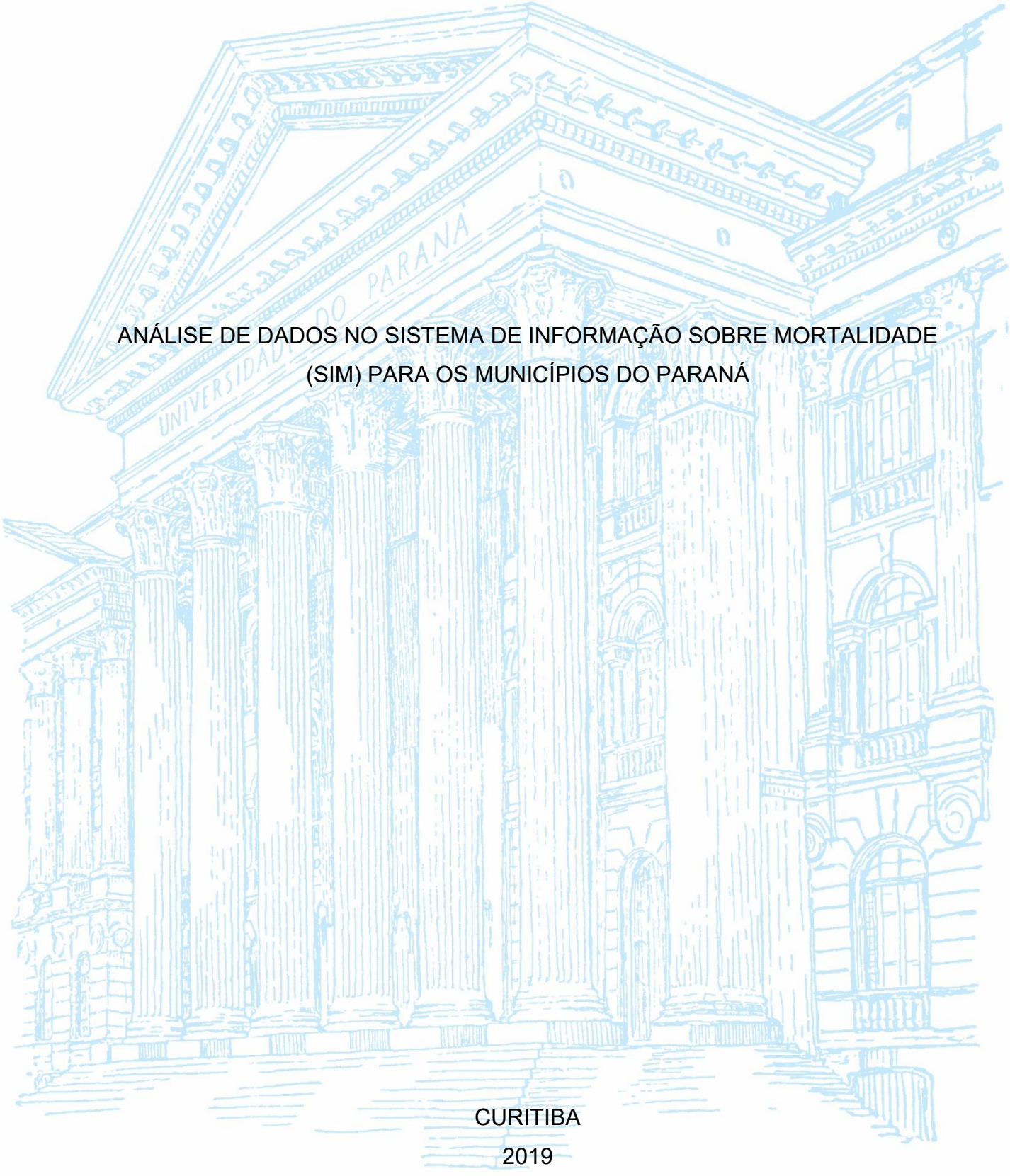
UNIVERSIDADE FEDERAL DO PARANÁ

JULIANO LUIZ DA SILVA



ANÁLISE DE DADOS NO SISTEMA DE INFORMAÇÃO SOBRE MORTALIDADE (SIM) PARA OS MUNICÍPIOS DO PARANÁ

CURITIBA

2019

JULIANO LUIZ DA SILVA

ANÁLISE DE DADOS NO SISTEMA DE INFORMAÇÃO SOBRE MORTALIDADE (SIM) PARA OS MUNICÍPIOS DO PARANÁ

Trabalho de conclusão de curso apresentado ao curso de Graduação em Gestão da Informação, departamento de Ciência e Gestão da Informação, do Setor de Ciências Sociais Aplicadas da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Bacharel em Gestão da Informação.

Orientadora: Profª Drª Denise Fukumi Tsunoda

CURITIBA

2019

TERMO DE APROVAÇÃO

JULIANO LUIZ DA SILVA

ANÁLISE DE DADOS NO SISTEMA DE INFORMAÇÃO SOBRE MORTALIDADE (SIM) PARA OS MUNICÍPIOS DO PARANÁ

Trabalho de Conclusão de Curso apresentado ao curso de Graduação em Gestão da Informação Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Bacharel em Gestão da Informação.

_____

Profa. Dra. Denise Fukumi Tsunoda - Orientadora
Departamento de Ciência e Gestão da Informação
UFPR

_____

Prof. Dr. Cicero Aparecido Bezerra
Departamento de Ciência e Gestão da Informação
UFPR

_____

Profa. Dra. Deborah Ribeiro Carvalho
Programa de Pós-graduação em Tecnologia em Saúde
PUCPR

Curitiba, 24 de junho de 2019.

AGRADECIMENTOS

Agradeço a todos que me auxiliaram nesta caminhada, à professora Denise pela paciência e cuidado, a Luana por sempre me incentivar, aos amigos e familiares pela força nos momentos em que tudo não fazia sentido. Aos colegas do grupo de estudos de estatística espacial da PUC pela oportunidade.

O que não me faz morrer me torna mais forte. (Friedrich Nietzsche, 2004, p. 10)

RESUMO

A saúde infantil destaca-se como um dos objetivos do milênio da Organização das Nações Unidas, a qual descreve a redução da mortalidade infantil como ponto crítico para desenvolvimento dos países. A saúde infantil pode ser mensurada através de índices e taxas, como a taxa de mortalidade infantil e materna. Nos últimos dez anos a taxa de mortalidade infantil apresenta redução no mundo e o Brasil acompanha essa tendência, assim como o estado do Paraná que apresentou em 2016 a taxa de 11,93‰. Pretende-se, a partir desse estudo, verificar as variáveis que influenciam a taxa de mortalidade infantil a partir da estatística, estatística espacial e mineração de dados. Para isso foi utilizada a base de dados SIM (Sistema de Informação sobre Mortalidade) e SINASC (Sistema de Informação sobre Nascidos Vivos) além de variáveis demográficas disponibilizadas pelo IBGE e indicador de cobertura de atenção básica dos municípios. Com a utilização de métodos de mineração de dados para a descoberta de padrões, verifica-se uma grande quantidade de padrões reportados nas árvores de decisão. Obteve com o algoritmo J48 a taxa de classificação corretas dos municípios do Paraná de 45,36% para a classificação a partir do triênio, 45,61% no ano de 2014, 48,62% para o ano de 2015 e 44,36% para o ano de 2016. Sob a perspectiva geográfica destaca-se a verificação de padrões de taxas similares entre municípios próximos e a verificações de mudanças ao longo do triênio (2014-2016) em especial na redução da taxa de mortalidade infantil na região sul do estado.

Palavras-chave: Mortalidade infantil. Sistema de Informação sobre Mortalidade. Mineração de dados. Estatística espacial. Brasil. Paraná.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A saúde infantil é um tema de constante preocupação das instituições de saúde ao redor do mundo. Esta preocupação é materializada pela elaboração das metas do milênio pela Organização das Nações Unidas, estabelece a ampliação do cuidado materno infantil, mensurado por índices e indicadores como a taxa de mortalidade infantil e a taxa de mortalidade materna.

Nos últimos dez anos, a taxa de mortalidade infantil apresenta redução no mundo, o Brasil acompanha essa tendência caminhando para números próximos aos dos países desenvolvidos, reduzindo sua taxa de mortalidade infantil de 30,4 ‰ nos anos 2000 para 14,5 ‰ em 2014. Entretanto, há disparidades entre regiões dentro do território brasileiro, por exemplo: a região norte do país apresenta uma taxa de 17,6 ‰ contra 9,4 ‰ na região sul. Esta diferença pode ser atribuída à multiplicidade de fatores, dentre os quais, destaca-se o nível de atuação da rede de atenção básica local, concentração e distribuição de renda, poder aquisitivo familiar e nível de saneamento básico.

A mortalidade infantil é definida como o óbito de crianças entre zero a um ano de idade, dentro deste composto é possível segmentar em dois componentes: mortalidade neonatal e mortalidade pós-neonatal (BRASIL, 2005).

A situação da mortalidade infantil vem sendo monitorada por gestores de saúde do estado do Paraná, disso promovem políticas de atenção e cuidado a população. A partir do ano de 2011 houve a implantação do Programa Rede Mãe Paranaense cujo objetivo é acompanhamento materno infantil, da fase pré-natal até a criança completar um ano de idade. Dentre as ações estabelecidas pela Secretária de Saúde com o programa estão: “estabelecimento de no mínimo de sete consultas pré-natal, realização de exames, estratificação e atendimento especializado dos pacientes alvos do programa de acordo com o risco e garantia de parto em hospital de acordo com o nível de risco” (PARANÁ, 2012, p. 13). O programa tem como objetivos a redução da mortalidade materno-infantil e manutenção do funcionamento da rede de atenção materno-infantil no estado do Paraná.

Segundo Paraná (2012) O programa teve início oriundo do sucesso do programa Mãe Curitibana (focado no público materno-infantil da capital do estado do Paraná) e da análise dos dados realizada nas bases de dados do DATASUS pela Secretária de Saúde entre os anos de 2006 a 2010. Por certo a análise de dados sob

este aspecto, auxiliou na tomada de decisão e na construção desta política pública. Para efetuar a análise deste tipo é fundamental a existência de tratamento e de transformação dos dados, a fim de atribuir a estes valor e conclusões diferentes das encontradas inicialmente.

Portanto, o presente estudo se propõe aplicar métodos de análise de dados, para identificar padrões que possam estar associados à fatores relacionados para entender a variações da taxa de mortalidade infantil no Estado do Paraná.

1.1 PROBLEMATIZAÇÃO

Políticas públicas auxiliam no aumento da qualidade de saúde. No Paraná o programa Rede Mãe Paranaense permite políticas assistenciais relacionadas ao cuidado materno-infantil. Netto et al. (2017) avalia o programa Rede Mãe Paranaense sob ponto de vista da 9ª regional de saúde do Paraná, nessa análise é possível acompanhar os óbitos sob ponto de vista de sua das causas evitáveis, ou seja, óbitos que poderiam ser evitados caso houvesse um adequado cuidado a mãe e ao recém-nascido. Como desfecho do estudo apresenta que houve “discreta redução nos óbitos após a implementação do programa na regional analisada. Assim sendo, a identificação do grau de evitabilidade do óbito se traduz em ações mais assertivas, para traçar com eficácia os investimentos e direcionar o atendimento a atenção materno-infantil.

Diante desses fatos, é válido mencionar que há poucos estudos relacionados a avaliação do programa Rede Mãe Paranaense e sua eficácia. Foi encontrado seis resultados com a busca pelos termos “mortalidade infantil” e “mãe paranaense”, logo sendo um assunto relevante a abordar com estudos direcionados. Desta forma, a mortalidade infantil a partir da rede assistencial otimizada necessita de acompanhamento e possivelmente de sugestões de melhorias.

Nesse contexto, procura-se identificar o panorama dos óbitos infantis no Paraná, a partir de um estudo de seus municípios respondendo o seguinte problema de pesquisa: **quais os variáveis que influenciam na variação da taxa de mortalidade infantil entre os municípios do Paraná, entre os anos de 2014 a 2016, utilizando a base de dados SIM (Sistema de Informação sobre Mortalidade)?**

1.2 OBJETIVOS

A partir do problema de pesquisa foram elaborados os objetivos gerais e os específicos, reduzindo escopo deste trabalho à resposta de seu problema. Este estudo se propõe a aplicar métodos de análise de dados, para identificar padrões associados a fatores relacionados a variações da taxa de mortalidade infantil entre os municípios do Estado do Paraná.

Os objetivos específicos estão diretamente ligados com o objetivo geral, propondo um detalhamento deste, são objetivos específicos:

- analisar a base de dados de mortalidade infantil por meio da estatística descritiva;
- utilizar métodos de mineração de dados a fim de descobrir padrões relevantes na base de dados do estado do Paraná no período estudado;
- descrever geograficamente a mortalidade infantil no Paraná.

1.3 JUSTIFICATIVA ACADÊMICA

Destaca-se então o papel do analista de dados junto ao profissional de saúde na construção do cerne de conhecimento necessário para uma análise efetiva. O analista de dados é uma das possíveis atribuições que o Gestor da Informação pode praticar, o qual possui conhecimento das áreas estatísticas e de aprendizado de máquina, o que propicia sua capacidade de para construir a rede de conhecimento a partir de dados. Ademais, possui as competências necessárias para ser o elo entre a área estratégica e o profissional de saúde, tal qual este profissional, multidisciplinar que é está acostumado a trabalhar. Visto que a Gestão da Informação, conforme constituída em seu curso da Universidade Federal do Paraná, possui como tripé as áreas de Tecnologia da Informação, Administração e Ciência da Informação.

1.4 JUSTIFICATIVA CIENTÍFICA

A seção de justificativa busca descrever a contribuição de determinado estudo para a produção científica de sua área. Uma forma de analisar a contribuição é revisitar os estudos relacionados publicados. Portanto, foi realizado um levantamento

bibliográfico de artigos publicados nos últimos cinco anos (2015 a 2019) a partir do portal de periódicos da CAPES, o portal é uma biblioteca virtual que reúne e disponibiliza às instituições de ensino e pesquisa no Brasil o melhor da produção científica internacional.

A estratégia deste levantamento foi fundamentada na combinação de termos a partir de operadores de booleanos (AND e OR), em conjunto com os filtros disponíveis no portal de periódicos CAPES. Os termos de buscas utilizados foram exatos, ou seja, termos compostos deveriam aparecer sequencialmente. Foram utilizados os filtros:

- por idioma (português e inglês);
- local onde o termo aparece no artigo (por título);
- por período (últimos 5 anos)

No Quadro 1 é possível verificar a descrição dos termos e quantidade de resultados encontrados, tanto em inglês como em português, a partir da busca efetuada em 06 de abril de 2019.

