Identificação de Perfis de Óbitos infantis em Santa Maria/RS no Período de 2000 a 2017

Márian O. Pires[1], Sylvio Andre Garcia Vieira [1]

[1]Curso Sistemas de Informação - Universidade Franciscana (UFN)
Caixa Postal 151 – 97010-030 – Santa Maria – RS – Brasil

piresoleques@gmail.com, sylvio@ufn.edu.br

Abstract. *Child mortality is an indicator used to set public policy priorities, plan actions and services, monitor health status, and evaluate health system performance in a region. Annually the Department of Informatics of the Unified Health System (DATASUS) publishes health-related data from all regions of Brazil, providing the use of computational techniques for data analysis. In this context, the present work aims to use association rules and artificial neural networks in order to detect patterns of causes of death in infant deaths in the city of Santa Maria.*

Resumo. *Mortalidade infantil é um indicador utilizado para definir prioridades em políticas públicas, planejar ações e serviços, monitorar a situação de saúde e avaliar o desempenho do sistema de saúde de uma região. Anualmente o Departamento de Informática do Sistema Único de Saúde (DATASUS) publica dados relacionados à saúde de todas as regiões do brasil, propiciando a utilização de técnicas computacionais para análise de dados. Neste contexto, o trabalho apresentado tem como objetivo utilizar de regras de associação e redes neurais artificiais, no intuito de detectar padrões de causas de mortes nos óbitos infantis na cidade de Santa Maria.*

1. Introdução

Em 2017, foram estimadas cerca de 4,1 milhões de mortes de crianças antes de completar um ano de idade no mundo [World Health Organization 2019] . Comparado a anos anteriores, estes níveis de mortalidade infantil (MI) tenderam a diminuir. A queda se torna expressiva quando equipara-se os números atuais com os de décadas passadas, em que, em 1990, o número anual de mortes de crianças menores de um ano chegou a aproximadamente 8,8 milhões [World Health Organization 2019].

A taxa de mortalidade infantil (TMI) é um tópico mencionado com frequência em debates da Organização das Nações Unidas (ONU). Há um empenho para que os países-membros assumam a responsabilidade de diminuir a TMI, entre outros fatores, em prol da eliminação da extrema pobreza e da fome no planeta [World Health Organization 2019].

A relação entre a MI com a fome e situações de extrema pobreza, se deve por uma combinação de fatores biológicos, sociais, culturais e de falhas do sistema de saúde [França et al. 2017]. É uma temática que, por sua abrangência, se torna um instrumento indispensável para: definir prioridades nas políticas públicas, planejar ações e serviços,

monitorar a situação de saúde, bem como avaliar o desempenho do sistema de saúde de uma região [Paiz et al. 2018].

Entre 1982 e 2005, a TMI no Brasil apresentou uma queda de aproximadamente 80,4%, de 71,3 crianças mortas antes de um ano de idade (a cada 1000 nascidos vivos) para 14 por 1000 nascidos vivos [United Nations Children's Fund 2017]. Este declínio foi acompanhado por medidas públicas e diversas ações estratégicas relacionadas à melhoria das condições de vida e assistência à saúde. No entanto, as taxas de mortalidade em crianças que vivem em favelas urbanas, crianças indígenas e algumas regiões ainda permanecem com a TMI acima do recomendado pela ONU [Menezes et al. 2019].

No intuito de produzir dados de qualidade para análise, planejamento e avaliações de ações de saúde, o Ministério da Saúde, nos últimos anos, tem realizado investimentos específicos nos sistemas de informação nacionais [França et al. 2017]. O Departamento de Informática do Sistema Único de Saúde (DATASUS), criado em 1991, é o responsável por organizar, validar e publicar anualmente os Indicadores e Dados Básicos para a Saúde (IDB) em todas as regiões do país [Rede Integrada de Informações para a Saúde 2019]. Entre estes indicadores, se encontra o Sistema de Informações de Mortalidade (SIM), que obtém regularmente dados sobre a mortalidade brasileira a partir de declarações de óbitos.

O incentivo a coleta e armazenamento regular de dados verídicos, propicia uma melhor análise de indicadores pelos gestores de saúde. Entretanto, quanto maior a quantidade de dados, maior a dificuldade de aplicar técnicas tradicionais de análise de dados [Tan et al. 2009]. Neste contexto, a aplicação de técnicas computacionais focadas em descoberta de informações em grandes bases de dados, como a mineração de dados, poderiam auxiliar a responder questões pertinentes destes gestores. Estas técnicas já são usadas na área da saúde para prever probabilidades de doenças, auxiliar médicos em diagnóstico e tomadas de decisão clínicas [Jothi and Husain 2015].

Compreendendo a extensão, complexidade e importância do tema de mortalidade infantil, este trabalho tem como objetivo geral identificar padrões de óbitos infantis na cidade de Santa Maria/RS. Desta forma, são identificados como específicos:

- Obter os dados referentes a mortalidade infantil do Sistema DATASUS;
- Preparar a base de dados de mortalidade do DATASUS para que o conhecimento possa ser extraído;
- Definir técnicas computacionais adequadas com os dados utilizados;
- Utilizar a linguagem de programação Python e suas bibliotecas na aplicação dessas técnicas;
- Utilizar a ferramenta Weka para identificar regras de associação que possam auxiliar na identificação dos perfis.
- Avaliar entre os algoritmos utilizados qual expõe melhores resultados na base trabalhada.
- Analisar os resultados e identificar os perfis das crianças que faleceram com até um ano de idade.

2. Referencial Teórico

Neste tópico serão apontados conceitos e ferramentas empregados, de modo a auxiliar a compreensão das temáticas a serem abordadas no decorrer do trabalho.

2.1. Mortalidade Infantil

A Mortalidade Infantil consiste nos óbitos ocorridos no primeiro ano de vida de uma criança [Dias et al. 2017]. Sua taxa é calculada a partir da divisão entre o número de óbitos de menores de um ano dividido pelo número total de nascidos vivos, multiplicado por 1000.

$$TMI = \frac{Noi}{Nnv} * 1000 \tag{1}$$

O resultado da Equação 1 gera estimativas de mortes de crianças a cada mil nascidos vivos. Por meio deste, são delimitados valores para classificar o nível de mortalidade de uma população estudada em determinada região. De acordo com [Rede Integrada de Informações para a Saúde 2000], as taxas de mortalidade são classificadas entre alta, média e baixa: alta acima de 50 por mil; média, entre 20 a 40 por mil; e baixa com valores de 20 por mil nascidos vivos ou menos.

Segundo [Ministério da Saúde 2000], com base na TMI, é possível realizar uma variedade de interpretações em relação a população analisada como: condições de desenvolvimento socioeconômico, infraestrutura ambiental, acesso a qualidade dos recursos disponíveis para atenção à saúde materna e infantil, variações populacionais, identificação de situações de desigualdade social, entre outros fatores. As interpretações que surgem decorrentes a MI se devem a uma criança menor de um ano carecer de maiores cuidados por possuir um sistema fisiológico em desenvolvimento, sistema imunológico fragilizado e ainda estar vulnerável a fatores ocorridos no período de gestação, parto e pós-parto [Martins 2018].

Internacionalmente, os debates em relação a MI giram em torno de metas e objetivos para melhorar as condições de vida da população. Planos como os Objetivos de Desenvolvimento do Milênio (ODM) e Objetivo de Desenvolvimento Sustentável (ODS) utilizam da análise da TMI como método de progredir, rapidamente, rumo à eliminação da extrema pobreza e da fome do planeta. Tais fatores afetam especialmente as populações mais pobres dos países menos desenvolvidos.

No Brasil, apesar do declínio contínuo da TMI nos últimos anos, como mostrado na Figura 1, ainda há uma grande variabilidade dos valores das taxas entre os estados brasileiros, principalmente entre crianças que vivem em favelas urbanas e crianças indígenas [Menezes et al. 2019].

Segundo [França et al. 2017], a variância da TMI entre os estados torna a avaliação de seu desempenho indispensável para definir as prioridades nas políticas públicas, planejar ações e serviços, monitorar a situação de saúde, bem como avaliar o desempenho do sistema de saúde em cada região. Neste sentido, o Ministério da Saúde tem incentivado metas relacionadas a qualidade de informações inseridas das bases de óbitos, visto que a MI é um instrumento efetivo na avaliação da situação de saúde e qualidade de vida de uma população [Ministério da Saúde 2000].

2.2. Departamento de Saúde Pública

Na intenção de disponibilizar ferramentas que auxiliem no desenvolvimento, pesquisa e incorporação de tecnologias de informática voltadas à saúde, mediante a implementação

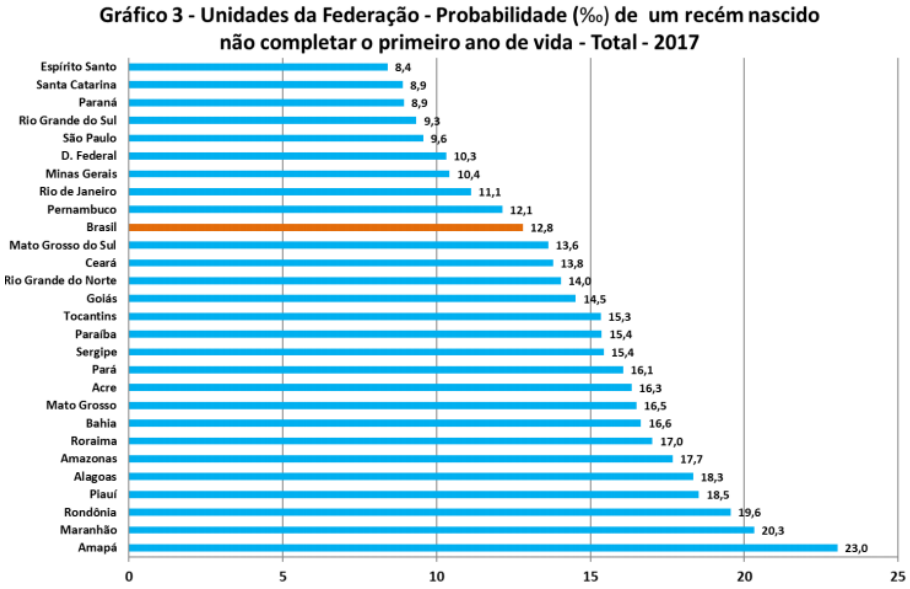

Figura 1. IBGE, Projeção da População do Brasil - 2017

de sistemas que auxiliem na disseminação de informações, em 1991, foi criado o Departamento de Saúde Pública (DATASUS) [Datasus C 2019]. Este setor é responsável por organizar, validar e publicar anualmente os Indicadores e Dados Básicos para a Saúde (IDB) em todas as regiões do país, tanto em versão web, quanto em folhetos impressos [Datasus C 2019].

No DATASUS, esta inserido o Sistema de Informação de Mortalidade (SIM), dados estes que foram utilizados neste trabalho. Com o objetivo de obter de maneira regular os dados sobre mortalidade do país e proporcionar produção de estatísticas de mortalidade foi criado o SIM, dispondo dados de certidão de óbitos completos, com informações referentes à saúde da mãe e da criança para todos os níveis do Sistema Único de Saúde (SUS)[Datasus B 2019]. As bases de dados dos anos 2000 até 2017 (último ano que foram lançadas as informações em domínio público) são de grande relevância para o presente estudo.

2.3. Mineração de Dados

A mineração de dados é um processo de descoberta de informações que auxilia a extração de dados relevantes em grandes bases de dados [Tan et al. 2009]. Se trata da aplicação de técnicas implementadas por meio de algoritmos computacionais, capazes de receber como entrada um conjunto de fatos ocorridos no mundo real, e de devolver, como saída, um padrão comportamental que pode ser descoberto c por meio de diversos métodos - regras de associação, funções de mapeamento ou modelagens de um perfil. [Silva 2017]

Para cumprir seu objetivo, o processo de mineração de dados utiliza de três tarefas: associação, agrupamento e predição[Silva 2017]. Como exibido na —Figura 2, tarefas de predição analisam um conjunto de dados a fim de associar corretamente a seu rótulo. Associações observam um conjunto de dados a procura de regras e tarefas de agrupamento verificam conjuntos de dados a fim de associar perfis similares, organizando-os em grupos.

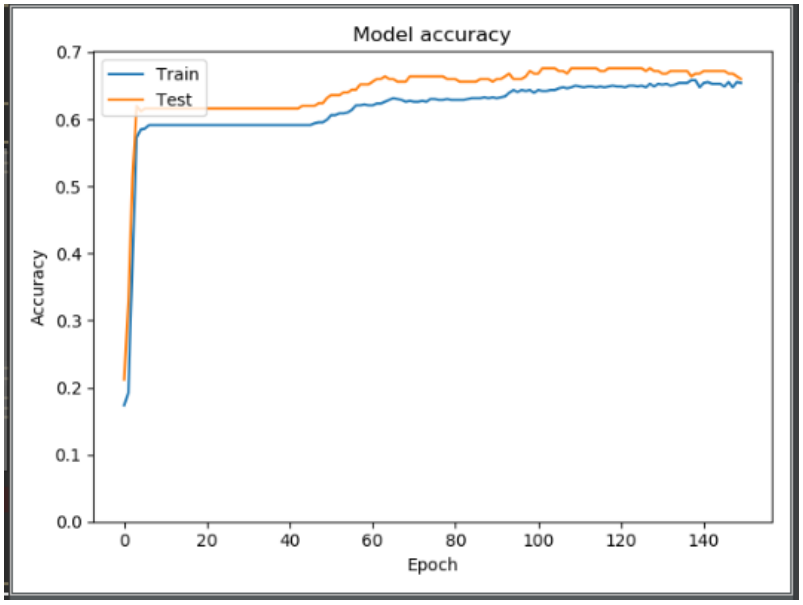

Figura 2. Acurácia

Devido a área da saúde gerar grandes quantidades de dados - registros médicos eletrônicos, relatórios administrativos, entre outros - utilizar tais técnicas computacionais contribuiriam principalmente a prever probabilidades de doenças, procura de diagnósticos e tomadas de decisões clínicas [Jothi and Husain 2015].

2.4. Weka

Weka (*(Waikato Environment for Knowledge Analysis*) é um software open source, desenvolvido na Universidade de Waikato, com o objetivo de disponibilizar técnicas de aprendizado de máquina e mineração de dados de maneira a resolver problemas práticos importantes da indústria da Nova Zelândia [Weka A 2019]. O software possui ferramentas para preparação, classificação, regressão, clusterização, mineração de regras de associação e visualização de dados aplicados a aprendizado de máquina para tarefas de mineração de dados. [Weka B 2019]

Dos algoritmos disponibilizados pela ferramenta, destacam-se a aplicação do J48 e RepTREE, algoritmos que melhor se adequam aos dados desta pesquisa, na sua montagem a partir de um conjunto de dados de treinamento, que tem como finalidade gerar árvores de decisão baseadas em um conjunto de dados de treinamento [Librelotto and Mozzaquatro 2014].E a regra *cross validation*, uma técnica baseada em amostragem de dados, permite que os dados estudados sejam aleatoriamente divididos em partições de tamanhos iguais para treinamentos de hipótese de indução [Monard and José 2003].

Ambos os algoritmos classificam os valores para atributos numéricos, dividindo os valores restantes em instâncias correspondentes em grupos [Weka C 2019]. A distinção destes está relacionada ao REPtree, que irá desenvolver uma árvore de decisão/regressão usando a variância da informação. O J48 implementa uma classe para gerar uma árvore de decisão podendo esta ser formada com podas ou não [Weka C 2019]. Com a aplicação desses algoritmos, a ferramenta utilizada fornece informações de precisão, bem como

meios gráficos para visualizar o desempenho do classificador e a possibilidade de visualizar as previsões de um modelo de classificação [Frank et al. 2004].

2.5. Redes Neurais Artificiais

Redes Neurais Artificiais são um subcampo específico de aprendizado de máquina que a partir de dados de entrada, são realizados sucessivos processos de aprendizados em camadas de representações cada vez mais significativas [Chollet 2018]. Um modelo de aprendizado de máquina transforma dados de entrada em saídas a partir de um processo de aprendizagem, quando expostos a exemplos conhecidos de entradas e saídas [Chollet 2018].

Por meio do reconhecimento de modelos de aprendizado de máquina, os dados assimilados inseridos geram representações que se aproximam do resultado esperado. A estrutura para este reconhecimento utiliza de um número de células simples, denominadas de "neurônios artificiais" recebendo as entradas e pesos respectivos [Nascimento 2017]. Em redes neurais artificiais, os dados são assimilados por meio de sequências de camadas que aprendem de acordo com os exemplos dados como entrada. Para definir o que cada camada assimila, os pesos são utilizados como parâmetros na intenção para que a rede mapeie corretamente entradas de exemplo para seus destinos associados.

Para controlar a saída de uma rede neural artificial é preciso medir o quanto essas previsões estão de acordo com os dados reais.Para isso, ambos os dados são submetidos a uma função de perda (*loss function*), que calcula o desempenho do algoritmo. Conforme o resultado da função de perda, é feito um ajuste dos pesos para que os resultados saiam com o mínimo de acurácia esperada. Esse ajuste é realizado por meio do auxílio de um otimizador. O modelo de Redes neurais é uma ferramenta poderosa que complementa maneiras de criar aprendizado dos dados, permitindo que os computadores produzam conhecimento de maneira inteligente, podendo ser implementados como ferramenta de apoio à área da saúde [Miotto 2017].

2.6. Python

Python é uma linguagem de programação de alto nível, interpretada, orientada a objetos e *open source* que disponibiliza um conjunto de pacotes e bibliotecas voltadas à ciência de dados. Dentre as bibliotecas e pacotes que foram utilizados neste trabalho estão:

- Pandas: Biblioteca *open source*, que fornece estruturas de dados de alto desempenho e ferramentas para leitura e manipulação de dados [Pandas 2019].
- Scikit-learn: Biblioteca de aprendizado de máquina de código aberto que oferece ferramentas para mineração de dados e análise de dados [Scikit-learn 2019].
- Keras: API de redes neurais de alto nível, desenvolvido com foco em permitir a experimentação rápida, sendo capaz de ir da ideia ao resultado com o menor atraso possível [Keras 2019].
- NumPy: Pacote voltado para computação cientifica com uma coleção de funções matemáticas para trabalhar com estrutura de dados [Numpy 2019].
- Matplotlib: Biblioteca de plotagem para a linguagem de programação Python que produz gráficos de qualidade de publicação [Matplotlib 2019].

3. Trabalhos Correlatos

No decorrer deste tópico, serão apresentados artigos que relatam aplicações de métodos de mineração de dados voltados na área da saúde, que estão servindo de base no desenvolvimento desta pesquisa.

3.1. Fatores que contribuem para a mortalidade infantil utilizando mineração de dados

Neste estudo, a autora Sartorelli(2017) utiliza dados das bases do SIM e SINASC de em um município de médio porte do Estado do Paraná com a finalidade de otimizar o processo decisório e planejamento de ações de gestores municipais em relação a causas de óbitos infantis. Foi utilizado o processo de extração de conhecimento (KDD - *knowledge Discovery in Databases*), e como primeira fase de pré-processamento, foram associadas as duas bases do SINASC e SIM em um único conjunto totalizando 266 registros com 15 variáveis constantes: idade da mãe, estado civil, escolaridade da mãe, número de nascidos vivos, número de nascidos mortos, duração da gestação, tipo de gravidez, tipo de parto, número de consultas de pré-natal, Apgar do primeiro minuto, Apgar do quinto minuto, peso ao nascer, presença de anomalia, dias de vida e evitabilidade além da obtenção de dois perfis a ser registrados referentes a óbito infantil por meio de frequência absoluta (FA) e frequência relativa (FR).

Por meio da ferramenta WEKA, foram aplicados os algoritmos J48 e NPP selecionando as variáveis: duração da gestação - menos de 22 semanas, de 22 a 27 semanas, de 28 a 31 semanas, de 32 a 36 semanas e de 37 a 41 semanas e evitabilidade, ocorrendo após um recorte na variável evitabilidade extraindo as características FA e FR.

Concluiu-se que os fatores: baixo peso ao nascer, idade da gestante e presença de anomalias aumentam o risco de mortalidade de crianças antes de completar seu primeiro ano. Além disso, enfatizou-se a necessidade de políticas públicas associadas a serviços de saúde pré-natal e atenção médica que se refere à nascidos prematuros e evitabilidade de partos prematuros.

3.2. Applying data mining techniques to improve diagnosis in neonatal jaundice

A icterícia é uma doença comum em recém-nascidos, embora seja benigna na maioria dos casos, um diagnóstico incorreto ou tardio pode colocar recém-nascidos em risco de desenvolver hiperbilirrubinemia grave e kernicterus.

Retirando dados do Departamento de Obstetrícia do Centro Hospitalar Tamega e Souza,foram analisados dados de recém-nascidos saudáveis com 35 ou mais semanas de gestação, a fim de fornecer tornar mais preciso o diagnóstico de icterícia neonatal utilizando a versão 3.6 da ferramenta WEKA para mineração de dados. No total 72 variáveis foram coletadas e analisadas, incluindo informações da família do recém nascido, gestacionais, parto, exame físico do recém-nascido, além de níveis de bilirrubina transcutânea medidas do nascimento até a alta hospitalar com intervalos de tempo máximo de 8 horas entre as medições.

Dos algoritmos de classificação aplicados neste estudo, os com melhores resultados selecionados foram: J48 (implementando algoritmo C4.5), simple CART, naïve-Bayes, multilayer perceptron, SMO e simple logistic. Os subgrupos foram divididos em três situações: fatores de risco utilizando dados logo após o nascimento, níveis de bilirrubina transcutânea sem fatores de risco obtidos antes 24 horas do nascimento e combinação de níveis de bilirrubina transcutânea com fatores de risco incluídos.

Constatou-se, neste trabalho,que os modelos de árvore de decisão J48 ou Simple Cart obtiveram a vantagem de ser mais aceitos pela comunidade médica.É recomendado

que sejam exploradas novas tecnologias para a aplicação de mineração de dados na tomada de decisões médicas.

4. Metodologia

Este trabalho de viés epidemiológico reúne características quantitativas e qualitativas, em razão de ter como base números e classificações e também fornecer uma interpretação na qual pretende verificar a relação da realidade com o objeto de estudo [Boente and Braga 2004]. Para a sua realização, foi utilizada uma amostra de 1560 crianças mortas antes de completar um ano de idade. O processo de realização foi dividido em três etapas: Extração de Dados, Pré-Processamento de Dados e Aplicação dos Algoritmos de Mineração e Redes Neurais.

4.1. Extração de Dados

As bases utilizadas neste trabalho foram extraídas do portal de serviços do DATASUS, referentes dos arquivos disseminâveis para tabulação do SIM. A partir da página de arquivos, foram extraídos os dados de declarações de óbitos infantis no período de 2000 até o ano mais recente publicado em domínio público (2017), do estado do Rio Grande do Sul. Estes dados foram convertidos para um arquivo de formato csv (*comma separated value* e para dar inicio a segunda etapa, foram filtrados apenas os casos da cidade de Santa Maria.

4.2. Pré-Processamento de Dados

Segundo Tan et al.(2009) a etapa referente ao pré-processamento de dados inclui procedimentos de: fusão de dados de múltiplas fontes, limpeza de dados para remoção de ruídos, observações duplicadas, seleção de registros e de características relevantes para o estudo [Tan et al. 2009].

A base contendo o conjunto de dados de mortalidade possuía inicialmente 40 atributos referentes a base de 2000. Ao longo dos anos, a cada atualização provinda do SIM, novos foram acrescentados e modificados até o ano de 2017, último ano a ser publicado, contendo 92 atributos na base. Dentre os incluídos nas bases ao longo dos anos, podemos citar: Escolaridade da Mãe a partir de 2010, CRM[1] do médico que atendeu este paciente no estabelecimento, atestante do óbito e necropsia[2] (caso tenha sido realizada ou não).

Por meio da biblioteca *pandas* as bases de todos os anos de MI em Santa Maria foram unidas, e por meio desta, foi realizada a primeira análise para normalização dos dados. Os atributos que possuíam relevância segundo [Sartorelli 2017] e [Ferreira et al. 2012] foram selecionados como fixos na base a ser analisada correspondendo aos atributos: idade da mãe, escolaridade materna, filhos nascidos vivos e mortos, tipo de parto, tipo de gravidez, peso ao nascer, idade gestacional e tempo de vida do recém nascido (idade).

Após a seleção dos atributos de acordo com a literatura, utilizando a ferramenta WEKA, ocorreu a preparação dos dados para detectar outros atributos relevantes e a normalização destes. Com base em [Ferreira et al. 2012], foram estipuladas as seguintes regras para o pré-processamento de dados:

[1]Número que um profissional da saúde adquire após realizar a inscrição no Conselho Regional de Medicina.

[2]Procedimento médico que examina um cadáver para determinar a causa e modo de morte.

- Eliminação: Atributos nos quais possuíam mais variáveis ignoradas ou incompletas (sem registros inseridos por enfermeiras ou médicos) do que completas e identificadores foram removidos da base.
- