CAMILA NASCIMENTO SILVA

ANÁLISE DE FATORES SOCIOECONÔMICOS ASSOCIADOS À QUEDA DE COBERTURA VACINAL DE POLIOMIELITE NO BRASIL POR MEIO DE APRENDIZADO NÃO-SUPERVISIONADO

PROJETO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO
APRESENTADO AO DEPARTAMENTO DE ENGENHARIA INDUSTRIAL
DA PUC-RIO, COMO PARTE DOS REQUISITOS PARA OBTENÇÃO
DO TÍTULO DE ENGENHEIRO DE PRODUÇÃO

Orientador: Leonardo dos Santos Lourenço Bastos

Departamento de Engenharia Industrial
Rio de Janeiro, 16 de novembro de 2022.

RESUMO

A diminuição da cobertura vacinal no Brasil tem ficado cada vez mais expressiva, trazendo diversas preocupações para os órgãos de saúde pública. Com a redução da cobertura vacinal, doenças já erradicadas no país estão sujeitas a retornarem, que é o caso da poliomielite. Com essa ameaça presente, é necessário entender os aspectos socioeconômicos associados com a baixa cobertura vacinal. Dessa forma, este trabalho teve como objetivo analisar a relação entre a baixa cobertura vacinal para poliomielite e os aspectos socioeconômicos e de desenvolvimento nas regiões e municípios brasileiros. Com isso, utilizou-se técnicas de análises descritivas dos dados e modelagem de aprendizado não-supervisionada de machine learning (clusterização). Através das análises, notou-se que aparentemente existe uma associação de alguns fatores socioeconômicos, como O IDH e Gini, com a redução da cobertura vacinal no Brasil, o que demonstra que municípios e regiões mais subdesenvolvidos estão apresentando um menor público-alvo imunizado. Portanto, os resultados encontrados sugerem que há a necessidade da aplicação de políticas e ações públicas, principalmente nas regiões menos desenvolvidas, para que a cobertura vacinal brasileira retorne ao seu patamar desejável, acima de 95%.

PALAVRAS-CHAVE

Cobertura. Poliomielite. Socioeconômicos. Vacinal.

ABSTRACT

The decrease in vaccination coverage in Brazil has become increasingly significant, bringing several concerns to public health agencies. With the reduction in vaccination coverage, diseases that have already been eradicated in the country are subject to a return, which is the case of poliomyelitis. With this threat present, it is necessary to understand the socioeconomic aspects associated with low vaccination coverage. Thus, this study aimed to analyze the relationship between low vaccination coverage for poliomyelitis and socioeconomic and development aspects in Brazilian regions and municipalities. With this, descriptive data analysis techniques and unsupervised machine learning modeling (clustering) were used. Through the analyses, it was noted that apparently there is an association of some socioeconomic factors, such as the HDI and Gini, with the reduction of vaccination coverage in Brazil, which demonstrates that more underdeveloped municipalities and regions are presenting a smaller immunized target audience. Therefore, the results found suggest that there is a need to apply public policies and actions, especially in less developed regions, so that the Brazilian vaccination coverage returns to its desirable level, above 95%.

KEY WORDS

Coverage. Polio. Socioeconomic. Vaccine.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

O programa nacional de imunização (PNI) do Brasil tem sido responsável por controlar e em alguns casos até erradicar diversas doenças imunopreviníveis. Com início em 1973, o PNI é encarregado pela política de imunizantes no país, desde a aquisição das doses vacinais até sua distribuição para o público-alvo apropriado. Através desse programa são disponibilizadas mais de 20 vacinas para a população brasileira, que contemplam bebês, crianças, adolescentes, adultos, idosos e gestantes (UNASUS, 2022).

Esse programa de imunização é de extrema importância para o país, com ele evita-se que doenças como o sarampo, a rubéola, a difteria e a poliomielite voltem a se manifestar no Brasil (Instituto Butantan, 2022). Porém com o passar dos anos a cobertura vacinal contra esses agentes têm caído e o mérito de possuir o certificado de erradicação dessas quatro doenças começou a ser ameaçado (Laboissière, 2018).

Tal ameaça tem gerado uma grande preocupação para diversos órgãos da área da saúde, acarretando a realização de estudos e pesquisas para compreender os seus motivos. Desde o ano de 2015, a cobertura vacinal no Brasil vem se mostrando abaixo da média recomendada pela Organização Mundial da Saúde (OMS) de 95% de abrangência (Instituto Butantan, 2022). Com isso, apenas dois anos depois do recebimento do certificado pela Organização Pan-Americana de Saúde (OPAS), o retorno do vírus do sarampo começou a se manifestar, gerando novos casos da doença em diversos estados brasileiros (Laboissière, 2018).

Portanto, a evolução recente com o retorno de doenças consideradas erradicadas não se trata apenas de uma preocupação, mas sim de uma real ameaça. Esse cenário não se limita ao vírus do sarampo, e pode ocorrer com outras doenças, como a poliomielite, que também tem apresentado uma baixa cobertura vacinal ao longo dos últimos 7 anos. A poliomielite é uma doença séria que pode infectar de maneira silenciosa, mas com grande gravidade, provocando paralisias irreversíveis e fatais. Em 2015, a vacina contra a poliomielite atingiu a meta de 95% do público-alvo vacinado pela última vez. Desde então, a taxa de cobertura da vacina manteve-se abaixo da meta estabelecida, chegando pela primeira vez abaixo dos 70% em 2021(Conselho Nacional de Saúde, 2022).

Em 2022, a meta da campanha de vacinação contra a poliomielite não foi alcançada em diversos estados, ocasionando a prorrogação da campanha em alguns deles. Até o término

da campanha, apenas cerca de 54% do público-alvo foi vacinado. Nenhuma unidade federativa atingiu a meta dos 95%, e a região Norte verificou o menor percentual, com 43,22% (G1, 2022). Em outubro de 2022, a porcentagem de crianças brasileiras vacinadas aumentou para 70,15%. A região Sul liderou a estatística com a maior cobertura, 76,26%, seguida da região Nordeste (74,96%), Sudeste (64,38%), Centro-Oeste (61,38%), e por fim a região Norte (57,06%) (Gov.br, 2022).

Uma baixa cobertura vacinal pode ser ocasionada por diferentes fatores, principalmente aqueles relacionados ao acesso à vacina e inequidade (Ledford, 2022). A nível municipal, as condições de moradia, como a escassez de saneamento básico presente na localidade, estão entre os possíveis fatores de propensão à doença. Além disso, outros fatores socioeconômicos podem representar obstáculos e dificultar o acesso a vacinação. Esses fatores podem ser medidos através do Índice de Desenvolvimento Humano (IDH) e do índice Gini. Estudos prévios mostraram que cidades com o baixo desenvolvimento socioeconômico apresentaram baixa taxa de vacinação para COVID-19 – essa relação também pode estar presente na imunização de pólio (Bastos et al., 2022). Outro possível fator é a hesitação vacinal visto que, com a erradicação da poliomielite, é possível que o interesse na proteção e medo da doença tenham decaído ao longo dos anos (Conselho Nacional de Saúde, 2022).

Ainda há incertezas sobre a relação entre fatores socioeconômicos e a cobertura vacinal contra a poliomielite. Dessa forma, este trabalho tem como objetivo investigar possíveis aspectos socioeconômicos associados com a cobertura vacinal contra da poliomielite no Brasil. Para isso, foram avaliados dados de cobertura vacinal entre os anos de 2010 e 2020, em mais de e 5.000 municípios brasileiros. Este trabalho contribui com um panorama da evolução da cobertura vacinal contra a poliomielite no país, bem como para a identificação de perfis de municípios associados com a baixa cobertura vacinal utilizando técnicas de aprendizado não-supervisionado (clustering).

Com intuito de entender a correlação desses fatores com a queda da cobertura vacinal, o presente estudo foi dividido em mais quatro capítulos. O segundo capítulo inicia com uma abordagem do contexto da poliomielite no Brasil nos últimos anos, passando por uma breve descrição do andamento de um projeto de Data Science e finalizando com as técnicas utilizadas no estudo. Já o terceiro capítulo apresenta o problema, os dados e métodos utilizados na análise. O quarto capítulo apresenta os resultados das análises e sua interpretação. Por fim, o quinto capítulo conclui o estudo com um resumo das análises, resultados obtidos, e repercussões.

2. FUNDAMENTAÇÃO TEÓRICA

2.1. Contexto da poliomielite no Brasil

A poliomielite é uma doença que não registra mais nenhum caso de contágio desde a última ocorrência registrada em 1989 na Paraíba. O Brasil recebeu o certificado de eliminação da doença no ano de 1994 e desde então vem incentivando a sua imunização, a fim de evitar novos casos de contaminação (UNASUS, 2021).

O vírus da poliomielite é transmitido via oral-fecal, podendo estar presente em água e alimentos contaminados, acarretando o favorecimento à contaminação aqueles sem hábitos higiênicos. As principais vítimas da poliomielite são crianças de até quatro anos de idade que, devido à idade, ainda estão na fase de aprendizado sobre hábitos de higiene e cuidados contra doenças infectocontagiosas. Quando infecta um indivíduo, o poliovírus se prolifera inicialmente nas suas vias de entrada como boca, garganta e intestino, seguindo para corrente sanguínea, onde pode causar sérios problemas. Uma vez na corrente sanguínea, a doença pode atingir o sistema nervoso, o que ocasiona uma de suas famosas consequências, a paralisia. Em casos extremos, a doença pode levar seu paciente a óbito (Fiocruz, 2022).

Não existe um tratamento para a poliomielite. Consequentemente, a prevenção realizada através de programas de saneamento básico e principalmente através da vacinação. Atualmente existem dois tipos de vacina, uma delas a injetável, que é aplicada aos 2, 4 e 6 meses de vida da criança, seguida por dois reforços, no período de 15 a 18 meses e de 4 a 5 anos de idade. O outro tipo de vacina é a oral, que é aplicada como dose de reforço aos 15 meses e aos 4 anos de idade da criança (Fiocruz, 2022).

Portanto, campanhas de vacinação são de extrema importância para que a população se imunize, pois somente desta forma a doença poderá se manter extinta ou ser erradicada nos locais que ainda possuem casos. Embora o Brasil seja um dos países livre do vírus, a baixa cobertura de vacinação contra a doença tem trazido o risco de retorno da mesma. Desde 2016 a taxa de cobertura do público-alvo brasileiro se manteve abaixo de 95%, que é a média estabelecida para que se tenha uma imunidade coletiva. Em 2020, a taxa de cobertura atingiu apenas 75% da cobertura do público-alvo (Nunes, 2021).

Contudo, a situação pode ser mais problemática em regiões específicas. As regiões Norte e Nordeste são as que apresentam as menores taxas de imunização. O Nordeste chegou a

atingir apenas cerca de 79,5% de cobertura em 2016 – entretanto o Norte ainda se demonstrou inferior, atingindo cerca de 76% de cobertura em 2016, além de ter tido uma redução de mais de 10% em 2021 (Conselho Nacional de Saúde, 2022). A pandemia de COVID-19 pode ter motivado a queda da cobertura de poliomielite em 2021, pois o receio de se infectar pelo vírus SARS-CoV2 ao frequentar uma unidade de saúde fez com que a adesão às demais vacinas diminuíssem. Além disso, em alguns locais as campanhas e as medidas preventivas de vacinação de rotina foram suspensas. Porém, a pandemia não demonstrou ser o principal problema das diminuições na cobertura de vacinação. A falta de imunização dos brasileiros tem provocado uma taxa inferior a 90% de cobertura nos últimos 6 anos. Essa taxa reduzida pode ser oriunda da redução das campanhas de vacinação e do desconhecimento da necessidade de vacinação por parte da população (Fujita et al., 2022).

2.2. Ciclo de vida projeto de dados

Para extrair informações e apoiar a tomada de decisão, a manipulação de dados se tornou de extrema importância. A alta circulação e armazenamento de dados traz a necessidade de análises com uma maior extração de conhecimento e insights das bases de dados. Desta maneira, o *Data Science* possui o propósito de transformar números dispersos em informações mais evidentes e de fácil visualização (Bichler et al., 2017). A técnica conta com um campo interdisciplinar, que compreende a ciência da computação, matemática, estatística e *Machine Learning* (Cao, 2016).

Para isso existem diversos *frameworks* que propõem orientar o desenvolvimento de métodos por meio de técnicas de mineração de dados, com o intuito de dar um sentido aos dados brutos ao estabelecer padrões. O CRoss Industry Standard Process for Data Mining (CRISP-DM) é um dos exemplos mais populares, e pode ser definido como uma metodologia de mineração de dados que abrange um plano completo para conduzir um projeto de Data Science (Manresa, 2020).

O CRISP-DM possui um caráter cíclico e é composto por seis fases que apresentam interações entre si, como apresentado na Figura 1.

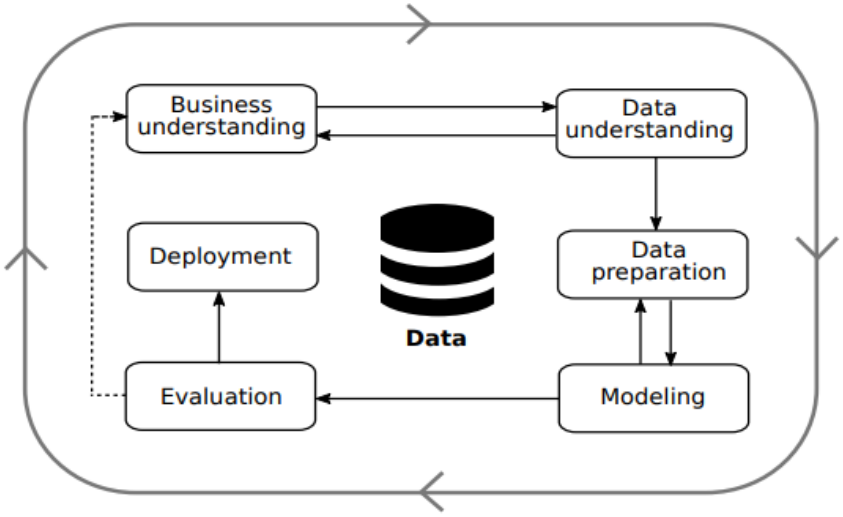

Figura 1 - Ciclo de vida de um projeto de ciência de dados

Fonte: Manresa, 2022.

A primeira etapa do ciclo é a compreensão do negócio, que também pode ser entendida como a percepção do problema. Ela se concentra em assegurar que as soluções auxiliem no processo de tomada de decisão com o interesse de uma perspectiva de negócios. Para isso, a fase conta com três tarefas: a determinação dos objetivos do negócio, explicitando quais são as questões almejadas; a avaliação da situação, que verifica os dados disponíveis; e o objetivo da mineração de dados, que determina como será realizado o objetivo e com quais dados (Manresa, 2020).

Já a segunda etapa abrange o entendimento dos dados, incluindo a coleta das bases disponíveis, a percepção de suas características e sua qualidade. Essa fase dialoga com a fase anterior ao possibilitar a reformulação do objetivo do projeto após o conhecimento dos dados. Assim, essa fase pode ser dividida em três sub etapas: a coleta e descrição dos dados, a sua exploração dos dados e a avaliação de sua qualidade (Manresa, 2020).

Na terceira fase é realizada a preparação dos dados, que consiste na sua seleção, limpeza, construção e formatação. Desta maneira, forma-se um conjunto de dados final que atenda aos requisitos necessários na próxima fase. A modelagem dos dados é considerada a quarta fase do ciclo, composta pela escolha das técnicas de modelagem, criação e avaliação de modelos. Essa fase também tem uma ligação com a anterior, pois diferentes algoritmos podem exigir distintos procedimentos na formatação e preparação dos dados (Manresa, 2020).

A quinta fase é composta pela avaliação. Nessa fase, após a seleção do melhor modelo, os resultados serão interpretados e avaliados, seguidos de uma revisão do processo de aprendizagem e da determinação dos próximos passos. Por fim, na última etapa do ciclo é realizada a implementação modelo, isto é, a solução adquirida é introduzida no contexto prático e é feita uma devolutiva de sua implementação (Manresa, 2020).

2.3. Técnicas utilizadas

2.3.1. Análise descritiva

Cada conjunto de dados possui um tipo de comportamento, e é de extrema importância que esse comportamento seja compreendido durante o seu estudo, para que se possa obter bons resultados. Sendo assim, a análise descritiva é um importante procedimento inicial para o entendimento de um conjunto de informações, pois através dela são detalhadas singularidades e características, identificando padrões e tendências no conjunto de dados. Para a realização dessa análise são utilizadas técnicas de síntese da informação e ferramentas de visualização (Maçaira, 2022).

O primeiro passo da análise é identificar as variáveis, que são os atributos observados na base de dados, podendo estas ser de dois tipos: qualitativas ou quantitativas. As variáveis qualitativas são aquelas que representam uma característica do objeto definido por uma categoria, ao invés de um valor numérico, e podem ser nominais ou ordinais. Já as variáveis quantitativas representam características mensuráveis, que possuem um valor numérico, Existem dois tipos de variáveis quantitativas, as contínuas e as discretas (Costa e Leo, 2020).

Após a definição das variáveis existentes, podem ser utilizadas ferramentas de medidas de tendência central e de posição para se entender mais sobre o seu comportamento. As medidas de tendência central são uma forma de resumir a informação apresentada, de modo a se produzir parâmetros comparáveis. Por outro lado, as medidas de posição buscam entender como os dados estão posicionados em relação a outros (Costa e Leo, 2020; Maçaira, 2022).

Além disso, a análise gráfica faz com que se tenha uma visualização na análise dos dados, por exemplo, ao facilitar comparações. Dentre os diferentes tipos de gráficos existentes, alguns dos mais utilizados para a representação de um conjunto de dados são os gráficos de linhas, de colunas e de dispersão (Maçaira, 2022).

Os gráficos de linhas são ideais para a representação de séries temporais, mostrando o comportamento da variável de interesse ao longo do tempo. Já os gráficos de dispersão são utilizados para fazer uma comparação e entender a relação entre duas variáveis quantitativas de um indivíduo do conjunto de dados. Por fim, os gráficos de colunas fazem uma comparação entre variáveis quantitativas e qualitativas, mostrando a quantidade de um atributo (Reis e Reis, 2002).

Além dos gráficos, também foram utilizadas matrizes de correlação entre as variáveis, a correlação entre duas variáveis pode variar no intervalo [-1, 1]. Quando o valor se aproxima dos extremos, -1 ou 1, significa que existe uma forte correlação entre as variáveis, essa correlação pode ser crescente, no caso de valores positivos, ou decrescente, quando os valores são negativos.

2.3.2. Clusterização

A clusterização de dados é uma técnica que divide um conjunto de dados heterogêneos em subgrupos com objetos que possuem similaridades entre si, denominados clusters. Os dados são agrupados por semelhanças, sem que seja feita uma pré-classificação das informações antes do agrupamento. O principal objetivo da técnica é que sejam criados grupos com objetos de alta semelhança, mas de maneira que exista uma ampla distinção entre um cluster e outro (Cassiano, 2014).

Encontrar a melhor maneira de decompor os dados em grupos é uma tarefa difícil. Por isso, a técnica de clusterização conta com diferentes métodos para a realização do agrupamento. Cada método tem sua peculiaridade, uns são melhores para grandes grupos de dados, outros já trabalham melhor com menos informações. Por outro lado, alguns precisam que seja pré-definido uma quantidade de clusters, enquanto outros não precisam dessa imposição (Doni, 2004).

Dentre os diferentes métodos existentes, dois deles possuem maior visibilidade: o método hierárquico e o método particional ou não hierárquico. Além de se diferenciarem uns dos outros, os métodos de clusterização também apresentam técnicas que possuem suas particularidades. Na Tabela 1 pode-se visualizar dois algoritmos de cada um dos métodos descritos acima (Bastos, 2018).

Tabela 1 - Descrição de técnicas de clusterização

Método	Técnica	Descrição
Hierárquico	Aninhamento Aglomerativo (ANGES)	Clustering aglomerativo é uma técnica na qual nós únicos são mesclados até que todo o conjunto de dados seja um único cluster. Este procedimento constrói um dendrograma no qual uma hierarquia é calculada em relação ao número de clusters. O procedimento de mesclagem depende da função de ligação usada para estimar a distância entre dois clusters e dos critérios para combinar dois clusters calculando sua dissimilaridade. Existem quatro populares funções de ligação: - Ligação simples: considera a distância mínima entre os componentes de dois clusters; - Ligação completa: inversamente, funde dois clusters que possuem a distância máxima entre eles; - Ligação média: une dois clusters com a menor distância média; - Distância da ala: considera a variância mínima dentro do cluster.
	Análise Divisória (DIANA)	Ao contrário do AGNES, o DIANA começa com todo o conjunto de dados como um único cluster. Em seguida, ele se divide iterativamente em clusters até que todos os pontos de dados sejam clusters considerando a distância mínima entre os pontos de dados dentro de um cluster. Ele também calcula uma hierarquia que serve como base para a escolha do melhor número de clusters.
Particionais	K-Means	Um número inicial de k clusters deve ser definido. Ele seleciona k pontos aleatoriamente como centróides e atribui aos clusters os pontos de dados mais próximos do respectivo centróide. No K-Means, o centróide corresponde à média das coordenadas dos pontos dentro do mesmo cluster. Assim, em cada iteração, os centróides são recalculados e os pontos de dados são reatribuídos aos clusters para minimizar a distância média total com relação ao centróide.
	K-Medoids	Um número inicial de k clusters deve ser definido. Ele seleciona k pontos de dados como centróides aleatoriamente e atribui aos clusters os pontos de dados mais próximos do respectivo centróide. Em K-Medoids, os centróides são pontos de dados. Assim, em cada iteração os centróides são recalculados e os pontos de dados são reatribuídos aos clusters para minimizar a distância média total com relação ao centróide.

Fonte: Bastos, 2018

Embora todos esses métodos sejam utilizados, o K-Means é um método tradicional, sendo um dos mais utilizado devido sua simplicidade. Após a especificação da quantidade de clusters e da definição dos centróides de cada um dos clusters, o algoritmo calcula a distância

entre os centróides e os demais pontos, atribuindo os dados ao cluster mais próximo. Para isso, o método segue a seguinte formulação matemática (Dabbura, 2018):

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \left\| x^i - \mu_k \right\|^2$$

Onde $w_{ik} = 1$ para o ponto x^i se ele pertencer ao cluster k; caso contrário, o ponto não pertencer ao cluster k, então $w_{ik} = 0$. Além disso, μ_k é o centróide do cluster de x^i.

Além dos diversos tipos de clusterização, também existem variadas métricas de avaliação de clusters para definir a quantidade ideal de clusters a ser utilizada. O objetivo geral da análise das métricas é entender se os agrupamentos estão coerentes, se os membros de um cluster são mais semelhantes entre si ou aos membros de outros clusters (Scikit-learn, 2022).

Algumas das métricas de avaliação mais utilizadas são o coeficiente de silhueta, o índice Calinski Harabasz, e o índice Davies-Bouldin. Todas as três costumam ser utilizadas quando o verdadeiro conjunto de agrupamento dos dados é desconhecido.

O coeficiente da silhueta de um cluster é calculado através da média da distância intra-cluster e do cluster mais próximo para todos os membros do conjunto. Com uma variabilidade de [-1;1], um valor alto indica que os clusters resultantes são densos e bem separados (Mehta, 2022). Definido para cada amostra, o coeficiente da silhueta é composto por duas pontuações e é calculado da seguinte maneira (Rousseeuw, 1986):

$$s = \frac{b - a}{\max(a, b)}$$

Onde a é adistancia média entre uma amostra e os demais pontos de um cluster, e b é a distância média de uma amostra e o todos os outros pontos do cluster mais próximo.

O índice de Calinski Harabasz, que também é conhecido como critério de razão da variância, calcula-se como a razão entre a soma da dispersão entre os clusters e a dispersão dentro do cluster para todos os clusters. Quanto maior o valor do índice, melhor é o agrupamento e melhor separados estão os clusters. Para um conjunto de dados E de tamanho n_E que foi agrupado em k clusters, a pontuação de Calinski-Harabasz s é dado pela seguinte formulação (Caliński e Harabasz, 1974):

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k - 1}$$

Onde:

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T$$

De maneira que C_q é o conjunto de pontos no cluster q, c_q é o centro do cluster q, c_E é o centro de E e n_q é o número de pontos no cluster q. Além disso, tr representa o traço da matriz da variável.

Por fim, o índice de Davies-Bouldin mede a semelhança média entre cada cluster C_i para $i = 1, \dots, k$, e seu mais similar C_j. Essa similaridade de comparação entre as distâncias dentro dos clusters e entre os clusters é dada pela medida R_{ij}. Um menor valor do índice é preferível, pois um menor valor indica baixas medidas de dispersão intragrupo e grandes distâncias intergrupos. O seu valor mínimo é zero, que indica uma divisão ideal dos membros (Halkidi, Batistakis, e Vazirgiannis, 2001). Então o índice de Davies-Bouldin é definido como:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij}$$

Onde:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

De maneira que s_i é a distância média entre cada ponto do cluster i e o seu centroide, e d_{ij} é a distância entre os centroides dos clusters i e j.

3. METODOLOGIA

O presente estudo tem o propósito de analisar e identificar os diversos fatores associados com a queda da cobertura vacinal de poliomielite no Brasil. Essa queda pode estar associada a fatores como a falta de acesso à vacina, o subdesenvolvimento de diferentes regiões brasileiras, ou diferenças socioeconômicas dos seus variados estados e municípios, dentre outras questões.

Para a realização deste objetivo, as análises dos dados disponíveis sobre o assunto foram realizadas se espelhando no ciclo de vida de um *Data Science*, seguindo os seguintes passos ilustrados nesse capítulo.

3.1. Compreensão do problema

Durante os últimos anos, a cobertura vacinal brasileira vem caindo, fazendo com que menos pessoas estejam imunizadas. Essa queda traz a preocupação de que doenças já erradicadas, como a poliomielite, possam ressurgir com novos casos no país. Diante deste cenário, se faz necessário avaliar quais fatores podem estar associados a essa redução, realizando uma análise a nível nacional e municipal, para identificar possíveis diferenças.

Algumas hipóteses podem ser estudadas para investigar os possíveis fatores associados às quedas das imunizações contra a doença. Uma delas é de que municípios em desvantagem socioeconômica, como um baixo índice de desenvolvimento e uma baixa cobertura de atenção primária, estejam associados a uma baixa cobertura vacinal. Outra hipótese é que municípios com uma alta cobertura de plano de saúde possam apresentar uma errônea baixa cobertura na análise dos dados disponibilizados, visto que esses são originados de sistemas do setor público, como o DATASUS.

Desta forma, o estudo é norteado por duas perguntas: (i) Como tem sido o progresso da cobertura vacinal contra poliomielite no Brasil e suas regiões?; e (ii) Existe alguma relação entre características socioeconômicas e a queda da cobertura vacinal? Para responder a essas perguntas, foram realizados três blocos de análises. O primeiro busca avaliar a situação da cobertura vacinal no país e seus estados, enquanto os dois outros blocos têm como objetivo identificar uma possível associação entre fatores socioeconômicos e a cobertura vacinal.

3.2. Entendimento dos dados

Para a análise feita nesse estudo, foram obtidas informações de 7 fontes de dados integrados na plataforma do IEPS Data. As informações sobre a cobertura vacinal de poliomielite são obtidas originalmente do sistema do Programa Nacional de Imunização (PNI). Já os dados sobre a cobertura de atenção primária vêm do sistema e-Gestor. Essas variáveis e as demais, junto com suas correspondentes fontes, estão descritas na Tabela 2 (IEPS, 2022).

Tabela 2 - Variáveis presentes nas bases de dados retiradas do IEPS Data

Variável	Nome dos Indicadores	Bloco	Fonte
cob_ab	Cobertura da Atenção Básica (%)	Atenção Primária	e-Gestor
cob_vac_polio	Cobertura Vacinal de Poliomielite (%)	Atenção Primária	PNI, TabNet/DATASUS
cob_priv	Cobertura de Planos de Saúde (%)	Saúde Suplementar	ANS
idhm	Índice de Desenvolvimento Humano Municipal (2010)	Indicadores Socioeconômicos	Atlas Brasil
pct_pop0a4	População de 0 a 4 Anos (%)	Demografia	TabNet DATASUS
pop0a4	População de 0 a 4 Anos	Demografia	TabNet DATASUS
num_familias_bf	Número de Beneficiários do Bolsa Família	Indicadores Socioeconômicos	MDS
pct_san_adeq	Saneamento Básico Adequado (Censo 2010, %)	Indicadores Socioeconômicos	Censo

Fonte: IEPS Data, 2022

As variáveis expostas derivam de três bases de dados que possuem informações dos anos de 2010 a 2020, com exceção do Índice de Desenvolvimento Humano (IDH) e da porcentagem de saneamento básico, que são valores coletados do Censo de 2010. Duas das bases possuem a cobertura de atenção básica, cobertura vacinal, cobertura de planos de saúde e o IDH como variáveis. Elas se diferenciam com relação à granularidade, pois a base “DadoEstados” analisa as variáveis apenas para as regiões e estados do Brasil, enquanto a base “MunRegioes” possui informações detalhadas para os municípios brasileiros. A terceira base, “PopMun”, também é caracterizada por municípios. Porém, diferente das demais, ela contém apenas as variáveis restantes que as outras duas não possuem, população de 0 a 4 anos e seu

percentual, número de beneficiários do bolsa família e a estatística sobre a existência de saneamento básico adequado.

Além das três bases retiradas do IEPS Data, também foi utilizada outra base de dados derivada do Instituto Brasileiro de Geografia e Estatística (IBGE), "InfMun", com a finalidade de complementar as análises (IBGE, 2022). Esse conjunto de dados também apresenta as informações por municípios. As variáveis da base de dados do IBGE utilizadas encontram-se descritas na Tabela 3.

Tabela 3 - Variáveis da base de dados do IBGE

Variável	Nome dos Indicadores	Bloco	Fonte
gini	Índice Gini para Medição da Desigualdade	Indicadores Socioeconômicos	IBGE
idhm_renda	Componente do IDH Referente à Renda	Indicadores Socioeconômicos	IBGE
populacao_2020	População Estimada do Município em 2020	Demografia	IBGE
dens_pop_2020	Densidade Populacional do Município (população/km²)	Demografia	IBGE
tipo_urbano	Tipo de Urbanização do Município (2017)	Demografia	IBGE
area_km2	Área Territorial do Município	Demografia	IBGE

Fonte: IBGE

3.3. Preparo dos dados

Após a obtenção das informações deu-se início à manipulação dos dados, iniciando pelo primeiro passo, a realização da limpeza e formatação dos dados. Todas as bases possuem dados dos anos 2010 a 2020. A primeira base de dados manipulada foi a "DadosEstados", que continha apenas as regiões e os estados brasileiros. Nesse conjunto de dados não foi necessário realizar nenhuma limpeza na base, já que não existiam variáveis com valores faltantes.

No entanto, as bases com maior granularidade precisaram passar por um tratamento de dados devido a existência de variáveis com valores faltantes para alguns municípios. A primeira limpeza foi realizada na base “MunRegioes” através da utilização do filtro para cada uma das variáveis na ferramenta Microsoft Excel.

Ao filtrar para que fosse mostrado apenas os valores em branco das variáveis, foi possível saber quais municípios possuíam dados incompletos. Após essa descoberta, todos eles, juntamente com os demais dados, foram retirados da base, com intuito de que permanecessem apenas os municípios que possuíssem informações de todas as variáveis em todos os anos de interesse. Depois de conhecer e remover quais municípios possuíam valores em branco, foi utilizada a linguagem Python para realizar a limpeza das duas bases restantes.

Com a limpeza de dados realizada, foi possível dar início a formatação dos mesmos. Para isso foram realizados dois procedimentos: o primeiro deles foi o cálculo da média anual das variáveis analisadas e o segundo foi a determinação do valor da variação anual de cada variável. Para a obtenção desse resultado foi calculada a diferença entre o valor do ano em análise e o valor do ano anterior, dividida pelo valor do ano anterior, iniciando as análises no ano de 2011. O primeiro procedimento de formatação foi aplicado para todas as bases de dados, já o segundo procedimento foi aplicado exclusivamente para as bases “DadosEstados” e “MunRegioes”.

Depois que todas as bases de dados municipais foram limpas e formatadas, por questão de praticidade, viu-se a necessidade de criar apenas um conjunto de dados com todas as informações municipais. Assim, todas as três bases municipais, “MunRegioes”, “InfMun” e “PopMun”, foram unificadas criando uma única base de dados municipais, a base “DadosCompletos”.

3.4. Modelagem

Em um primeiro passo, para avaliar as distribuições das variáveis no trabalho, foi utilizado o intervalo interquartil (IQR) e a mediana para variáveis quantitativas e frequência absoluta e proporção para as variáveis qualitativas. Além disso, as relações entre as variáveis foram avaliadas por meio de gráficos de dispersão e por matrizes de correlação.

Para entender a associação entre a cobertura vacinal de poliomielite e aspectos socioeconômicos a nível municipal, foi utilizado um modelo de clusterização. Embora existam

diversos métodos de clusterização, considerando que as variáveis socioeconômicas não apresentam uma hierarquia clara entre si, a modelagem do estudo baseou-se em dois métodos de clusterização particionais: o K-Means e o K-Medoids. Além disso, estes métodos realizam o agrupamento de maneira global, isto é, as distâncias calculadas são feitas sob todas as observações, e não localmente como outros métodos.

Para a implementação dos métodos foi utilizada a linguagem Python, portanto foi necessário a importar algumas bibliotecas e seus módulos. Nas análises numéricas foram utilizadas as bibliotecas Numpy, Pandas e alguns módulos das bibliotecas Scikit-learn e Scikit-learn-extra. Já para as análises gráficas foram utilizadas as bibliotecas Seaborn e a Matplotlib.

Os dados empregados na realização do estudo foram retirados da base de dados final, a base “DadosCompletos”. Posteriormente à sua importação, foi feita uma normalização dos dados para que fosse utilizada uma escala comum, já que as variáveis apresentavam unidades diferentes. A normalização de cada um dos valores foi obtida através da subtração do valor mínimo da amostra e da divisão pela amplitude.

A normalização fez com que todos os valores amostrais ficassem em uma mesma escala, que varia de zero a um. Porém, antes se dar início à clusterização, ainda é neceessário definir a quantidade ideal de grupos em que os valores serão segmentados. Para essa decisão foram utilizados três métricas de avaliação: o coeficiente de silhueta, o índice Calinski Harabasz e o índice Davies-Bouldin. A partir da codificação e execução das métricas, foi criada uma tabela comparativa com os resultados para uma variação de dois a vinte sete categorias de agrupamento. Após definir a quantidade de grupos com melhores métricas, o próximo passo foi descrever os perfis de municípios encontrados.

4. ANÁLISE E DISCUSSÃO DOS RESULTADOS

4.1. Análise descritiva da cobertura no domínio nacional

A primeira ação realizada foi a montagem de um gráfico para entender o comportamento da vacinação de poliomielite no Brasil no período dos últimos dez anos, apresentado na Figura 2. Para sua produção, foram utilizadas as médias anuais da cobertura de vacinação nos estados brasileiros. Além deste, também foi feito um gráfico mostrando as variações anuais das coberturas, exposto na Figura 3, para verificar o quão expressiva foi a variação de um ano para o ano seguinte.