Quadro 1 - Quantidade de artigos resultantes no portal de periódicos da CAPES nos últimos 5 anos (2015 a 2019) de acordo com idioma.

termo	**PT-BR**	**EN**	**termo**
SEM FILTRO: (mortalidade infantil)	379	17.984	**QUALQUER**: (infant mortality)
TÍTULO: (mortalidade infantil)	36	622	**TÍTULO**: (infant mortality)
TÍTULO: (mortalidade infantil) AND **QUALQUER**: (brasil)	27	58	**TÍTULO**: (infant mortality) AND **QUALQUER**: (brazil)
TÍTULO:(mortalidade infantil) **AND** SIM	11	20	**TÍTULO**:(infant mortality) **AND** SIM
TÍTULO:(mortalidade infantil) **AND** SIM **AND** paraná	2	4	**TÍTULO**:(infant mortality) **AND** SIM **AND** paraná

FONTE: O Autor (2019).

Foram analisados os 31 artigos resultantes das buscas que continham os termos mortalidade infantil e a base de dados SIM, e selecionadas as que atenderam aos critérios:

- tema principal sendo a mortalidade infantil no Brasil;
- não deve ser Meta-Análise ou revisão de bibliografia;

- apresentar estatística descritiva dos dados utilizados;
- apresentar a fonte dos dados;
- apresentar o método e a ferramenta utilizada para a sua aplicação;
- apresentar resultados relevantes (conclusões tais como “precisamos melhorar as políticas públicas” foram suprimidas).

Destes, somente nove atenderam a todos os critérios, excluídos oito artigos duplicados e demais por não conformidade. A partir dos artigos restantes será detalhado o estudo de cinco, como critério de escolha foi a pertinência e relação com o presente estudo.

Kropiwiec et. al. (2017) procuram identificar os fatores associados à mortalidade infantil no município de Joinville no estado de Santa Catarina. A partir das bases de dados SIM e SINASC, CNES (cadastro nacional de estabelecimentos de saúde) e Relatório Anual para a categorização de dois modelos de atenção básica. A partir de variáveis demográficas da mãe, das variáveis: etnia, escolaridade, ocupação, situação conjugal materna, nascimento em outro município, modelo de atenção básica, local do nascimento, tipo de estabelecimento, complexidade do hospital, número de filhos mortos, paridade materna, mês do início do pré-natal, número de consultas pré-natal, tipo de parto, idade materna, tipo de gestão, idade gestacional, sexo e peso do recém-nascido, índice de Agpar no 1º e 5º minuto e presença da má formações. A partir disso foi realizada a análise estatística com a utilização de modelos de regressão logística. Como resultado, apresenta que fatores que constituem risco para os óbitos são: Mãe adolescente, duração da gestação menor que 32 semanas, peso ao nascer menor que 1.500 g, Apgar no 1º e no 5º minuto de vida menor que 7.

Leal et al. (2017) estudam os determinantes sociais, demográficos, da saúde reprodutiva e utilização dos serviços de saúde e sua associação com a mortalidade infantil em 75 munícipios de pequeno porte no Vale do Jequitinhonha e nas regiões Norte e Nordeste do Brasil. Por meio de um estudo caso-controle, verificam variáveis demográficas maternas, de saúde reprodutiva, adequação de pré-natal através de um método estatístico regressão logística. Apresentam fatores já então conhecidos, como ocorrência de óbito entre mães com história de perda fetal e infantil e que não fizeram pré-natal adequado. Destacam o não encaminhamento de mulheres de alto risco durante o pré-natal para serviços especializados, o grande número de partos em domicílios que ocorrem em municípios de pequeno e médio porte por conta de

dificuldade de acesso à maternidades. Por fim, reforçam maior adesão dos profissionais de saúde aos protocolos assistenciais do Ministério da Saúde.

Lima et al. (2017) estudam fatores relacionados à mortalidade infantil nas mães residentes de Cuiabá no estado do Mato Grosso. Utilizam da estatística através da regressão logística e obtém como resultado a taxa de mortalidade: mães sem companheiro, baixo número de consultas de pré-natal, baixo preso ao nascer, prematuridade, Agpar <= 7 no 1º minuto, malformação congênita e sexo masculino. Apresentam ainda, reflexos positivos do programa Bolsa Alimentação e Bolsa Família na redução da mortalidade infantil em Cuiabá.

Sanders et al. (2017) analisam fatores associados a mortalidade infantil no município de Fortaleza no estado do Ceará através de um estudo de caso-controle. Utilizam da regressão logística e tem como resultado variáveis associadas à mortalidade infantil, sendo estas: gestação gemelar, idade gestacional <= 36 semanas.

Rodrigues et al. (2014) apresentam um estudo utilizando análise espacial da mortalidade infantil e a adequação das informações vitais no estado de Pernambuco. Utilizam dados secundários do SIM e SINASC para o desenvolvimento de cinco (5) indicadores por município: coeficiente de mortalidade geral padronizado por idade, desvio médio relativo do coeficiente de mortalidade geral, razão entre nascidos vivos informados e estimados, desvio médio relativo da taxa de natalidade e proporção de óbitos sem definição de causa básica. Com isso, foi utilizado a estatística espacial a partir do índice de Moran Local, para identificar agregados espaciais de mortalidade infantil. Como resultado relatam que 76,6% dos municípios apresentam informações vitais consolidadas. Concluem descrevendo a formação de cluster para a mortalidade infantil em 34 municípios, formando três agregados espaciais.

1.5 DELIMITAÇÕES DA PESQUISA

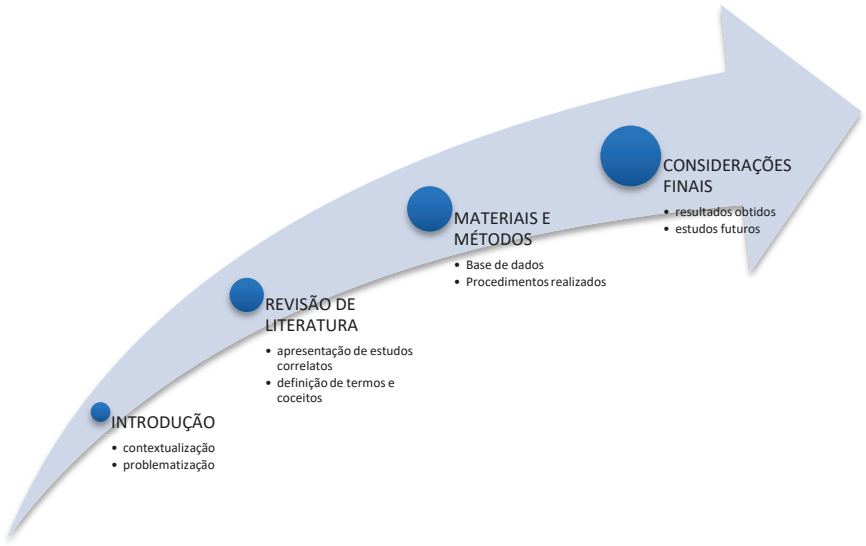
Trata-se de um estudo com crianças que vieram à óbito, de idade de zero a um ano, residentes no estado do Paraná no Brasil, no período entre os anos de 2014 a 2016. Sendo este período o triênio (período de três anos) mais recente completo disponível na base de dados SIM (Sistema de Informação sobre Mortalidade).

O presente estudo não abordará questões relacionadas a subnotificação de óbitos, ou seja, a possibilidade da existência de óbitos além dos registrados na base SIM.

1.6 ESTRUTURA DO DOCUMENTO

A figura apresenta um compilado das seções principais (Figura 1) e dos principais tópicos abordados em cada etapa. Esta é uma esquematização visual que tem como objetivo representar fielmente todas as etapas relacionadas em cada fase.

Figura 1 - Etapas do Estudo



FONTE: O autor (2019).

A figura apresenta distinção entre as quatro principais etapas do desenvolvimento deste estudo, todavia a forma de sua construção não foi procedural, muitas vezes alternando entre as etapas ou então realizando-as em paralelo. A revisitação das seções já trabalhadas é comum e busca alinhar o estado atual do artigo com o que já foi desenvolvido, esta etapa sendo por vezes árdua e repetitiva.

2 REVISÃO DE LITERATURA

A revisão de literatura tem como tarefa de familiarizar o leitor com o conteúdo que está sendo apresentado. Portanto, será exposto a contextualização sobre termos como a Análise de dados, atrelada em seu uso na tomada de decisão na área de saúde. Em seguida será apresentado o papel fundamental do Sistema Único de Saúde para a população brasileira, bem como seu histórico.

2.1 ANÁLISE DE DADOS

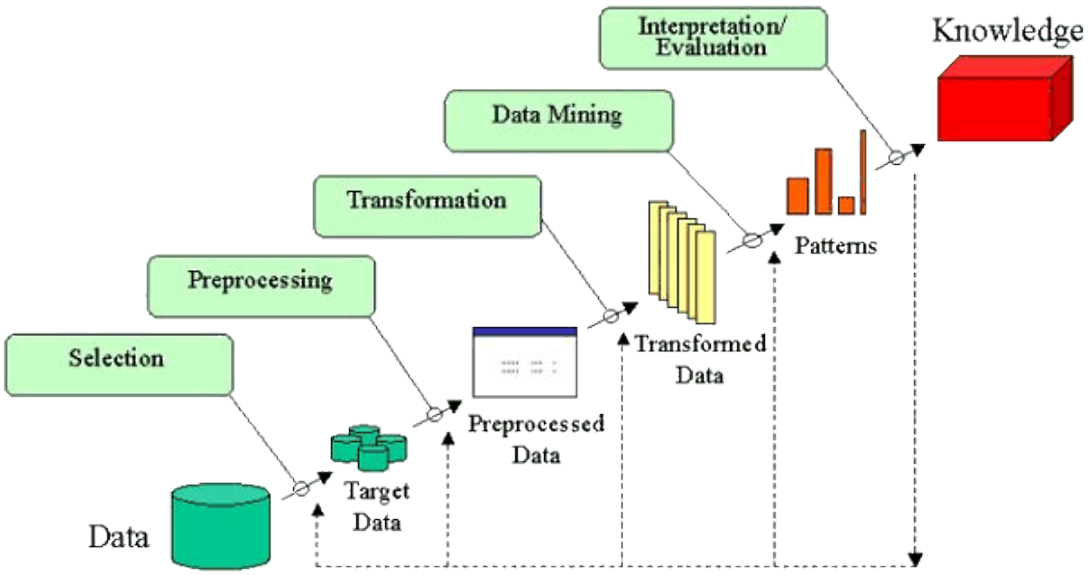
O termo análise vem do grego *análusis,eōs*, que significa dissolução, separação em partes. Já o termo dado, pode ser definido como registro sobre algum fato observado. Neste contexto, a análise de dados é o processo de quebra dos dados em unidades menores, a fim de entender algum fenômeno. Ou então sob uma ótica de termo genérico guarda-chuva, ao qual é possível entender como um aglomerado das áreas como mineração de dados e a estatística. Teixeira descreve análise de dados sob ponto de vista da metodologia científica como:

> O processo de formação de sentido além dos dados, e esta formação se dá consolidando, limitando e interpretando o que as pessoas disseram e o que o pesquisador viu e leu, isto é, o processo de formação de significado (TEIXEIRA, 2003, p. 191-192).

Diante disso, a análise de dados como ferramenta de entendimento de algum fenômeno perpassa na socialização deste conhecimento para evolução da ciência, ao qual só é possível a partir de uma metodologia passível de reprodução dos resultados. Serão utilizados conceitos e métodos da mineração de dados, estatística descritiva e estatística espacial.

O processo de mineração de dados é um dos componentes da descoberta de conhecimento em bases de dados (KDD), e tem como objetivo a produção de conhecimento a partir dos dados (Figura 2).

Figura 2 - Processo de KDD



Fonte: Fayyad (1996).

Dentre as etapas estão a seleção de dados, ou seja, a partir de determina população se faz necessário selecionar uma amostra para trabalho, além do estabelecimento de determinado *problema* a ser solucionado. Pré-processamento ocupa-se com a limpeza e adequação da base, como o tratamento de *outliers*, valores incorretos ou faltantes. A seguir se dá a etapa de transformação de dados, pois em alguns algoritmos é necessário alterar escala em determinadas variáveis. Perpassa pela etapa de mineração de dados que ocupa na aplicação dos algoritmos e por último a análise interpretação dos resultados, nesta etapa o resultado da mineração é analisado levando em conta fatores como: taxa de acerto, velocidade de processamento, adequação ao problema proposto.

O algoritmo a ser implementado é o J48, que é a implementação na linguagem de programação Java do algoritmo escrito em C denominado C 4.5 que por sua vez tem origem no algoritmo ID3 (CHENG 2008). DUTRA (2008) relata que o método C4.5 tem sido largamente empregado para construir árvores de decisão que são muito comumente utilizadas para descoberta de padrões. O algoritmo tem como resultado a porcentagem de classificação de registros correto e incorretos, a árvore de decisão e a matriz de confusão HAY (1988).

A outra dimensão da análise de dados a ser tralhada é a abordagem da estatística descritiva. Esta pode ser entendida a partir da organização e descrição dos dados, a partir de sumários e indicadores segundo Silvestre (p. 4, 2007). As variáveis das bases de dados constituem-se conforme descrito por Gonçalves (p. 13, 1978) a partir de dois tipos: variáveis qualitativas e quantitativas.

2.2 SISTEMA ÚNICO DE SAÚDE BRASILEIRO

O Departamento de Informática do Sistema Único de Saúde (DATASUS) é o ator que promove e mantém os dados vitais sobre mortalidade, dito isso o papel fundamental do Sistema Único de Saúde (SUS), seu funcionamento e histórico auxiliam no entendimento da base de dados a ser estudada.

A partir do relatório publicado pelo Ministério da Saúde é possível entender o SUS como um sistema de abrangência nacional, democrático apoiado pela participação social (BRASIL, 2001, p. 5). Criado no período da redemocratização pós ditadura (1985), momento de grandes mudanças estruturais políticas (ASSUNÇÃO et al. 2014). Convocada pela Presidência da República da época, a CNS (Conferência Nacional de Saúde) aprova demandas da sociedade a favor da reforma da saúde. Dentre as pautas aprovadas estão o fortalecimento do setor público de saúde, expansão da cobertura a todos os cidadãos e unificação da medicina previdenciária à saúde pública, constituindo assim um sistema único. Paim aponta os conceitos e concepções levados para a CNS como:

> Determinação social do processo saúde-doença, organização social dos serviços de saúde, com matriz teórica marxista, mas também a promoção da saúde, consciência sanitária, políticas públicas e a noção de sistema de saúde (PAIM, 2008, p. 99).

Portanto, o SUS tem como papel muito além da cura de agravos, mas na melhoria da qualidade de vida de determinada população (BRASIL, 2001, p. 5) e apresenta como princípios a universalidade, integridade e equidade.

2.3 EPIDEMIOLOGIA

A epidemiologia trata entre outras coisas o componente de saúde de determinada população, logo, engloba o tema mortalidade infantil. Pode-se entender epidemiologia como: “conceitos, métodos e teorias que permitam estudar, conhecer e transformar o processo saúde-doença na dimensão coletiva” (FRANCO e PASSOS, 2005, p. 31). Estes métodos perpassam no processo de construção do diagnostico coletivo conforme explica Soares et al. (2019, p. 184) como: levantamento de dados da população, de doenças encontradas nesta população, da contabilização da morbidade e por fim de registros do serviço de saúde local. Demonstra que é necessário dimensionar estas variáveis a partir de três vertentes: pessoas, tempo e espaço. Retrata a origem da epidemiologia, que até o início do século XX focava mais fortemente nas doenças infecciosas. Atualmente, relata uma mudança no perfil epidemiológico da população e com isso a mudança do foco da epidemiologia. A partir de métodos do século XX os estudos focam em outros tipos de doenças como doenças não infeciosas, agravos de causas externas ou então desvios nutricionais.

Os estudos de epidemiologia atualmente conforme conta Soares et al. (2019, p. 186) adere a métodos da epidemiologia analítica que atua na descoberta das causas para a ocorrência das doenças. A descoberta das causas perpassa na utilização de dados e em sua transformação em indicadores, para que possam ser comparados com outros locais ou momentos do tempo. Destaca que não é comum elaborar comparações a partir de números absolutos, portanto elabora-se frequência relativas por meio de proporção ou coeficientes (Quadro 2).

Quadro 2 - Razões utilizadas para construção de Indicadores na Epidemiologia

Nome	Descrição
Proporção	Representa a importância desses casos ou mortes no conjunto total
Coeficiente	Representa o risco de determinado evento ocorrer na população

Fonte: Adaptado de Soares (2019).

Relata a diferença do índice, que diferentemente de coeficiente não expressa probabilidade, porque o numerador não apresenta a ocorrência do evento. Esclarece

então que o termo mais indicado para a mortalidade infantil é índice de mortalidade infantil não coeficiente.

2.4 MORTALIDADE INFANTIL

A mortalidade infantil é estudada a partir das mais variadas visões como: social, epidemiológica e estatística. Dentre as diversas abordagens, o ponto de partida é o significado de mortalidade e mortalidade infantil e suas implicações. Logo, buscando aprofundar a análise, será adotado uma das definições do Ministério da Saúde brasileiro sobre o tema, que traz a definição de mortalidade como: “desaparecimento de qualquer sinal de vida em qualquer momento após o nascimento, sem possibilidade de ressuscitação (MINISTÉRIO DA SAÚDE, 2007, p. 9). A morte sob esta visão é um dado a ser registrado com intuito de elaborar ações para a melhoria da qualidade de vida da população. No Brasil, a mortalidade é registrada pelo departamento de informática do SUS a partir da declaração de óbito, documento base do Sistema de Informação sobre Mortalidade, composta de três vias de igual conteúdo distribuídas pelas secretárias estaduais e municipais (BRASIL, 2009). Seu conteúdo apresenta informações relativas à identificação do indivíduo, bem como causa de morte e informações demográficas utilizadas para ações do âmbito de saúde pública. As causas de morte são agrupadas em evitáveis e não evitáveis, sendo as evitáveis “definidas como uma morte que pode ser evitável com adequação do cuidado de saúde, a partir de serviços efetivos de saúde (MALTA, 2007).

Ainda é possível estratificar a mortalidade infantil em dois grupos etários, mortalidade neonatal e mortalidade pós-neonatal, onde a mortalidade neonatal conta desde o nascimento até 28 dias incompletos e a mortalidade pós neonatal que abrange 28 dias após o nascimento até 365 dias incompletos. A taxa de mortalidade infantil ou TMI, por isso, é calculado utilizando duas variáveis, conforme a fórmula (1):

$$TMI = \frac{Óbitos\ por\ residência}{Nascidos\ vivos\ por\ residência} * 1000 \quad (1)$$

Com a taxa de mortalidade infantil é possível comparar e verificar a qualidade de vida da população e com isso promover a criação de políticas públicas coerentes com o estado de saúde desta população.

Como os casos de mortalidade serão analisados a partir da unidade federativa do Paraná, é essencial conhecer a geografia e as características demográficas do estado, para então inferir se os padrões encontrados a partir dos algoritmos possuem relevância. Portanto a próxima seção tratará de estabelecer este elo.

2.5 DIVISÃO GEOGRÁFICA E ANÁLISE ESPACIAL

Um dos pontos importantes na utilização dos indicadores conforme explanado na seção de epidemiologia, é a possibilidade de comparação temporal e geográfica, dito isso, é essencial entender o contexto em que se estabelece os dados de saúde infantil, no Paraná, a partir de uma análise de seu território. O paraná constitui-se como unidade federativa dentro do território brasileiro (27º estado mais populoso do país), possui 399 municípios e extenso território e uma população de 10.444.526 pessoas de acordo com o último censo do IBGE em 2010. Faz fronteira com estados de São Paulo, Santa Catarina, Mato Grosso do Sul e com os países Argentina e Paraguai. Dentre as características dos municípios estão o PIB (Produto Interno Bruto) que é a soma de todas as riquezas de determinado município.

2.5.1 ESTATÍSTICA ESPACIAL

A estatística espacial será uma das visões utilizadas para a identificação de padrões, conforme definição de Andrade (2019, p.17) estatística espacial é:

> O ramo da estatística que permite analisar a localização espacial de eventos. Ou seja, além de identificar, localizar e visualizar a ocorrência de fenômenos que se materializam no espaço, tarefas possibilitadas pelo uso dos SIG (Sistema de Informação Geográfica), utilizando-se a estatística espacial é possível modelar a ocorrência destes fenômenos, incorporando, por exemplo, os fatores determinantes, a estrutura de distribuição espacial ou a identificação de padrões (ANDRADE, 2007, p. 17).

Constata-se a partir o potencial da estatística espacial na identificação de padrões, conforme estudado por BEZERRA FILHO (2007) et al. Dos determinantes

da taxa de mortalidade infantil no Ceará e MORAIS NETO et al. (2001) para identificar as áreas de risco no município de Goiânia.

Como insumos para a utilização dos métodos de estatística espacial estão alguns conceitos chave como: ponto, áreas e dependência espacial. Sendo o ponto, explicado por Andrade et al. (2007) como o uma localização pontual (coordenadas). Descreve que o termo área como áreas geográficas com limites definidos, utiliza-se na literatura o conceito de polígonos.

A estatística espacial apresenta variados meios de análise, dentre os diversos algoritmos existentes dentre deste campo será utilizado o *box-map* para visualização dos quadrantes conforme NASCIMENTO et al. (2014) realizaram no estudo com a mortalidade neonatal em São Paulo.

2.6 RELAÇÕES ENTRE A MORTALIDADE INFANTIL E SISTEMA DE INFORMAÇÃO SOBRE MORTALIDADE

A partir de um detalhamento da tabela apresentada na justificativa (Tabela 5) é possível verificar que: 5 (cinco) dos 9 (nove) trabalhos recuperados utiliza regressão (logística ou polinomial) em suas análises. Ainda, 4 (quatro) explicitam o uso da ferramenta SPSS e 2 (dois) o software STATA. Apesar de apenas 2 (dois) artigos mencionarem o Excel, é provável que todos os trabalhos o utilizem, visto que é uma ferramenta bastante conhecida para a tabulação e visualização de dados.

Visualizando os estudos a partir de sua geografia é possível observar que não foram encontrados estudos dentro dos filtros propostos que tem como objeto o estado do Paraná ou então as suas cidades.

Tabela 1 - Artigos resultantes das buscas no portal de periódicos CAPES que atenderam aos critérios de seleção a partir dos termos mortalidade infantil e SIM.

Método Comum	**Quantidade de Artigos**	**Título**	**FERRAMENTAS**
Regressão Logística	4	FATORES ASSOCIADOS À MORTALIDADE INFANTIL EM MUNICÍPIO COM ÍNDICE DE DESENVOLVIMENTO HUMANO ELEVADO. (KROPIWIEC et. al. 2017).	SPSS
		DETERMINANTES DO ÓBITO INFANTIL NO VALE DO JEQUITINHONHA E NAS REGIÕES NORTE E NORDESTE DO BRASIL. (LEAL et. al. 2017)	R
		ESTUDO DE BASE POPULACIONAL SOBRE MORTALIDADE INFANTIL. (LIMA et. al. 2017)	Registry Plus Link Plus, STATA
		FATORES ASSOCIADOS À MORTALIDADE INFANTIL EM UMA CAPITAL DO NORDESTE BRASILEIRO. (SANDERS et. al. 2017)	SPSS, EXCEL
Estatística Descritiva	2	MORTALIDADE INFANTIL SEGUNDO COR OU RAÇA COM BASE NO CENSO DEMOGRÁFICO DE 2010 E NOS SISTEMAS NACIONAIS DE INFORMAÇÃO EM SAÚDE NO BRASIL. (CALDAS et. al. 2017)	SPSS, EXCEL
		MORTALIDADE INFANTIL POR COR OU RAÇA EM RONDÔNIA, AMAZÔNIA BRASILEIRA (GAVA, et. al 2017)	SPSS
Joinpoint	1	TENDÊNCIA DA MORTALIDADE INFANTIL NO MUNICÍPIO DE RIO BRANCO, AC, 1999 A 2015. (RAMALHO et. al. 2018)	Joinpoint
Índice de auto correlação espacial Moran Local	1	ANÁLISE ESPACIAL DA MORTALIDADE INFANTIL E ADEQUAÇÃO DAS INFORMAÇÕES VITAIS: UMA PROPOSTA PARA DEFINIÇÃO DE ÁREAS PRIORITÁRIAS. (RODRIGUES et. al. 2014)	GEODA
Regressão Polinomial	1	TENDÊNCIA SECULAR DA MORTALIDADE INFANTIL, COMPONENTES ETÁRIOS E EVITABILIDADE NO ESTADO DE SÃO PAULO – 1996 A 2012. (ARECO et. al. 2012)	STATA
Total	9		

FONTE: O Autor (2019).

Após a explanação dos conceitos fundamentais, a próxima seção apresenta os encaminhamentos metodológicos, que compreende a caracterização da pesquisa, definição das ferramentas de trabalho utilizadas, construção da base de dados e definição dos métodos utilizados.

3 ENCAMINHAMENTOS METOLÓGICOS

Esta seção apresenta a caracterização da pesquisa, a elaboração da base dados e a seleção das ferramentas e os métodos utilização na condução da pesquisa.

3.1 CARACTERIZAÇÃO DA PESQUISA

A presente pesquisa pode ser descrita como quantitativa, de caráter descritivo experimental (GIL, p. 52, 2008), identificado dentro da área da saúde como estudo ecológico retrospectivo, onde o foco é analisar grupos de pessoas ao invés de indivíduos (BEARGLEHOLE, p. 41-42, 2003). O estudo caracteriza-se como experimental e quantitativo, do ponto de vista que as variáveis apresentam caráter numérico discreto e descritivo por se tratar de um estudo sobretudo de estatística descritiva, o qual visa entender o cenário a partir de métodos de visualização comuns na estatística.

Sob o ponto de vista epidemiológico o estudo é ecológico, pois tem como foco uma área geográfica e estes dados são comparados temporalmente com a mesma região e calculados a partir de taxas de agregados de dados.

3.2 MATERIAIS E MÉTODOS

A seção materiais e métodos visa apresentar o caminho na elaboração da pesquisa sob ponto de vista ferramental e de implementação. A seguir, serão apresentados as ferramentas e o método para sua escolha.

3.2.1 FERRAMENTAS

A partir da escolha das ferramentas e sua utilização, é possível aferir o nível de domínio do autor em determinada área de estudo, visto que, algumas áreas possuem ferramentas que proporcionam agilidade e grau de confiabilidade dos resultados singulares. Para a escolha das ferramentas deve-se considerar:

- custos de licença da ferramenta;
- prazo permitido de uso;

- quantidade de dados processada pela ferramenta;
- visualização esperada das informações.

Foi realizado um levantamento de alguma das ferramentas mais comuns utilizadas nos estudos sobre área da saúde por profissionais de análise de dados, conforme Quadro 3.

Quadro 3 - Ferramentas mais presentes nos artigos de Mortalidade Infantil

Áreas	**Ferramentas**
Estatísticas	SPSS, R, Stata, Matlab
Epidemiológicas	TabWin, Epiinfo
Geoespaciais	MapInfo, Terraview e ArcGis, QGIS, Geoda
Linguagem de Programação	R e Python
Planilhas Eletrônicas	Microsoft Office Excel, OpenOffice Calc
Mineração de Dados	Weka (Waikato Environment for Knowledge Analysis)

FONTE: O Autor (2019).

Foram selecionadas as ferramentas deste estudo a partir do Quadro 3 com os critérios: formatos de arquivo, personalização dos resultados, tempo de processamento da análise, familiaridade das ferramentas, sendo elas:

- Sistema TabWin para coletar os dados das bases disponibilizadas pelo DATASUS;
- Software QGIS 3.2 para preparação de dados de mapa;
- Preparação da base de dados tabulares utilizando o Microsoft Office Excel;
- Preparação da base de dados foi utilizada a linguagem de programação Python versão 3.5;
- Mineração de dados com a ferramenta Weka;
- Análise espacial dos dados com o software Geoda.

Com a descrição das ferramentas a serem utilizadas, serão descritos os detalhes destas ferramentas e seu propósito neste estudo. Para coleta de dados, a ferramenta online TabWin auxiliará na tarefa de capturar dados epidemiológicos

relacionados a mortalidade infantil. Após isso, foi utilizado o software QGIS para agrupar os polígonos dos municípios com os dados de tabulação simples, esta união tem como objetivo propiciar a análise espacial. E então com o software Geoda foi aplicado ferramentas de estatística espacial. A seguir, será utilizado estatística com a linguagem de programação R, para criação de *box-plots*, testes de normalidade e correlação. Por fim, o *software* Weka será utilizado para a realização do método de mineração de dados através do algoritmo J48 e PART.

3.2.2 CONSTRUÇÃO DA BASE DE DADOS

A construção da base de dados foi realizada conforme Quadro 4, obtidas por meio de sistemas online abertos do DATASUS (Departamento de Informática do SUS), IBGE (Instituto Brasileiro de Geografia e Estatística) e do sistema de Atenção Básica E-Gestor.

Quadro 4 - Variáveis utilizadas na construção da base de dados

Variáveis	**Descrição**	**Fonte**	**Endereço Eletrônico**
Óbitos	Quantidade de óbitos por ano/munícipio	SIM (DATASUS)	http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/inf10pr.def
Nascidos Vivos	Quantidade de nascidos vivos por ano/munícipio	SINASC (DATASUS)	http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sinasc/cnv/nvpr.def
IDHM	Dimensão renda e escolaridade do índice de desenvolvimento humano municipal	Atlas Brasil	http://www.atlasbrasil.org.br/2013/pt/consulta/
Taxa de Desemprego	Proporção (%) da população Residente economicamente ativa de 16 anos e mais que se encontra sem trabalho na semana de referência, em determinado espaço geográfico, no ano considerado[1]	IBGE	http://www2.datasus.gov.br/DATASUS/index.php?area=0206&id=7401000&VObj=http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/censo/cnv/desempr

[1] http://tabnet.datasus.gov.br/cgi/ibge/censo/Taxa_Desemprego.pdf

(continua)

(continuação)

Variáveis	Descrição	Fonte	Endereço
Cobertura de Atenção Básica	Percentual da população coberta por equipes da Estratégia da Saúde da Família e por equipes de Atenção Básica tradicional e padronizadas em relação a estimativa populacional[2]	EGESTOR	https://egestorab.saude.gov.br/paginas/acessoPublico/relatorios/relHistoricoCoberturaAB.xhtml
Municípios do Paraná	Polígonos dos municípios do Brasil	IPEA	http://www.ipea.gov.br/ipeageo/malhas.html
Demográficas	População por município	IBGE	http://www2.datasus.gov.br/DATASUS/index.php?area=0206&id=7401000&VObj
PIB por município	Produto Interno Bruno do município	IBGE	
Taxa de Saneamento Básico Adequado	Leva em conta escoto, abastecimento de agua e coleta de lixo conforme nota técnica https://biblioteca.ibge.gov.br/visualizacao/livros/liv54598.pdf	IBGE	ftp://ftp.ibge.gov.br/Censos/Censo_Demografico_2010/indicadores_sociais_municipais/Unidades_da_Federacao/parana.zip
Taxa de Saneamento Básico semiadequado			
Taxa de Saneamento Básico inadequado			

FONTE: O Autor (2019).

Destaca-se a escolha da variável do Índice de Desenvolvimento Humano Municipal, este índice é capaz de identificar o desenvolvimento humano a partir de três dimensões, longevidade, educação e renda. No presente estudo será utilizado somente duas das três dimensões, pois a dimensão longevidade leva em conta a expectativa de vida, sendo este índice influenciado pela taxa de mortalidade infantil, portanto, para evitar relações errôneas.

Em geral, a união de diversas fontes de dados é realizada por meio de uma variável comum entre todas estas bases como o caso do *linkage* (ALMEIDA et al. 1996). Neste caso foi utilizado o código do município fornecido pelo IBGE, conforme explica a nota técnica do IBGE:

---

[2] https://egestorab.saude.gov.br/paginas/acessoPublico/relatorios/nota_tecnica/nota_tecnica_relatorio_de_cobertura_AB.pdf

> Os códigos dos municípios no IBGE são construídos de maneira que os dois primeiros dígitos representam a Unidade da Federação a que pertencem, os quatro dígitos seguintes são a própria identificação do município em ordem alfabética e o último é um dígito verificador formado a partir dos anteriores (IBGE, 2010, p. 17).

O código do munícipio aparecia sem o digito verificador no mapa, portando este foi retirado em todas as outras bases para possibilitar a junção, esta tarefa foi realizada com apoio do software QGIS 3.2.6. O próximo passo é agregar os dados oriundos das bases de dados SIM e SINASC, disponibilizadas pelo DATASUS também em formato *.csv* no TABNET[3], que é o tabulador de domínio público disponibilizado pelo governo brasileiro.

Em seguida, foi realizado o calculo o indicador da taxa de mortalidade infantil, este indicador será calculado pela função do QGIS a seguir:

```
round(
        (coalesce("obito_2014",0) + coalesce("obito_2015",0) + coalesce("obito_2016" ,0))
                                                                /
        (coalesce("nasc_2014",0) + coalesce("nasc_2015",0) + coalesce("nasc_2016",0))
* 1000 ,3) (3)
```

Com a construção deste indicador, a descoberta de padrões e inferências entre os municípios se torna mais acertada, pois o uso do indicador reduz o impacto dos óbitos em municípios de pequeno porte.

Após a construção da base, será realizado a seguir, a esquematização da aplicação dos métodos estatísticos, de mineração de dados e estatística.

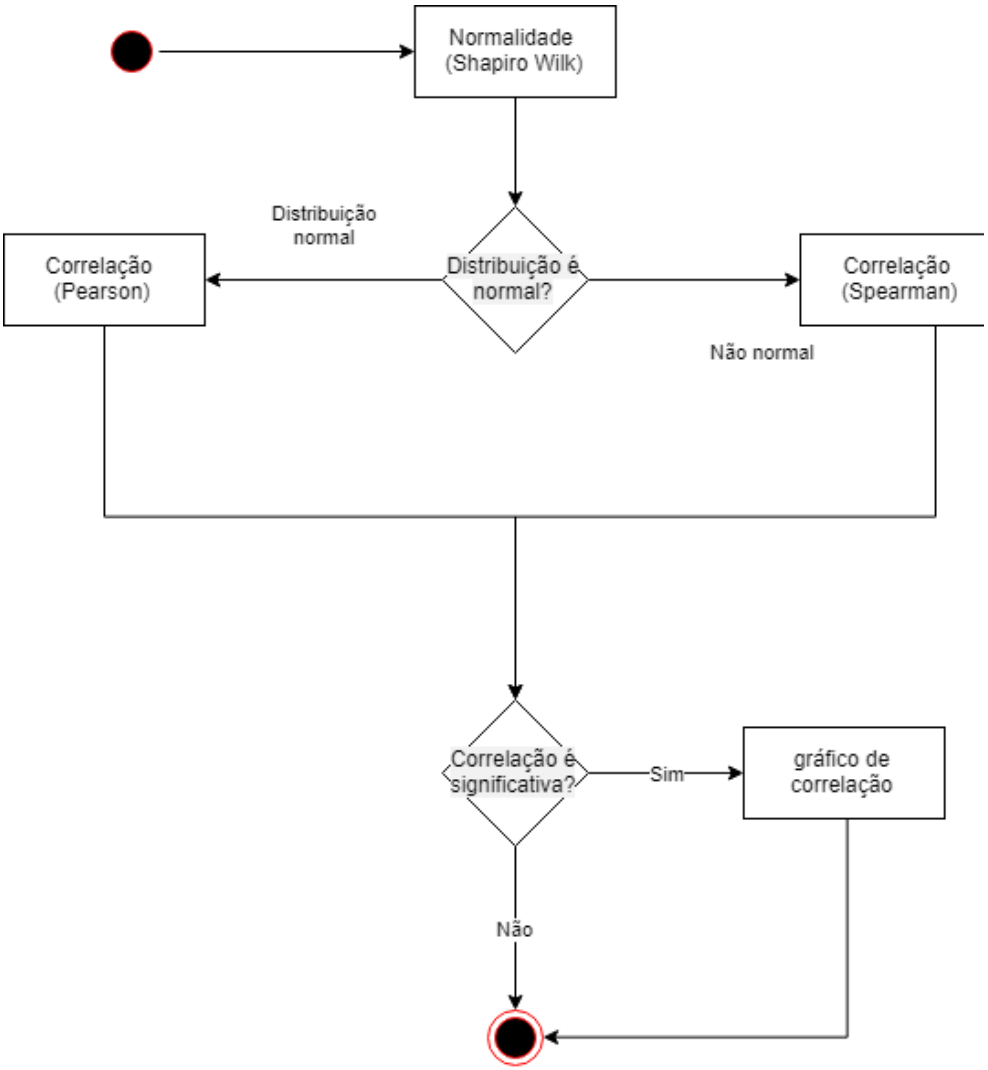
3.2.3 MÉTODOS ESTATÍSTICOS

Os métodos estatísticos atuam como componente chave de grandes partes dos estudos acadêmicos, sobretudo ferramentas que utilizam a estatística descritiva. Portanto, será realizado uma abordagem a partir da estatística, a fim de verificar e tentar encontrar padrões de acordo com fluxo da Figura 3.

---

[3] http://www2.datasus.gov.br/DATASUS/APRESENTACAO/TABNET/Tutorial_tabNet_FINAL.pptx_html/html/index.html#8

Figura 3 - Fluxo dos testes estatísticos



FONTE: O Autor (2019).

O primeiro passo foi a aplicação do teste de normalidade Shapiro Wilk (ROYSTON, 1992), o qual foi adotado o nível de confiança de 95%. O teste tem como as seguintes hipóteses:

$$\boldsymbol{H_0} = \textit{Dados Normais}$$
$$\boldsymbol{H_1} = \textit{Dados não normais}$$

Sendo a hipótese nula $\boldsymbol{H_0}$ admite que a distribuição é normal, e $\boldsymbol{H_1}$ ($p < 0,05$) rejeita-se a hipótese nula, logo a distribuição não é normal. A partir da normalidade de determinado conjunto de dados é então efetuado o teste de correlação. Para dados

que se aproximam da curva normal utiliza-se o teste de correlação Spearman (SPEARMAN, 1904), caso contrário, utiliza-se o teste de correlação Pearson (PEARSON, 1896).

Figueiredo Filho (2009, p. 119) explica a variação do resultado do teste de correlação de Pearson que varia de -1 a 1. Relata que o sinal indica direção positiva ou negativa do relacionamento e o valor sugere a força da relação entre as variáveis, ou seja, em qual grau uma variável *varia* em função da outra. Uma correlação com valor de (-1 ou 1) indica que o valor de uma variável pode ser determinado exatamente ao se saber o valor da outra. E um valor zero (0) ou próximo a zero indica que não a relação linear entre as variáveis. Esta relação será verificada a partir do seguinte teste de hipótese:

$$\boldsymbol{H_0} = \textit{Não existe correlação linear para esta população}$$
$$\boldsymbol{H_1} = \textit{Existe correlação para esta população}$$

Este teste será avaliado de acordo com Quadro 5 descrito por Cohen, no qual $\boldsymbol{H_0}$ representará valores pequenos e $\boldsymbol{H_1}$ se dará a partir de valores médios ou grandes.

Quadro 5 - Classificação dos valores encontrados na correlação de Spearman

Valor	Intervalo
Pequenos	0,10 e 0,29
Médios	0,30 e 0,49
Grandes	0,50 e 1

FONTE: Adaptado de Cohen (1988).

Este quadro apresenta uma das propostas de interpretação do valor de correlação de Spearman, dentre as diversas existentes.

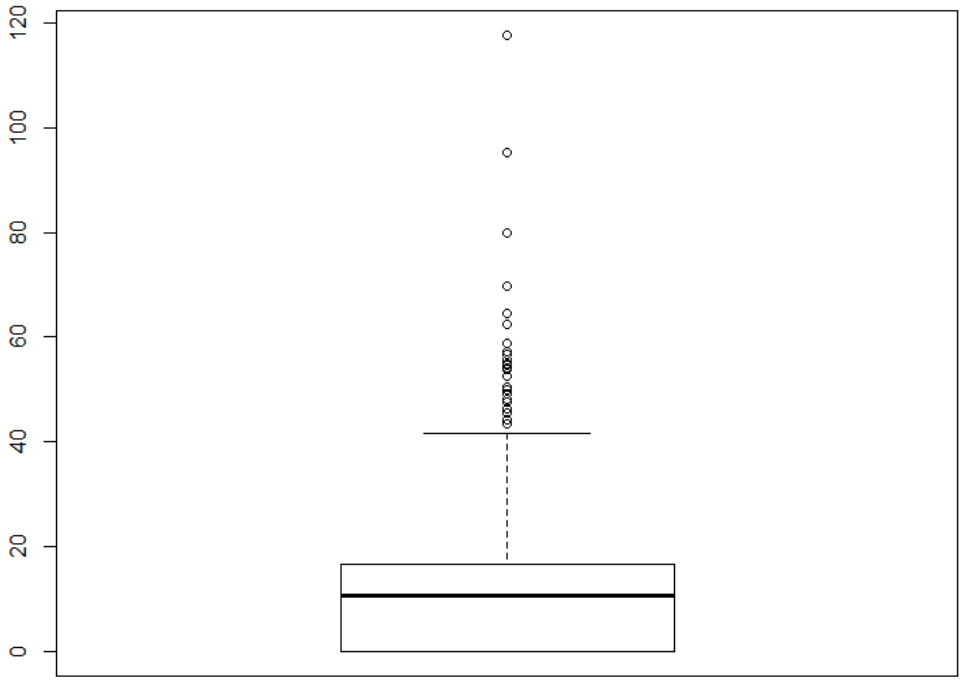
Concluída a preparação da área estatística, inicia-se então a etapa de mineração de dados. A qual será realizada de acordo com a seção de análise de dados.

3.2.4 MINERAÇÃO DE DADOS

.

A mineração de dados inicia-se pela etapa de pré-processamento, esta foi realizada com objetivo de adequar a base de dados aos softwares utilizados, levando em conta os requisitos do algoritmo em questão. O algoritmo a ser utilizado será o J48 conforme descrito na seção de revisão de literatura, disponível no software Weka. Um dos requisitos para aplicação do algoritmo é que seu atributo meta (resposta) necessita ser do tipo discreto. Todavia, o campo taxa de mortalidade infantil foi transformado de numérico para variável discreta com a utilização dos quartis com o software R-Studio a partir do comando *boxplot(variavel)$stats*[4]. Este comando apresenta o boxplot conforme Figura 4 e quartis.

Figura 4- *Box-plot* da taxa de mortalidade infantil no Paraná entre os anos de 2014 a 2016



FONTE: O Autor (2019).

Verifica-se que 75% dos municípios possuem uma taxa de mortalidade infantil abaixo de 16,66‰ e que há *outliers* com taxa de mortalidade infantil elevadíssimas. Foi definido a aplicação de dois testes distintos, o primeiro utilizando a taxa de mortalidade infantil por ano e o segundo por triênio, ambos utilizando o intervalo dos

---

[4] https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/boxplot.html

anos de 2014 a 2015. Ambos os testes terão utilizarão os agrupamentos do Quadro 6, estes dados oriundos do comando do *box-plot*.

Quadro 6 - Regra de Agrupamento da variável Taxa de Mortalidade Infantil

Grupo	Intervalo	Quantidade de Municípios
Grupo 1	0 - 10,5945	176
Grupo 2	10,5945 - 16,6670	146
Grupo 3	16,6670 - 41,6670	75
Grupo 4	41,6670 - 117,6470	2

FONTE: O Autor (2019).

Concluída a etapa de preparação da base de dados foi dado, início a etapa de aplicação do algoritmo J48, o qual apresenta como resultado um texto de informações relativas a classificação além da árvore de decisão. Os parâmetros do algoritmo foram deixados como padrão (validação cruzada), como atributo meta foi utilizado campo taxa de mortalidade infantil por grupos.

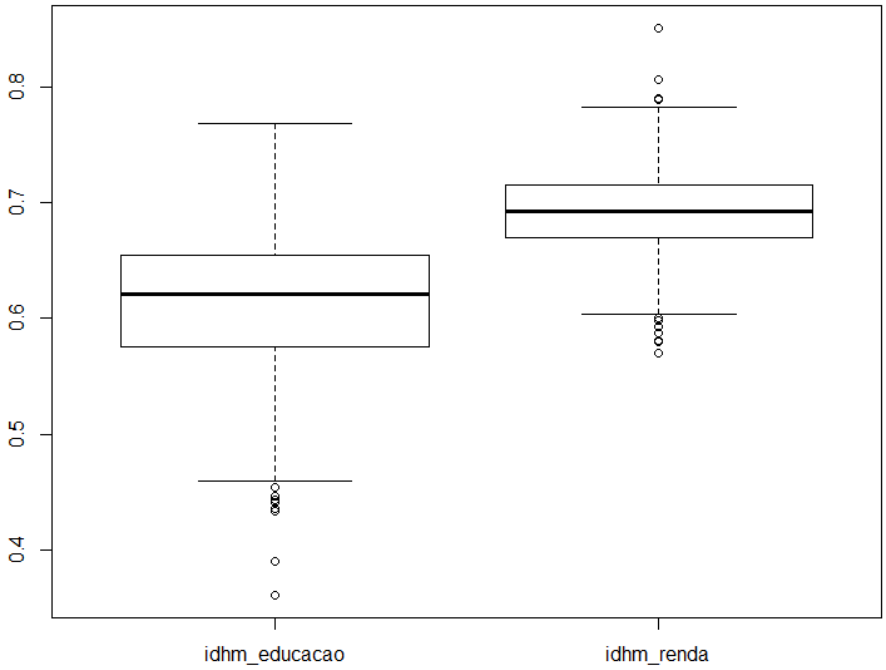
4 APRESENTAÇÃO DOS RESULTADOS

Nesta seção serão apresentados os resultados, estes serão divididos em três componentes estatísticos gerais, de estatística espacial e mineração de dados.

4.1 ESTATÍSTICA

Para os resultados estatísticos inicialmente foi apresentado as variáveis utilizando *box-plots*. Possibilitam sumarização dos dados e verificação de conglomerados significativos de dados e valores fora da curva, os chamados outliers, conforme Figura 5.

Figura 5 - *Box-plot* das variáveis de IDHM dimensões renda e educação



FONTE: O Autor (2019).

Observa-se que a maior quantidade de municípios da dimensão renda do indicador de IDHM renda está acima da dimensão educação. A figura ainda apresenta *outliers* especialmente abaixo do primeiro quartil. A variável IDHM na dimensão educação apresentam 2 (dois) munícipios (Doutor Ulysses e de Cerro Azul) com seu índice abaixo de 0,4, sendo os munícipios de. Apesar de seu índice pequeno o município de Doutor Ulysses apresenta a taxa de mortalidade infantil do triênio

próxima a média nacional (10‰ contra 12,70‰), já o município que Cerro Azul possui uma elevada taxa de mortalidade infantil 34‰. Esta diferença de 24 pontos entre os municípios, pode revelar que este indicador não é um bom candidato a ser utilizado para descobrir padrões sobre mortalidade.

A seguir, a tabela 2 apresenta os descritos do quartil da variável cobertura de atenção básica a partir de sua agregação por ano e pelo triênio

Tabela 2 - Descritivo da variável cobertura de atenção básica e suas subdivisões

	Valor mínimo	**1º quartil**	**Mediana**	**Média**	**3º quartil**	**Valor máximo**
Cobertura Atenção Básica do triênio	29,25	88,82	99,06	92,18	100	100
Cobertura Atenção Básica 2014	25,70	87,56	99,68	91,25	100	100
Cobertura Atenção Básica 2015	20,02	89,12	100	92,61	100	100
Cobertura Atenção Básica 2016	32,66	88,65	100	92,69	100	100
Variância	8,42	0,46	0,20	0,44	100	100

FONTE: O Autor (2019).

Os registros de atenção básica conforme apresentam significativos outliers abaixo do primeiro quartil, não há grandes variações entre na média, mediana e primeiro quartil das variáveis apresentadas, com exceção do valor mínimo que apresenta a variância de 8,42 pontos.

A variável população apresentada na Tabela 3, demonstra significativa variação entre a população dos municípios do estado do Paraná, o que pode indicar perfis sociodemográficos distintos.

Tabela 3 - Descritivo da variável população

Valor mínimo	1º quartil	Mediana	Média	3º quartil	Valor máximo
1409	5037	9026	26177	17274	1751907

FONTE: O Autor (2019).

Já a Tabela 4 apresenta o descritivo do PIB (Produto Interno Bruto) dos munícipios. Demonstra da mesma forma que a variável população apresenta grande diferença entre o perfil produtivo dos munícipios do estado.

Tabela 4 - Descritivo do PIB

Valor mínimo	1º quartil	Mediana	Média	3º quartil	Valor máximo
19955	60335	120887	544586	262626	53106497

FONTE: O Autor (2019).

O PIB apresenta grande variabilidade dentre os munícipios do estado, tanto que há *outliers* significativos acima no terceiro quartil. Por fim, apresenta-se a Tabela 5 com as variáveis de adequação saneamento básico.

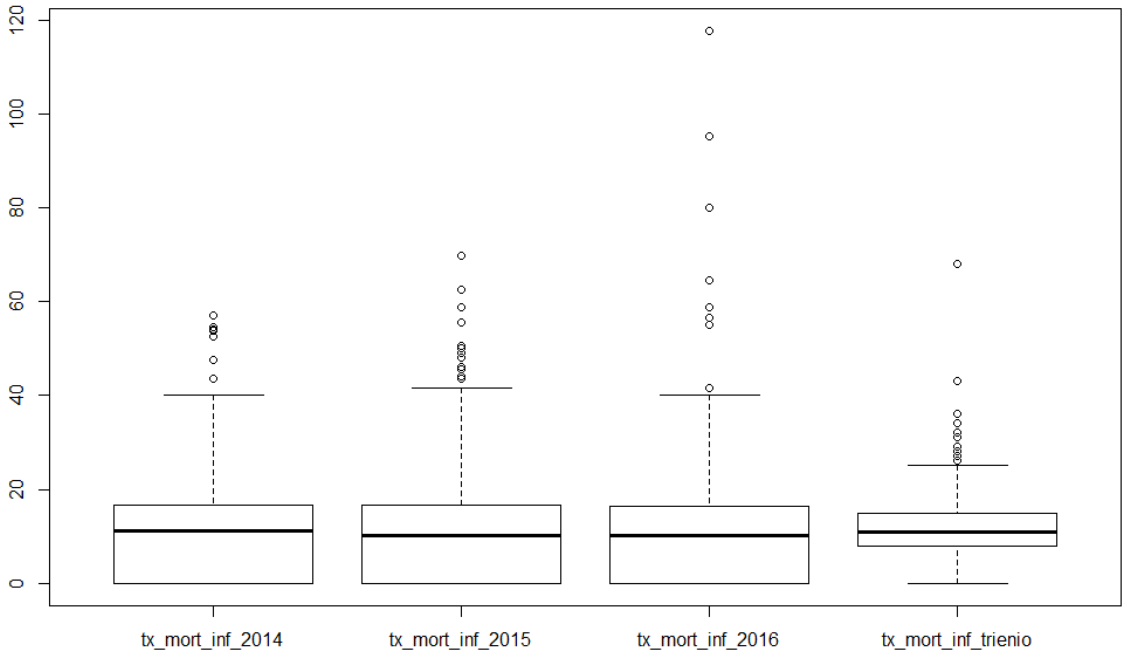
Tabela 5 - Descritivo das variáveis de adequação do saneamento básico

	Valor mínimo	1º quartil	Mediana	Média	3º quartil	Valor máximo
Taxa de saneamento adequado	0,240	4,975	22,525	29,302	46,767	96,020
Taxa de saneamento semiadequado	2,97	37,53	54,89	54,34	74,06	96,43
Taxa de saneamento inadequado	0,01	7,01	13,14	16,43	23,61	65,67

FONTE: O Autor (2019).

A seguir, é apresentado o *box-plot* (Figura 6) com os dados da mortalidade infantil. É possível verificar que o indicador do triênio suaviza os *outilers* extremos como acontece no ano de 2016.

Figura 6 - *Boxplot* mortalidade infantil 2014 a 2016 e triênio



FONTE: O Autor (2019).

Grande parte dos munícipios do estado do Paraná possuem sua taxa abaixo de 20 pontos, um valor próximo da média nacional. Porém, há algumas exceções, como o caso do munícipio Rancho Alegre D'Oeste que apresenta a taxa de mortalidade de 117‰ no ano de 2014.

A seguir, será utilizado o teste estatístico de Shapiro (Quadro 7) visando verificar se os dados se aproximam de uma distribuição normal, conforme explicado na seção de metodologia.

Quadro 7 - Teste estatístico de normalidade Shapiro-Wilk (p-valor com 5 casas decimais)

Variável	**p-valor**	**Resultado do Teste**
IDHM renda	0,02813	Aceita-se a hipótese nula
DHM educação	0	
Atenção básica	0	
Taxa de mortalidade infantil triênio	0	
População	0	
PIB	0	
Saneamento adequado	0	
Saneamento semiadequado	0	
Saneamento inadequado	0	

FONTE: O Autor (2019).

A partir do resultado do teste de normalidade, foi utilizado o teste de correlação Spearman, pois foi aceita a hipótese nula no teste de normalidade. A correlação foi realizada a partir da variável dependente taxa de mortalidade infantil, a partir de outras variáveis disponíveis, conforme quadro 8.

Quadro 8 - Correlação Spearman das variáveis com taxa de mortalidade infantil triênio Paraná

Variável	Correlação	p-valor	Resultado
IDHM Renda	-0,06348814	0,2057	Não existe correlação linear para esta população (Pequeno)
IDHM Educação	-0,03153197	0,5300	
População	0,06265366	0,2117	
PIB	0,06573179	0,1901	
Saneamento Adequado	0,06174722	0,2190	
Saneamento Semiadequado	0,05729965	0,2535	
Saneamento inadequado	0,03101261	0,5368	

FONTE: O Autor (2019).

A partir da correlação, é constado que apesar de existir certo grau de correlação entre as variáveis e a taxa de mortalidade infantil, esta correlação é

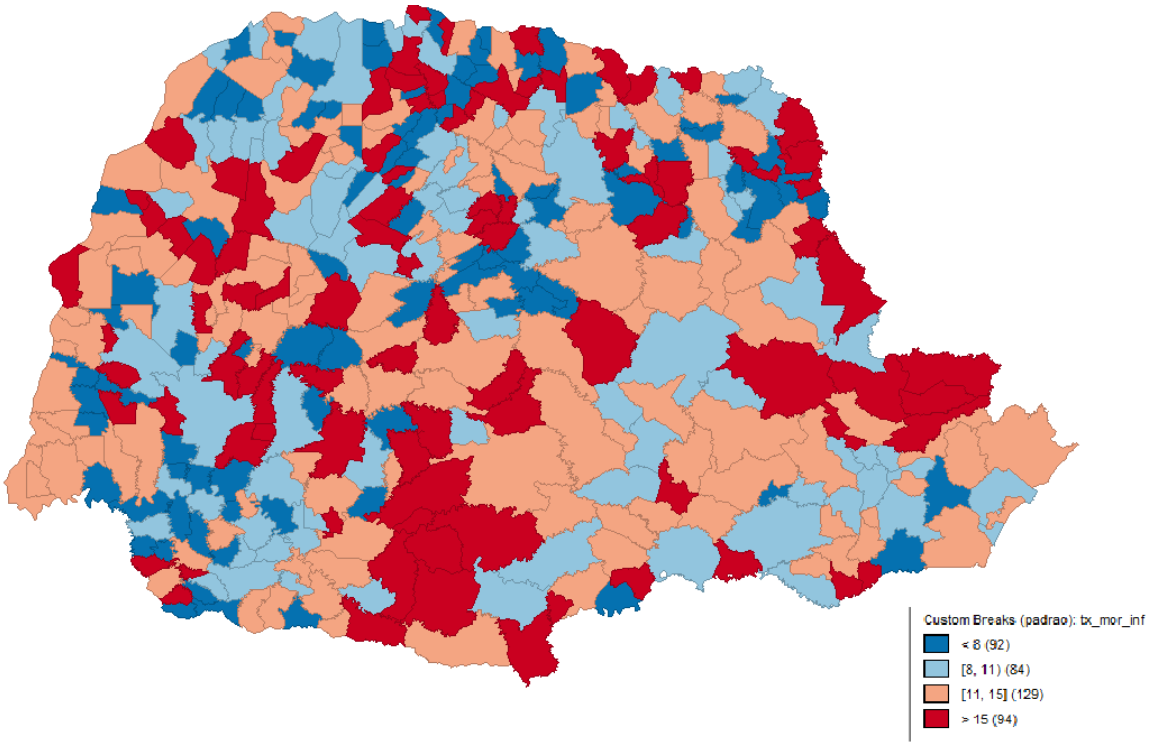
pequena, portanto, adota-se a hipótese H_0, que tem como proposição a não existência existe correlação linear para variáveis testadas.

A seguir, foi utilizado a estatística espacial para verificar a relação espacial na taxa de mortalidade infantil.

4.2 ESTATÍSTICA ESPACIAL

Com a ferramenta Geoda foi gerado um box-map, nele visualiza-se nas cores mais escuras a taxa de mortalidade infantil mais elevada, (Figura 7), encontram-se principalmente na região centro e sudeste do estado do Paraná.

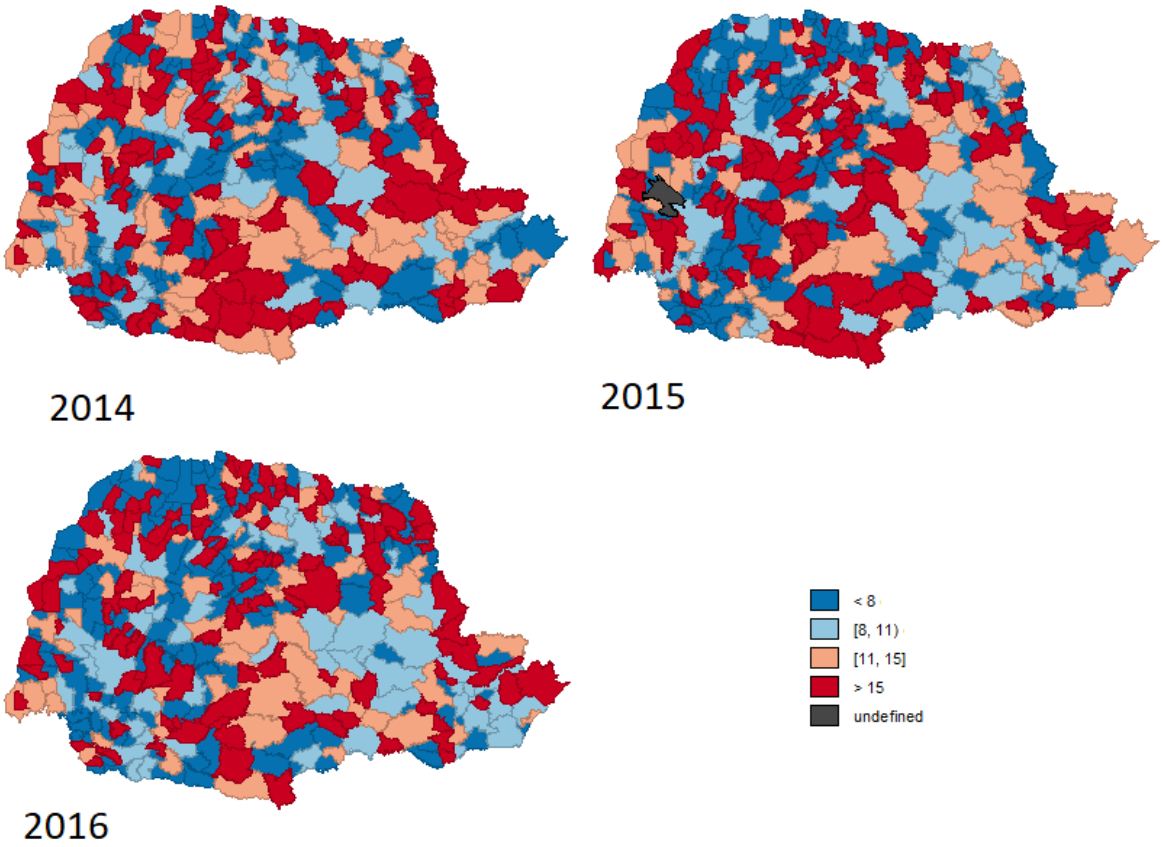
Figura 7 - *Box-map* da taxa de mortalidade infantil do Triênio



FONTE: O Autor (2019).

Verifica-se o agrupamento de municípios próximos em conglomerados dentro do mesmo grupo classificado pelo mapa, em espacial ao grupo que apresenta a taxa de mortalidade entre 11 e 15. Com a visão por ano, é possível descrever e visualizar a dimensão temporal nas relações especiais de saúde do estado conforme Figura 8.

Figura 8 - *Box-map* da taxa de mortalidade infantil 2014-2016



FONTE: O Autor (2019).

Verifica-se que o perfil espacial ao longo do triênio é modificado, a região sul por exemplo, apresenta diminuição da taxa de mortalidade infantil, da mesma forma que a região sudeste.

4.3 MINERAÇÃO DE DADOS

Dentre as três aplicações definidas na metodologia, inicia-se a primeira com a execução algoritmo J48 utilizando as seguintes variáveis: taxa de mortalidade infantil do triênio (2014-2016), IDHM dimensão renda, IDHM dimensão escolaridade, população (censo 2010), cobertura de atenção básica, produto interno bruto municipal, taxa de adequação de saneamento, taxa de saneamento semi-adequado, taxa de saneamento inadequado. Foi aplicado na base de forma integral (399 registros) e teve como tempo processamento do algoritmo de menos que um segundo, e resultou na seguinte matriz de confusão:

Quadro 9 - Matriz de confusão atributo meta mortalidade infantil triênio (2014-2016)

Classificado como				
a	b	c	d	Valor de referência
60	78	8	0	a = GRUPO 2
53	113	9	1	b = GRUPO 1
23	44	8	0	c = GRUPO 3
0	2	0	0	d = GRUPO 4

FONTE: O Autor (2019).

O quadro apresenta os grupos e a respectiva classificação realizada pelo algoritmo, todos os municípios do grupo 4 (de 41,6670 a 117,6470) foram classificados incorretamente. Dos municípios do grupo 3 (16,6670 - 41,6670) apenas seis (6) dos municípios foram classificados corretamente, já o grupo 2 apenas 60 dos 146 municípios foram classificados corretamente. Por fim, o grupo 1 (0 - 10,5945) apresentou a maioria dos municípios (113 dos 176) classificados corretamente. O quadro 10 condessa o resultado da classificação em instâncias classificadas corretamente e as classificadas incorretamente.

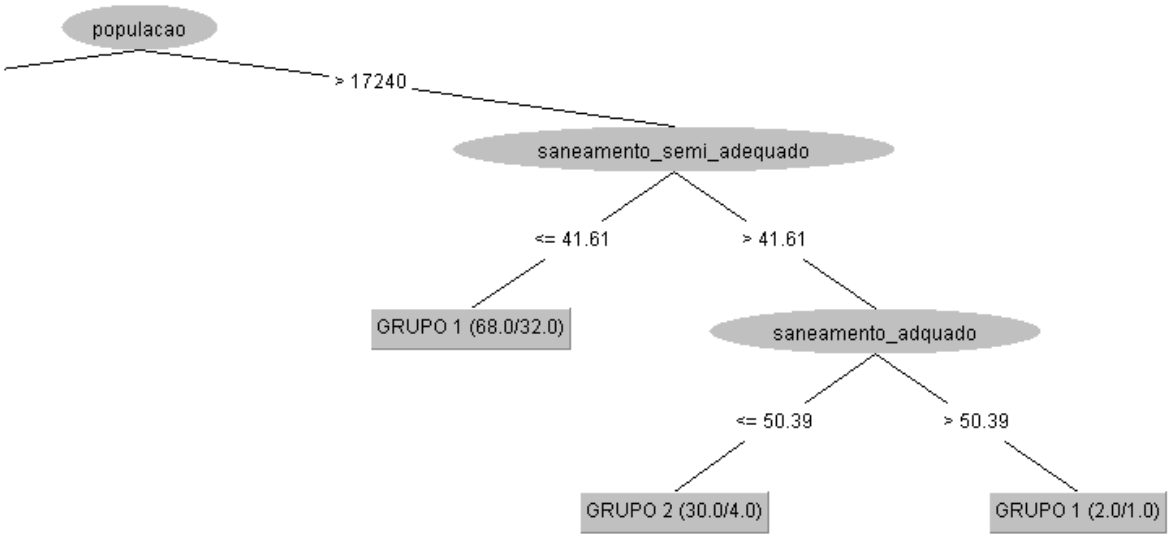
Quadro 10 - Resultados da classificação do algoritmo J48 com atributo meta taxa de mortalidade infantil no Paraná (triênio)

Instâncias	Registros	Porcentagem (%)
Classificadas corretamente	167	45,36
Classificadas incorretamente	232	54,64

FONTE: O Autor (2019).

Mais da metade (54,64) dos municípios foram classificados incorretamente. Dentre as instâncias classificadas corretamente foram encontradas as regras, das quais destacam-se a apresenta na Figura 9, as demais regras estão descritas no apêndice.

Figura 9 - Recorte da árvore de decisão geradas partir do algoritmo J48 com atributo meta taxa de mortalidade infantil triênio



FONTE: O Autor (2019).

A adequação do saneamento revela-se como ponto de importância na criação de regras para municípios com mais de 17.240 habitantes. Pois, conforme diminui a taxa de saneamento adequado do município e aumenta a taxa de saneamento semiadequado, pode acarretar o aumento da taxa de mortalidade infantil.

A seguir, foi realizado o teste agrupando a taxa de mortalidade infantil por ano, neste caso procura-se observar alterações sazonais em decorrência de fatores como migração ou epidemias, por exemplo. Foram utilizadas as mesmas variáveis do primeiro teste. A aplicação do algoritmo demorou menos de 1 (um) segundo em cada um dos três testes e teve como resultado a classificação do Quadro 11.

Quadro 11 - Resultados da classificação por triênio do algoritmo J48

Instâncias	2014 (%)	2015 (%)	2016 (%)
Instâncias classificadas corretamente	45,61	48,62	44,36
Instâncias classificadas incorretamente	54,39	51,38	55,64

FONTE: O Autor (2019).

Observa-se baixa oscilação das instâncias classificadas corretamente ao longo dos anos, porém, nota-se um aumento discreto na taxa de acerto no ano de 2015. Detalhando os resultados apresentados no quadro de classificação dos acertos, a matriz de confusão de todos nos anos 2014 a 2016 verifica-se respectivamente os quadros 12, 13 e 14.

Quadro 12 - Matriz de confusão usando atributo meta taxa de mortalidade infantil 2014

Classificado como				2014
a	b	c	d	Valor de referência
18	9	64	2	a = GRUPO 3
16	37	61	0	b = GRUPO 2
30	25	127	2	c = GRUPO 1
2	1	5	0	d = GRUPO 4

FONTE: O Autor (2019).

As instâncias classificadas com a taxa de mortalidade infantil de 2014 seguem o padrão de classificação do triênio, ou seja, o grupo 4 apresentando todas as classificações incorretas. Para o grupo 3 foram somente 18 dos 75 municípios foram classificados, já o grupo 2 apresentou 37 dos seus munícipios classificados corretamente, valor bem abaixo do classificado corretamente no triênio (60).

O quadro 10 apresenta a matriz de confusão a partir da taxa de mortalidade infantil de 2015, nota-se que o grupo 2 apresenta a maioria dos seus registros classificados corretamente, novamente nenhum registro é classificado corretamente para o grupo 4.

Quadro 13 - Matriz de confusão usando atributo meta taxa de mortalidade infantil 2015

Classificado como				2015
a	b	c	d	Valor de referência
3	86	1	0	a = GRUPO 3
5	184	16	2	b = GRUPO 2
1	81	7	0	c = GRUPO 1
0	12	1	0	d = GRUPO 4

FONTE: O Autor (2019).

Além disso, do ano de 2014 para este o erro de classificação acentua consideravelmente para o grupo 1, passando de 57 para 82. Já para o ano de 2016, constata-se redução de erros do grupo 1 e aumento de erros no grupo 2 em relação ao ano de 2015.

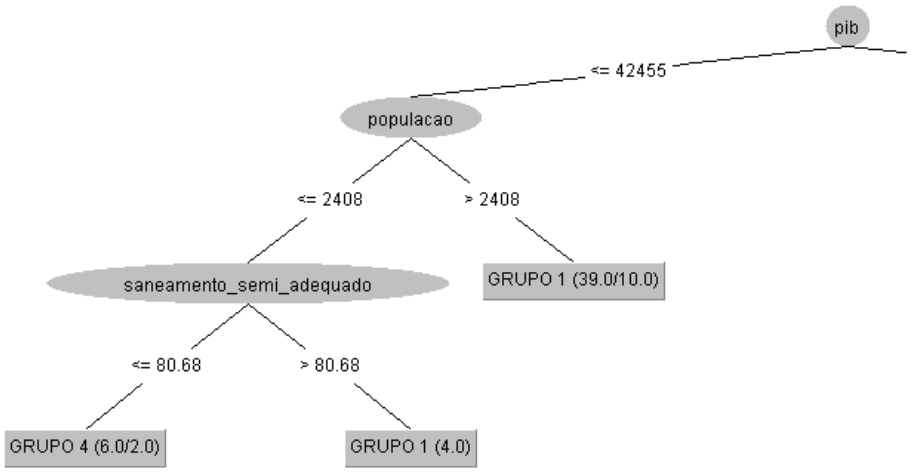
Quadro 14 - Matriz de confusão usando atributo meta taxa de mortalidade infantil 2016

Classificado como				2016
a	b	c	d	Valor de referência
21	59	14	0	a = GRUPO 3
34	139	32	3	b = GRUPO 2
9	63	17	0	c = GRUPO 1
0	7	1	0	d = GRUPO 4

FONTE: O Autor (2019).

Após a análise da matriz de confusão, verifica-se a árvore de decisão gerada para a taxa de mortalidade infantil, referente ao ano de 2014 gerou-se uma árvore, da qual destaca-se algumas das suas folhas (Figura 10).

Figura 10 - Recorte da árvore de decisão gerada a partir do algoritmo J48 com atributo meta taxa de mortalidade infantil 2014



FONTE: O Autor (2019).

Já a execução do algoritmo com a taxa de mortalidade do ano de 2015, resultou em uma árvore com somente um nó e nenhuma regra. Logo, foi utilizada uma aplicação alternativa do algoritmo C4.5 denominado PART, disponível também no software Weka. Este algoritmo tem como resultado regra em formato textual, não disponibilizando uma árvore de decisão. Sua execução durou menos de um segundo e classificou corretamente 187 das 399 instâncias. Dentre as regras apresentadas estão:

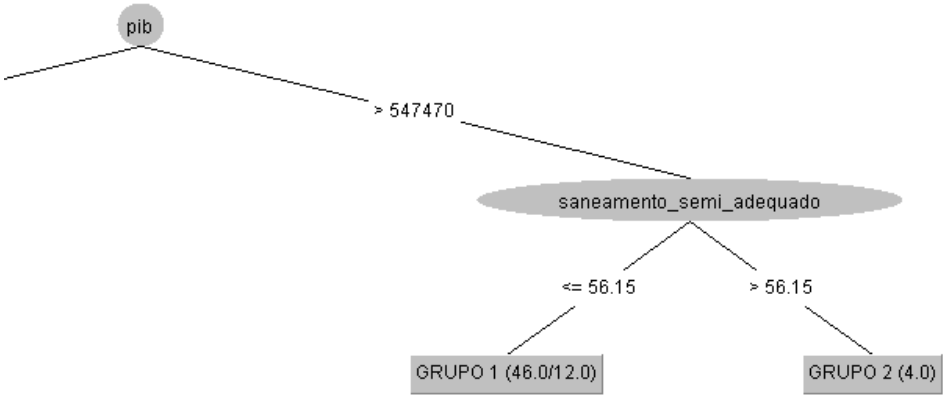
saneamento_adquado <= 6,15 E populacao > 4568 = GRUPO 3
saneamento_adquado <= 1,22 E idhm_educacao <= 0,649 = GRUPO 4

Estas regras são em certo grau complementares, a primeira regra descreve que a partir de uma adequação de saneamento baixa e uma população acima de 4568 pessoas a taxa de mortalidade infantil tende a ser alta (16,6670 - 41,6670). A segunda regra especifica que em casos que a taxa de saneamento adequado seja muito baixa (<= 1,22) e o IDHM dimensão educação seja abaixo de 0,649 o município tende a pertencer ao grupo 4 (41,6670 - 117,6470), ou seja, apresenta ter uma taxa de mortalidade infantil extremamente alta. Todavia, vale lembrar que nenhum dos testes

conseguiu acertar resultados para o grupo 4 conforme mostra as respectivas matrizes de confusões.

Por fim, o algoritmo aplicado para a taxa de mortalidade infantil no ano de 2016, gerou uma árvore de tamanho 103, com 52 duas folhas, na Figura 11 é possível verificar um recorde desta árvore.

Figura 11 - Recorte da árvore de decisão gerada a partir do algoritmo J48 com atributo meta taxa de mortalidade infantil 2016



FONTE: O Autor (2019).

Conforme já apresentado nos outros anos, o indicador saneamento semi-adequado apresentou potencial para a construção de regras, em conjunto com o PIB do município. Entanto, devido a taxa de acerto desse experimento ser baixa, não é possível generalizar este resultado.

5 CONSIDERAÇÕES FINAIS

A taxa de mortalidade infantil está reduzindo ao longo dos anos no Brasil, o Paraná apresenta uma taxa de 11,93‰ entre 2014 a 2016. Isso se deve a melhoria no cuidado materno infantil, através da promoção de políticas públicas como o programa Rede Mãe Paranaense.

O presente estudo alcançou o objetivo específico da análise estatística da base de dados, a partir da visualização descritiva dos dados, a exibição dos quartis das variáveis, além disso a aplicação do teste correlação de Spearman. Com a aplicação da correlação verificou-se que as variáveis do triênio não são capazes de explicar a taxa de mortalidade infantil no Paraná, tampouco identificar padrões relacionada a taxa.

Como atendimento ao objetivo da utilização de métodos de mineração de dados para a descoberta de padrões, verifica-se uma grande quantidade de padrões encontrados a partir das árvores de decisões. Dentre as quais, obteve com os algoritmos J48 e PART uma taxa de acerto entre os 399 municípios do Paraná de 45,36% para a classificação a partir do triênio, 45,61% no ano de 2014, 48,62% para o ano de 2015 e 44,36% para o ano de 2016. Das árvores de decisões geradas observa-se como padrões relevantes, dentre os quais a relação entre a diminuição da taxa de saneamento adequado nos municípios junto com o aumento a taxa de saneamento semiadequado, com o aumento da taxa de mortalidade infantil para o triênio (2014-2016). Já para o ano de 2015 revela-se um padrão com potencial a relação entre taxa de saneamento adequado menor que 6,15 em populações com mais de 4.568 uma tendência de se encontrar com uma taxa de mortalidade entre 16,6670‰ a 41,6670‰.

Alinhado ao objetivo da descrição da mortalidade infantil a partir de sua dimensão geográfica, destaca-se a verificação de padrões de taxas similares entre municípios próximos e a verificações de mudanças ao longo do triênio (2014-2016) em especial na redução da taxa de mortalidade infantil na região sul.

5.1 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Recomenda-se para trabalhos futuros utilizar indicadores da taxa de mortalidade infantil sob a perspectiva de óbitos evitáveis, a fim de aferir relações que realmente possam impactar em decisões sob o âmbito de políticas públicas. E ainda a utilização de indicadores de correlação espacial, inclusive para verificar se a relação entre espacialidade e taxa de mortalidade infantil destacada na seção de estatística espacial, é válida e generalizável ao território brasileiro, por exemplo.

REFERÊNCIAS

ALMEIDA, Marcia Furquim de et al. O uso da técnica de" linkage" de sistemas de informação em estudos de coorte sobre mortalidade neonatal. **Revista de Saúde Pública**, v. 30, p. 141-147, 1996. Disponível em: < http://www.scielo.br/pdf/rsp/v30n2/5055.pdf >. Acesso em: 26 jun. 2019.

ANDRADE, C. L. T. de; SZWARCWALD, C. L. Desigualdades sócio-espaciais da adequação das informações de nascimentos e óbitos do Ministério da Saúde, Brasil, 2000-2002. **Caderno Saúde Pública** 2007, v. 23, n.5. Disponível em: <http://dx.doi.org/10.1590/S0102-311X2007000500022>. Acesso em: 19 mar. 2019.

ARECO, K. C. N.; KONSTANTYNER, T.; TADDEI, J. A. C.. Tendência secular da mortalidade infantil, componentes etários e evitabilidade no Estado de São Paulo–1996 a 2012. **Revista Paulista de Pediatria**, v. 34, n. 3, p. 263-270, 2016.

BEAGLEHOLE, R., BONITA, R., KJELLSTROM, T. Epidemiologia básica atual. São Paulo: Santos, 2003. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/4338974/mod_resource/content/1/BONITA%20et%20al%20-%20cap%203.pdf> Acesso em: 22 abr. 2019.

BEZERRA FILHO, J. G. et al. Distribuição espacial da taxa de mortalidade infantil e principais determinantes no Ceará, Brasil, no período 2000-2002. **Cadernos de Saúde Pública**, v. 23, p. 1173-1185, 2007.

BRAZIL. MINISTÉRIO DA SAÚDE. SECRETARIA EXECUTIVA. **SUS--princípios e conquistas**. Ministério da Saúde, 2001. Disponível em <http://bvsms.saude.gov.br/bvs/publicacoes/sus_principios.pdf>. Acesso em: 19 mar. 2019.

BRASIL. Ministério da Saúde. **Indicadores de mortalidade: C.1 Taxa de mortalidade infantil.** 2005. Disponível em: <http://tabnet.datasus.gov.br/cgi/idb2005/c01.htm>. Acesso em: 25 jun. 2019.

BRASIL. Ministério da Saúde. A declaração de óbito: documento necessário e importante / Ministério da Saúde, Conselho Federal de Medicina, Centro Brasileiro de Classificação de Doenças. – 3. ed. – Brasília: Ministério da Saúde, 2009.

CALDAS, A. D. R. et al. Mortalidade infantil segundo cor ou raça com base no Censo Demográfico de 2010 e nos sistemas nacionais de informação em saúde no Brasil. **Cadernos de Saúde Pública**, v. 33, p. e00046516, 2017.

CHENG, Jie et al. Improved decision trees: a generalized version of id3. In: **Machine Learning Proceedings 1988**. Morgan Kaufmann, 1988. p. 100-106.

COHEN, J. Statistical power analysis for the behaviors science.(2nd). **New Jersey: Laurence Erlbaum Associates, Publishers, Hillsdale**, 1988.

DUTRA, Rogerio Garcia; MARTUCCI, Moacyr. Adaptive Fuzzy Neural Tree Network. **IEEE Latin America Transactions**, v. 6, n. 5, p. 453-460, 2008.

FIGUEIREDO FILHO, Dalson Britto; SILVA JÚNIOR, José Alexandre da. Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r). 2009.

FRANCO, Laércio Joel; PASSOS, Afonso Dinis Costa. Fundamentos de epidemiologia. In: **Fundamentos de epidemiologia**. 200

GAVA, Caroline; CARDOSO, Andrey Moreira; BASTA, Paulo Cesar. Mortalidade infantil por cor ou raça em Rondônia, Amazônia Brasileira. **Revista de Saúde Pública**, v. 51, p. 1-9, 2017.

GIL, Antônio Carlos. Métodos e técnicas de pesquisa social. 6. ed. Editora Atlas SA, 2008.

HAY, A. M. The derivation of global estimates from a confusion matrix. **International Journal of Remote Sensing**, v. 9, n. 8, p. 1395-1398, 1988.

IBGE (Instituto Brasileiro de Geografia e Estatística). Cidades. 2014 Disponível em: https://cidades.ibge.gov.br/brasil/pr/santo-antonio-do-paraiso/pesquisa/39/30279?tipo=ranking. Acesso em: 12 junho 2019.

KROPIWIEC, Maria Volpato; FRANCO, Selma Cristina; AMARAL, Augusto Randüz do. Fatores associados à mortalidade infantil em município com índice de desenvolvimento humano elevado. **Revista Paulista de Pediatria**, v. 35, n. 4, p. 391-398, 2017.

LEAL, Maria do Carmo et al. Determinantes do óbito infantil no Vale do Jequitinhonha e nas regiões Norte e Nordeste do Brasil. **Revista de Saúde Pública**, v. 51, p. 1-9, 2017.

LIMA, Jaqueline Costa et al. Estudo de base populacional sobre mortalidade infantil. **Revista Ciência & Saúde Coletiva**, v. 22, p. 931-939, 2017.
SANDERS, Lídia Samara de Castro et al. Mortalidade infantil: análise de fatores associados em uma capital do Nordeste brasileiro. Cadernos Saúde Coletiva, v. 25, n. 1, 2017.

MALTA, D. C., DUARTE, E C. Causas de mortes evitáveis por ações efetivas dos serviços de saúde: uma revisão da literatura. **Ciência Saúde Coletiva**, Rio de Janeiro, v. 12, n. 3, p. 765-776, Junho 2007 . Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-81232007000300027>. Acesso em: 06 mai. 2019.

Ministério da Saúde. Indicadores de mortalidade: C.1 Taxa de mortalidade infantil. 2005. [Acesso em 3/5/2007]. Disponível: <
em http://tabnet.datasus.gov.br/cgi/idb2005/c01.htm

MORAIS NETO, Otaliba Libânio de et al. Diferenças no padrão de ocorrência da mortalidade neonatal e pós-neonatal no Município de Goiânia, Brasil, 1992-1996:

análise espacial para identificação das áreas de risco. **Cadernos de Saúde Pública**, v. 17, p. 1241-1250, 2001.

NASCIMENTO, Luiz Fernando Costa; ALMEIDA, Milena Cristina da Silva; GOMES, Camila de Moraes Santos. Causas evitáveis e mortalidade neonatal nas microrregiões do estado de São Paulo. **Revista Brasileira de Ginecologia e Obstetrícia**. Rio de Janeiro, v. 36, n. 7, p. 303-309, julho de 2014. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-72032014000700303 >. Acesso em. 16 jun 2019.

NETTO, Amanda et al. Mortalidade infantil: avaliação do programa Rede Mãe Paranaense em Regional de Saúde do Paraná. **Cogitare enferm**, v. 22, n. 1, p. 01-08, 2017.

PARANÁ. Secretaria de Estado da Saúde (SESA). Linha guia rede mãe paranaense. Paraná, 2012. Disponível em: <http://www.saude.pr.gov.br/arquivos/File/ACS/linha_guia_versao_final.pdf>. Acesso em: 31 mai. 2019.

PEARSON, Karl. VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. **Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character**, n. 187, p. 253-318, 1896.

RAMALHO, Alanderson Alves et al. Tendência da mortalidade infantil no município de Rio Branco, AC, 1999 a 2015. **Revista Saúde Pública**, v. 52, p. -, 2018.

RODRIGUES, Mirela et al. Análise espacial da mortalidade infantil e adequação das informações vitais: uma proposta para definição de áreas prioritárias**. Ciência & Saúde Coletiva**, v. 19, p. 2047-2054, 2014.

RUTSTEIN, David D. et al. Measuring the quality of medical care: a clinical method. New England **Journal of Medicine**, v. 294, n. 11, p. 582-588, 1976.

ROYSTON, Patrick. Approximating the Shapiro-Wilk W-Test for non-normality. Statistics and computing, v. 2, n. 3, p. 117-119, 1992.

SOARES, Darli Antônio; ANDRADE, Selma Maffei de; CAMPOS, João José Batista de. Epidemiologia e indicadores de saúde. **Bases da saúde coletiva**. Londrina: Ed.UEL, p. 183-210, 2001. Disponível em: <http://www.epsjv.fiocruz.br/pdtsp/includes/header_pdf.php?id=266&ext=.pdf&titulo= EPIDEMIOLOGIA >. Acesso em: 1 jun. 2019.

SPEARMAN, Charles. The proof and measurement of association between two things. **American journal of Psychology**, v. 15, n. 1, p. 72-101, 1904.

TEIXEIRA, Enise Barth. A análise de dados na pesquisa científica: importância e desafios em estudos organizacionais. **Desenvolvimento em questão**, v. 1, n. 2, p. 177-201, 2003.

WORLD BANK. Data: Taxa de Mortalidade Infantil, 2014. Disponível em: <https://data.worldbank.org/indicator/SP.DYN.IMRT.IN>. Acesso em: 12 mai. 2019.

APÊNDICE TRATAMENTO DA VARIÁVEL COBERTURA DE ATENÇÃO BÁSICA

In [10]:
```
import pandas as pd
import numpy as np
```

In [28]:
```
a = 'atencao_basica_pr/Cobertura-AB-TODOS OS MUNICÍPIOS - PR-Abril de 2014.csv'
```

In [31]:
```
meses = ['Janeiro', 'Fevereiro', 'Março', 'Abril', 'Maio', 'Junho', 'Julho', 'Agosto', 'Setembro','Outubro', 'Novembro', 'Dezembro']
anos = ['2014','2015','2016']
```

In [133]:
```
files = []

base = pd.read_csv('atencao_basica_pr/'+ caminho, sep=";", engine='python', skiprows=[0,1,2,3,4,5,6,7,410,411,412,413,414])
base = base[['IBGE']]
base = base.set_index('IBGE')
cobertura_atencao_basica = None
df = None
for ano in anos:
    for mes in meses:
        caminho = 'Cobertura-AB-TODOS OS MUNICÍPIOS - PR-{} de {}.csv'.format(mes, ano)
        df = pd.read_csv('atencao_basica_pr/'+ caminho, sep=";", engine='python', skiprows=[0,1,2,3,4,5,6,7,410,411,412,413,414])
        cobertura_atencao_basica = df[['IBGE', 'Cobertura AB']]
        cobertura_atencao_basica.columns = ['IBGE', 'cab{}_{}'.format(mes, ano)]
        base = base.join(cobertura_atencao_basica.set_index('IBGE'))
```

In [134]:
```
base = base.replace('%','', regex=True)
```

In [136]:
```
base.to_csv('ab_pr_2014_2016.csv',sep=';')
```

APÊNDICE ÁRVORE DE DECISÃO COM TAXA DE MORTALIDADE INFANTIL DE 2014

```
=== Run information ===
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
------------------
pib <= 42455
|   populacao <= 2408
|   |   saneamento_semi_adequado <= 80.68: GRUPO 4 (6.0/2.0)
|   |   saneamento_semi_adequado > 80.68: GRUPO 1 (4.0)
|   populacao > 2408: GRUPO 1 (39.0/10.0)
pib > 42455
|   populacao <= 32095
|   |   saneamento_adquado <= 7.25
|   |   |   idhm_renda <= 0.725
|   |   |   |   pib <= 43549: GRUPO 2 (3.0)
|   |   |   |   pib > 43549
|   |   |   |   |   pib <= 60795
|   |   |   |   |   |   idhm_renda <= 0.654
|   |   |   |   |   |   |   saneamento_semi_adequado <= 70.88: GRUPO 1 (5.0)
|   |   |   |   |   |   |   saneamento_semi_adequado > 70.88: GRUPO 2 (3.0)
|   |   |   |   |   |   idhm_renda > 0.654: GRUPO 1 (14.0)
|   |   |   |   |   pib > 60795
|   |   |   |   |   |   saneamento_semi_adequado <= 69.01
|   |   |   |   |   |   |   ab_2014 <= 94.778
|   |   |   |   |   |   |   |   populacao <= 7425: GRUPO 1 (2.0)
|   |   |   |   |   |   |   |   populacao > 7425: GRUPO 3 (4.0/1.0)
|   |   |   |   |   |   |   ab_2014 > 94.778
|   |   |   |   |   |   |   |   idhm_renda <= 0.65: GRUPO 2 (4.0)
|   |   |   |   |   |   |   |   idhm_renda > 0.65: GRUPO 3 (3.0/1.0)
|   |   |   |   |   |   saneamento_semi_adequado > 69.01
|   |   |   |   |   |   |   populacao <= 7541
|   |   |   |   |   |   |   |   idhm_renda <= 0.696
|   |   |   |   |   |   |   |   |   saneamento_semi_adequado <= 80.95: GRUPO 1 (10.0/2.0)
```

```
|  |  |  |  |  |  |  |  saneamento_semi_adequado > 80.95
|  |  |  |  |  |  |  |  |  populacao <= 4903: GRUPO 3 (3.0)
|  |  |  |  |  |  |  |  |  populacao > 4903: GRUPO 1 (3.0/1.0)
|  |  |  |  |  |  |  |  idhm_renda > 0.696: GRUPO 2 (6.0/1.0)
|  |  |  |  |  |  |  populacao > 7541: GRUPO 1 (17.0/1.0)
|  |  |  idhm_renda > 0.725: GRUPO 3 (4.0)
|  |  saneamento_adquado > 7.25
|  |  |  populacao <= 4145
|  |  |  |  idhm_renda <= 0.652: GRUPO 1 (2.0)
|  |  |  |  idhm_renda > 0.652
|  |  |  |  |  idhm_renda <= 0.677: GRUPO 3 (8.0)
|  |  |  |  |  idhm_renda > 0.677
|  |  |  |  |  |  saneamento_adquado <= 18.05: GRUPO 1 (4.0/1.0)
|  |  |  |  |  |  saneamento_adquado > 18.05: GRUPO 3 (3.0)
|  |  |  populacao > 4145
|  |  |  |  saneamento_inadequado <= 19.65: GRUPO 2 (142.0/89.0)
|  |  |  |  saneamento_inadequado > 19.65
|  |  |  |  |  saneamento_semi_adequado <= 32.62
|  |  |  |  |  |  saneamento_semi_adequado <= 28.33: GRUPO 3 (7.0/1.0)
|  |  |  |  |  |  saneamento_semi_adequado > 28.33: GRUPO 2 (3.0)
|  |  |  |  |  saneamento_semi_adequado > 32.62
|  |  |  |  |  |  pib <= 272552
|  |  |  |  |  |  |  saneamento_semi_adequado <= 65.98: GRUPO 1 (43.0/19.0)
|  |  |  |  |  |  |  saneamento_semi_adequado > 65.98: GRUPO 2 (3.0/1.0)
|  |  |  |  |  |  pib > 272552: GRUPO 3 (3.0)
|  populacao > 32095
|  |  idhm_renda <= 0.755
|  |  |  populacao <= 48198: GRUPO 1 (16.0/6.0)
|  |  |  populacao > 48198: GRUPO 2 (26.0/4.0)
|  |  idhm_renda > 0.755: GRUPO 1 (9.0/1.0)

Number of Leaves  :         30

Size of the tree :      59
```

APÊNDICE ÁRVORE DE DECISÃO COM TAXA DE MORTALIDADE INFANTIL DE 2015

PART decision list

---

populacao <= 4802 AND
populacao <= 3891 AND
populacao > 2578: GRUPO 1 (41.0/8.0)

populacao <= 4802 AND
populacao <= 2491 AND
saneamento_inadequado > 11.26: GRUPO 1 (11.0/1.0)

pib <= 43143 AND
saneamento_adquado > 3.7: GRUPO 3 (7.0)

populacao > 4677 AND
saneamento_adquado <= 12.96 AND
saneamento_inadequado <= 49.34: GRUPO 1 (75.0/33.0)

saneamento_adquado <= 6.15 AND
populacao > 4568: GRUPO 3 (8.0/1.0)

saneamento_adquado <= 5.07 AND
saneamento_semi_adequado <= 76.66 AND
saneamento_adquado <= 1.4: GRUPO 1 (3.0)

saneamento_adquado <= 5.07 AND
populacao <= 3913: GRUPO 4 (5.0/1.0)

populacao > 4568: GRUPO 1 (228.0/119.0)

populacao <= 3926: GRUPO 1 (3.0/1.0)

populacao <= 4020: GRUPO 3 (3.0)

saneamento_adquado <= 1.22 AND
idhm_educacao <= 0.649: GRUPO 4 (2.0)

saneamento_semi_adequado > 54.04 AND
saneamento_adquado > 1.22 AND
saneamento_semi_adequado > 62.01: GRUPO 1 (4.0)

saneamento_adquado <= 28.1: GRUPO 3 (5.0/1.0)

: GRUPO 1 (4.0/1.0)

Number of Rules : 14

APÊNDICE ÁRVORE DE DECISÃO COM TAXA DE MORTALIDADE INFANTIL DE 2016

J48 pruned tree

---

pib <= 264918
| populacao <= 3380
| | saneamento_adquado <= 1.59
| | | pib <= 33258
| | | | idhm_renda <= 0.674: GRUPO 3 (3.0)
| | | | idhm_renda > 0.674
| | | | | idhm_renda <= 0.68: GRUPO 1 (2.0)
| | | | | idhm_renda > 0.68: GRUPO 3 (3.0/1.0)
| | | pib > 33258: GRUPO 1 (5.36)
| | saneamento_adquado > 1.59
| | | saneamento_semi_adequado <= 76.04: GRUPO 1 (13.64)
| | | saneamento_semi_adequado > 76.04
| | | | saneamento_adquado <= 6.15: GRUPO 1 (8.0/2.0)
| | | | saneamento_adquado > 6.15: GRUPO 4 (2.0)
| populacao > 3380
| | idhm_educacao <= 0.585
| | | populacao <= 6532
| | | | idhm_renda <= 0.649
| | | | | saneamento_semi_adequado <= 39.07: GRUPO 3 (3.0)
| | | | | saneamento_semi_adequado > 39.07
| | | | | | pib <= 52258
| | | | | | | idhm_educacao <= 0.507: GRUPO 1 (5.0)
| | | | | | | idhm_educacao > 0.507
| | | | | | | | pib <= 43901: GRUPO 2 (2.0)
| | | | | | | | pib > 43901: GRUPO 1 (2.0)
| | | | | | pib > 52258: GRUPO 2 (4.0)
| | | | idhm_renda > 0.649: GRUPO 1 (16.0/2.0)
| | | populacao > 6532

```
|   |   |   saneamento_inadequado <= 49.9
|   |   |   |   ab_2016 <= 79.427
|   |   |   |   |   saneamento_semi_adequado <= 68.54: GRUPO 3 (6.0/1.0)
|   |   |   |   |   saneamento_semi_adequado > 68.54: GRUPO 2 (2.0)
|   |   |   |   ab_2016 > 79.427
|   |   |   |   |   ab_2016 <= 86.407
|   |   |   |   |   |   ab_2016 <= 82.55: GRUPO 1 (3.0)
|   |   |   |   |   |   ab_2016 > 82.55: GRUPO 3 (4.0/1.0)
|   |   |   |   |   ab_2016 > 86.407: GRUPO 2 (45.0/22.0)
|   |   |   saneamento_inadequado > 49.9: GRUPO 1 (6.0/3.0)
|   |   idhm_educacao > 0.585
|   |   |   saneamento_inadequado <= 28.93
|   |   |   |   ab_2016 <= 98.91
|   |   |   |   |   saneamento_inadequado <= 7.25
|   |   |   |   |   |   saneamento_adquado <= 79.21: GRUPO 1 (14.0/3.0)
|   |   |   |   |   |   saneamento_adquado > 79.21: GRUPO 2 (3.0/2.0)
|   |   |   |   |   saneamento_inadequado > 7.25
|   |   |   |   |   |   saneamento_adquado <= 30.57
|   |   |   |   |   |   |   saneamento_adquado <= 21.28
|   |   |   |   |   |   |   |   saneamento_adquado <= 14.73: GRUPO 1 (20.0/13.0)
|   |   |   |   |   |   |   |   saneamento_adquado > 14.73: GRUPO 3 (7.0/1.0)
|   |   |   |   |   |   |   saneamento_adquado > 21.28
|   |   |   |   |   |   |   |   idhm_renda <= 0.694: GRUPO 2 (3.0/1.0)
|   |   |   |   |   |   |   |   idhm_renda > 0.694: GRUPO 1 (3.0)
|   |   |   |   |   |   saneamento_adquado > 30.57: GRUPO 3 (7.0/1.0)
|   |   |   |   ab_2016 > 98.91
|   |   |   |   |   saneamento_inadequado <= 11.35
|   |   |   |   |   |   saneamento_inadequado <= 2.52: GRUPO 1 (5.0/1.0)
|   |   |   |   |   |   saneamento_inadequado > 2.52
|   |   |   |   |   |   |   idhm_renda <= 0.712
|   |   |   |   |   |   |   |   idhm_educacao <= 0.616: GRUPO 1 (6.0)
|   |   |   |   |   |   |   |   idhm_educacao > 0.616
|   |   |   |   |   |   |   |   |   populacao <= 4095: GRUPO 1 (2.0)
|   |   |   |   |   |   |   |   |   populacao > 4095
```

```
|  |  |  |  |  |  |  |  |  populacao <= 11875
|  |  |  |  |  |  |  |  |  |  populacao <= 5692
|  |  |  |  |  |  |  |  |  |  |  populacao <= 4903: GRUPO 3 (2.0)
|  |  |  |  |  |  |  |  |  |  |  populacao > 4903: GRUPO 2 (2.0)
|  |  |  |  |  |  |  |  |  |  populacao > 5692: GRUPO 3 (5.0)
|  |  |  |  |  |  |  |  |  populacao > 11875
|  |  |  |  |  |  |  |  |  |  idhm_renda <= 0.702: GRUPO 2 (2.0/1.0)
|  |  |  |  |  |  |  |  |  |  idhm_renda > 0.702: GRUPO 1 (2.0)
|  |  |  |  |  |  |  idhm_renda > 0.712
|  |  |  |  |  |  |  |  idhm_renda <= 0.73: GRUPO 3 (6.0/1.0)
|  |  |  |  |  |  |  |  idhm_renda > 0.73
|  |  |  |  |  |  |  |  |  idhm_renda <= 0.739: GRUPO 2 (2.0)
|  |  |  |  |  |  |  |  |  idhm_renda > 0.739: GRUPO 3 (3.0/1.0)
|  |  |  |  |  saneamento_inadequado > 11.35
|  |  |  |  |  |  saneamento_inadequado <= 25.5: GRUPO 1 (56.0/16.0)
|  |  |  |  |  |  saneamento_inadequado > 25.5: GRUPO 3 (4.0/1.0)
|  |  |  saneamento_inadequado > 28.93: GRUPO 1 (12.0)
pib > 264918
|  pib <= 547470
|  |  idhm_educacao <= 0.639
|  |  |  ab_2016 <= 77.816: GRUPO 2 (7.0/1.0)
|  |  |  ab_2016 > 77.816
|  |  |  |  ab_2016 <= 88.965: GRUPO 3 (2.0)
|  |  |  |  ab_2016 > 88.965
|  |  |  |  |  populacao <= 13169: GRUPO 3 (4.0/1.0)
|  |  |  |  |  populacao > 13169
|  |  |  |  |  |  idhm_renda <= 0.73
|  |  |  |  |  |  |  populacao <= 25172: GRUPO 1 (5.0/1.0)
|  |  |  |  |  |  |  populacao > 25172: GRUPO 2 (7.0/1.0)
|  |  |  |  |  |  idhm_renda > 0.73: GRUPO 1 (2.0/1.0)
|  |  idhm_educacao > 0.639
|  |  |  saneamento_inadequado <= 12.01
|  |  |  |  idhm_educacao <= 0.643: GRUPO 1 (2.0)
|  |  |  |  idhm_educacao > 0.643
```

| | | | saneamento_adquado <= 84.61: GRUPO 2 (12.0/1.0)
| | | | saneamento_adquado > 84.61: GRUPO 1 (2.0)
| | | saneamento_inadequado > 12.01: GRUPO 1 (5.0)
| pib > 547470
| | saneamento_semi_adequado <= 56.15: GRUPO 1 (46.0/12.0)
| | saneamento_semi_adequado > 56.15: GRUPO 2 (4.0)

Number of Leaves : 52

Size of the tree : 103

APÊNDICE ÁRVORE DE DECISÃO COM TAXA DE MORTALIDADE INFANTIL DO TRIÊNIO (2014-2016)

J48 pruned tree

---

populacao <= 17240
| pib <= 60583
| | populacao <= 2408
| | | idhm_educacao <= 0.618: GRUPO 3 (3.0/1.0)
| | | idhm_educacao > 0.618
| | | | populacao <= 1818: GRUPO 1 (2.0/1.0)
| | | | populacao > 1818: GRUPO 2 (5.0/1.0)
| | populacao > 2408: GRUPO 1 (91.0/33.0)
| pib > 60583
| | saneamento_semi_adequado <= 76.86
| | | idhm_renda <= 0.728
| | | | idhm_renda <= 0.644
| | | | | saneamento_adquado <= 27.61
| | | | | | saneamento_semi_adequado <= 50.7
| | | | | | | populacao <= 11337
| | | | | | | | idhm_renda <= 0.615: GRUPO 3 (3.0)
| | | | | | | | idhm_renda > 0.615
| | | | | | | | saneamento_semi_adequado <= 41.17: GRUPO 2 (2.0)

```
|   |   |   |   |   |   |   |   saneamento_semi_adequado > 41.17: GRUPO 3 (2.0)
|   |   |   |   |   |   |   populacao > 11337: GRUPO 2 (4.0)
|   |   |   |   |   |   saneamento_semi_adequado > 50.7
|   |   |   |   |   |   |   saneamento_inadequado <= 25.03: GRUPO 2 (4.0)
|   |   |   |   |   |   |   saneamento_inadequado > 25.03: GRUPO 1 (2.0)
|   |   |   |   |   saneamento_adquado > 27.61: GRUPO 3 (4.0)
|   |   |   |   idhm_renda > 0.644
|   |   |   |   |   saneamento_semi_adequado <= 18.69
|   |   |   |   |   |   idhm_educacao <= 0.62: GRUPO 2 (4.0)
|   |   |   |   |   |   idhm_educacao > 0.62: GRUPO 3 (5.0)
|   |   |   |   |   saneamento_semi_adequado > 18.69: GRUPO 1 (119.0/67.0)
|   |   |   idhm_renda > 0.728
|   |   |   |   idhm_educacao <= 0.692: GRUPO 2 (8.0)
|   |   |   |   idhm_educacao > 0.692: GRUPO 3 (2.0)
|   |   saneamento_semi_adequado > 76.86
|   |   |   ab <= 69.243: GRUPO 1 (2.0/1.0)
|   |   |   ab > 69.243
|   |   |   |   idhm_educacao <= 0.664
|   |   |   |   |   idhm_renda <= 0.699
|   |   |   |   |   |   idhm_educacao <= 0.573: GRUPO 2 (5.0)
|   |   |   |   |   |   idhm_educacao > 0.573
|   |   |   |   |   |   |   saneamento_inadequado <= 9.22: GRUPO 2 (3.0)
|   |   |   |   |   |   |   saneamento_inadequado > 9.22
|   |   |   |   |   |   |   |   ab <= 99.837
|   |   |   |   |   |   |   |   |   saneamento_semi_adequado <= 81.9: GRUPO 2 (2.0)
|   |   |   |   |   |   |   |   |   saneamento_semi_adequado > 81.9: GRUPO 1 (3.0)
|   |   |   |   |   |   |   |   ab > 99.837: GRUPO 1 (6.0)
|   |   |   |   |   idhm_renda > 0.699: GRUPO 1 (10.0)
|   |   |   |   idhm_educacao > 0.664: GRUPO 2 (8.0/1.0)
populacao > 17240
|   saneamento_semi_adequado <= 41.61: GRUPO 1 (68.0/32.0)
|   saneamento_semi_adequado > 41.61
|   |   saneamento_adquado <= 50.39: GRUPO 2 (30.0/4.0)
|   |   saneamento_adquado > 50.39: GRUPO 1 (2.0/1.0)
```

Number of Leaves : 27

Size of the tree : 53