,853959	0,039932	0,827770	0,857257	0,884066	0,714031	0,950625
2021	0,847683	0,039871	0,819503	0,851192	0,879415	0,718854	0,931759

- **Anos Finais do Ensino Fundamental**

Média	Desvio	Q0	Mediana	Q1	Mín	Máx
0,677977	0,067319	0,635115	0,677443	0,719893	0,440981	0,996032

Mínimos:		
Cidade	**Ano**	**Anos Finais EF**
Jacuizinho	2021	0,440981

Máximos:		
Cidade	**Ano**	**Anos Finais EF**
Tupanci do Sul	2021	0,996032

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,627087	0,056947	0,594816	0,628288	0,659669	0,457131	0,845291
2008	0,627087	0,056947	0,594816	0,628288	0,659669	0,457131	0,845291
2009	0,666950	0,057313	0,629548	0,668693	0,704800	0,465643	0,881856
2010	0,666950	0,057313	0,629548	0,668693	0,704800	0,465643	0,881856
2011	0,668279	0,058913	0,631371	0,670439	0,706251	0,479936	0,840960
2012	0,668279	0,058913	0,631371	0,670439	0,706251	0,479936	0,840960
2013	0,664231	0,064694	0,624064	0,664613	0,707392	0,449835	0,843413
2014	0,664231	0,064694	0,624064	0,664613	0,707392	0,449835	0,843413
2015	0,675310	0,054781	0,646213	0,672085	0,706626	0,490560	0,903019
2016	0,675310	0,054781	0,646213	0,672085	0,706626	0,490560	0,903019
2017	0,711041	0,065400	0,673621	0,709803	0,755925	0,504341	0,926229
2018	0,711041	0,065400	0,673621	0,709803	0,755925	0,504341	0,926229
2019	0,727099	0,063352	0,689259	0,725760	0,758826	0,504555	0,931264
2020	0,727099	0,063352	0,689259	0,725760	0,758826	0,504555	0,931264
2021	0,689375	0,063080	0,649813	0,683627	0,729749	0,440981	0,996032

- **Anos Iniciais do Ensino Fundamental**

Média	Desvio	Q0	Mediana	Q1	Mín	Máx
0,760785	0,087277	0,700944	0,762126	0,821106	0,410818	1,0

Mínimos:		
Cidade	**Ano**	**Anos Iniciais EF**
Cacique Doble	2009	0,410818
Cacique Doble	2010	0,410818

Máximos:	
Cidade - Ano	**Anos Iniciais EF**
Centenário 2019, Centenário 2020, Coronel Pilar 2019, Coronel Pilar 2020, Dois Lajeados 2017, Dois Lajeados 2018, Dois Lajeados 2019, Dois Lajeados 2020, Fagundes Varela 2017, Fagundes Varela 2018, Ipiranga do Sul 2021, Mormaço 2017, Mormaço 2018, Picada Café 2019, Picada Café 2020, São Valentim 2017, São Valentim 2018, Severiano de Almeida 2019, Severiano de Almeida 2020, Taquaruçu do Sul 2017, Taquaruçu do Sul 2018, Vista Alegre do Prata 2011, Vista Alegre do Prata 2012	1,0

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,653269	0,061733	0,615541	0,651232	0,690376	0,445334	0,957336
2008	0,653269	0,061733	0,615541	0,651232	0,690376	0,445334	0,957336
2009	0,706341	0,065156	0,668176	0,704983	0,745213	0,410818	0,956770
2010	0,706341	0,065156	0,668176	0,704983	0,745213	0,410818	0,956770
2011	0,736161	0,064077	0,698621	0,736138	0,776971	0,506498	1,000000
2012	0,736161	0,064077	0,698621	0,736138	0,776971	0,506498	1,000000
2013	0,773521	0,064644	0,734032	0,774529	0,810434	0,543971	0,988238
2014	0,773521	0,064644	0,734032	0,774529	0,810434	0,543971	0,988238
2015	0,795265	0,064064	0,753889	0,791042	0,830120	0,602110	0,998398
2016	0,795265	0,064064	0,753889	0,791042	0,830120	0,602110	0,998398
2017	0,813722	0,071927	0,762790	0,813448	0,861559	0,600548	1,000000
2018	0,813722	0,071927	0,762790	0,813448	0,861559	0,600548	1,000000
2019	0,828402	0,070731	0,779803	0,829938	0,872430	0,573703	1,000000
2020	0,828402	0,070731	0,779803	0,829938	0,872430	0,573703	1,000000
2021	0,797662	0,063856	0,754477	0,799523	0,843906	0,618364	1,000000

- **Ensino Fundamental (Geral)**

Média	Desvio	Q0	Mediana	Q1	Mín	Máx
0,719381	0,069216	0,672587	0,718233	0,766274	0,470991	0,937536

Mínimos:		
Cidade	**Ano**	**Ensino Fundamental**
Capão Bonito do Sul	2007	0,470991
Capão Bonito do Sul	2008	0,470991

Máximos:		
Cidade	**Ano**	**Ensino Fundamental**
Picada Café	2019	0,937536
Picada Café	2020	0,937536

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,640178	0,051316	0,609650	0,641069	0,671388	0,470991	0,826393
2008	0,640178	0,051316	0,609650	0,641069	0,671388	0,470991	0,826393
2009	0,686645	0,053145	0,653785	0,686343	0,723432	0,539557	0,894451
2010	0,686645	0,053145	0,653785	0,686343	0,723432	0,539557	0,894451
2011	0,702220	0,052084	0,669101	0,700914	0,735141	0,540527	0,879032
2012	0,702220	0,052084	0,669101	0,700914	0,735141	0,540527	0,879032
2013	0,718876	0,057056	0,680445	0,718590	0,757103	0,560084	0,915826
2014	0,718876	0,057056	0,680445	0,718590	0,757103	0,560084	0,915826
2015	0,735288	0,052098	0,703499	0,734385	0,771905	0,576345	0,924917
2016	0,735288	0,052098	0,703499	0,734385	0,771905	0,576345	0,924917
2017	0,762381	0,059442	0,720449	0,761154	0,804211	0,572850	0,928635
2018	0,762381	0,059442	0,720449	0,761154	0,804211	0,572850	0,928635
2019	0,777750	0,058214	0,738822	0,777401	0,818693	0,627724	0,937536
2020	0,777750	0,058214	0,738822	0,777401	0,818693	0,627724	0,937536
2021	0,743518	0,054267	0,704095	0,743044	0,781322	0,588804	0,888837

- **Ensino Médio**

Média	Desvio	Q0	Mediana	Q1	Mín	Máx
0,760749	0,150442	0,651381	0,760083	0,874024	0,307213	1,0

Mínimos:		
Cidade	**Ano**	**Ensino Médio**
São Valério do Sul	2019	0,307213

Máximos:		
Cidade	**Ano**	**Ensino Médio**
Numerosos registros com valor 1,0		

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,822438	0,136513	0,719835	0,833691	0,939647	0,421822	1,0
2008	0,806122	0,137001	0,697546	0,811070	0,912004	0,416896	1,0
2009	0,775159	0,137689	0,680708	0,784856	0,874188	0,416613	1,0
2010	0,737673	0,135264	0,642313	0,746344	0,832776	0,400898	1,0
2011	0,731253	0,136527	0,639887	0,733660	0,821767	0,388045	1,0
2012	0,739313	0,141897	0,640202	0,736080	0,839485	0,391446	1,0
2013	0,764523	0,146470	0,663585	0,768813	0,866935	0,380966	1,0
2014	0,783828	0,148684	0,679895	0,786326	0,892626	0,391344	1,0
2015	0,753353	0,153348	0,640514	0,754199	0,861194	0,389968	1,0
2016	0,718170	0,156036	0,613784	0,702488	0,811605	0,334184	1,0
2017	0,722304	0,155558	0,614980	0,710139	0,833186	0,307928	1,0
2018	0,742359	0,154984	0,630336	0,738150	0,849886	0,319757	1,0
2019	0,760118	0,156424	0,644941	0,759057	0,878563	0,307213	1,0
2020	0,770147	0,158439	0,659515	0,767034	0,898667	0,320446	1,0
2021	0,784565	0,156866	0,667464	0,785234	0,914654	0,335367	1,0

- **Escolaridade Adulta**

Média	Desvio	Q0	Mediana	Q1	Mín	Máx
0,45375	0,099487	0,382641	0,449233	0,523215	0,177169	0,765011

Mínimos:		
Cidade	**Ano**	**Escolaridade Adulta**
Sério	2007	0,177169

Máximos:		
Cidade	**Ano**	**Escolaridade Adulta**
Porto Alegre	2021	0,765011

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,381412	0,093740	0,313430	0,372352	0,441849	0,177169	0,722588
2008	0,394575	0,094849	0,324316	0,384757	0,457032	0,181339	0,732295
2009	0,408114	0,095438	0,336174	0,399619	0,469877	0,194574	0,740561
2010	0,421824	0,095907	0,350425	0,412300	0,483025	0,204800	0,747800
2011	0,431048	0,095004	0,359789	0,423343	0,490636	0,214660	0,750170
2012	0,440500	0,093897	0,370157	0,433271	0,498963	0,216615	0,752302
2013	0,450649	0,092843	0,382239	0,445294	0,509979	0,222545	0,754367
2014	0,457663	0,091986	0,390339	0,450943	0,517305	0,224258	0,755879
2015	0,468471	0,090850	0,403605	0,463060	0,526798	0,227300	0,757978
2016	0,477966	0,089748	0,411916	0,474822	0,537123	0,234389	0,759981
2017	0,479830	0,089810	0,411541	0,474892	0,539705	0,235093	0,758465
2018	0,487889	0,089010	0,419437	0,483435	0,547191	0,237478	0,760226
2019	0,495302	0,088368	0,429282	0,492511	0,554970	0,239314	0,761513
2020	0,502208	0,087825	0,436523	0,502067	0,563547	0,240123	0,763388
2021	0,508309	0,087336	0,442306	0,507356	0,568052	0,242181	0,765011

- **Pré-Escola**

Média	Desvio	Q0	Mediana	Q1	Mín	Máx
0,831681	0,19753	0,727027	0,902778	1,0	0,0	1,0

Mínimos:	
Cidade - Ano	**Pré-Escola**
Barão do Triunfo 2007, Forquetinha 2010, Forquetinha 2011, Porto Vera Cruz 2008, São Valério do Sul 2009, São Valério do Sul 2010, São Valério do Sul 2011	0,0

Máximos:		
Cidade	**Ano**	**Pré-Escola**
Numerosos registros com valor 1,0		

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,648499	0,238551	0,476340	0,650601	0,852700	0,000000	1,0
2008	0,632213	0,226384	0,473232	0,621825	0,800087	0,000000	1,0
2009	0,673399	0,222219	0,506296	0,668142	0,862675	0,000000	1,0
2010	0,693181	0,216083	0,537948	0,694871	0,867657	0,000000	1,0
2011	0,752457	0,204323	0,615024	0,790687	0,914055	0,000000	1,0
2012	0,800618	0,178356	0,679985	0,830039	0,976002	0,167438	1,0
2013	0,831016	0,162857	0,722222	0,860902	1,000000	0,183066	1,0
2014	0,886125	0,135074	0,809272	0,928571	1,000000	0,197427	1,0
2015	0,891567	0,129467	0,828255	0,935484	1,000000	0,194308	1,0
2016	0,940840	0,093331	0,905405	0,992042	1,000000	0,269231	1,0
2017	0,954812	0,078080	0,931732	1,000000	1,000000	0,312293	1,0
2018	0,955373	0,071708	0,928375	1,000000	1,000000	0,370786	1,0
2019	0,947620	0,075697	0,903293	0,994389	1,000000	0,440517	1,0
2020	0,940962	0,076388	0,897071	0,974191	1,000000	0,454117	1,0
2021	0,924938	0,087360	0,876126	0,945229	1,000000	0,368353	1,0

- **Apropriação da Renda**

Média	Desvio	Q0	Mediana	Q1	Mín	Máx
0,664954	0,122063	0,583178	0,665653	0,750275	0,207288	1,0