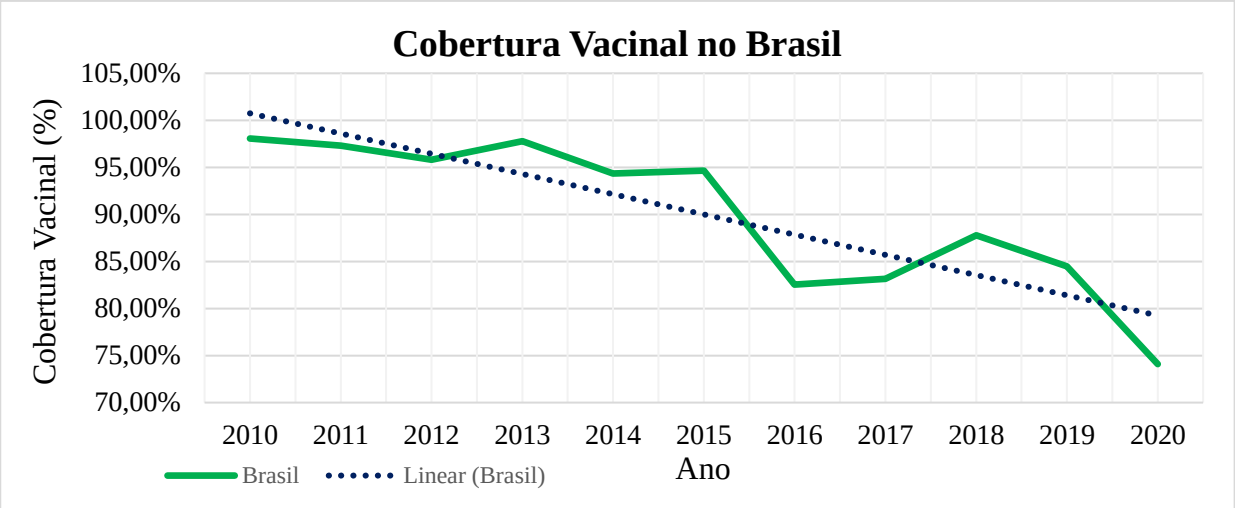

Figura 2 - Cobertura vacinal no Brasil

Fonte: Autora, 2022

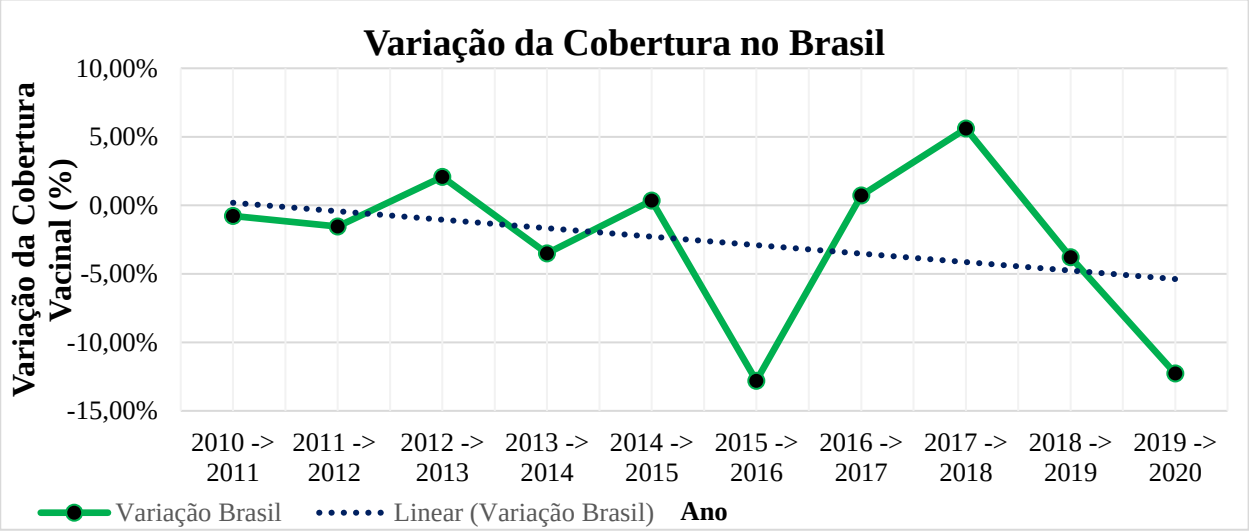

Figura 3 - Variação da cobertura vacinal no Brasil

Fonte: Autora, 2022

Inicialmente, é possível notar que a cobertura vacinal brasileira contra a pólio sofreu uma queda nos últimos anos. O valor ideal para que se tenha uma imunização coletiva é uma cobertura acima de 95%. No entanto, desde 2013 o Brasil parece não alcançar tal meta. Entre 2013 e 2014 a queda foi de cerca de 3%, ficando ligeiramente abaixo do valor idealizado. Já entre 2014 e 2015 a cobertura teve uma pequena subida, quase alcançando a meta novamente. No entanto, a queda de cobertura vacinal se intensificou entre 2015 e 2016, quando o país teve uma queda brusca na porcentagem de indivíduos imunizados.

A redução de 2015 para 2016 foi maior nos demais anos, podendo ser comparada com a queda no ano de 2020. Contudo, a queda no ano de 2020 pode estar associada à pandemia de COVID-19, enquanto em 2016 não ocorreu nenhum evento atípico. Sendo assim, sua queda pode esta correlacionada a diversos fatores, como a carência de campanhas de vacinação, a insuficiência de vacinas ou até mesmo a hesitação vacinal da população. Porém, devido à limitação da obtenção de dados sobre distribuição das vacinas e a eficiência das campanhas em cada município municípios, não foi possível investigar tal suposição.

Além do gráfico para a análise da vacinação no país, também foram produzidos dois gráficos regionais, um da cobertura vacinal, apresentado na Figura 4, e um da variação da cobertura, exposto na Figura 5. Esses gráficos foram esboçados a partir dos valores da série temporal da cobertura de vacinação para os estados. Neles, as linhas representam as unidades de federação e as cores representam as regiões brasileiras.

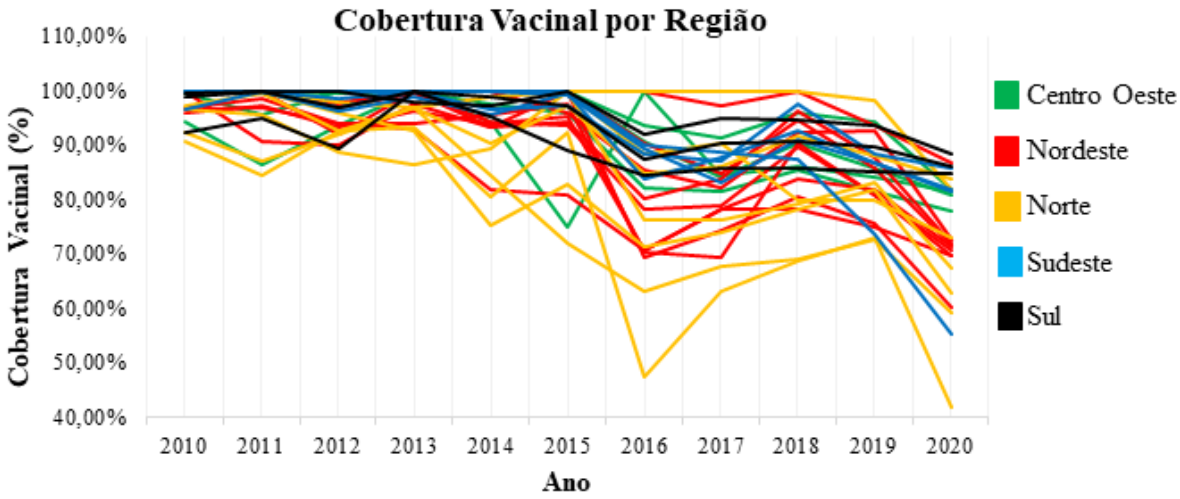

Figura 4 - Cobertura vacinal nos estados e regiões

Fonte: Autora, 2022

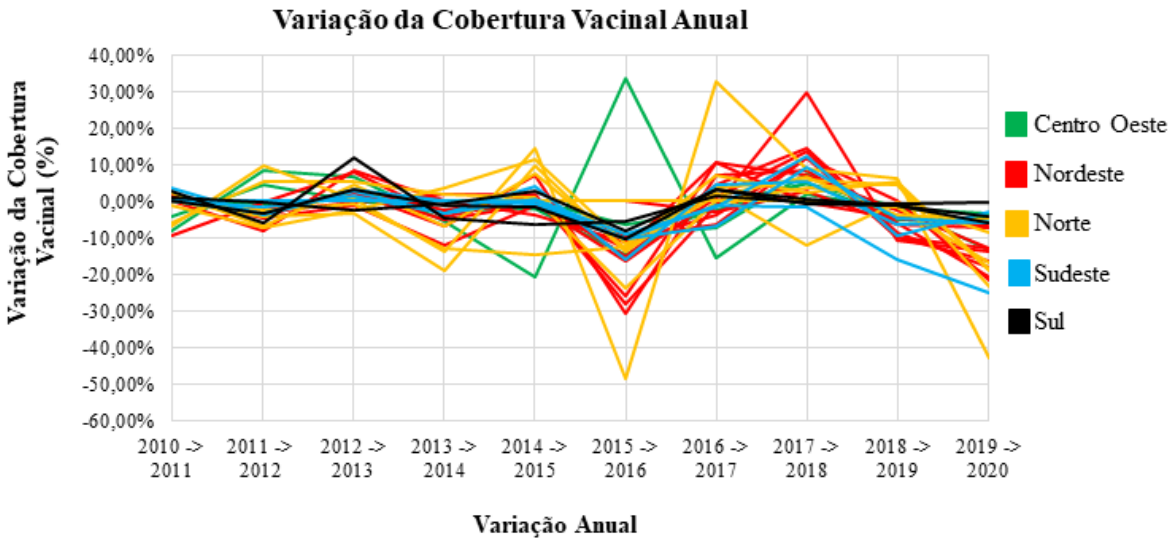

Figura 5 - Variação anual da cobertura vacinal nos estados

Fonte: Autora, 2022

Ao analisar a cobertura vacinal de maneira regional, é possível notar uma aparente correlação entre as regiões e a taxa de cobertura. Até o ano de 2015, grande parte das regiões – com exceção de alguns estados do Norte e do Nordeste – possuíam mais de 90% da população alvo imunizada. Já entre 2015 e 2016, poucos foram os estados que se mantiveram acima dessa taxa, a maioria oscilou entre 80% e 90% de cobertura. No ano seguinte, diversos estados conseguiram amenizar a queda e aumentar a sua cobertura vacinal. No entanto, muitos deles não conseguiram retornar para o percentual histórico, entre 90% e 100%. Em 2020 também é notório que as unidades de federação tiveram outra queda brusca na cobertura, o que ocasionou em uma cobertura média vacinal brasileira abaixo dos 75%.

Embora quase todos os estados brasileiros tenham sofrido essa queda, é perceptível que os estados localizados nas regiões Norte e Nordeste sofreram uma queda mais significativa resultando, quase que predominantemente, em uma cobertura média inferior à dos estados das regiões Sudeste e Sul. Alguns desses estados obtiveram uma cobertura vacinal abaixo de 80% a partir de 2015. Esse padrão é demonstrado na variação da cobertura exibida na Figura 5, principalmente entre 2015 e 2016, período em que os estados do Nordeste e Norte tiveram uma maior variação negativa em comparação ao do Sudeste e de Sul. É interessante notar que nesse mesmo período a única variação positiva foi do Distrito Federal, que representa a capital do país.

Essa aparente correlação entre as regiões e a taxa de cobertura pode ser consequência de diversos fatores. Além dos já citados anteriormente, como uma possível distribuição desigual de vacinas, a ineficácia de campanhas e potencial hesitação vacinal da população alvo, questões socioeconômicas também podem estar associadas a essa possível correlação. A próxima seção apresenta uma análise mais aprofundada sobre a possível relação entre fatores socioeconômicos e a cobertura vacinal.

4.2. Análise descritiva da cobertura vacinal no domínio municipal

Para lançar luz sobre os possíveis fatores associados com a variação da cobertura vacinal, a preparação dos dados iniciou-se com a limpeza das observações com dados em branco, que eliminou 387 municípios da base. Em seguida, foi realizada uma análise mais exploratória para os 5.183 municípios restantes, iniciada com a geração de gráficos de dispersão relacionando a cobertura vacinal dos municípios brasileiros às demais variáveis de interesse, apresentados na Figura 6.

O primeiro gráfico no canto superior direito da Figura 6 tenta identificar se existe alguma relação entre a cobertura vacinal e a cobertura de planos de saúde nos municípios brasileiros. A hipótese é de que, quanto maior a cobertura de plano saúde, menor seria a cobertura vacinal oficial, pois parte da população poderia vacinar-se preferencialmente na rede privada, cujos dados não são referidos no DATASUS. Portanto, a vacinação na rede privada não é representada nos dados oficiais utilizados nesse estudo. No entanto, ao observar o gráfico não é notada nenhuma correlação aparente entre as duas variáveis, não corroborando essa hipótese para os municípios analisados.

Já no gráfico no canto superior esquerdo é analisada a relação entre a cobertura vacinal e a cobertura da atenção primária média nos municípios. Como a atenção primária é, entre outros, encarregada das atividades de prevenção e proteção da população, ela é a responsável por garantir uma vacinação efetiva. Neste caso, uma baixa cobertura da atenção primária poderia estar associada à uma baixa cobertura vacinal. Embora o gráfico não demonstre uma correlação forte entre as variáveis, é possível averiguar que alguns municípios possuem uma baixa cobertura vacinal e de atenção primária. Essa relação aparenta ser mais presente na região Norte do que nas demais.

Nos dois próximos gráficos é possível notar uma associação positiva entre a cobertura vacinal e o IDH. Da mesma forma, o gráfico seguinte demonstra uma leve correlação negativa

entre a cobertura vacinal e o índice Gini, sugerindo que uma maior desigualdade econômica pode estar correlacionada com uma baixa cobertura vacinal em um município.

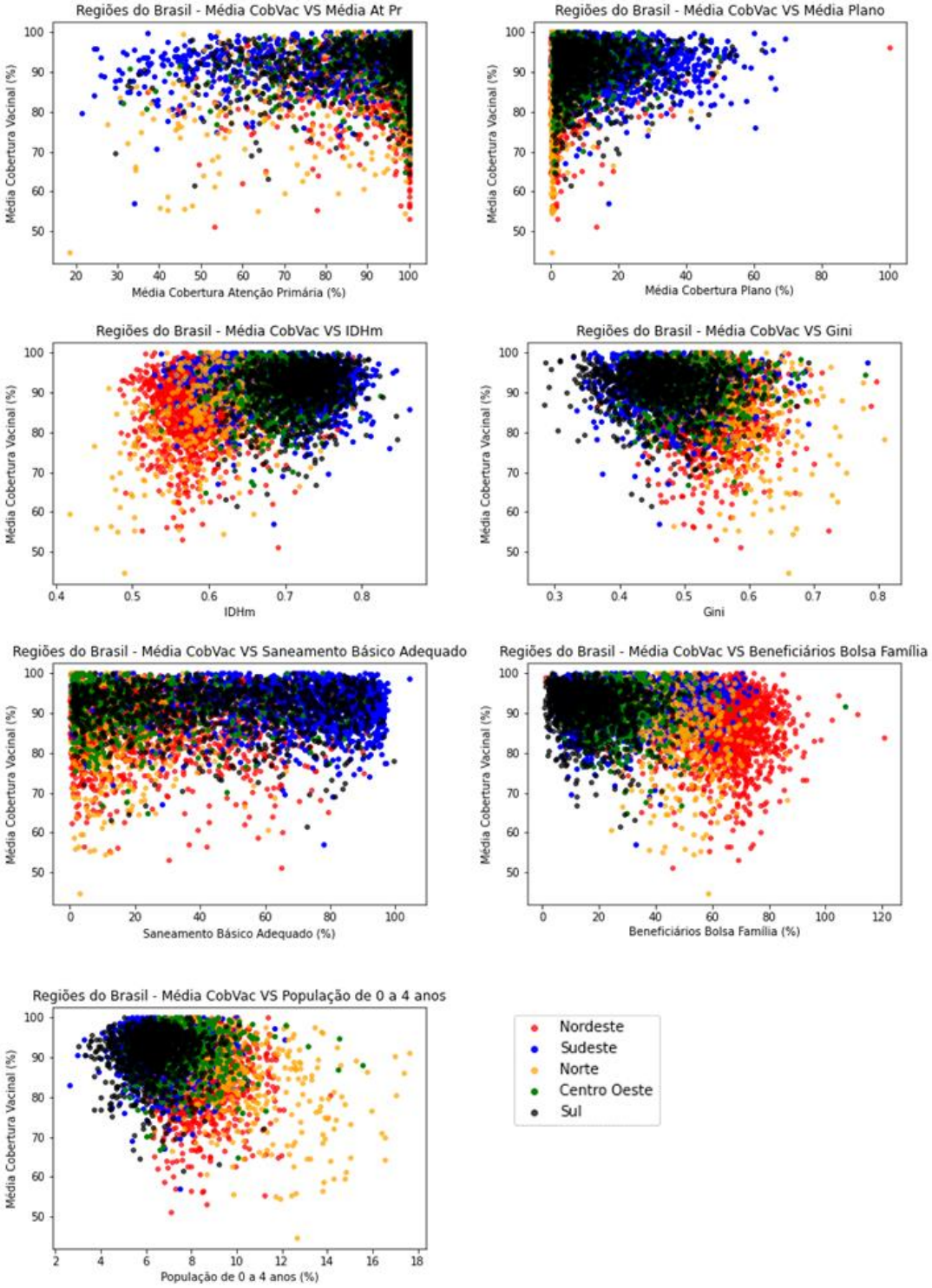

Figura 6 - Relação entre a cobertura vacinal e variáveis de interesse nas regiões e municípios brasileiros

Fonte: Autora, 2022

A análise da relação entre o acesso a saneamento básico adequado e a cobertura vacinal não indica uma associação direta. Embora alguns municípios possuam baixos valores para ambas as variáveis, não existe uma correlação aparente. Já na análise entre a cobertura vacinal e o número de beneficiários do bolsa família, é possível notar uma leve tendência entre as variáveis, que pode estar relacionada a outras questões socioeconômicas associadas a ambas as variáveis.

Por fim, na análise da população de 0 a 4 anos, pode-se reparar que a cobertura vacinal parece ser menor com o aumento do público-alvo, e é especialmente notável para municípios da região Norte. Portanto, esse fenômeno pode estar associado à dificuldade de acesso da população aos locais de vacinação ou à entraves na logística de distribuição e aplicação das vacinas. No entanto, não é possível analisar se esses fatores estão associados com a queda da cobertura, já que não foram encontrados dados que explicitassem como decorreu a distribuição de imunizantes ao longo do período estudado.

Além da análise visual das relações, também foi criada uma matriz de correlação (método Spearman) para quantificar o grau de associação monotônica entre as variáveis, exposta na Figura 7. A cobertura vacinal não possui uma forte correlação com as outras variáveis, mas apresenta uma pequena correlação com algumas delas, como o IDH (0.2) e o índice Gini (-0.2), indicando uma leve relação positiva e negativa, respectivamente. É importante notar que essa análise não exclui a possibilidade de existirem relações não-monotônicas entre essas variáveis ou efeitos conjuntos.

A matriz também permite analisar a correlação entre as outras variáveis de interesse. Nessa análise, é possível destacar alguns indicadores socioeconômicos que possuem forte correlações entre si: o IDH, o índice Gini, a cobertura de plano de saúde e o percentual de beneficiários do bolsa família. Municípios com um maior IDH apresentam, em média, uma maior cobertura de plano de saúde e uma menor taxa de beneficiários do Bolsa Família. Esses estados também apresentam, em média, um menor índice Gini, sugerindo uma menor desigualdade social.

Figura 7 - Matriz de correlação Spearman entre a cobertura vacinal e variáveis de interesse (nível municipal)

Fonte: Autora, 2022

As análises já realizadas demonstram a existência de uma associação entre a queda da cobertura vacinal e fatores socioeconômicos. Porém, para um melhor entendimento sobre a distribuição vacinal entre os municípios e regiões, foi montada a Tabela 4, que correlaciona a taxa de cobertura no município com o seu tipo de urbanização e a região brasileira a que ele pertence.

Para a criação da Tabela 4 foram calculados os tercis da média de cobertura vacinal, dividindo-a em três categorias: baixa cobertura, média cobertura e alta cobertura. Os municípios que possuem uma cobertura vacinal menor que 88,02% (primeiro tercil) foram inseridos no grupo de baixa cobertura. Já os com uma cobertura vacinal média maior ou igual a 88,02%, mas inferior a 93,74%, foram classificados no grupo de média cobertura. Por fim, os demais municípios com média vacinal superior ou igual a 93,74% (segundo tercil) foram categorizados no grupo de alta cobertura vacinal. Cada faixa de cobertura abrange cerca de 1.700 municípios.

A Figura 8 mostra como estão distribuídos os municípios brasileiros de acordo com a cobertura vacinal. É possível notar que os municípios com baixa cobertura vacinal (vermelho) estão mais presentes nas regiões Norte e Nordeste. Já as regiões Sul e Sudeste aparentam ter mais municípios de alta cobertura, enquanto o Centro-Oeste apresenta uma mistura das três categorias de cobertura.

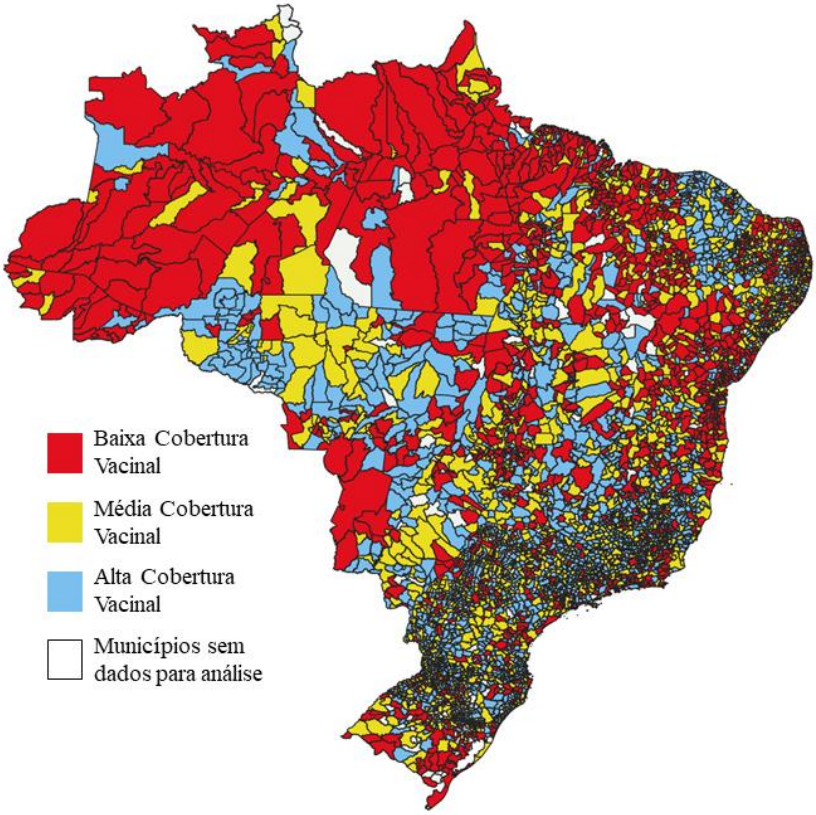

Figura 8 - Mapa de distribuição da cobertura vacinal por municípios

Fonte: Autora, 2022

Para permitir uma melhor compreensão sobre a distribuição das faixas de cobertura vacinal entre as regiões brasileiras, é importante entender quantos municípios existem em cada uma das regiões e quantos deles são do tipo urbano, semiurbano ou rural. Essa análise permite verificar a predominância de determinadas categorias de cobertura vacinal em uma região, ou de um determinado tipo urbano em uma categoria de cobertura.

Ao analisar o número total de observações, o Nordeste e o Sudeste são regiões com a maior quantidade de municípios, abrangendo mais de 60% do total de 5.183 municípios analisados. Essa dominância pode ocasionar uma participação significativa destas regiões nas categorias de cobertura, mesmo que o total de municípios da região não seja um número muito alto em relação às demais regiões também presentes no grupo. Além disso, os dados indicam que os municípios brasileiros são predominantemente do tipo rural.

Tabela 4 - Número de municípios por região, tipo urbano e categoria de cobertura vacinal

Características	Total	Baixa Cobertura Vacinal	Média Cobertura Vacina	Alta Cobertura Vacinal
Região N(%)	**5.183 (100%)**	**1.711 (100%)**	**1.710 (100%)**	**1.762 (100%)**
Centro Oeste	448 (8,64%)	136 (7,95%)	150 (8,77%)	162 (9,19%)
Nordeste	1.742 (33,61%)	785 (45,88%)	551 (32,22%)	406 (23,04%)
Norte	427 (8,24%)	240 (14,03%)	91 (5,32%)	96 (5,44%)
Sudeste	1.532 (29,55%)	317 (18,53%)	517 (30,23%)	698 (39,59%)
Sul	1.034 (19,95%)	233 (13,61%)	401 (23,45%)	400 (22,69%)
Tipo Urbano N(%)	**5.183 (100%)**	**1.711 (100%)**	**1.710 (100%)**	**1.762 (100%)**
Urbano	1.414 (27,28%)	476 (27,82%)	514 (30,06%)	424 (24,06%)
Rural	3.066 (59,16%)	995 (58,15%)	964 (56,37%)	1.107 (62,83%)
Semiurbano	703 (13,56%)	240 (14,03%)	232 (13,57%)	231 (13,11%)

Fonte: Autora, 2022

Já ao explorar a distribuição das regiões pelas faixas de cobertura nota-se que algumas categorias parecem são dominadas por uma região. Por exemplo, quase 50% dos municípios com baixa cobertura vacinal pertencem à região Nordeste. Embora a região Norte não lidere a estatística do total de municípios na categoria de baixa cobertura vacinal, os 240 municípios presentes nessa categoria representam mais de 50% do total de municípios da região Norte.

A faixa de alta cobertura vacinal é liderada por municípios da região Sudeste, com aproximadamente 40% do total de municípios da categoria, e cerca de 45% do total de municípios da região. Enquanto isso, a faixa de média cobertura não aparenta ter a predominância de uma região brasileira. Ela apresenta um grande percentual de municípios das regiões Nordeste e Sudeste, seguidas da região Sul.

Para investigar uma possível relação entre as faixas de cobertura e fatores socioeconômicos, a próxima análise avalia ambos os aspectos em conjunto, considerando as medianas e o intervalo interquartil das variáveis por categoria de cobertura vacinal. Os resultados estão disponíveis na Tabela 5.

Tabela 5 - Indicadores socioeconômicos por categoria de cobertura vacinal

Z	**Total (N = 5.183)**	**Baixa Cobertura Vacinal (N = 1.711)**	**Média Cobertura Vacinal (N = 1.710)**	**Alta Cobertura Vacinal (N = 1.762)**
	Mediana (IQR)	Mediana (IQR)	Mediana (IQR)	Mediana (IQR)
Cobertura de atenção primária à saúde (%)	98,87 (88,7; 100)	99,09 (85,9; 100)	98,91 (88,2; 100)	98,74 (91,1; 100)
Cobertura de Plano de Saúde (%)	3,81 (1,20; 11,32)	2,12 (0,81; 7,71)	4,64 (1,43; 13,02)	4,97 (1,67; 13,17)
Índice de Desenvolvimento Humano Municipal (IDHm)	0,662 (0,598; 0,717)	0,623 (0,576; 0,697)	0,676 (0,605; 0,723)	0,677 (0,622; 0,722)
Índice Gini	0,505 (0,462; 0,547)	0,524 (0,483; 0,564)	0,500 (0,458; 0,542)	0,490 (0,451; 0,532)
Índice de Desenvolvimento Humano Municipal por renda (IDHm_renda)	0,649 (0,571; 0,706)	0,604 (0,552; 0,685)	0,666 (0,579; 0,713)	0,667 (0,592; 0,714)
População alvo (%) (0 a 4 anos)	35,19 (18,42; 60,10)	49,88 (25,53; 65,39)	31,14 (16,74; 58,95)	28,87 (16,20; 52,81)
Densidade Populacional (km^2)	25,85 (11,98; 59,14)	22,60 (10,00; 56,26)	27,74 (13,52; 61,43)	27,05 (13,30; 58,82)

Fonte: Autora, 2022

Os resultados sugerem que a cobertura de atenção primária não se diferenciam substancialmente entre as três categorias de cobertura vacinal. As medianas das três faixas de cobertura e o intervalo interquartil são bem similares. Já no caso da cobertura de plano de saúde, aparentemente acontece o contrário do esperado. Municípios com uma maior cobertura vacinal apresentam uma maior cobertura mediana de plano de saúde. No entanto, essa variável parece ter uma associação com outras duas, a cobertura de atenção primária e o IDH. Os municípios que apresentam uma maior cobertura de atenção primária aparentam ter uma menor cobertura de saúde suplementar, o que pode ser correlacionado a uma falta de necessidade da obtenção de um plano de saúde. Por outro lado, a cobertura de plano de saúde também pode estar relacionada à uma questão econômica e não de necessidade, já que municípios com maior IDH são os que possuem uma maior presença da cobertura de saúde suplementar.

O IDH e índice Gini são duas variáveis que aparentam obter uma maior associação com cobertura vacinal do que as demais. Os resultados na Tabela 5 sugerem que os municípios com

um menor desenvolvimento e uma maior desigualdade social também possuem uma menor cobertura vacinal. Em conjunto, esses resultados sugerem uma possível relação entre os níveis de cobertura vacinal e questões socioeconômicas nos municípios brasileiros.

Para melhor compreender essas relações, a análise na próxima seção aplica o método de clusterização. Essa aplicação tem o objetivo de separar os municípios em agrupamentos, de maneira que sejam formados grupos por municípios que possuem características semelhante entre si, mas diferentes dos demais.

4.3. Perfis de cobertura vacinal - Clusterização

4.3.1. Definição do número de clusters

Para avaliar a qualidade dos clusters gerados pelos métodos K-Means e K-Medoids e determinar o número ideal de clusters, foram utilizadas três métricas de avaliação para diferentes números de clusterse métodos de clusterização aplicados. A análise avalia resultados para agrupamentos entre 2 e 27 clusters, que representam o limite natural de até um cluster por estado. Como métricas de avaliação foram utilizados o coeficiente de silhueta, o índice Calinski Harabasz e o índice Davies-Bouldin. Os resultados estão disponíveis na Tabela 6.

Os valores em negrito representam o número de clusters que apresenta a melhor performance para cada método e métrica. Os resultados indicam que o agrupamento ideal deve conter dois clusters, para ambos os métodos. Como o método de K-Means possui métricas superiores às do K-Medoids, a análise de clusterização baseia-se no método K-Means com dois clusters.

Tabela 6 - Avaliação dos resultados de clusterização por método e número de clusters

Nº de Clusters	K-Means			K-Medoids		
	Silhueta	**Davies-Bouldin**	**Calinski-Harabasz**	**Silhueta**	**Davies-Bouldin**	**Calinski-Harabasz**
2	**0,392**	**1,017**	**4263,811**	**0,382**	**1,038**	**4165,089**
3	0,295	1,253	3226,219	0,296	1,257	3212,898
4	0,293	1,213	2912,849	0,267	1,319	2792,238
5	0,288	1,157	2745,672	0,210	1,575	2393,124
6	0,298	1,134	2541,238	0,210	1,471	2332,665
7	0,262	1,250	2359,373	0,174	1,621	2072,936
8	0,240	1,347	2177,038	0,188	1,564	1933,988
9	0,237	1,354	2029,229	0,175	1,592	1781,525
10	0,220	1,380	1919,496	0,153	1,615	1655,384
11	0,214	1,396	1830,829	0,158	1,638	1551,038
12	0,204	1,384	1753,960	0,170	1,553	1608,453
13	0,194	1,403	1678,911	0,166	1,526	1545,218
14	0,200	1,376	1614,224	0,151	1,592	1454,250
15	0,196	1,358	1561,202	0,154	1,537	1403,605
16	0,191	1,377	1496,544	0,148	1,529	1349,860
17	0,177	1,423	1454,282	0,153	1,531	1338,757
18	0,179	1,411	1400,160	0,143	1,565	1306,197
19	0,175	1,442	1359,007	0,139	1,624	1262,324
20	0,175	1,423	1317,713	0,133	1,660	1208,693
21	0,165	1,441	1278,247	0,149	1,560	1185,916
22	0,170	1,406	1248,451	0,147	1,519	1177,080
23	0,165	1,461	1209,578	0,145	1,558	1139,783
24	0,162	1,448	1180,707	0,139	1,600	1088,458
25	0,167	1,421	1158,405	0,133	1,604	1061,843
26	0,163	1,465	1127,356	0,130	1,640	1032,415
27	0,157	1,440	1105,399	0,127	1,646	1003,637

Fonte: Autora, 2022

4.3.2. Análise dos clusters

O método K-Means com dois clusters gera dois grupos contendo 3.085 municípios (Cluster 1) e 2.098 municípios (Cluster 2).

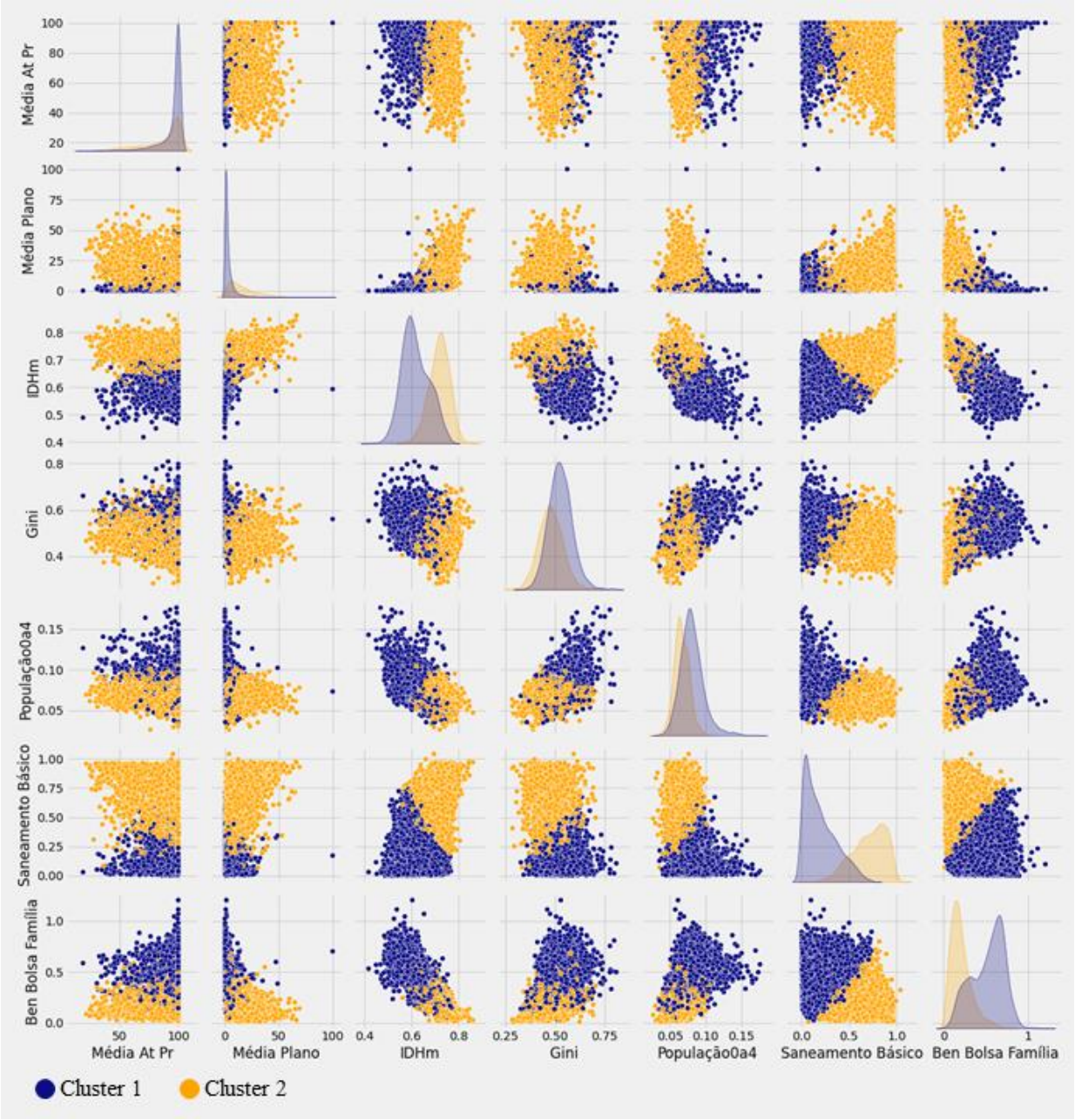

Figura 9 - Gráficos de dispersão e distribuição das variáveis de interesse em cada cluster (K-Means, dois clusters)

Fonte: Autora, 2022

Os gráficos na Figura 9 apresentam a relação entre as variáveis socioeconômicas de interesse para os dois clusters identificados, incluindo também a distribuição de cada variável no grupo de municípios formados por cada cluster. Em alguns gráficos, nota-se uma clara divisão entre os dois clusters. Por exemplo, no gráfico do IDH versus saneamento básico, é notável que os municípios com maior IDH possuem um maior saneamento básico e estão agrupados no Cluster 2, enquanto os de menor IDH possuem um menor saneamento básico e

estão presentes no Cluster 1. Para entender melhor as características e analisar o perfil de cada cluster, a Tabela 7 apresenta o percentual das regiões brasileiras em cada um dos clusters.

Tabela 7- Distribuição dos municípios analisados com relação aos clusters e regiões

Características	**Cluster 1 (N = 3.085)**	**Cluster 2 (N = 2.098)**
Região	**N(%)**	**N(%)**
Centro Oeste	354 (11,47%)	94 (4,48%)
Nordeste	1.613 (52,29%)	129 (6,15%)
Norte	408 (13,22%)	19 (0,91%)
Sudeste	317 (10,28%)	1.215 (57,91%)
Sul	393 (12,74%)	641 (30,55%)

Fonte: Autora, 2022

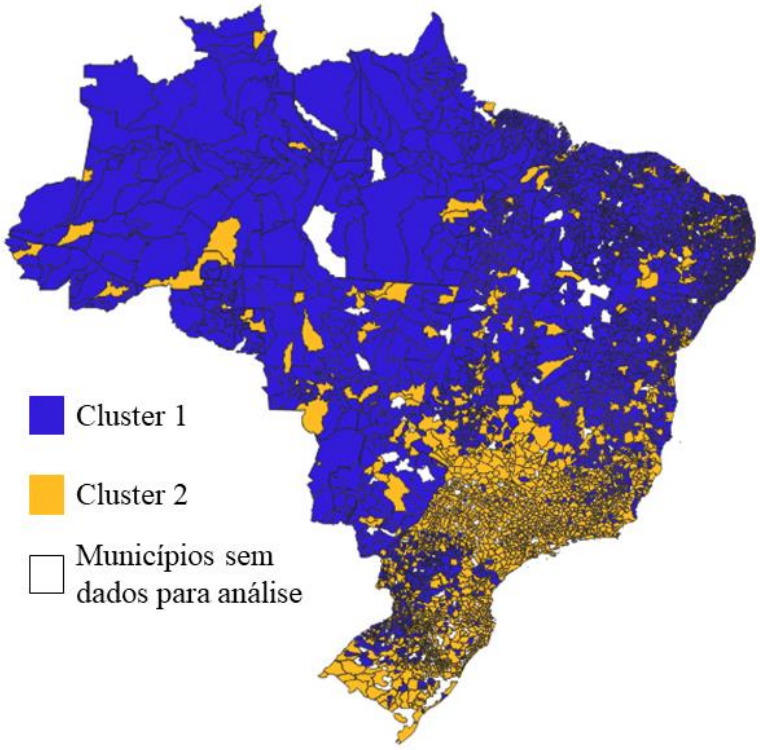

Figura 10 - Mapa municípios divididos pelos clusters obtidos (K-Means, dois clusters)

Fonte: Autora, 2022

Os resultados indicam que o Cluster 1 é dominado pela região Nordeste, seguida pelas regiões Centro-Oeste e Norte – juntas, essas três regiões representam aproximadamente 77% dos municípios no Cluster 1 e 46% do total dos 5.183 municípios analisados. Em contrapartida, o Cluster 2 é dominado pelos municípios das regiões Sudeste e Sul, que juntas totalizam cerca

de 89% dos municípios nesse cluster. Esses resultados sugerem que a clusterização dividiu o Brasil em dois ao considerar apenas as variáveis socioeconômicas, sem levar em consideração a cobertura vacinal. É possível visualizar com mais clareza essa divisão na Figura 10.

A Tabela 8 aprofunda essa análise, apresentando o comportamento detalhado das variáveis socioeconômicas em cada um dos clusters. Para entender a distribuição das variáveis nos clusters, foram calculados as medianas e o intervalo interquartil para todas as variáveis em cada cluster. O Cluster 1, que é constituído principalmente por municípios das regiões Nordeste, Norte e Centro Oeste, apresenta valores expressivos para alguns indicadores socioeconômicos, mostrando ser uma parte do Brasil com menor desenvolvimento. Essa premissa é apoiada pelas variáveis IDH, índice Gini, proporção de beneficiários do bolsa família e acesso ao saneamento básico.

Tabela 8 - Relação entre os clusters gerados (K-Means, dois clusters) e variáveis socioeconômicas

Características	**Cluster 1 (N = 3.085)**	**Cluster 2 (N = 2.098)**
	Mediana (IQR)	Mediana (IQR)
Cobertura Vacinal (%)	90,17 (84,45; 94,53)	92,27 (87,99; 95,53)
Cobertura de Atenção Primária à Saúde (%)	100,0 (94.99; 100,0)	95,70 (79.55; 100,0)
Cobertura de Plano de Saúde (%)	1,42 (0,66; 3,31)	11,31 (5,54; 21,03)
Índice de Desenvolvimento Humano Municipal (IDHm)	0,606 (0,573; 0,656)	0,716 (0.680; 0,745)
Índice Gini	0,526 (0,488; 0,564)	0,478 (0,438; 0,518)
Proporção de Beneficiários do Bolsa Família (%)	56,91 (38,15; 67,27)	18,61 (11,93; 28,86)
Saneamento Básico Adequado (%)	13,81 (5,05; 26,71)	70,57 (55,78; 83,75)
População Alvo (%) (0 a 4 anos)	7,97 (7,04; 8,96)	6,56 (5,90; 7,35)

Fonte: Autora, 2022

O índice Gini e a porcentagem de beneficiários de bolsa família possuem valores altos, sugerindo uma maior desigualdade social e maior dependência da população de auxílios governamentais nos municípios que formam esse cluster. Além disso, o IDH não tem um valor muito alto, principalmente quando comparado ao valor obtido no Cluster 2. Já o saneamento básico adequado apresenta um valor muito baixo, refletindo o baixo desenvolvimento dos

municípios presentes no Cluster 1. Em contrapartida, o Cluster 2 é formado por municípios com um maior desenvolvimento socioeconômico, expresso pelas diversas variáveis apresentadas na análise.

Os dois clusters gerados apresentam caraterísticas homogêneas em diversas dimensões socioeconômicas. A análise subsequente tem como objetivo identificar a possível presença de comportamentos similares na queda de cobertura vacinal nos municípios que formam cada cluster.

Essa análise se inicia com uma avaliação do progresso da cobertura vacinal média nos últimos anos para cada cluster, apresentada na Figura 11. Essa média foi calculada a partir da média anual dos municípios presentes nos respectivos clusters. Os resultados indicam que a cobertura vacinal média no Cluster 2, o cluster de maior desenvolvimento socioeconômico, está sempre acima da curva que representa o Cluster 1. Embora uma curva seja superior à outra, ambas demonstram um desenvolvimento similar, com aumentos ou reduções nos mesmos períodos.

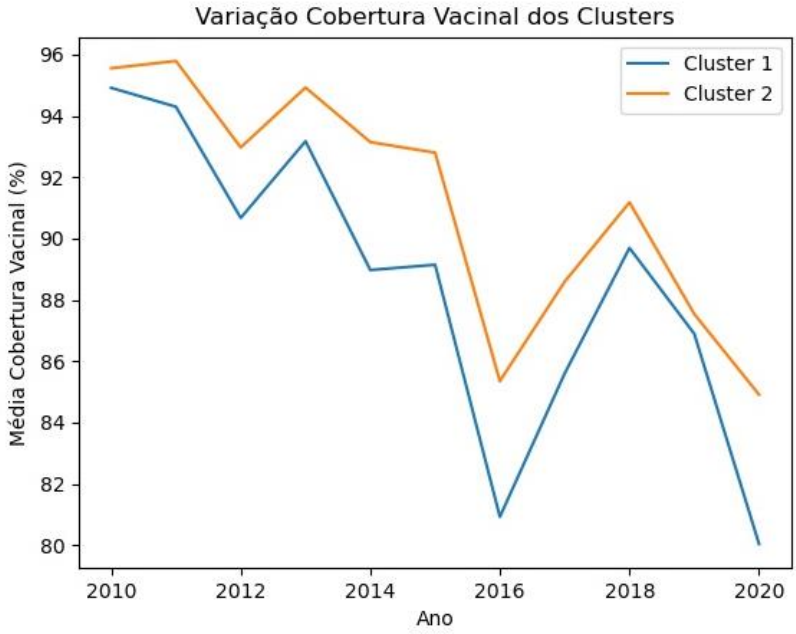

Figura 11 - Média da cobertura vacinal anual em cada cluster

Fonte: Autora

Para entender a possível relação entre a cobertura vacinal e as demais variáveis socioeconômicas, ambas são representadas na Figura 12. Os gráficos de dispersão são similares aos gráficos da seção 4.2, porém as cores representam os dois cluster ao invés das regiões.

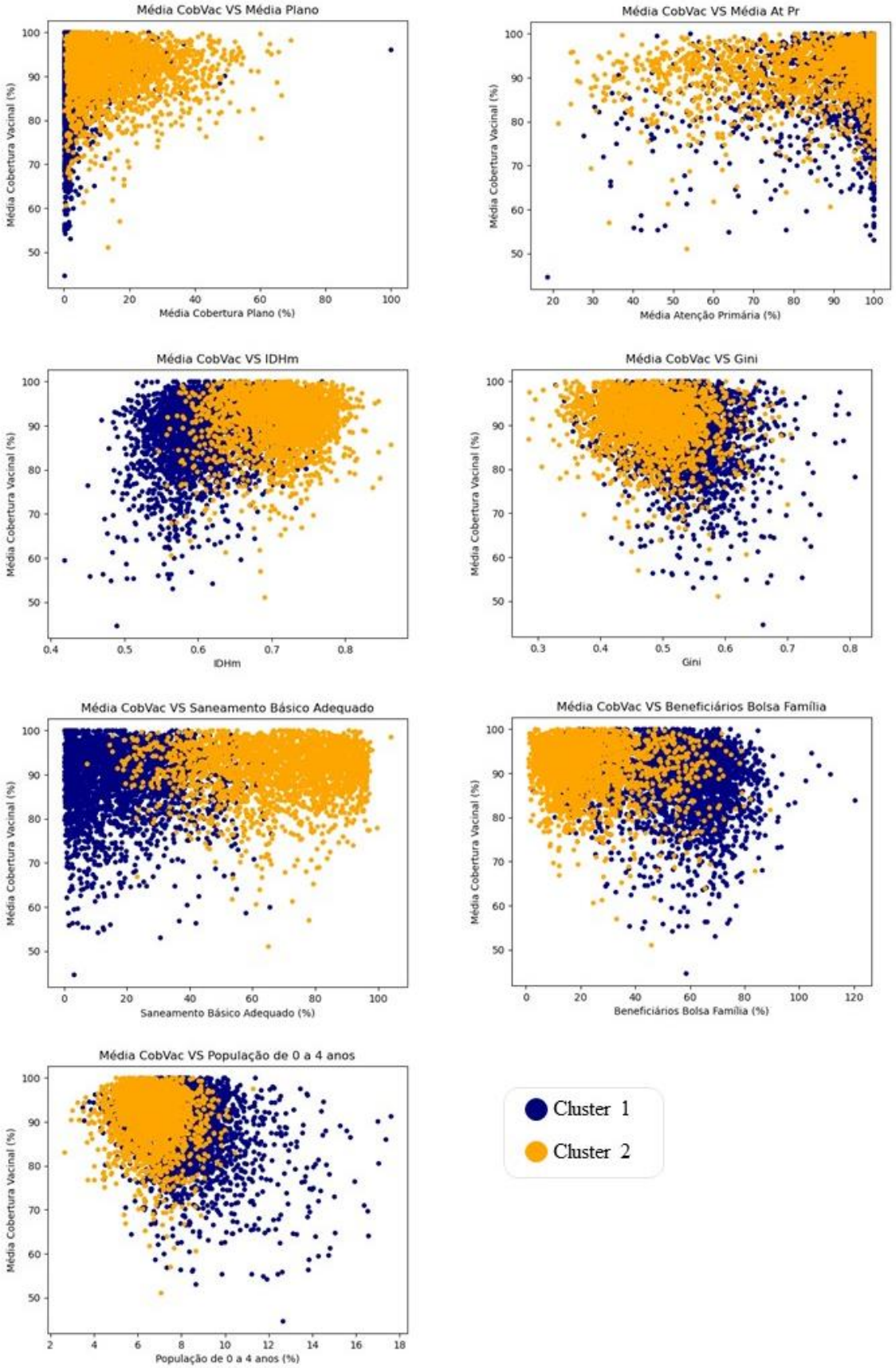

Figura 12 - Gráficos de dispersão dos clusters versus as variáveis de interesse

Fonte: Autora, 2022

De maneira semelhante à análise por regiões, os gráficos de cobertura de atenção primária e de cobertura de plano de saúde não indicam uma relação direta com a cobertura vacinal em ambos os clusters. Em ambos os gráficos, existe uma grande dispersão entre os municípios em cada cluster e pode-se verificar, por exemplo, múltiplos municípios que apresentam uma alta ou baixa cobertura vacinal, embora tenham distintos perfis de atenção primária.

Novamente, é possível perceber uma leve associação entre o IDH a nível municipal e a cobertura média vacinal. No entanto quando se olha somente para o conjunto de pontos do Cluster 2, as variáveis não demonstram uma tendência comportamental entre o IDH e a cobertura vacinal, o que é exposto na matriz de correlação do Cluster 2 (0,0), na Figura 14. O que pode ser perceber nesse gráfico é uma clara divisão dos dois clusters, com o Cluster 2 estando mais presente na esquerda do gráfico, com menores valores de IDH, enquanto o Cluster 1 encontra-se mais à direita. Além desta divisão, o Cluster 1 apresenta mais municípios com uma baixa cobertura vacinal do que o Cluster 2, refletindo a existência de uma correlação entre as duas variáveis ao se analisar o todo. Essa relação se dá pelo fato de o Cluster 1 possuir um menor IDH e mais municípios com baixa cobertura vacinal. Por outro lado, o Cluster 2 apresenta um maior IDH e menos municípios com baixa cobertura vacinal.

No gráfico demonstrando a relação entre o índice Gini e a cobertura vacinal também se visualiza uma leve correlação negativa para ambos os clusters, indicando que, embora não pareça muito forte, o índice Gini tem uma correlação com a cobertura vacinal. Essa pequena associação também pode ser visualizada nas Figura 13 e Figura 14, que representam matrizes de correlação de cada um dos cluster. Em ambos os clusters, a correlação entre a cobertura vacinal e o índice Gini é de -0,2.

