A Inteligência Artificial aplicada para o estudo da mortalidade infantil e sua relação com o saneamento básico no Brasil

Maria Paula Henriques Prandt
Unifesp
São José dos Campos - SP
maria.prandt@unifesp.br

Leonardo Silva Pinto
Unifesp
São José dos Campos - SP
leonardo.pinto@unifesp.br

Resumo

A Taxa de Mortalidade Infantil é um indicador que mensura a proporção de mortes de crianças antes de um ano. Nesse contexto, o presente trabalho busca aplicar técnicas de Inteligência Artificial, principalmente regressão, para analisar quais fatores socioeconômicos influenciam mais o crescimento dessa taxa em comparação com outros. Diante disso, pode-se tirar conclusões acerca de quais fatores econômicos, educacionais e sanitários podem ser usados para melhor predizer a TMI dos municípios brasileiros.

I. Introdução

A Taxa de Mortalidade Infantil (TMI) é um dos principais índices socioeconômicos que revelam a condição de saúde das cidades brasileiras. Essa métrica é calculada anualmente e indica a proporção de quantas crianças morrem antes de completarem um ano de vida. Tal parâmetro é de suma importância, pois pode ser usado como indicador da infraestrutura de cada município e da sua capacidade em prevenir e tratar doenças que ocasionam essas mortes.

A motivação para este estudo consiste na existência de trabalhos que já relacionam fatores socioeconômicos, sanitários e educacionais com a taxa de mortalidade infantil, nos quais provam uma relação entre eles. Sendo assim, é importante fazer um estudo de forma a analisar esses fatores de forma a corroborar com os estudos já realizados e também adquirir novas informações.

Diante do contexto supracitado, muitos estudos buscam entender a correlação da TMI com a infraestrutura dos municípios [2,3,9,11]. Desse modo, o presente trabalho busca investigar a hipótese de que os índices de saneamento básico (tais como o percentual de esgoto tratado e o percentual de acesso à água tratada) são os principais ofensores da TMI em comparação com os índices de escolaridade e índices econômicos.

O estudo desses índices ao longo do tempo pode fornecer informações valiosas para o entendimento de quais decisões e iniciativas precisam ser tomadas para melhorar a qualidade de vida das populações. Além disso, busca-se, nesse estudo, encontrar padrões e exceções que representem um conhecimento útil, seja numa perspectiva nacional, por estado ou macrorregião.

Para a execução do estudo, será utilizado, como fonte de dados, os registros administrativos obtidos pela Atlas Brasil, de 2012 até 2017. Com isso, será possível comparar os resultados obtidos com os de outros trabalhos relacionados que utilizam de outras fontes, além de entender os padrões e as exceções existentes nos dados.

II. Conceitos Fundamentais

A. Taxa de Mortalidade Infantil

A Taxa de Mortalidade Infantil é métrica calculada para cada município e é definida pela proporção entre o número de óbitos de crianças com menos de um ano de idade e o número de crianças nascidas vivas, tudo isso multiplicado por mil, assim como indicado pela Figura 1 [1].

$$\frac{\text{Número de óbitos de crianças com menos de 1 ano de vida}}{\text{Número de nascidos vivos}} \times 1.000$$

Figura 1: Fórmula para calcular a TMI

Alguns autores apontam a necessidade de acompanhamento dessa taxa dado que muitos óbitos podem ser evitados quando as mães e seus filhos possuem acesso a serviços de saúde [2] e a saneamento básico [3].

B. Inteligência Artificial

A inteligência artificial (IA) pode ser descrita de diversas maneiras, mas, em suma, se trata de um ramo da ciência da

computação que busca automatizar o processo de pensamento e tomada de decisão. Para o presente trabalho será usado como base a seguinte definição de Inteligência Artificial:

"[Automatização de] atividades que associamos ao pensamento humano, atividades como a tomada de decisões, a resolução de problemas, o aprendizado..." (Bellman, 1978)

Essa área de tecnologia tem diversos segmentos, alguns buscam criar IAs para simular como humanos pensam e agem, já outras buscam resultados baseados na racionalidade [4].
Independente da definição atribuída à essa tecnologia, fato é que ela pode e deve ser utilizada como ferramenta para encontrar dados novos, úteis e relevantes a partir de bases de dados.

C. Clusterização

Diante da possibilidade de utilização da IA como ferramenta para mineração de dados, um método de aprendizado muito aplicado é a Clusterização ou Clustering.
Esse método consiste em um aprendizado não supervisionado, ou seja, um aprendizado que utiliza dados não rotulados. A partir dessas amostras, o algoritmo criará grupos a fim de, posteriormente, o especialista analisar e entender o comportamento de cada cluster [5].
Com a utilização de clusterização, é possível definir similaridades e diferenças entre os grupos de dados e extrair conclusões úteis sobre eles. Um dos algoritmos mais utilizados nesse processo é o K-means que é um método de clusterização particional e tem como métrica a distância euclidiana entre as amostras de dados. Neste algoritmo, são definidos k-centróides iniciais e, a partir deles, as distâncias entre os demais pontos são calculados e os grupos são formados. Em seguida, os centróides são recalculados e os grupos refeitos até que mais nenhuma amostra mude de grupo [5].

D. Regras de Associação

As Regras de Associação (RAs) são caracterizadas por encontrar elementos que co-ocorrem numa determinada base de transacções [6]. Em outras palavras, trata-se de encontrar declarações do tipo "se-então", que ajudam a mostrar a probabilidade de relacionamentos entre itens de dados em grandes conjuntos de dados em vários tipos de bancos de dados

E. Regressão

Imagine que existe um conjunto de amostras e busca-se uma função h(x) que se aproxima da função f(x) que define as amostras. Nesse caso, a tarefa pode ser definida como uma regressão[4]. Um exemplo didático é uma regressão linear, na qual busca-se encontrar a melhor h(x) tal que:

$$h(x) = w_1 x + w_0$$

Tarefas desse tipo podem ser realizadas por Redes Neurais Artificiais (RNAs) que consistem em um modelo computacional de rede neural capaz de resolver problemas de regressão e classificação. Esses modelos funcionam "disparando" quando as combinações lineares das entradas ultrapassam um limiar - análogo a um funcionamento de um neurônio biológico [4].

F. Validação Cruzada

A Validação Cruzada é uma técnica utilizada nos modelos de Inteligência Artificial para detectar o overfitting dos dados, ou seja, se os dados estão muito ajustados aos seus dados de treino, e para isso ele particiona a base de dados em conjuntos de treino e teste.
Um dos métodos de Validação Cruzada é o K-fold, que consiste em criar K conjuntos de dados, de forma que um folds são os testes, enquanto o restante de treino, fazendo uma rotação até que todos os conjuntos tenham sido de teste. Dessa forma, detectar e tratar casos de overfitting se tornam bem mais simples, como por exemplo, utilizando a média das acurácias obtidas com cada fold de teste.

G. Medidas de Avaliação de Regressão Linear

Existem algumas formas de realizar medidas de avaliação na tarefa de Regressão Linear. Duas delas, são:

Erro Médio Absoluto (MAE): Essa métrica indica a média das diferenças entre o valor predito pelo modelo e o valor real. Pode-se aplicar essa medida no conjunto de teste e ela é calculada pela fórmula indicada na figura 2 [12] onde *y* é o valor predito na regressão, *p* o valor real da amostra e *n* o número de amostras.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - p_i|$$

Figura 2: Fórmula para calcular a MAE

R^2 Ajustado: Se trata de uma medida estatística que pode ser usada para comparar diferentes regressões lineares. Ela é calculada penalizando variáveis que não contribuem para o modelo e leva em consideração quantidade de parâmetros e quantidade de variáveis preditas. Pode ser calculada pela fórmula indicada pela figura 3, onde p é o número de variáveis independentes e n o número de observações [13].

$$R^2_{Ajustado} = 1 - \frac{(1-R^2)(n-1)}{(n-p-1)}$$

Figura 3: Fórmula para calcular o R^2_{Ajustado}

$$R^2 = 1 - \frac{\Sigma(y_{i\,real} - y_{i\,predito})^2}{\Sigma(y_{i\,real} - y_{médio\ dos\ valores\ preditos})^2}$$

Figura 4: Fórmula para calcular o R^2

Diante dessas métricas, pretende-se, então, realizar diferentes regressões lineares utilizados diferentes índices e comparar essas regressões.

III. Trabalhos Relacionados

Alguns estudos têm sido feitos para entender os fatores que influenciam o aumento da TMI no Brasil, porém a maioria dos artigos encontrados não aplicaram as técnicas de Inteligência

Artificial para realizar suas análises. No entanto, vale ressaltar alguns estudos realizados:

A. Diferenciais nos fatores de risco para a mortalidade infantil em cinco cidades brasileiras: um estudo de caso-controle com base no SIM e no SINASC

O estudo estabelece uma relação entre a TMI e as condições socioeconômicas do município. O mesmo não utiliza de métricas referente ao saneamento básico, porém, relaciona condições socioeconômicas mais baixas com o aumento da TMI e, de forma lógica, pode-se pressupor uma relação entre cidades mais pobres e piores condições de saneamento básico [2].

B. Mortalidade infantil e saneamento básico : sua incidência nas regiões brasileiras. Lume Repositório Digital

O estudo realizado por Fernanda Rutkovski utiliza dados históricos disponibilizados na internet para o público e relaciona a TMI com saneamento básico. Os resultados mostram que as regiões Norte e Nordeste do Brasil são as que apresentam maior TMI e também as que apresentam piores condições de saneamento básico, coleta de lixo e tratamento de água [3].

C. Impactos do saneamento sobre saúde e educação: uma análise espacial

A tese discorre sobre o efeito do saneamento na saúde e educação, utilizando de dados do Brasil, conseguiu estabelecer um nível de relação entre o saneamento básico e a educação, também conseguiu verificar que a população mais nova é a mais atingida pelas consequências de um saneamento básico precário [9].

D. Regression Model to Evaluate the Impact of Basic Sanitation Services in Households and Schools on Child Mortality in the Municipalities of the State of Alagoas, Brazil

O artigo consiste de uma pesquisa no estado de Alagoas, relacionando o impacto do serviço de saneamento básico em casas e escolas com a taxa de mortalidade infantil. Em seu estudo, ele trabalha com dados do IBGE de diferentes épocas e utiliza de modelos de regressão para estimar os valores de taxa de mortalidade, o estudo concluiu que uma melhora apenas no esgoto de casas já diminuiria consideravelmente a taxa de mortalidade infantil.

IV. Objetivo

Busca-se com o presente trabalho, principalmente, encontrar parâmetros socioeconômicos que se relacionam com a Taxa de Mortalidade Infantil e, de modo mais específico, realizar uma análise comparativa entre a influência dos índices sanitários, educacionais e econômicos na TMI.

Com essas análises feitas e com ajuda de um especialista, será possível interpretar o quanto a estrutura de saneamento básico do município influencia a TMI.

V. Metodología Experimental

Nesta seção será comentado acerca do procedimento que será feito a fim de encontrar parâmetros relevantes para o estudo da TMI, sendo que será utilizado um pipeline de mineração de dados na forma de abordar o problema do presente projeto [10].

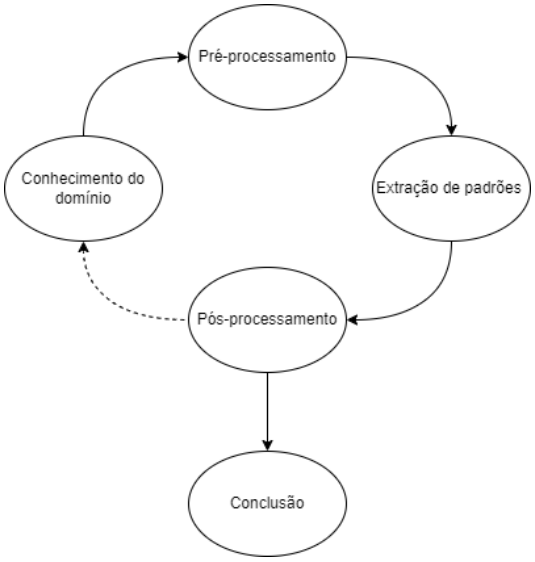

Figura 5: Pipeline da metodologia experimental

De modo mais específico, será seguido o pipeline experimental apresentado na Figura 6.

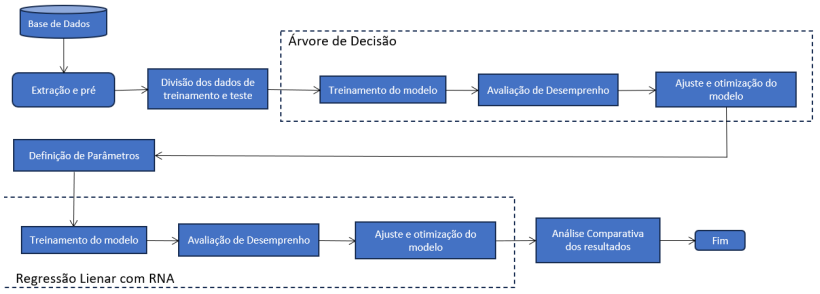

Figura 6: Pipeline Experimental

A. Conhecimento do Domínio

Para entender melhor os fatores relacionados a TMI, contamos com a ajuda de uma especialista para nos fornecer assistência, tanto no uso da base de dados que iremos utilizar nesse artigo, como também em fatores que podemos focar a fim de adquirir resultados mais relevantes.

Dessa forma, iremos utilizar de uma base de dados pública, com dados de municípios do Brasil contendo diversos indicadores, dentre eles a TMI, para encontrar fatores relevantes e relacionados a TMI.

B. Base de Dados

A base de dados utilizada é o Registro Administrativo Total de 2012 a 2017 e se encontra no acervo do site Atlas Brasil, ele é disponibilizado em um arquivo .xlsx, contendo informações de diversos fatores coletados pelo IBGE, possuindo a menor granularidade como município, porém tendo informações sumarizadas em outras quatro granularidades.

Para execução das comparações propostas pelo trabalho vale ressaltar que será necessário dividir a base de dados em 3 bases menores. Essas tabelas irão conter todas as amostras, porém com parâmetros (colunas) diferentes. Sendo assim, as bases de dados irão conter os seguintes parâmetros:

1. Base com informações de saneamento básico:

- NIS_AGUA (Percentual da população urbana residentes em domicílios ligados a rede de abastecimento de água)
- SNIS_PESGOTO (Percentual da população urbana residente em domicílios ligados à rede de esgotamento sanitário)
- SNIS_PESGTRA (Percentual de esgoto tratado)
- PDEFSAN (Percentual de pessoas inscritas no Cadastro Único sem abastecimento de água, esgotamento sanitário e coleta de lixo adequadas)

2. Base com informações sobre educação:

- TTREVA_EF_TOTAL (Taxa de evasão no ensino fundamental)
- TTREVA_EM_TOTAL (Taxa de evasão no ensino médio)
- DOCSUP_EF_TOTAL (Porcentagem de docentes do ensino fundamental com formação adequada)
- DOCSUP_EM_TOTAL (Porcentagem de docentes do ensino médio com formação adequada).