Mínimos:		
Cidade	**Ano**	**Apropriação Renda**
Passa Sete	2007	0,207288

Máximos:	
Cidade - Ano	**Apropriação Renda**
Água Santa 2019, Ijuí 2012, Ijuí 2013, Ijuí 2014, Ijuí 2015, Ijuí 2016, Ipiranga do Sul 2019, Porto Alegre 2014, Porto Alegre 2019, Selbach 2014, Três Arroios 2013, Três Arroios 2014, Três Arroios 2019, Vista Alegre do Prata 2020	1,0

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,574374	0,109412	0,506192	0,571416	0,646054	0,207288	0,927015
2008	0,586884	0,109551	0,511826	0,580854	0,657637	0,235380	0,932110
2009	0,615274	0,111738	0,533557	0,612915	0,685913	0,271394	0,941777
2010	0,607817	0,111598	0,534658	0,607571	0,686668	0,297318	0,970827
2011	0,625626	0,111919	0,551672	0,628832	0,706572	0,315405	0,963677
2012	0,650593	0,116660	0,571816	0,652400	0,734030	0,228461	1,000000
2013	0,688445	0,115139	0,613819	0,687423	0,770420	0,284017	1,000000
2014	0,712630	0,116897	0,636957	0,712569	0,798398	0,282395	1,000000
2015	0,689775	0,115132	0,615663	0,687980	0,766330	0,286285	1,000000
2016	0,696404	0,114210	0,621505	0,690577	0,773871	0,295428	1,000000
2017	0,700714	0,110250	0,629700	0,698591	0,776703	0,309118	0,985327
2018	0,705719	0,109698	0,633560	0,704528	0,783317	0,281783	0,980571
2019	0,726934	0,108852	0,654535	0,725682	0,800755	0,282959	1,000000
2020	0,698766	0,110428	0,626748	0,694495	0,773517	0,233763	1,000000
2021	0,693695	0,111309	0,620645	0,682481	0,767415	0,260637	0,994803

- **Geração da Renda**

Média	Desvio	Q0	Mediana	Q1	Mín	Máx
0,631379	0,134203	0,533577	0,616324	0,720454	0,278784	1,0

Mínimos:		
Cidade	**Ano**	**Geração Renda**
Caraá	2009	0,278784

Máximos:		
Cidade	**Ano**	**Geração Renda**
Numerosos registros com valor 1,0		

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,585874	0,131762	0,487532	0,572110	0,669068	0,291195	1,0
2008	0,602015	0,135553	0,496918	0,585079	0,695269	0,292593	1,0
2009	0,582467	0,137037	0,482205	0,563330	0,669580	0,278784	1,0
2010	0,603923	0,131450	0,502761	0,588223	0,682987	0,329673	1,0
2011	0,622000	0,131944	0,521324	0,603828	0,712452	0,343747	1,0
2012	0,596517	0,130224	0,502366	0,579248	0,681135	0,289772	1,0
2013	0,661963	0,129205	0,567985	0,649634	0,746547	0,374943	1,0
2014	0,659802	0,123763	0,569348	0,645684	0,740592	0,377900	1,0
2015	0,648086	0,125569	0,558180	0,632715	0,731436	0,365241	1,0
2016	0,645584	0,126821	0,555762	0,629813	0,733000	0,358747	1,0
2017	0,642130	0,124667	0,553509	0,624362	0,724846	0,346166	1,0
2018	0,646539	0,134609	0,550824	0,629479	0,735849	0,352830	1,0
2019	0,647284	0,127387	0,551858	0,633987	0,735216	0,358531	1,0
2020	0,633715	0,127435	0,541693	0,627455	0,719478	0,368073	1,0
2021	0,692396	0,143394	0,592032	0,681820	0,783399	0,355025	1,0

- **Mortes por Causas Evitáveis**

Média	Desvio	Q0	Mediana	Q1	Mín	Máx
0,618712	0,088103	0,559871	0,623466	0,683218	0,22119	0,860508

Mínimos:		
Cidade	**Ano**	**Mortes Evitáveis**
Lavras do Sul	2010	0,22119

Máximos:		
Cidade	**Ano**	**Mortes Evitáveis**
Água Santa	2020	0,860508

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,631516	0,081470	0,574932	0,634018	0,695083	0,377904	0,835231
2008	0,631627	0,080850	0,576065	0,633807	0,691308	0,399450	0,840265
2009	0,628065	0,083077	0,572379	0,637467	0,687849	0,295236	0,852430
2010	0,629399	0,082088	0,572427	0,630422	0,686847	0,221190	0,836977
2011	0,626131	0,084717	0,569637	0,630173	0,687696	0,299376	0,834152
2012	0,628109	0,085766	0,566101	0,632250	0,694894	0,313917	0,848331
2013	0,621104	0,087985	0,559382	0,623751	0,687298	0,358360	0,844771
2014	0,623096	0,087087	0,567256	0,629330	0,682815	0,312579	0,849730
2015	0,621880	0,088104	0,571136	0,626206	0,683691	0,358077	0,852340
2016	0,622454	0,089254	0,563317	0,629089	0,689948	0,331362	0,828264
2017	0,611271	0,087870	0,556510	0,615962	0,671840	0,334142	0,843275
2018	0,599595	0,090275	0,540935	0,603132	0,664586	0,309931	0,849340
2019	0,596463	0,091057	0,537090	0,597988	0,664119	0,310188	0,821857
2020	0,604508	0,094364	0,539924	0,608281	0,676848	0,362539	0,860508
2021	0,605591	0,095681	0,538697	0,611069	0,676913	0,313598	0,839612

- **Condições Gerais de Saúde**

Média	Desvio	Q0	Mediana	Q1	Mín	Máx
0,760937	0,053842	0,724709	0,764171	0,799851	0,507235	0,916685

Mínimos:		
Cidade	**Ano**	**Condições Saúde**
Pedro Osório	2018	0,507235

Máximos:		
Cidade	**Ano**	**Condições Saúde**
Água Santa	2020	0,916685

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,760529	0,052585	0,723606	0,761582	0,800840	0,561771	0,883851
2008	0,762814	0,051515	0,726044	0,765038	0,800302	0,598096	0,881149
2009	0,762634	0,052833	0,727638	0,766991	0,800452	0,590956	0,879340
2010	0,764000	0,051817	0,729147	0,766596	0,799172	0,576495	0,892847
2011	0,760923	0,052679	0,728067	0,763481	0,798688	0,566279	0,882083
2012	0,761979	0,052689	0,725962	0,764151	0,800208	0,555159	0,896675
2013	0,760056	0,052819	0,723003	0,765237	0,797403	0,549338	0,888941
2014	0,762808	0,052386	0,726837	0,765867	0,800187	0,583239	0,888680
2015	0,764307	0,051408	0,731288	0,768920	0,801022	0,599409	0,910942
2016	0,764300	0,053018	0,727110	0,769086	0,804632	0,557000	0,897289
2017	0,760455	0,054140	0,724584	0,761929	0,798766	0,552195	0,905127
2018	0,755878	0,056675	0,716707	0,757966	0,796549	0,507235	0,898411
2019	0,756006	0,057213	0,719431	0,756338	0,799398	0,532815	0,883879
2020	0,759521	0,058119	0,720514	0,763219	0,801636	0,532296	0,916685
2021	0,757863	0,056825	0,720006	0,760905	0,799700	0,542128	0,908342

- **Óbitos por Causas Mal Definidas**

Média	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
0,903162	0,054932	0,877176	0,91179	0,941984	0,583341	0,995426

Mínimos:		
Cidade	**Ano**	**Óbitos Mal Definidos**
Uruguaiana	2016	0,583341

Máximos:		
Cidade	**Ano**	**Óbitos Mal Definidos**
Bento Gonçalves	2014	0,995426

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,889541	0,061788	0,862357	0,903254	0,930498	0,621654	0,987919
2008	0,894002	0,061015	0,872686	0,905032	0,932393	0,639727	0,984358
2009	0,897203	0,059380	0,874504	0,908944	0,935729	0,653339	0,985901
2010	0,898601	0,055147	0,877540	0,906740	0,936996	0,672739	0,984980
2011	0,895714	0,050805	0,868275	0,902724	0,934552	0,643192	0,985063
2012	0,895849	0,051219	0,870851	0,902728	0,930724	0,667801	0,987968
2013	0,899009	0,049463	0,877633	0,906519	0,930207	0,668794	0,991138
2014	0,902519	0,050560	0,878204	0,912100	0,934749	0,602950	0,995426
2015	0,906734	0,050592	0,880204	0,914803	0,940982	0,607937	0,994099
2016	0,906145	0,054332	0,875997	0,916402	0,945968	0,583341	0,994413
2017	0,909639	0,053874	0,877830	0,916553	0,952809	0,592921	0,990214
2018	0,912160	0,055445	0,885090	0,921268	0,955080	0,604040	0,991256
2019	0,915548	0,053626	0,890832	0,926791	0,954811	0,633445	0,988835
2020	0,914535	0,054674	0,888956	0,927223	0,954859	0,639857	0,989164
2021	0,910134	0,052464	0,883142	0,919459	0,946448	0,683372	0,987132

- **Longevidade**

Média	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
0,898181	0,053517	0,860884	0,899166	0,936633	0,678646	1,0

Mínimos:		
Cidade	**Ano**	**Longevidade**
Muliterno	2011	0,678646

Máximos:		
Cidade	**Ano**	**Longevidade**
Numerosos registros com valor 1,0		

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,876314	0,054123	0,839643	0,876215	0,910746	0,721779	1,0
2008	0,879699	0,051594	0,845950	0,879279	0,915175	0,732895	1,0
2009	0,879539	0,051246	0,845269	0,879490	0,914943	0,727425	1,0
2010	0,883778	0,052078	0,849654	0,883108	0,917414	0,698267	1,0
2011	0,884898	0,053150	0,850111	0,888464	0,921272	0,678646	1,0
2012	0,889868	0,052696	0,854523	0,889631	0,926577	0,702556	1,0
2013	0,890788	0,052059	0,851089	0,891031	0,925157	0,715956	1,0
2014	0,897359	0,049595	0,864838	0,896722	0,930634	0,727200	1,0
2015	0,904081	0,049805	0,868328	0,906328	0,938967	0,736477	1,0
2016	0,906371	0,051408	0,869034	0,911136	0,947040	0,742758	1,0
2017	0,915475	0,048381	0,882649	0,916762	0,949692	0,765027	1,0
2018	0,914736	0,049406	0,879666	0,915501	0,949982	0,783428	1,0
2019	0,919449	0,048214	0,886915	0,922751	0,955741	0,773832	1,0
2020	0,923710	0,050246	0,886254	0,928131	0,963135	0,751841	1,0
2021	0,906464	0,055302	0,864190	0,915022	0,947489	0,748551	1,0

- **Consultas Pré-Natal**

Média	Desvio	Q0	Mediana	Q1	Mín	Máx
0,772631	0,080911	0,724622	0,783005	0,833079	0,431154	0,959339

Mínimos:		
Cidade	**Ano**	**Consultas Pré-Natal**
Bom Jesus	2007	0,431154

Máximos:		
Cidade	**Ano**	**Consultas Pré-Natal**
Três de Maio	2009	0,959339

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,716680	0,095861	0,655033	0,725971	0,780709	0,431154	0,950329
2008	0,741524	0,090431	0,683921	0,750433	0,798359	0,469802	0,958453
2009	0,752595	0,085039	0,697423	0,762886	0,807509	0,461813	0,959339
2010	0,763240	0,085553	0,710509	0,772776	0,820687	0,438086	0,959007
2011	0,764571	0,085022	0,708157	0,775201	0,822973	0,459514	0,958224
2012	0,766526	0,083963	0,712727	0,778039	0,827855	0,449303	0,934824
2013	0,762692	0,081145	0,710946	0,775429	0,822644	0,441438	0,919311
2014	0,765162	0,075103	0,719176	0,780930	0,821009	0,451137	0,904424
2015	0,769618	0,072832	0,723578	0,782584	0,822210	0,481664	0,905599
2016	0,774370	0,071599	0,728145	0,785363	0,827700	0,516662	0,917647
2017	0,783626	0,070480	0,736593	0,788945	0,840923	0,541736	0,937697
2018	0,795599	0,065562	0,749220	0,799076	0,849684	0,564479	0,940434
2019	0,807732	0,060057	0,768658	0,812498	0,855690	0,562624	0,944370
2020	0,811786	0,058123	0,774497	0,818415	0,856682	0,574896	0,944300
2021	0,813480	0,057828	0,774003	0,817906	0,859691	0,609927	0,948477