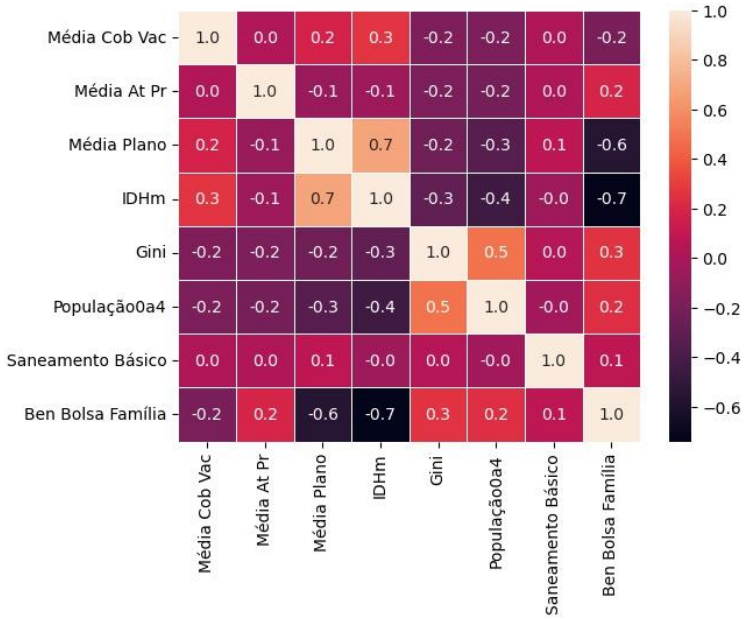

Figura 13 - Matriz de correlação Spearman Cluster 1

Fonte: Autora, 2022

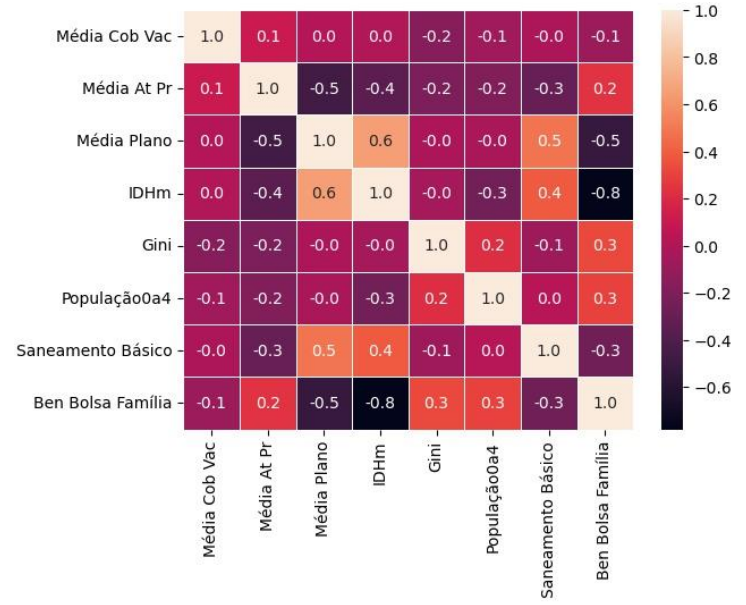

Figura 14 - Matriz de correlação Spearman Cluster 2

Fonte: Autora, 2022

Portanto, através das análises, foi possível identificar que alguns fatores socioeconômicos estão associados com a cobertura vacinal em municípios brasileiros. Municípios mais desenvolvidos, com maiores valores de IDH, apresentam maiores valores de cobertura vacinal e estão predominantemente localizados nas regiões Sudeste e Sul. Por outo lado, municípios com menor desenvolvimento, maior desigualdade e maior concentração de

renda possuem menores valores de cobertura vacinal. Esses são, em geral, municípios das regiões Centro Oeste, Nordeste e Norte, que apresentam valores baixos para o IDH e altos para o percentual de beneficiários do bolsa família.

Sendo assim, o modelo de clusterização e as análises descritivas indicam que existe uma divisão entre regiões mais desenvolvidas e com melhores indicadores socioeconômicos e as regiões que são menos favorecidas. Esses fatores também estão associados à queda de cobertura vacinal. Embora todo o país tenha tido quedas na cobertura vacinal, especialmente nos anos de 2016 e de 2020, os municípios que compõem o Cluster 1 apresentam piores indicadores socioeconômicos e apresentaram uma maior queda na cobertura vacinal nos períodos em questão.

5. CONCLUSÃO

Este trabalho teve como objetivo compreender o comportamento da cobertura vacinal de poliomielite no Brasil e examinar a possível relação entre fatores socioeconômicos e a queda de cobertura vacinal no país. Entre 2010 e 2020 a cobertura de poliomielite no país diminuiu em mais de 20%, apresentando em 2020 um nível de cobertura um pouco abaixo de 75%. Ao analisar de maneira descritiva os municípios brasileiros, percebe-se que os que pertencem a regiões com um menor nível de desenvolvimento, como Nordeste e Norte, em média vacinaram menos que os demais.

Além disso, através da modelagem de clusterização foi possível visualizar uma clara divisão no país em dois grandes grupos de municípios. Um grupo é composto por municípios menos desenvolvidos, das regiões Centro Oeste, Nordeste e Norte, enquanto o outro é composto pelos municípios que possuem um maior nível de desenvolvimento, presentes nas regiões Sudeste e Sul. Nesses dois grupos também existem diferenças com relação à cobertura vacinal contra poliomielite. O grupo mais desenvolvido, predominantemente nas regiões Sudeste e Sul, possui um percentual de vacinação maior que o grupo composto por Centro Oeste, Nordeste e Norte. Com isso, esse trabalho demonstra a associação de fatores socioeconômicos com relação a queda da cobertura vacinal.

Através dos resultados desse estudo foi possível ver que o Brasil apresenta uma desigualdade de desenvolvimento e socioeconômica, que apresentam relações com a queda de cobertura vacinal. Portanto, é importante que medidas como o direcionamento de políticas e ações de saúde pública, sejam aplicadas de maneira distinta para os diferentes grupos de municípios, com intuito de melhorar a cobertura vacinal nas diversas regiões brasileiras. Desta maneira, será possível aumentar a taxa de vacinação na população, promovendo uma imunização coletiva e evitando que doenças já erradicadas apresentem um risco de retorno.

Embora o estudo tenha demonstrado que fatores socioeconômicos estão associados à queda da cobertura vacinal contra a poliomielite, algumas limitações encontradas podem ser exploradas em trabalhos futuros. Entre elas, existe a limitação da obtenção de dados de outras variáveis que também podem estar relacionadas com a queda de cobertura vacinal, como informações sobre a distribuição das vacinas, a dificuldade de acesso da população aos locais de aplicação e a eficiência das campanhas de vacinação, que podem ser coletadas e analisadas pela metodologia deste trabalho. Além disso, sugere-se aplicar uma modelagem estatística,

como modelos de regressão, para se estimar o efeito individual de cada uma das variáveis no aumento e diminuição da cobertura vacinal. Por fim, recomenda-se explorar o perfil de vacinação dos municípios, por exemplo, municípios com baixa cobertura vacinal para poliomielite também podem apresentar quedas de cobertura vacinal para outras doenças

6. BIBLIOGRAFIA

Bastos, L. (2018). *Analysis of Performance in Intensive Care Units*. Dissertação (Mestrado). Departamento de Engenharia Industrial. Pontifícia Universidade Católica do Rio de Janeiro. Disponível em www.maxwell.vrac.puc-rio.br/35727/35727.PDF. Acessado em: 24 de out.de 2022.

Bastos, L., Aguilar, S., Rache, B., Maçaira, P., Baião, F., Cerbino-Neto, J., . . . Bozza, F. A. (2022). Primary Healthcare Protects Vulnerable Populations from Inequity in COVID-19 Vaccination: An Ecological Analysis of Nationwide Data from Brazil. *The Lancet Regional Health - Americas, 14*, 100335. doi: https://doi.org/10.1016/j.lana.2022.100335

Bichler, M., Heinzl, A., & van der Aalst, W. M. P. (2017). Business Analytics and Data Science: Once Again? *Business & Information Systems Engineering, 59*(2), 77-79. doi: 10.1007/s12599-016-0461-1

Cao, L. (2016). Data Science and Analytics: A New Era. *International Journal of Data Science and Analytics, 1*(1), 1-2. doi: 10.1007/s41060-016-0006-1

Cassiano, K. M. (2014). *Análise de Séries Temporais Usando Análise Espectral Singular (SSA) e Clusterização de Suas Componentes Baseada em Densidade*. Dissertação (Doutorado). Departamento de Engenharia Elétrica. Pontifícia Universidade Católica do Rio de Janeiro. Disponível em http://meusite.mackenzie.com.br/rogerio/tgi/2004Cluster.PDF. Acessado em: 24 de out.de 2022.

Conselho Nacional de Saúde (2022). Vacina Inativada da Pólio Completa 10 Anos com Baixa Adesão no Brasil. Disponível em http://conselho.saude.gov.br/ultimas-noticias-cns/2581-vacina-inativada-da-polio-completa-10-anos-com-baixa-adesao-no-brasil#. Acessado em: 21 de out. de 2022.

Costa, A., & Leo, A. (2020). Estatística Descritiva: Principais Conceitos *Data Science e Direito*. Disponível em https://dsd.arcos.org.br/estatistica-descritiva-principais-conceitos/. Acessado em: 24 de out.de 2022.

Dabbura I. (2018). K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. Disponível em https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Doni, M. V. (2004). *Análise de Cluster: Métodos Hierárquicos e de Particionamento*. Trabalho de Conclusão de Curso (Bacharel em Sistemas de Informação). Universidade Presbiteriana Mackenzie. Disponível. Acessado em: 24 de out.de 2022.

Fiocruz (2022). Poliomelite: Sintomas, Transmissão e Prevenção. Disponível em https://www.bio.fiocruz.br/index.php/br/poliomielite-sintomas-transmissao-e-prevencao. Acessado em: 22 de out. de 2022.

Fujita, D. M., Gomes da Cruz, T. C., Ferreira, E. M., & Henrique da Silva Nali, L. (2022). The Continuous Decrease in Poliomyelitis Vaccine Coverage in Brazil. *Travel Medicine and Infectious Disease, 48*, 102352. doi: https://doi.org/10.1016/j.tmaid.2022.102352

G1 (2022). Brasil Vacinou Apenas 54% das Crianças Contra a Poliomielite; Campanha Termina na Sexta. Disponível em https://g1.globo.com/saude/noticia/2022/09/29/brasil-vacinou-apenas-54percent-das-criancas-contra-a-poliomielite-campanha-termina-na-sexta.ghtml. Acessado em, 21 de out. de 2022.

Gov.br (2022). Disponível em https://www.gov.br/saude/pt-br/composicao/se/demas/campanhas-de-vacinacao. Acessado em, 21 de out. de 2022.

Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2001). On Clustering Validation Techniques. Journal of Intelligent Information Systems 17, 107–145. doi: https://doi.org/10.1023/A:1012801612483

IBGE (2022). Site do IBGE. Disponível em https://www.ibge.gov.br/pt/inicio.html. Acessado em: 18 de out. de 2022.

IEPS (2022). Instituto de Estudos para Políticas de Saúde. IEPS Data. Disponível em https://gitlab.com/ieps-data/indicadores/-/tree/main. Acessado em: 18 de out. de 2022.

Instituto Butantan (2022). Doenças Erradicadas Podem Voltar: Conheça Quatro Consequências Graves da Baixa Imunização Infantil. Disponível em https://butantan.gov.br/noticias/doencas-erradicadas-podem-voltar-conheca-quatro-consequencias-graves-da-baixa-imunizacao-infantil. Acessado em, 21 de out. de 2022.

Laboissière, P. (2018). Saiba Quais Doenças Voltaram a Ameaçar o Brasil. *Agência Brasil*. Disponível em https://agenciabrasil.ebc.com.br/saude/noticia/2018-07/saiba-quais-doencas-voltaram-ameacar-o-brasil. Acessado em: 21 de out. de 2022.

Ledford, H. (2022). COVID Vaccine Hoarding Might Have Cost More than a Million Lives Disponível em https://www.nature.com/articles/d41586-022-03529-3?WT.ec_id=NATURE-202211&sap-outbound-id=58FBF4C74186B35AB624C577D3CC92B8B396AF45. Acessado em: 14 de nov. de 2022.

Maçaira, P. B., L.; Aguilar, S. & Peres, I. . (2022). *Inferência Estatística com R* (1ª edição). Rio de Janeiro, Brasil.

Manresa, A. P. (2020). *Machine Learning to Predict High-Cost Hospitalizations*. Dissertação (Mestrado). Departamento de Engenharia Industrial. Pontifícia Universidade Católica do Rio de Janeiro. Disponível em [https://www.maxwell.vrac.puc-rio.br/49137/49137.PDF&ved=2ahUKEwjC1c7iyPb6AhWdA7kGHamfBM4QFnoECBMQAQ&usg=AOvVaw1HuWe8r5upztgyAljV4UjS. Acessado em: 23 de out. de 2022.

Mehta, S. (2022). A Tutorial on Various Clustering Evaluation Metrics. Disponível em https://analyticsindiamag.com/a-tutorial-on-various-clustering-evaluation-metrics/. Acessado em.

Nunes, L. (2021). Cobertura Vacinal no Brasil 2020. Instituto de Estudos para Políticas de Saúde. Disponível em https://ieps.org.br/wp-content/uploads/2021/05/Panorama_IEPS_01.pdf. Acessado em: 22 de out. de 2022.

Reis, E. A., & Reis, I. A. (2002). Análise Descritiva de Dados. Relatório Técnico do Departamento de Estatística da UFMG 1. Disponível em http://www.est.ufmg.br/portal/arquivos/rts/rte0202.pdf Acessado em: 24 de out. de 2022.

Rousseeuw, P. J. (1986). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65. doi: https://doi.org/10.1016/0377-0427(87)90125-7

Scikit-learn (2022). Clustering Performance Evaluation. Disponível em https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis, Communications in Statistics, 3:1, 1-27. doi: 10.1080/03610927408827101

UNASUS (2021). Ministério da Saúde Promove Webinário para Reforçar Importância do Combate à Poliomielite. Disponível em https://www.unasus.gov.br/noticia/ministerio-da-saude-promove-webinario-para-reforcar-importancia-do-combate-a-poliomielite. Acessado em: 22 de out. de 2022.

UNASUS (2022). PNI: Entenda Como Funciona um dos Maiores Programas de Vacinação do Mundo. Disponível em https://www.unasus.gov.br/noticia/pni-entenda-como-funciona-um-dos-maiores-programas-de-vacinacao-do-mundo. Acessado em, 20 de out. de 2022.