3. Base com informações economicas:

- REN_PIBPC_D (Produto interno bruto per capita)
- REN_TRPCBF_D (Transferência per capita do Bolsa Família)
- PIND_POS (Percentual de extremamente pobres no Cadastro Único pós Bolsa Família).

Vale salientar que ao longo do desenvolvimento do projeto novas colunas podem ser inseridas nessas bases a fim de encontrar informações cada vez mais relevantes.

C. Pré Processamento

Diante dos dados extraídos, foi realizado verificado que haviam muitos dados faltantes, em diversas colunas, o que se fez necessário alguns tratamentos antes de utilizar das informações, sendo assim foi feito inicialmente uma filtragem dos dados que seriam trabalhados, ou seja, das colunas/características que iremos utilizar em nossa análise para depois ser realizado a limpeza.

Com os dados filtrados verificamos a quantidade de valores faltantes, como a quantidade era um valor considerável, optamos por fazer uma interpolação dos dados ordenando por nome da cidade, ano de pesquisa, população total e PIB, de forma que a distribuição dos dados não fosse muito prejudicada. Após feita a interpolação, restaram alguns dados faltantes, porém como a quantia era mínima, optou-se por retirá-los da análise. Também foi feita uma limpeza de forma a retirar outliers que poderiam vir a prejudicar a análise, sendo feito com base na coluna de mortalidade infantil utilizando tanto de um limite inferior como superior.

Com os dados tratados, para uma análise comparativa, foi feita uma divisão em três sub bases, cada uma com dados de diferentes parâmetros, sendo eles sanitários(base de dados 1), educacionais(base de dados 2) e econômicos(base de dados 3), de forma a relacionarmos eles com a taxa de mortalidade infantil de forma separada para verificar quais viriam a se relacionar mais com esta taxa.

D. Protocolo de Validação

O protocolo de validação utilizado para a regressão é o de validação cruzada K-fold, usando da biblioteca do sklearn para sua criação, foram utilizados cinco folds, sendo que assim visamos na análise considerando os diferentes resultados obtidos, para dessa forma termos um resultado menos enviesado e mais consistente.

E. Extração de Padrões

Para encontrarmos as relações entre os dados, utilizamos algoritmos de IA de árvore de decisão e regressão. Sendo que primeiramente utilizamos da árvore de decisão para podermos ter uma melhor compreensão sobre a relevância de cada parâmetro da sub base com a taxa de mortalidade infantil, e depois foi utilizada de regressão para predizer ela, utilizando ou não das informações obtidas da árvore de decisão para filtrar os parâmetros mais relevantes.

F. Métricas de Avaliação

Para medir a acurácia dos algoritmos, utilizamos do MAE, tanto na árvore de decisão como na regressão, como o parâmetro que estamos querendo predizer é a taxa de mortalidade infantil, ela já está normalizada e é uma medida de fácil compreensão, sendo assim, não vimos a necessidade de procurar métricas mais complexas para avaliar os modelos.

G. Pós Processamento

Com os resultados obtidos na etapa anterior iremos, junto a um especialista, procurar entender se o conhecimento adquirido é novo e útil [10], e se ele é satisfatório para o objetivo do projeto, e caso não seja, será reiniciado o processo, como mostrado na figura 2.

VI. Resultados e Discussão

Após feito o processo de mineração de dados, foram retiradas algumas informações acerca das métricas retiradas dos algoritmos usados nos experimentos.

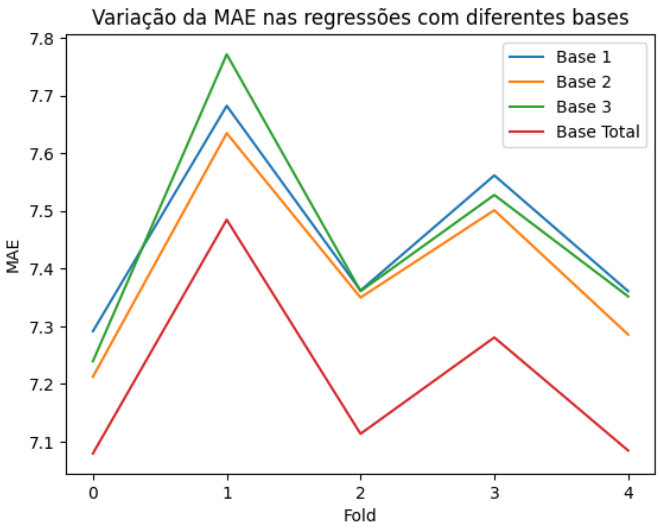

Figura 7 - Variação MAE na base de dados

Conforme a figura acima, vemos que as sub bases de dados variaram de maneira semelhante, porém verificamos que com exceção da base de dados total, a base de dados 2 foi a que obteve menor erro, que foi a base com dados educacionais.

Em nossa análise, fizemos também certas hipóteses a fim de verificar a relação entre o saneamento básico com a taxa de mortalidade infantil, entre elas a verificação de cidades com menos população e menor PIB tem os menores valores do parâmetro SNIS_PESGOTO (Percentual da população urbana residente em domicílios ligados à rede de esgotamento sanitário). Porém ao fazer a análise verificamos que essa hipótese não é válida, sendo assim, pegamos uma das amostras em que essa hipótese era verdadeira para verificarmos mais profundamente.

Sendo assim, escolhemos a cidade Rancho Queimado e verificamos tanto o SNIS_PESGOTO como a TMI ao longo dos anos analisados, conforme as duas figuras abaixo.

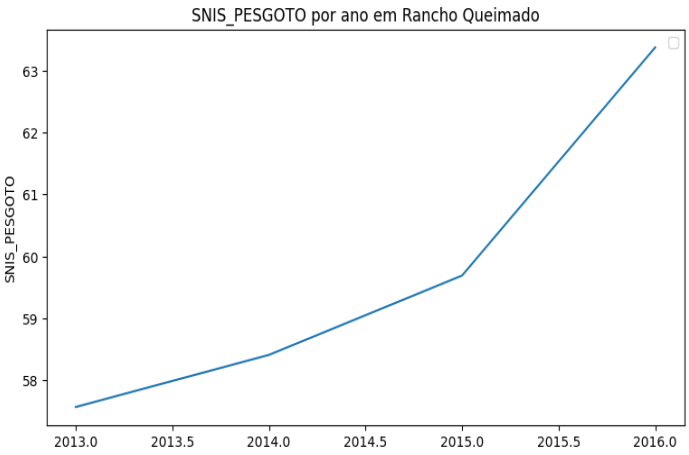

Figura 8 - SNIS_PESGOTO ao longo do tempo

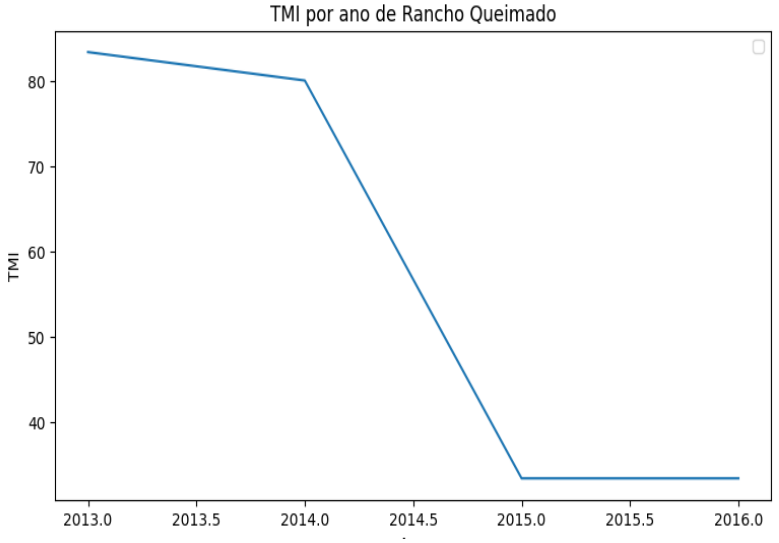

Figura 8 - SNIS_PESGOTO ao longo do tempo

A partir desses gráficos, podemos verificar que com o passar do tempo, conforme a taxa de residências com esgoto tratado aumentou, diminuiu a taxa de mortalidade infantil.

VII. Conclusões e Trabalhos Futuros

VIII. Entrega Final

A entrega final deste projeto consistirá em uma análise dos resultados obtidos através do pipeline apresentado na metodologia experimental, em que constituirá de resultados satisfatórios de pelo menos uma das técnicas de IA apresentadas.

Ademais, pretende-se entregar uma análise comparativa do impacto de diferentes tipos de índices na TMI. Essa análise pode ser feita num nível de região ou estado e podem ser feitas ao longo do tempo. As especificações serão definidas conforme os direcionamentos do especialista, sempre focando em achar informações relevantes e úteis.

References

[1] Boletim epidemiológico - Ministério da Saúde | Outubro de 2021. <Disponível em: https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/boletins/epidemiologicos/edicoes/2021/boletim_epidemiologico_svs_37_v2.pdf>

[2] Maia, L. T. de S., Souza, W. V. de ., & Mendes, A. da C. G.. (2012). Diferenciais nos fatores de risco para a mortalidade infantil em cinco cidades brasileiras: um estudo de caso-controle com base no SIM e no SINASC. Cadernos De Saúde Pública, 28(11), 2163–2176. https://doi.org/10.1590/S0102-311X2012001100016

[3] Rutkovski, Fernanda (2019). Mortalidade infantil e saneamento básico : sua incidência nas regiões brasileiras. Lume Repositório Digital UFRGS. http://hdl.handle.net/10183/201848

[4] NORVIG, Peter. Inteligência Artificial. [Digite o Local da Editora]: Grupo GEN, 2013. E-book. ISBN 9788595156104. Disponível em:

https://integrada.minhabiblioteca.com.br/#/books/9788595156104/. Acesso em: 29 mai. 2023.

[5] Aula de Aprendizagem não supervisionada - Agrupamento (Clustering). Faria, Fabio Augusto (UNIFESP).

[6] Vasconcelos, Lívia Maria Rocha de (2004). Aplicação de Regras de Associação para Mineração de Dados na Web. Instituto de Informática - Universidade Federal de Goiás. Disponível em: https://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_004-04.pdf

[7] Caracterizando a Mortalidade Infantil utilizando técnicas de Machine Learning: um Estudo de Caso em dois Estados Brasileiros - Santa Catarina e Amapá

[8] Oliveir,a Ivana Corrê (2001a. Aplicação de Data Mining na Busca de um Modelo de Prevenção da Mortalidade Infantil. <Disponível em: https://repositorio.ufsc.br/xmlui/handle/123456789/81803>

[9] SCRIPTORE, Juliana Souza. **Impactos do saneamento sobre saúde e educação: uma análise espacial**. 2016. Tese (Doutorado em Teoria Econômica) - Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, 2016. doi:10.11606/T.12.2016.tde-02082016-165540. Acesso em: 2023-06-05.

[10] Aula de Mineração de Dados. Faria, Fabio Augusto (UNIFESP).

[11] Cavalcanti, A.; Teixeira, A.; Pontes, K. Regression Model to Evaluate the Impact of Basic Sanitation Services in Households and Schools on Child Mortality in the Municipalities of the State of Alagoas, Brazil. Sustainability 2019, 11, 4150. https://doi.org/10.3390/su11154150

[12] Aula MIT. Introduction to Deep Learning. Alexander Amini. January 9,2023.

[13] Flávia Chein. Introdução aos modelos de regressão linear. Brasília - DF – Enap.