- **Mortalidade de Menores de 5 Anos**

Média	Desvio	Q0	Mediana	Q1	Mín	Máx
0,926922	0,039689	0,904707	0,932209	0,954918	0,706377	1,0

Mínimos:		
Cidade	**Ano**	**Mortalidade <5 anos**
Herval	2008	0,706377

Máximos:		
Cidade	**Ano**	**Mortalidade <5 anos**
Numerosos registros com valor 1,0		

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,900462	0,044718	0,873701	0,906021	0,932861	0,708597	0,989754
2008	0,902138	0,042891	0,879426	0,907021	0,930166	0,706377	0,990898
2009	0,905308	0,040145	0,884287	0,911786	0,934471	0,753389	1,000000
2010	0,915550	0,039062	0,896116	0,920967	0,943794	0,760775	1,000000
2011	0,920060	0,036674	0,901238	0,923819	0,944324	0,780830	0,998871
2012	0,925036	0,038216	0,904312	0,929749	0,952975	0,764637	1,000000
2013	0,924705	0,037656	0,905249	0,928727	0,951823	0,748199	1,000000
2014	0,930213	0,033766	0,912225	0,932682	0,954242	0,796597	1,000000
2015	0,933776	0,034514	0,913095	0,938497	0,957113	0,779848	1,000000
2016	0,937057	0,034184	0,918935	0,940090	0,961375	0,771869	1,000000
2017	0,939439	0,032685	0,920592	0,943380	0,963187	0,808870	1,000000
2018	0,940984	0,034541	0,921231	0,945520	0,966114	0,826295	1,000000
2019	0,939438	0,034223	0,919964	0,943750	0,963159	0,778238	1,000000
2020	0,945506	0,032878	0,929570	0,949270	0,969300	0,777834	1,000000
2021	0,943967	0,034390	0,924485	0,949016	0,969227	0,792525	1,000000

- **Saúde Materno-Infantil**

Média	Desvio	Q0	Mediana	Q1	Mín	Máx
0,849776	0,050563	0,820833	0,856228	0,885969	0,604558	0,968822

Mínimos:		
Cidade	**Ano**	**Saúde Materno-Infantil**
Barra do Quaraí	2008	0,604558

Máximos:		
Cidade	**Ano**	**Saúde Materno-Infantil**
Teutônia	2019	0,968822

Métricas Anuais:							
Ano	**Média**	**Desvio**	**Q0**	**Mediana**	**Q1**	**Mín**	**Máx**
2007	0,808571	0,059489	0,770573	0,815947	0,851358	0,618914	0,942922
2008	0,821831	0,055567	0,787576	0,828007	0,856856	0,604558	0,941861
2009	0,828952	0,051619	0,798224	0,833719	0,861122	0,643580	0,958751
2010	0,839395	0,051302	0,810609	0,844700	0,872813	0,661767	0,955713
2011	0,842316	0,050104	0,813302	0,850505	0,878204	0,669836	0,948295
2012	0,845781	0,051210	0,815086	0,854427	0,882151	0,649698	0,955627
2013	0,843699	0,048461	0,817448	0,851547	0,879305	0,685846	0,947826
2014	0,847688	0,044496	0,823977	0,855279	0,879668	0,698140	0,943516
2015	0,851697	0,042414	0,827984	0,858128	0,880962	0,710080	0,940604
2016	0,855714	0,042133	0,829891	0,862706	0,885522	0,717659	0,942233
2017	0,861532	0,041713	0,835484	0,866810	0,894026	0,711188	0,956231
2018	0,868292	0,040338	0,843735	0,872900	0,896343	0,744941	0,961313
2019	0,873585	0,037573	0,848049	0,877918	0,899756	0,753069	0,968822
2020	0,878646	0,036921	0,856831	0,883348	0,904115	0,756669	0,962912
2021	0,878723	0,037773	0,857791	0,882001	0,905268	0,753528	0,968080

APÊNDICE B – APRESENTAÇÃO DOS RESULTADOS ENCONTRADOS

O presente apêndice está dividido em três seções. Na Seção B.1, são detalhados os tempos de execução para a construção dos métodos computacionais utilizados e a obtenção de seus resultados. Já nas Seções B.2 e B.3, apresentam-se os melhores resultados, ordenados pelo coeficiente de determinação R^2, obtidos para as combinações de variáveis e seus respectivos dados médios para um melhor entendimento de suas relações dos índices do IDESE e o atributo do PBF.

B.1 Registro de execução dos modelos

Está seção tem como objetivo apresentar os tempos de execução demandados para a construção e o processamento dos modelos *K-means*, visando identificar o melhor e o segundo melhor valor de *k* para a segmentação de cada uma das 55 combinações de variáveis possíveis. Adicionalmente, apresenta-se o tempo total de processamento após a execução de todos os métodos computacionais de aprendizado supervisionado selecionados para este trabalho.

Para a otimização da apresentação dos dados, utilizou-se a abreviação de determinados termos, conforme detalhado na lista de legendas abaixo; contudo, os logs e mensagens completas podem ser consultados integralmente no repositório remoto do trabalho Portella (2025).

- **C.S**: Coeficiente de silhueta;
- **1° K**: Melhor *k* (melhor divisão de grupos efetuada após execução do cálculo de coeficiênte de silhueta no método *K-means*);
- **2° K**: Segundo melhor *k* (melhor divisão de grupos efetuada após execução do cálculo de coeficiênte de silhueta no método *K-means*);
- **DF**: DataFrame (estrutura de dados da biblioteca Pandas da linguagem de programação Python);
- **(s)**: Segundos.

```
Total de beneficiários : Idese
Coeficiente de silhueta:: 8.672011 (s)
1° K: 3 - C.S.: 0.537618, 2° K: 2 - C.S.: 0.512693, DF:: (7434, 2)
Construção do modelo K-Means para o 1° K foi: 0.006380 (s)
```

```
1º K: 103.589631 (s)
Construção do modelo K-Means para o 2º K foi: 0.007773 (s)
2º K: 92.040118 (s)
Construção do modelo K-Means grupo único: 0.008263 (s)
Grupo único: 17.642823 (s)
Total de beneficiários : Blocos individuais
Coeficiente de silhueta:: 7.289757 (s)
1º K: 2 - C.S.: 0.360252, 2º K: 3 - C.S.: 0.271779, DF:: (7434, 4)
Construção do modelo K-Means para o 1º K foi: 0.006407 (s)
1º K: 88.044846 (s)
Construção do modelo K-Means para o 2º K foi: 0.016743 (s)
2º K: 101.976964 (s)
Construção do modelo K-Means grupo único: 0.008201 (s)
Grupo único: 33.334940 (s)
Total de beneficiários : Bloco educação individual
Coeficiente de silhueta:: 7.219162 (s)
1º K: 2 - C.S.: 0.552110, 2º K: 3 - C.S.: 0.547917, DF:: (7434, 2)
Construção do modelo K-Means para o 1º K foi: 0.004936 (s)
1º K: 90.224207 (s)
Construção do modelo K-Means para o 2º K foi: 0.010056 (s)
2º K: 104.448219 (s)
Construção do modelo K-Means grupo único: 0.009161 (s)
Grupo único: 17.966675 (s)
Total de beneficiários : Bloco educação resumido
Coeficiente de silhueta:: 7.532672 (s)
1º K: 2 - C.S.: 0.292187, 2º K: 3 - C.S.: 0.256660, DF:: (7434, 5)
Construção do modelo K-Means para o 1º K foi: 0.006640 (s)
1º K: 94.380264 (s)
Construção do modelo K-Means para o 2º K foi: 0.012664 (s)
2º K: 109.252652 (s)
Construção do modelo K-Means grupo único: 0.009633 (s)
Grupo único: 24.737214 (s)
Total de beneficiários : Bloco educação completo
Coeficiente de silhueta:: 7.215941 (s)
1º K: 2 - C.S.: 0.273269, 2º K: 3 - C.S.: 0.220940, DF:: (7434, 7)
Construção do modelo K-Means para o 1º K foi: 0.005756 (s)
1º K: 91.298117 (s)
Construção do modelo K-Means para o 2º K foi: 0.012513 (s)
2º K: 104.999424 (s)
Construção do modelo K-Means grupo único: 0.008097 (s)
Grupo único: 26.298754 (s)
Total de beneficiários : Bloco renda individual
Coeficiente de silhueta:: 7.514243 (s)
1º K: 2 - C.S.: 0.548154, 2º K: 3 - C.S.: 0.517728, DF:: (7434, 2)
Construção do modelo K-Means para o 1º K foi: 0.005711 (s)
1º K: 89.354148 (s)
Construção do modelo K-Means para o 2º K foi: 0.010247 (s)
2º K: 102.711304 (s)
Construção do modelo K-Means grupo único: 0.008371 (s)
Grupo único: 18.023085 (s)
Total de beneficiários : Bloco renda completo
Coeficiente de silhueta:: 8.981935 (s)
1º K: 3 - C.S.: 0.391819, 2º K: 2 - C.S.: 0.363108, DF:: (7434, 3)
Construção do modelo K-Means para o 1º K foi: 0.008612 (s)
1º K: 105.700778 (s)
Construção do modelo K-Means para o 2º K foi: 0.007882 (s)
2º K: 90.438956 (s)
Construção do modelo K-Means grupo único: 0.007179 (s)
Grupo único: 46.151376 (s)
Total de beneficiários : Bloco saúde individual
Coeficiente de silhueta:: 8.395892 (s)
1º K: 3 - C.S.: 0.529258, 2º K: 2 - C.S.: 0.514890, DF:: (7434, 2)
Construção do modelo K-Means para o 1º K foi: 0.006394 (s)
1º K: 103.443321 (s)
Construção do modelo K-Means para o 2º K foi: 0.007675 (s)
2º K: 94.572564 (s)
Construção do modelo K-Means grupo único: 0.008198 (s)
Grupo único: 17.724024 (s)
Total de beneficiários : Bloco saúde resumido
Coeficiente de silhueta:: 7.280465 (s)
```

```
1º K: 2 - C.S.: 0.356091, 2º K: 3 - C.S.: 0.267375, DF:: (7434, 4)
Construção do modelo K-Means para o 1º K foi: 0.006526 (s)
1º K: 86.207310 (s)
Construção do modelo K-Means para o 2º K foi: 0.009796 (s)
2º K: 106.244894 (s)
Construção do modelo K-Means grupo único: 0.008457 (s)
Grupo único: 24.832084 (s)
Total de beneficiários : Bloco saúde completo
Coeficiente de silhueta:: 7.322186 (s)
1º K: 2 - C.S.: 0.292225, 2º K: 3 - C.S.: 0.208053, DF:: (7434, 8)
Construção do modelo K-Means para o 1º K foi: 0.007535 (s)
1º K: 87.734703 (s)
Construção do modelo K-Means para o 2º K foi: 0.011200 (s)
2º K: 105.823318 (s)
Construção do modelo K-Means grupo único: 0.008535 (s)
Grupo único: 34.032537 (s)
Total de beneficiários : Todos atributos independentes
Coeficiente de silhueta:: 7.102133 (s)
1º K: 2 - C.S.: 0.242956, 2º K: 3 - C.S.: 0.158397, DF:: (7434, 20)
Construção do modelo K-Means para o 1º K foi: 0.008889 (s)
1º K: 92.744694 (s)
Construção do modelo K-Means para o 2º K foi: 0.013375 (s)
2º K: 105.287592 (s)
Construção do modelo K-Means grupo único: 0.010834 (s)
Grupo único: 37.375701 (s)
Valor total repassado : Idese
Coeficiente de silhueta:: 7.099563 (s)
1º K: 3 - C.S.: 0.537305, 2º K: 2 - C.S.: 0.520650, DF:: (7434, 2)
Construção do modelo K-Means para o 1º K foi: 0.006405 (s)
1º K: 100.778869 (s)
Construção do modelo K-Means para o 2º K foi: 0.007525 (s)
2º K: 91.389114 (s)
Construção do modelo K-Means grupo único: 0.007257 (s)
Grupo único: 18.373914 (s)
Valor total repassado : Blocos individuais
Coeficiente de silhueta:: 8.576737 (s)
1º K: 2 - C.S.: 0.360506, 2º K: 3 - C.S.: 0.273212, DF:: (7434, 4)
Construção do modelo K-Means para o 1º K foi: 0.006001 (s)
1º K: 90.757864 (s)
Construção do modelo K-Means para o 2º K foi: 0.015154 (s)
2º K: 106.834229 (s)
Construção do modelo K-Means grupo único: 0.008032 (s)
Grupo único: 34.074451 (s)
Valor total repassado : Bloco educação individual
Coeficiente de silhueta:: 7.073659 (s)
1º K: 2 - C.S.: 0.553856, 2º K: 3 - C.S.: 0.546758, DF:: (7434, 2)
Construção do modelo K-Means para o 1º K foi: 0.005716 (s)
1º K: 92.192519 (s)
Construção do modelo K-Means para o 2º K foi: 0.009691 (s)
2º K: 106.212622 (s)
Construção do modelo K-Means grupo único: 0.010397 (s)
Grupo único: 18.632992 (s)
Valor total repassado : Bloco educação resumido
Coeficiente de silhueta:: 7.048540 (s)
1º K: 2 - C.S.: 0.290802, 2º K: 3 - C.S.: 0.256799, DF:: (7434, 5)
Construção do modelo K-Means para o 1º K foi: 0.005767 (s)
1º K: 97.037477 (s)
Construção do modelo K-Means para o 2º K foi: 0.011667 (s)
2º K: 110.052137 (s)
Construção do modelo K-Means grupo único: 0.009166 (s)
Grupo único: 24.608425 (s)
Valor total repassado : Bloco educação completo
Coeficiente de silhueta:: 7.366034 (s)
1º K: 2 - C.S.: 0.272316, 2º K: 3 - C.S.: 0.221505, DF:: (7434, 7)
Construção do modelo K-Means para o 1º K foi: 0.006542 (s)
1º K: 95.048388 (s)
Construção do modelo K-Means para o 2º K foi: 0.012565 (s)
2º K: 109.436247 (s)
Construção do modelo K-Means grupo único: 0.009110 (s)
Grupo único: 26.583833 (s)
```

```
Valor total repassado : Bloco renda individual
Coeficiente de silhueta:: 6.944724 (s)
1° K: 2 - C.S.: 0.524568, 2° K: 3 - C.S.: 0.517998, DF:: (7434, 2)
Construção do modelo K-Means para o 1° K foi: 0.008589 (s)
1° K: 90.965175 (s)
Construção do modelo K-Means para o 2° K foi: 0.010862 (s)
2° K: 101.078344 (s)
Construção do modelo K-Means grupo único: 0.008088 (s)
Grupo único: 18.910074 (s)
Valor total repassado : Bloco renda completo
Coeficiente de silhueta:: 7.642691 (s)
1° K: 3 - C.S.: 0.393162, 2° K: 2 - C.S.: 0.352365, DF:: (7434, 3)
Construção do modelo K-Means para o 1° K foi: 0.005860 (s)
1° K: 105.097318 (s)
Construção do modelo K-Means para o 2° K foi: 0.007348 (s)
2° K: 92.020540 (s)
Construção do modelo K-Means grupo único: 0.008414 (s)
Grupo único: 17.605312 (s)
Valor total repassado : Bloco saúde individual
Coeficiente de silhueta:: 8.312186 (s)
1° K: 3 - C.S.: 0.528154, 2° K: 2 - C.S.: 0.516628, DF:: (7434, 2)
Construção do modelo K-Means para o 1° K foi: 0.006670 (s)
1° K: 104.381374 (s)
Construção do modelo K-Means para o 2° K foi: 0.008507 (s)
2° K: 91.210654 (s)
Construção do modelo K-Means grupo único: 0.008070 (s)
Grupo único: 17.623394 (s)
Valor total repassado : Bloco saúde resumido
Coeficiente de silhueta:: 7.213667 (s)
1° K: 2 - C.S.: 0.355962, 2° K: 3 - C.S.: 0.266236, DF:: (7434, 4)
Construção do modelo K-Means para o 1° K foi: 0.007608 (s)
1° K: 88.848030 (s)
Construção do modelo K-Means para o 2° K foi: 0.010901 (s)
2° K: 106.970232 (s)
Construção do modelo K-Means grupo único: 0.007880 (s)
Grupo único: 24.306758 (s)
Valor total repassado : Bloco saúde completo
Coeficiente de silhueta:: 6.936910 (s)
1° K: 2 - C.S.: 0.292778, 2° K: 3 - C.S.: 0.208436, DF:: (7434, 8)
Construção do modelo K-Means para o 1° K foi: 0.005052 (s)
1° K: 90.507273 (s)
Construção do modelo K-Means para o 2° K foi: 0.011281 (s)
2° K: 106.403633 (s)
Construção do modelo K-Means grupo único: 0.008744 (s)
Grupo único: 34.720691 (s)
Valor total repassado : Todos atributos independentes
Coeficiente de silhueta:: 6.885610 (s)
1° K: 2 - C.S.: 0.243331, 2° K: 3 - C.S.: 0.158536, DF:: (7434, 20)
Construção do modelo K-Means para o 1° K foi: 0.006783 (s)
1° K: 93.204648 (s)
Construção do modelo K-Means para o 2° K foi: 0.015273 (s)
2° K: 109.180118 (s)
Construção do modelo K-Means grupo único: 0.010256 (s)
Grupo único: 35.675564 (s)
População beneficiária : Idese
Coeficiente de silhueta:: 6.902123 (s)
1° K: 2 - C.S.: 0.477357, 2° K: 3 - C.S.: 0.397674, DF:: (7434, 2)
Construção do modelo K-Means para o 1° K foi: 0.005918 (s)
1° K: 85.725435 (s)
Construção do modelo K-Means para o 2° K foi: 0.008562 (s)
2° K: 108.142617 (s)
Construção do modelo K-Means grupo único: 0.008120 (s)
Grupo único: 35.612403 (s)
População beneficiária : Blocos individuais
Coeficiente de silhueta:: 8.356569 (s)
1° K: 2 - C.S.: 0.364928, 2° K: 3 - C.S.: 0.264881, DF:: (7434, 4)
Construção do modelo K-Means para o 1° K foi: 0.007727 (s)
1° K: 88.455571 (s)
Construção do modelo K-Means para o 2° K foi: 0.010648 (s)
2° K: 103.015092 (s)
```

```
Construção do modelo K-Means grupo único: 0.008472 (s)
Grupo único: 46.732973 (s)
População beneficiária : Bloco educação individual
Coeficiente de silhueta:: 6.735779 (s)
1° K: 2 - C.S.: 0.453976, 2° K: 4 - C.S.: 0.388037, DF:: (7434, 2)
Construção do modelo K-Means para o 1° K foi: 0.006074 (s)
1° K: 90.314925 (s)
Construção do modelo K-Means para o 2° K foi: 0.010379 (s)
2° K: 125.113823 (s)
Construção do modelo K-Means grupo único: 0.008249 (s)
Grupo único: 37.417499 (s)
População beneficiária : Bloco educação resumido
Coeficiente de silhueta:: 7.145893 (s)
1° K: 2 - C.S.: 0.259770, 2° K: 4 - C.S.: 0.218340, DF:: (7434, 5)
Construção do modelo K-Means para o 1° K foi: 0.007923 (s)
1° K: 92.267194 (s)
Construção do modelo K-Means para o 2° K foi: 0.010234 (s)
2° K: 125.147880 (s)
Construção do modelo K-Means grupo único: 0.008781 (s)
Grupo único: 39.041533 (s)
População beneficiária : Bloco educação completo
Coeficiente de silhueta:: 6.943757 (s)
1° K: 2 - C.S.: 0.253525, 2° K: 3 - C.S.: 0.192592, DF:: (7434, 7)
Construção do modelo K-Means para o 1° K foi: 0.007049 (s)
1° K: 91.757229 (s)
Construção do modelo K-Means para o 2° K foi: 0.012516 (s)
2° K: 104.740875 (s)
Construção do modelo K-Means grupo único: 0.008542 (s)
Grupo único: 31.657755 (s)
População beneficiária : Bloco renda individual
Coeficiente de silhueta:: 6.827287 (s)
1° K: 2 - C.S.: 0.466590, 2° K: 3 - C.S.: 0.383779, DF:: (7434, 2)
Construção do modelo K-Means para o 1° K foi: 0.005045 (s)
1° K: 87.733221 (s)
Construção do modelo K-Means para o 2° K foi: 0.010795 (s)
2° K: 106.563718 (s)
Construção do modelo K-Means grupo único: 0.007798 (s)
Grupo único: 34.642719 (s)
População beneficiária : Bloco renda completo
Coeficiente de silhueta:: 7.095607 (s)
1° K: 2 - C.S.: 0.377793, 2° K: 4 - C.S.: 0.308245, DF:: (7434, 3)
Construção do modelo K-Means para o 1° K foi: 0.005249 (s)
1° K: 88.064004 (s)
Construção do modelo K-Means para o 2° K foi: 0.011751 (s)
2° K: 124.400996 (s)
Construção do modelo K-Means grupo único: 0.009111 (s)
Grupo único: 39.030643 (s)
População beneficiária : Bloco saúde individual
Coeficiente de silhueta:: 6.639563 (s)
1° K: 2 - C.S.: 0.446328, 2° K: 3 - C.S.: 0.392931, DF:: (7434, 2)
Construção do modelo K-Means para o 1° K foi: 0.005457 (s)
1° K: 88.891999 (s)
Construção do modelo K-Means para o 2° K foi: 0.009717 (s)
2° K: 93.764634 (s)
Construção do modelo K-Means grupo único: 0.007585 (s)
Grupo único: 44.546107 (s)
População beneficiária : Bloco saúde resumido
Coeficiente de silhueta:: 7.123559 (s)
1° K: 2 - C.S.: 0.331665, 2° K: 3 - C.S.: 0.299971, DF:: (7434, 4)
Construção do modelo K-Means para o 1° K foi: 0.004916 (s)
1° K: 89.584626 (s)
Construção do modelo K-Means para o 2° K foi: 0.010545 (s)
2° K: 101.382140 (s)
Construção do modelo K-Means grupo único: 0.007711 (s)
Grupo único: 37.044296 (s)
População beneficiária : Bloco saúde completo
Coeficiente de silhueta:: 7.031290 (s)
1° K: 2 - C.S.: 0.275947, 2° K: 4 - C.S.: 0.203975, DF:: (7434, 8)
Construção do modelo K-Means para o 1° K foi: 0.006016 (s)
1° K: 89.920421 (s)
```

```
Construção do modelo K-Means para o 2° K foi: 0.009937 (s)
2° K: 125.518760 (s)
Construção do modelo K-Means grupo único: 0.009240 (s)
Grupo único: 32.069821 (s)
População beneficiária : Todos atributos independentes
Coeficiente de silhueta:: 7.353611 (s)
1° K: 2 - C.S.: 0.243200, 2° K: 3 - C.S.: 0.161977, DF:: (7434, 20)
Construção do modelo K-Means para o 1° K foi: 0.007710 (s)
1° K: 94.148044 (s)
Construção do modelo K-Means para o 2° K foi: 0.013876 (s)
2° K: 118.721153 (s)
Construção do modelo K-Means grupo único: 0.015601 (s)
Grupo único: 27.718568 (s)
Repasse por beneficiário : Idese
Coeficiente de silhueta:: 6.772759 (s)
1° K: 2 - C.S.: 0.464510, 2° K: 3 - C.S.: 0.382956, DF:: (7434, 2)
Construção do modelo K-Means para o 1° K foi: 0.006033 (s)
1° K: 89.904758 (s)
Construção do modelo K-Means para o 2° K foi: 0.010453 (s)
2° K: 107.102491 (s)
Construção do modelo K-Means grupo único: 0.007419 (s)
Grupo único: 41.551978 (s)
Repasse por beneficiário : Blocos individuais
Coeficiente de silhueta:: 7.124725 (s)
1° K: 2 - C.S.: 0.347588, 2° K: 3 - C.S.: 0.256742, DF:: (7434, 4)
Construção do modelo K-Means para o 1° K foi: 0.005888 (s)
1° K: 90.459869 (s)
Construção do modelo K-Means para o 2° K foi: 0.010566 (s)
2° K: 106.854983 (s)
Construção do modelo K-Means grupo único: 0.007622 (s)
Grupo único: 34.064634 (s)
Repasse por beneficiário : Bloco educação individual
Coeficiente de silhueta:: 6.753816 (s)
1° K: 2 - C.S.: 0.477976, 2° K: 3 - C.S.: 0.379446, DF:: (7434, 2)
Construção do modelo K-Means para o 1° K foi: 0.005079 (s)
1° K: 94.713740 (s)
Construção do modelo K-Means para o 2° K foi: 0.009918 (s)
2° K: 102.845120 (s)
Construção do modelo K-Means grupo único: 0.008223 (s)
Grupo único: 32.504367 (s)
Repasse por beneficiário : Bloco educação resumido
Coeficiente de silhueta:: 7.059442 (s)
1° K: 2 - C.S.: 0.285261, 2° K: 3 - C.S.: 0.252415, DF:: (7434, 5)
Construção do modelo K-Means para o 1° K foi: 0.009098 (s)
1° K: 97.471944 (s)
Construção do modelo K-Means para o 2° K foi: 0.009349 (s)
2° K: 113.961190 (s)
Construção do modelo K-Means grupo único: 0.009270 (s)
Grupo único: 34.796364 (s)
Repasse por beneficiário : Bloco educação completo
Coeficiente de silhueta:: 7.180530 (s)
1° K: 2 - C.S.: 0.269835, 2° K: 3 - C.S.: 0.219819, DF:: (7434, 7)
Construção do modelo K-Means para o 1° K foi: 0.008059 (s)
1° K: 99.034890 (s)
Construção do modelo K-Means para o 2° K foi: 0.010733 (s)
2° K: 115.952580 (s)
Construção do modelo K-Means grupo único: 0.008844 (s)
Grupo único: 39.114889 (s)
Repasse por beneficiário : Bloco renda individual
Coeficiente de silhueta:: 6.927564 (s)
1° K: 2 - C.S.: 0.445305, 2° K: 4 - C.S.: 0.376407, DF:: (7434, 2)
Construção do modelo K-Means para o 1° K foi: 0.004845 (s)
1° K: 88.287323 (s)
Construção do modelo K-Means para o 2° K foi: 0.010573 (s)
2° K: 125.876928 (s)
Construção do modelo K-Means grupo único: 0.008197 (s)
Grupo único: 38.169388 (s)
Repasse por beneficiário : Bloco renda completo
Coeficiente de silhueta:: 7.448970 (s)
1° K: 2 - C.S.: 0.370199, 2° K: 3 - C.S.: 0.293485, DF:: (7434, 3)
```

```
Construção do modelo K-Means para o 1° K foi: 0.006492 (s)
1° K: 88.307724 (s)
Construção do modelo K-Means para o 2° K foi: 0.009830 (s)
2° K: 109.101036 (s)
Construção do modelo K-Means grupo único: 0.007992 (s)
Grupo único: 37.311686 (s)
Repasse por beneficiário : Bloco saúde individual
Coeficiente de silhueta:: 6.960589 (s)
1° K: 2 - C.S.: 0.448701, 2° K: 3 - C.S.: 0.362311, DF:: (7434, 2)
Construção do modelo K-Means para o 1° K foi: 0.004507 (s)
1° K: 85.918139 (s)
Construção do modelo K-Means para o 2° K foi: 0.009488 (s)
2° K: 105.952016 (s)
Construção do modelo K-Means grupo único: 0.008489 (s)
Grupo único: 36.983808 (s)
Repasse por beneficiário : Bloco saúde resumido
Coeficiente de silhueta:: 7.108964 (s)
1° K: 2 - C.S.: 0.339176, 2° K: 3 - C.S.: 0.239870, DF:: (7434, 4)
Construção do modelo K-Means para o 1° K foi: 0.006957 (s)
1° K: 80.685475 (s)
Construção do modelo K-Means para o 2° K foi: 0.011404 (s)
2° K: 107.328115 (s)
Construção do modelo K-Means grupo único: 0.008269 (s)
Grupo único: 36.464365 (s)
Repasse por beneficiário : Bloco saúde completo
Coeficiente de silhueta:: 7.256271 (s)
1° K: 2 - C.S.: 0.283908, 2° K: 3 - C.S.: 0.205472, DF:: (7434, 8)
Construção do modelo K-Means para o 1° K foi: 0.007195 (s)
1° K: 92.014190 (s)
Construção do modelo K-Means para o 2° K foi: 0.009252 (s)
2° K: 113.257757 (s)
Construção do modelo K-Means grupo único: 0.009141 (s)
Grupo único: 44.013763 (s)
Repasse por beneficiário : Todos atributos independentes
Coeficiente de silhueta:: 7.313766 (s)
1° K: 2 - C.S.: 0.241115, 2° K: 3 - C.S.: 0.157836, DF:: (7434, 20)
Construção do modelo K-Means para o 1° K foi: 0.007083 (s)
1° K: 90.199626 (s)
Construção do modelo K-Means para o 2° K foi: 0.015307 (s)
2° K: 118.658577 (s)
Construção do modelo K-Means grupo único: 0.015815 (s)
Grupo único: 36.586316 (s)
Repasse por população : Idese
Coeficiente de silhueta:: 6.951428 (s)
1° K: 2 - C.S.: 0.448479, 2° K: 4 - C.S.: 0.391006, DF:: (7434, 2)
Construção do modelo K-Means para o 1° K foi: 0.004815 (s)
1° K: 86.610358 (s)
Construção do modelo K-Means para o 2° K foi: 0.011177 (s)
2° K: 109.865349 (s)
Construção do modelo K-Means grupo único: 0.008094 (s)
Grupo único: 27.720141 (s)
Repasse por população : Blocos individuais
Coeficiente de silhueta:: 6.957956 (s)
1° K: 2 - C.S.: 0.341260, 2° K: 3 - C.S.: 0.255171, DF:: (7434, 4)
Construção do modelo K-Means para o 1° K foi: 0.005247 (s)
1° K: 89.680499 (s)
Construção do modelo K-Means para o 2° K foi: 0.009943 (s)
2° K: 111.584706 (s)
Construção do modelo K-Means grupo único: 0.008163 (s)
Grupo único: 36.512341 (s)
Repasse por população : Bloco educação individual
Coeficiente de silhueta:: 6.849406 (s)
1° K: 3 - C.S.: 0.440320, 2° K: 2 - C.S.: 0.429449, DF:: (7434, 2)
Construção do modelo K-Means para o 1° K foi: 0.006950 (s)
1° K: 103.116719 (s)
Construção do modelo K-Means para o 2° K foi: 0.007793 (s)
2° K: 95.598165 (s)
Construção do modelo K-Means grupo único: 0.008111 (s)
Grupo único: 33.899424 (s)
Repasse por população : Bloco educação resumido
```

```
Coeficiente de silhueta:: 7.164501 (s)
1° K: 2 - C.S.: 0.273719, 2° K: 3 - C.S.: 0.237282, DF:: (7434, 5)
Construção do modelo K-Means para o 1° K foi: 0.016686 (s)
1° K: 95.427308 (s)
Construção do modelo K-Means para o 2° K foi: 0.011117 (s)
2° K: 109.140706 (s)
Construção do modelo K-Means grupo único: 0.009310 (s)
Grupo único: 25.207258 (s)
Repasse por população : Bloco educação completo
Coeficiente de silhueta:: 7.035523 (s)
1° K: 2 - C.S.: 0.253937, 2° K: 3 - C.S.: 0.208614, DF:: (7434, 7)
Construção do modelo K-Means para o 1° K foi: 0.007778 (s)
1° K: 91.923582 (s)
Construção do modelo K-Means para o 2° K foi: 0.011288 (s)
2° K: 108.868301 (s)
Construção do modelo K-Means grupo único: 0.008378 (s)
Grupo único: 29.381523 (s)
Repasse por população : Bloco renda individual
Coeficiente de silhueta:: 7.109470 (s)
1° K: 2 - C.S.: 0.431054, 2° K: 3 - C.S.: 0.389935, DF:: (7434, 2)
Construção do modelo K-Means para o 1° K foi: 0.006686 (s)
1° K: 89.023917 (s)
Construção do modelo K-Means para o 2° K foi: 0.008742 (s)
2° K: 108.295444 (s)
Construção do modelo K-Means grupo único: 0.008679 (s)
Grupo único: 36.472285 (s)
Repasse por população : Bloco renda completo
Coeficiente de silhueta:: 7.083111 (s)
1° K: 2 - C.S.: 0.357028, 2° K: 5 - C.S.: 0.274964, DF:: (7434, 3)
Construção do modelo K-Means para o 1° K foi: 0.007539 (s)
1° K: 89.111268 (s)
Construção do modelo K-Means para o 2° K foi: 0.014363 (s)
2° K: 143.392123 (s)
Construção do modelo K-Means grupo único: 0.007905 (s)
Grupo único: 21.451733 (s)
Repasse por população : Bloco saúde individual
Coeficiente de silhueta:: 7.314683 (s)
1° K: 3 - C.S.: 0.434294, 2° K: 2 - C.S.: 0.424579, DF:: (7434, 2)
Construção do modelo K-Means para o 1° K foi: 0.005579 (s)
1° K: 103.475652 (s)
Construção do modelo K-Means para o 2° K foi: 0.007150 (s)
2° K: 90.000892 (s)
Construção do modelo K-Means grupo único: 0.008008 (s)
Grupo único: 39.034781 (s)
Repasse por população : Bloco saúde resumido
Coeficiente de silhueta:: 7.463981 (s)
1° K: 2 - C.S.: 0.324487, 2° K: 4 - C.S.: 0.239905, DF:: (7434, 4)
Construção do modelo K-Means para o 1° K foi: 0.004688 (s)
1° K: 92.042374 (s)
Construção do modelo K-Means para o 2° K foi: 0.010554 (s)
2° K: 121.666253 (s)
Construção do modelo K-Means grupo único: 0.008461 (s)
Grupo único: 30.847444 (s)
Repasse por população : Bloco saúde completo
Coeficiente de silhueta:: 6.742161 (s)
1° K: 2 - C.S.: 0.277927, 2° K: 3 - C.S.: 0.197129, DF:: (7434, 8)
Construção do modelo K-Means para o 1° K foi: 0.006718 (s)
1° K: 90.678636 (s)
Construção do modelo K-Means para o 2° K foi: 0.011117 (s)
2° K: 112.836166 (s)
Construção do modelo K-Means grupo único: 0.008905 (s)
Grupo único: 35.457422 (s)
Repasse por população : Todos atributos independentes
Coeficiente de silhueta:: 7.133969 (s)
1° K: 2 - C.S.: 0.240114, 2° K: 3 - C.S.: 0.158106, DF:: (7434, 20)
Construção do modelo K-Means para o 1° K foi: 0.006062 (s)
1° K: 94.540501 (s)
Construção do modelo K-Means para o 2° K foi: 0.015470 (s)
2° K: 115.866849 (s)
Construção do modelo K-Means grupo único: 0.010867 (s)
```

```
Grupo único: 43.861195 (s)
```

B.2 Resultados da melhor divisão

```
Combinação:Total de beneficiários | Todos atributos independentes
 Modelo: XGBoost |  Grupo: 0
R² Médio: 0.813919 |  MSE Médio: 0.001896
Amostras no grupo: 3220
População média: 23876.88
Informações médias das amostras:
Total de beneficiários: 1134.43
 Equivalente a (37.50%) do valor médio: (825.03)
Características IDESE:
                                Grupo  Original  Erro(%)
Idese                          0.6670    0.7253    -8.04
Bloco Educação                 0.6264    0.6914    -9.40
Bloco Renda                    0.5698    0.6482   -12.09
Bloco Saúde                    0.8048    0.8363    -3.77
Anos Finais EF                 0.6418    0.6780    -5.34
Anos Iniciais EF               0.7030    0.7608    -7.60
Ensino Fundamental             0.6724    0.7194    -6.54
Ensino Médio                   0.7119    0.7607    -6.43
Escolaridade Adulta            0.4135    0.4538    -8.86
Pré Escola                     0.7077    0.8317   -14.90
Apropriação da Renda           0.5789    0.6650   -12.94
Geração da Renda               0.5607    0.6314   -11.19
Mortes por Causas Evitáveis    0.5850    0.6187    -5.44
Condições Gerais de Saúde      0.7319    0.7609    -3.81
Óbitos por Causas Mal Definidas  0.8788    0.9032    -2.70
Longevidade                    0.8684    0.8982    -3.31
Consultas Pré Natal            0.7180    0.7726    -7.07
Mortalidade de Menores de 5 anos  0.9101    0.9269    -1.81
Saúde Materno Infantil         0.8141    0.8498    -4.20

Combinação:Valor total repassado | Todos atributos independentes
 Modelo: XGBoost |  Grupo: 0
R² Médio: 0.793511 |  MSE Médio: 0.00196
Amostras no grupo: 3225
População média: 23846.97
Informações médias das amostras:
Valor total repassado: 1502126.83
 Equivalente a (27.92%) do valor médio: (1174245.23)
Características IDESE:
                                Grupo  Original  Erro(%)
Idese                          0.6671    0.7253    -8.03
Bloco Educação                 0.6264    0.6914    -9.40
Bloco Renda                    0.5699    0.6482   -12.08
Bloco Saúde                    0.8049    0.8363    -3.76
Anos Finais EF                 0.6418    0.6780    -5.33
Anos Iniciais EF               0.7030    0.7608    -7.59
Ensino Fundamental             0.6724    0.7194    -6.53
Ensino Médio                   0.7118    0.7607    -6.44
Escolaridade Adulta            0.4135    0.4538    -8.86
Pré Escola                     0.7080    0.8317   -14.87
Apropriação da Renda           0.5790    0.6650   -12.93
Geração da Renda               0.5608    0.6314   -11.17
Mortes por Causas Evitáveis    0.5851    0.6187    -5.43
Condições Gerais de Saúde      0.7320    0.7609    -3.81
Óbitos por Causas Mal Definidas  0.8788    0.9032    -2.70
Longevidade                    0.8685    0.8982    -3.30
Consultas Pré Natal            0.7181    0.7726    -7.06
Mortalidade de Menores de 5 anos  0.9102    0.9269    -1.81
Saúde Materno Infantil         0.8141    0.8498    -4.19

Combinação:Total de beneficiários | Todos atributos independentes
 Modelo: XGBoost |  Grupo: 1
```

R² Médio: 0.7388 | MSE Médio: 0.000649
Amostras no grupo: 4229
População média: 20896.35
Informações médias das amostras:
Total de beneficiários: 589.45
 Equivalente a (-28.55%) do valor médio: (825.03)
Características IDESE:

	Grupo	Original	Erro(%)
Idese	0.7697	0.7253	6.12
Bloco Educação	0.7409	0.6914	7.16
Bloco Renda	0.7078	0.6482	9.20
Bloco Saúde	0.8603	0.8363	2.87
Anos Finais EF	0.7056	0.6780	4.07
Anos Iniciais EF	0.8048	0.7608	5.79
Ensino Fundamental	0.7552	0.7194	4.98
Ensino Médio	0.7980	0.7607	4.89
Escolaridade Adulta	0.4844	0.4538	6.75
Pré Escola	0.9261	0.8317	11.35
Apropriação da Renda	0.7305	0.6650	9.85
Geração da Renda	0.6852	0.6314	8.52
Mortes por Causas Evitáveis	0.6443	0.6187	4.14
Condições Gerais de Saúde	0.7830	0.7609	2.90
Óbitos por Causas Mal Definidas	0.9217	0.9032	2.06
Longevidade	0.9208	0.8982	2.52
Consultas Pré Natal	0.8142	0.7726	5.38
Mortalidade de Menores de 5 anos	0.9397	0.9269	1.38
Saúde Materno Infantil	0.8770	0.8498	3.20

Combinação:Total de beneficiários | Bloco educação completo
 Modelo: XGBoost | Grupo: 0
R² Médio: 0.729089 | MSE Médio: 0.002964
Amostras no grupo: 2493
População média: 29772.75
Informações médias das amostras:
Total de beneficiários: 1265.94
 Equivalente a (53.44%) do valor médio: (825.03)
Características IDESE:

	Grupo	Original	Erro(%)
Anos Finais EF	0.6282	0.6780	-7.34
Anos Iniciais EF	0.6773	0.7608	-10.98
Ensino Fundamental	0.6527	0.7194	-9.27
Ensino Médio	0.7429	0.7607	-2.35
Escolaridade Adulta	0.4051	0.4538	-10.73
Pré Escola	0.6266	0.8317	-24.65

Combinação:Valor total repassado | Todos atributos independentes
 Modelo: XGBoost | Grupo: 1
R² Médio: 0.714004 | MSE Médio: 0.000877
Amostras no grupo: 4224
População média: 20915.65
Informações médias das amostras:
Valor total repassado: 923909.49
 Equivalente a (-21.32%) do valor médio: (1174245.23)
Características IDESE:

	Grupo	Original	Erro(%)
Idese	0.7697	0.7253	6.13
Bloco Educação	0.7410	0.6914	7.17
Bloco Renda	0.7079	0.6482	9.22
Bloco Saúde	0.8603	0.8363	2.87
Anos Finais EF	0.7056	0.6780	4.07
Anos Iniciais EF	0.8049	0.7608	5.80
Ensino Fundamental	0.7552	0.7194	4.98
Ensino Médio	0.7981	0.7607	4.91
Escolaridade Adulta	0.4845	0.4538	6.77
Pré Escola	0.9261	0.8317	11.36
Apropriação da Renda	0.7306	0.6650	9.87
Geração da Renda	0.6852	0.6314	8.53
Mortes por Causas Evitáveis	0.6443	0.6187	4.14
Condições Gerais de Saúde	0.7831	0.7609	2.91
Óbitos por Causas Mal Definidas	0.9218	0.9032	2.06

Longevidade	0.9208	0.8982	2.52
Consultas Pré Natal	0.8143	0.7726	5.39
Mortalidade de Menores de 5 anos	0.9397	0.9269	1.38
Saúde Materno Infantil	0.8770	0.8498	3.20

Combinação:Total de beneficiários | Todos atributos independentes
Modelo: MLP | Grupo: 0
R² Médio: 0.702486 | MSE Médio: 0.002955
Amostras no grupo: 3220
População média: 23876.88
Informações médias das amostras:
Total de beneficiários: 1134.43
Equivalente a (37.50%) do valor médio: (825.03)
Características IDESE:

	Grupo	Original	Erro(%)
Idese	0.6670	0.7253	-8.04
Bloco Educação	0.6264	0.6914	-9.40
Bloco Renda	0.5698	0.6482	-12.09
Bloco Saúde	0.8048	0.8363	-3.77
Anos Finais EF	0.6418	0.6780	-5.34
Anos Iniciais EF	0.7030	0.7608	-7.60
Ensino Fundamental	0.6724	0.7194	-6.54
Ensino Médio	0.7119	0.7607	-6.43
Escolaridade Adulta	0.4135	0.4538	-8.86
Pré Escola	0.7077	0.8317	-14.90
Apropriação da Renda	0.5789	0.6650	-12.94
Geração da Renda	0.5607	0.6314	-11.19
Mortes por Causas Evitáveis	0.5850	0.6187	-5.44
Condições Gerais de Saúde	0.7319	0.7609	-3.81
Óbitos por Causas Mal Definidas	0.8788	0.9032	-2.70
Longevidade	0.8684	0.8982	-3.31
Consultas Pré Natal	0.7180	0.7726	-7.07
Mortalidade de Menores de 5 anos	0.9101	0.9269	-1.81
Saúde Materno Infantil	0.8141	0.8498	-4.20

Combinação:Valor total repassado | Bloco educação completo
Modelo: XGBoost | Grupo: 0
R² Médio: 0.701422 | MSE Médio: 0.002801
Amostras no grupo: 2499
População média: 29535.92
Informações médias das amostras:
Valor total repassado: 1594735.09
Equivalente a (35.81%) do valor médio: (1174245.23)
Características IDESE:

	Grupo	Original	Erro(%)
Anos Finais EF	0.6281	0.6780	-7.35
Anos Iniciais EF	0.6773	0.7608	-10.98
Ensino Fundamental	0.6527	0.7194	-9.27
Ensino Médio	0.7423	0.7607	-2.42
Escolaridade Adulta	0.4050	0.4538	-10.74
Pré Escola	0.6277	0.8317	-24.53

Combinação:Total de beneficiários | Bloco educação completo
Modelo: XGBoost | Grupo: 1
R² Médio: 0.653863 | MSE Médio: 0.00117
Amostras no grupo: 4956
População média: 18367.79
Informações médias das amostras:
Total de beneficiários: 603.24
Equivalente a (-26.88%) do valor médio: (825.03)
Características IDESE:

	Grupo	Original	Erro(%)
Anos Finais EF	0.7030	0.6780	3.69
Anos Iniciais EF	0.8028	0.7608	5.52
Ensino Fundamental	0.7529	0.7194	4.66
Ensino Médio	0.7697	0.7607	1.18
Escolaridade Adulta	0.4782	0.4538	5.40
Pré Escola	0.9348	0.8317	12.40

Combinação:Total de beneficiários | Bloco educação resumido

```
 Modelo: XGBoost |  Grupo: 0
R² Médio: 0.615518 |  MSE Médio: 0.00425
Amostras no grupo: 2099
População média: 28479.52
Informações médias das amostras:
Total de beneficiários: 1218.78
 Equivalente a (47.73%) do valor médio: (825.03)
Características IDESE:
                     Grupo  Original  Erro(%)
Ensino Fundamental   0.6573    0.7194    -8.62
Ensino Médio         0.7365    0.7607    -3.19
Escolaridade Adulta  0.3970    0.4538   -12.52
Pré Escola           0.5726    0.8317   -31.16

Combinação:Total de beneficiários | Todos atributos independentes
 Modelo: MLP |  Grupo: 1
R² Médio: 0.608982 |  MSE Médio: 0.000994
Amostras no grupo: 4229
População média: 20896.35
Informações médias das amostras:
Total de beneficiários: 589.45
 Equivalente a (-28.55%) do valor médio: (825.03)
Características IDESE:
                                 Grupo  Original  Erro(%)
Idese                           0.7697    0.7253     6.12
Bloco Educação                  0.7409    0.6914     7.16
Bloco Renda                     0.7078    0.6482     9.20
Bloco Saúde                     0.8603    0.8363     2.87
Anos Finais EF                  0.7056    0.6780     4.07
Anos Iniciais EF                0.8048    0.7608     5.79
Ensino Fundamental              0.7552    0.7194     4.98
Ensino Médio                    0.7980    0.7607     4.89
Escolaridade Adulta             0.4844    0.4538     6.75
Pré Escola                      0.9261    0.8317    11.35
Apropriação da Renda            0.7305    0.6650     9.85
Geração da Renda                0.6852    0.6314     8.52
Mortes por Causas Evitáveis     0.6443    0.6187     4.14
Condições Gerais de Saúde       0.7830    0.7609     2.90
Óbitos por Causas Mal Definidas 0.9217    0.9032     2.06
Longevidade                     0.9208    0.8982     2.52
Consultas Pré Natal             0.8142    0.7726     5.38
Mortalidade de Menores de 5 anos 0.9397   0.9269     1.38
Saúde Materno Infantil          0.8770    0.8498     3.20

Combinação:Valor total repassado | Bloco educação resumido
 Modelo: XGBoost |  Grupo: 0
R² Médio: 0.596458 |  MSE Médio: 0.003816
Amostras no grupo: 5347
População média: 20533.75

Informações médias das amostras:
Valor total repassado: 1082588.82
 Equivalente a (-7.81%) do valor médio: (1174245.23)
Características IDESE:
                     Grupo  Original  Erro(%)
Ensino Fundamental   0.7437    0.7194     3.38
Ensino Médio         0.7706    0.7607     1.29
Escolaridade Adulta  0.4764    0.4538     4.98
Pré Escola           0.9333    0.8317    12.21

Combinação:Total de beneficiários | Bloco educação resumido
 Modelo: MLP |  Grupo: 0
R² Médio: 0.570605 |  MSE Médio: 0.004811
Amostras no grupo: 2099
População média: 28479.52
Informações médias das amostras:
Total de beneficiários: 1218.78
 Equivalente a (47.73%) do valor médio: (825.03)
Características IDESE:
                     Grupo  Original  Erro(%)
```

```
Ensino Fundamental   0.6573    0.7194    -8.62
Ensino Médio         0.7365    0.7607    -3.19
Escolaridade Adulta  0.3970    0.4538   -12.52
Pré Escola           0.5726    0.8317   -31.16

Combinação:População beneficiária | Todos atributos independentes
 Modelo: XGBoost |  Grupo: 1
R² Médio: 0.553404 |  MSE Médio: 0.008506
Amostras no grupo: 4108
População média: 21815.68
Informações médias das amostras:
População beneficiária: 0.03
 Equivalente a (-35.79%) do valor médio: (0.04)
Características IDESE:
                               Grupo  Original  Erro(%)
Idese                         0.7710    0.7253     6.30
Bloco Educação                0.7413    0.6914     7.22
Bloco Renda                   0.7106    0.6482     9.64
Bloco Saúde                   0.8610    0.8363     2.95
Anos Finais EF                0.7072    0.6780     4.31
Anos Iniciais EF              0.8059    0.7608     5.93
Ensino Fundamental            0.7565    0.7194     5.16
Ensino Médio                  0.7988    0.7607     5.00
Escolaridade Adulta           0.4864    0.4538     7.19
Pré Escola                    0.9237    0.8317    11.06
Apropriação da Renda          0.7342    0.6650    10.42
Geração da Renda              0.6871    0.6314     8.82
Mortes por Causas Evitáveis   0.6462    0.6187     4.44
Condições Gerais de Saúde     0.7844    0.7609     3.09
Óbitos por Causas Mal Definidas  0.9227  0.9032    2.16
Longevidade                   0.9209    0.8982     2.53
Consultas Pré Natal           0.8152    0.7726     5.51
Mortalidade de Menores de 5 anos  0.9398  0.9269   1.39
Saúde Materno Infantil        0.8775    0.8498     3.26

Combinação:Valor total repassado | Bloco educação completo
 Modelo: MLP |  Grupo: 0
R² Médio: 0.535534 |  MSE Médio: 0.004369
Amostras no grupo: 2499
População média: 29535.92
Informações médias das amostras:
Valor total repassado: 1594735.09
 Equivalente a (35.81%) do valor médio: (1174245.23)
Características IDESE:
                     Grupo  Original  Erro(%)
Anos Finais EF       0.6281    0.6780    -7.35
Anos Iniciais EF     0.6773    0.7608   -10.98
Ensino Fundamental   0.6527    0.7194    -9.27
Ensino Médio         0.7423    0.7607    -2.42
Escolaridade Adulta  0.4050    0.4538   -10.74
Pré Escola           0.6277    0.8317   -24.53

Combinação:Valor total repassado | Bloco educação resumido
 Modelo: MLP |  Grupo: 0
R² Médio: 0.528581 |  MSE Médio: 0.004348
Amostras no grupo: 5347
População média: 20533.75
Informações médias das amostras:
Valor total repassado: 1082588.82
 Equivalente a (-7.81%) do valor médio: (1174245.23)
Características IDESE:
                     Grupo  Original  Erro(%)
Ensino Fundamental   0.7437    0.7194     3.38
Ensino Médio         0.7706    0.7607     1.29
Escolaridade Adulta  0.4764    0.4538     4.98
Pré Escola           0.9333    0.8317    12.21
```

B.3 Resultados da segunda melhor divisão

```
Combinação:Total de beneficiários | Todos atributos independentes
 Modelo: XGBoost |  Grupo: 2
R² Médio: 0.80834 |  MSE Médio: 0.00171
Amostras no grupo: 1600
População média: 24364.16
Informações médias das amostras:
Total de beneficiários: 1249.43
 Equivalente a (51.44%) do valor médio: (825.03)
Características IDESE:
                                 Grupo  Original  Erro(%)
Idese                           0.6363    0.7253   -12.26
Bloco Educação                  0.5774    0.6914   -16.48
Bloco Renda                     0.5361    0.6482   -17.29
Bloco Saúde                     0.7955    0.8363    -4.88
Anos Finais EF                  0.6252    0.6780    -7.79
Anos Iniciais EF                0.6710    0.7608   -11.80
Ensino Fundamental              0.6481    0.7194    -9.91
Ensino Médio                    0.7083    0.7607    -6.90
Escolaridade Adulta             0.3862    0.4538   -14.88
Pré Escola                      0.5671    0.8317   -31.82
Apropriação da Renda            0.5419    0.6650   -18.50
Geração da Renda                0.5303    0.6314   -16.01
Mortes por Causas Evitáveis     0.5929    0.6187    -4.18
Condições Gerais de Saúde       0.7320    0.7609    -3.80
Óbitos por Causas Mal Definidas 0.8712    0.9032    -3.54
Longevidade                     0.8617    0.8982    -4.06
Consultas Pré Natal             0.6892    0.7726   -10.80
Mortalidade de Menores de 5 anos 0.8962   0.9269    -3.32
Saúde Materno Infantil          0.7927    0.8498    -6.72

Combinação:Valor total repassado | Todos atributos independentes
 Modelo: XGBoost |  Grupo: 2
R² Médio: 0.791457 |  MSE Médio: 0.001582
Amostras no grupo: 1601
População média: 24355.59
Informações médias das amostras:
Valor total repassado: 1528641.59
 Equivalente a (30.18%) do valor médio: (1174245.23)
Características IDESE:
                                 Grupo  Original  Erro(%)
Idese                           0.6364    0.7253   -12.26
Bloco Educação                  0.5775    0.6914   -16.48
Bloco Renda                     0.5361    0.6482   -17.29
Bloco Saúde                     0.7955    0.8363    -4.88
Anos Finais EF                  0.6252    0.6780    -7.79
Anos Iniciais EF                0.6711    0.7608   -11.79
Ensino Fundamental              0.6481    0.7194    -9.91
Ensino Médio                    0.7083    0.7607    -6.89
Escolaridade Adulta             0.3863    0.4538   -14.87
Pré Escola                      0.5672    0.8317   -31.80
Apropriação da Renda            0.5419    0.6650   -18.50
Geração da Renda                0.5302    0.6314   -16.02
Mortes por Causas Evitáveis     0.5929    0.6187    -4.18
Condições Gerais de Saúde       0.7321    0.7609    -3.79
Óbitos por Causas Mal Definidas 0.8713    0.9032    -3.53
Longevidade                     0.8617    0.8982    -4.06
Consultas Pré Natal             0.6892    0.7726   -10.80
Mortalidade de Menores de 5 anos 0.8962   0.9269    -3.31
Saúde Materno Infantil          0.7927    0.8498    -6.71

Combinação:Total de beneficiários | Todos atributos independentes
 Modelo: XGBoost |  Grupo: 0
R² Médio: 0.772811 |  MSE Médio: 0.002438
Amostras no grupo: 3085
População média: 22256.44
Informações médias das amostras:
Total de beneficiários: 881.41
```

```
 Equivalente a (6.83%) do valor médio: (825.03)
Características IDESE:
                                 Grupo  Original  Erro(%)
Idese                           0.7157    0.7253    -1.33
Bloco Educação                  0.6952    0.6914     0.54
Bloco Renda                     0.6272    0.6482    -3.23
Bloco Saúde                     0.8246    0.8363    -1.40
Anos Finais EF                  0.6686    0.6780    -1.38
Anos Iniciais EF                0.7527    0.7608    -1.07
Ensino Fundamental              0.7106    0.7194    -1.22
Ensino Médio                    0.7456    0.7607    -1.99
Escolaridade Adulta             0.4491    0.4538    -1.02
Pré Escola                      0.8752    0.8317     5.24
Apropriação da Renda            0.6401    0.6650    -3.74
Geração da Renda                0.6144    0.6314    -2.69
Mortes por Causas Evitáveis     0.5863    0.6187    -5.24
Condições Gerais de Saúde       0.7415    0.7609    -2.55
Óbitos por Causas Mal Definidas 0.8968    0.9032    -0.71
Longevidade                     0.8842    0.8982    -1.56
Consultas Pré Natal             0.7674    0.7726    -0.68
Mortalidade de Menores de 5 anos 0.9287   0.9269     0.20
Saúde Materno Infantil          0.8481    0.8498    -0.20

Combinação:Valor total repassado | Todos atributos independentes
 Modelo: MLP |  Grupo: 2
R² Médio: 0.75167 |  MSE Médio: 0.001855
Amostras no grupo: 1601
População média: 24355.59
Informações médias das amostras:
Valor total repassado: 1528641.59
 Equivalente a (30.18%) do valor médio: (1174245.23)
Características IDESE:
                                 Grupo  Original  Erro(%)
Idese                           0.6364    0.7253   -12.26
Bloco Educação                  0.5775    0.6914   -16.48
Bloco Renda                     0.5361    0.6482   -17.29
Bloco Saúde                     0.7955    0.8363    -4.88
Anos Finais EF                  0.6252    0.6780    -7.79
Anos Iniciais EF                0.6711    0.7608   -11.79
Ensino Fundamental              0.6481    0.7194    -9.91
Ensino Médio                    0.7083    0.7607    -6.89
Escolaridade Adulta             0.3863    0.4538   -14.87
Pré Escola                      0.5672    0.8317   -31.80
Apropriação da Renda            0.5419    0.6650   -18.50
Geração da Renda                0.5302    0.6314   -16.02
Mortes por Causas Evitáveis     0.5929    0.6187    -4.18
Condições Gerais de Saúde       0.7321    0.7609    -3.79
Óbitos por Causas Mal Definidas 0.8713    0.9032    -3.53
Longevidade                     0.8617    0.8982    -4.06
Consultas Pré Natal             0.6892    0.7726   -10.80
Mortalidade de Menores de 5 anos 0.8962   0.9269    -3.31
Saúde Materno Infantil          0.7927    0.8498    -6.71
Combinação:Valor total repassado | Todos atributos independentes
 Modelo: XGBoost |  Grupo: 0
R² Médio: 0.745555 |  MSE Médio: 0.002715
Amostras no grupo: 3086
População média: 22254.71
Informações médias das amostras:
Valor total repassado: 1286812.34
 Equivalente a (9.59%) do valor médio: (1174245.23)
Características IDESE:
                                 Grupo  Original  Erro(%)
Idese                           0.7157    0.7253    -1.32
Bloco Educação                  0.6952    0.6914     0.55
Bloco Renda                     0.6273    0.6482    -3.22
Bloco Saúde                     0.8246    0.8363    -1.40
Anos Finais EF                  0.6687    0.6780    -1.38
Anos Iniciais EF                0.7527    0.7608    -1.06
Ensino Fundamental              0.7107    0.7194    -1.21
Ensino Médio                    0.7456    0.7607    -1.99
```

```
Escolaridade Adulta               0.4492    0.4538    -1.00
Pré Escola                        0.8753    0.8317     5.24
Apropriação da Renda              0.6402    0.6650    -3.72
Geração da Renda                  0.6144    0.6314    -2.69
Mortes por Causas Evitáveis       0.5863    0.6187    -5.24
Condições Gerais de Saúde         0.7415    0.7609    -2.55
Óbitos por Causas Mal Definidas   0.8967    0.9032    -0.71
Longevidade                       0.8842    0.8982    -1.56
Consultas Pré Natal               0.7675    0.7726    -0.67
Mortalidade de Menores de 5 anos  0.9287    0.9269     0.19
Saúde Materno Infantil            0.8481    0.8498    -0.20

Combinação:Total de beneficiários | Todos atributos independentes
 Modelo: XGBoost |  Grupo: 1
R² Médio: 0.7388 |  MSE Médio: 0.000649
Amostras no grupo: 2764
População média: 20843.14
Informações médias das amostras:
Total de beneficiários: 516.42
 Equivalente a (-37.41%) do valor médio: (825.03)
Características IDESE:
                                  Grupo  Original  Erro(%)
Idese                             0.7875    0.7253     8.58
Bloco Educação                    0.7532    0.6914     8.93
Bloco Renda                       0.7364    0.6482    13.61
Bloco Saúde                       0.8730    0.8363     4.39
Anos Finais EF                    0.7190    0.6780     6.05
Anos Iniciais EF                  0.8218    0.7608     8.02
Ensino Fundamental                0.7704    0.7194     7.09
Ensino Médio                      0.8080    0.7607     6.22
Escolaridade Adulta               0.4980    0.4538     9.75
Pré Escola                        0.9363    0.8317    12.57
Apropriação da Renda              0.7639    0.6650    14.88
Geração da Renda                  0.7089    0.6314    12.27
Mortes por Causas Evitáveis       0.6699    0.6187     8.27
Condições Gerais de Saúde         0.7993    0.7609     5.05
Óbitos por Causas Mal Definidas   0.9288    0.9032     2.84
Longevidade                       0.9349    0.8982     4.09
Consultas Pré Natal               0.8268    0.7726     7.01
Mortalidade de Menores de 5 anos  0.9427    0.9269     1.70
Saúde Materno Infantil            0.8847    0.8498     4.11
Combinação:Total de beneficiários | Bloco educação completo
 Modelo: XGBoost |  Grupo: 2
R² Médio: 0.73691 |  MSE Médio: 0.000807
Amostras no grupo: 2035
População média: 33108.67
Informações médias das amostras:
Total de beneficiários: 1394.35
 Equivalente a (69.01%) do valor médio: (825.03)
Características IDESE:
                     Grupo  Original  Erro(%)
Anos Finais EF      0.6243    0.6780    -7.91
Anos Iniciais EF    0.6698    0.7608   -11.96
Ensino Fundamental  0.6471    0.7194   -10.05
Ensino Médio        0.7533    0.7607    -0.98
Escolaridade Adulta 0.4032    0.4538   -11.14
Pré Escola          0.5824    0.8317   -29.98

Combinação:Total de beneficiários | Todos atributos independentes
 Modelo: MLP |  Grupo: 0
R² Médio: 0.735842 |  MSE Médio: 0.003105
Amostras no grupo: 3085
População média: 22256.44
Informações médias das amostras:
Total de beneficiários: 881.41
 Equivalente a (6.83%) do valor médio: (825.03)
Características IDESE:
                                  Grupo  Original  Erro(%)
Idese                             0.7157    0.7253    -1.33
Bloco Educação                    0.6952    0.6914     0.54
```

Bloco Renda	0.6272	0.6482	-3.23
Bloco Saúde	0.8246	0.8363	-1.40
Anos Finais EF	0.6686	0.6780	-1.38
Anos Iniciais EF	0.7527	0.7608	-1.07
Ensino Fundamental	0.7106	0.7194	-1.22
Ensino Médio	0.7456	0.7607	-1.99
Escolaridade Adulta	0.4491	0.4538	-1.02
Pré Escola	0.8752	0.8317	5.24
Apropriação da Renda	0.6401	0.6650	-3.74
Geração da Renda	0.6144	0.6314	-2.69
Mortes por Causas Evitáveis	0.5863	0.6187	-5.24
Condições Gerais de Saúde	0.7415	0.7609	-2.55
Óbitos por Causas Mal Definidas	0.8968	0.9032	-0.71
Longevidade	0.8842	0.8982	-1.56
Consultas Pré Natal	0.7674	0.7726	-0.68
Mortalidade de Menores de 5 anos	0.9287	0.9269	0.20
Saúde Materno Infantil	0.8481	0.8498	-0.20

Combinação:Valor total repassado | Bloco educação completo
Modelo: XGBoost | Grupo: 0
R² Médio: 0.731375 | MSE Médio: 0.003856
Amostras no grupo: 2040
População média: 32128.18
Informações médias das amostras:
Valor total repassado: 1671179.29
Equivalente a (42.32%) do valor médio: (1174245.23)
Características IDESE:

	Grupo	Original	Erro(%)
Anos Finais EF	0.6244	0.6780	-7.90
Anos Iniciais EF	0.6699	0.7608	-11.94
Ensino Fundamental	0.6472	0.7194	-10.04
Ensino Médio	0.7520	0.7607	-1.15
Escolaridade Adulta	0.4026	0.4538	-11.28
Pré Escola	0.5830	0.8317	-29.90

Combinação:Valor total repassado | Todos atributos independentes
Modelo: XGBoost | Grupo: 1
R² Médio: 0.714004 | MSE Médio: 0.000877
Amostras no grupo: 2762
População média: 20848.26

Informações médias das amostras:
Valor total repassado: 843046.58
Equivalente a (-28.21%) do valor médio: (1174245.23)
Características IDESE:

	Grupo	Original	Erro(%)
Idese	0.7875	0.7253	8.58
Bloco Educação	0.7532	0.6914	8.94
Bloco Renda	0.7364	0.6482	13.62
Bloco Saúde	0.8730	0.8363	4.39
Anos Finais EF	0.7190	0.6780	6.05
Anos Iniciais EF	0.8218	0.7608	8.03
Ensino Fundamental	0.7704	0.7194	7.09
Ensino Médio	0.8080	0.7607	6.21
Escolaridade Adulta	0.4980	0.4538	9.74
Pré Escola	0.9363	0.8317	12.58
Apropriação da Renda	0.7639	0.6650	14.89
Geração da Renda	0.7090	0.6314	12.29
Mortes por Causas Evitáveis	0.6699	0.6187	8.27
Condições Gerais de Saúde	0.7994	0.7609	5.05
Óbitos por Causas Mal Definidas	0.9288	0.9032	2.84
Longevidade	0.9350	0.8982	4.10
Consultas Pré Natal	0.8267	0.7726	7.00
Mortalidade de Menores de 5 anos	0.9427	0.9269	1.70
Saúde Materno Infantil	0.8847	0.8498	4.11

Combinação:Total de beneficiários | Bloco educação completo
Modelo: XGBoost | Grupo: 0
R² Médio: 0.701133 | MSE Médio: 0.005095
Amostras no grupo: 2686
População média: 19586.40
Informações médias das amostras:

```
Total de beneficiários: 668.36
 Equivalente a (-18.99%) do valor médio: (825.03)
Características IDESE:
                     Grupo  Original  Erro(%)
Anos Finais EF       0.6973    0.6780     2.85
Anos Iniciais EF     0.7892    0.7608     3.73
Ensino Fundamental   0.7432    0.7194     3.32
Ensino Médio         0.6331    0.7607   -16.78
Escolaridade Adulta  0.4713    0.4538     3.86
Pré Escola           0.9194    0.8317    10.55
Combinação:Valor total repassado | Bloco educação completo
 Modelo: XGBoost |  Grupo: 2
R² Médio: 0.667676 |  MSE Médio: 0.001108
Amostras no grupo: 2733
População média: 20807.30
Informações médias das amostras:
Valor total repassado: 1171160.33
 Equivalente a (-0.26%) do valor médio: (1174245.23)
Características IDESE:
                     Grupo  Original  Erro(%)
Anos Finais EF       0.6983    0.6780     3.00
Anos Iniciais EF     0.7906    0.7608     3.92
Ensino Fundamental   0.7444    0.7194     3.48
Ensino Médio         0.6359    0.7607   -16.40
Escolaridade Adulta  0.4736    0.4538     4.37
Pré Escola           0.9196    0.8317    10.57

Combinação:Total de beneficiários | Todos atributos independentes
 Modelo: MLP |  Grupo: 2
R² Médio: 0.662266 |  MSE Médio: 0.003018
Amostras no grupo: 1600
Populaç
Informações médias das amostras:
Total de beneficiários: 1249.43
 Equivalente a (51.44%) do valor médio: (825.03)
Características IDESE:
                                Grupo  Original  Erro(%)
Idese                          0.6363    0.7253   -12.26
Bloco Educação                 0.5774    0.6914   -16.48
Bloco Renda                    0.5361    0.6482   -17.29
Bloco Saúde                    0.7955    0.8363    -4.88
Anos Finais EF                 0.6252    0.6780    -7.79
Anos Iniciais EF               0.6710    0.7608   -11.80
Ensino Fundamental             0.6481    0.7194    -9.91
Ensino Médio                   0.7083    0.7607    -6.90
Escolaridade Adulta            0.3862    0.4538   -14.88
Pré Escola                     0.5671    0.8317   -31.82
Apropriação da Renda           0.5419    0.6650   -18.50
Geração da Renda               0.5303    0.6314   -16.01
Mortes por Causas Evitáveis    0.5929    0.6187    -4.18
Condições Gerais de Saúde      0.7320    0.7609    -3.80
Óbitos por Causas Mal Definidas  0.8712    0.9032    -3.54
Longevidade                    0.8617    0.8982    -4.06
Consultas Pré Natal            0.6892    0.7726   -10.80
Mortalidade de Menores de 5 anos  0.8962    0.9269    -3.32
Saúde Materno Infantil         0.7927    0.8498    -6.72
Combinação:Total de beneficiários | Bloco educação completo
 Modelo: XGBoost |  Grupo: 1
R² Médio: 0.653863 |  MSE Médio: 0.00117
Amostras no grupo: 2728
População média: 16594.20
Informações médias das amostras:
Total de beneficiários: 554.58
 Equivalente a (-32.78%) do valor médio: (825.03)
Características IDESE:
                     Grupo  Original  Erro(%)
Anos Finais EF       0.6990    0.6780     3.10
Anos Iniciais EF     0.8007    0.7608     5.25
Ensino Fundamental   0.7498    0.7194     4.24
Ensino Médio         0.8920    0.7607    17.25
```

```
Escolaridade Adulta  0.4742    0.4538     4.52
Pré Escola           0.9313    0.8317    11.97
Combinação:Valor total repassado | Todos atributos independentes
 Modelo: MLP |  Grupo: 0
R² Médio: 0.628872 |  MSE Médio: 0.004033
Amostras no grupo: 3086
População média: 22254.71
Informações médias das amostras:
Valor total repassado: 1286812.34
 Equivalente a (9.59%) do valor médio: (1174245.23)
Características IDESE:
                                 Grupo  Original  Erro(%)
Idese                           0.7157    0.7253    -1.32
Bloco Educação                  0.6952    0.6914     0.55
Bloco Renda                     0.6273    0.6482    -3.22
Bloco Saúde                     0.8246    0.8363    -1.40
Anos Finais EF                  0.6687    0.6780    -1.38
Anos Iniciais EF                0.7527    0.7608    -1.06
Ensino Fundamental              0.7107    0.7194    -1.21
Ensino Médio                    0.7456    0.7607    -1.99
Escolaridade Adulta             0.4492    0.4538    -1.00
Pré Escola                      0.8753    0.8317     5.24
Apropriação da Renda            0.6402    0.6650    -3.72
Geração da Renda                0.6144    0.6314    -2.69
Mortes por Causas Evitáveis     0.5863    0.6187    -5.24
Condições Gerais de Saúde       0.7415    0.7609    -2.55
Óbitos por Causas Mal Definidas 0.8967    0.9032    -0.71
Longevidade                     0.8842    0.8982    -1.56
Consultas Pré Natal             0.7675    0.7726    -0.67
Mortalidade de Menores de 5 anos 0.9287   0.9269     0.19
Saúde Materno Infantil          0.8481    0.8498    -0.20
```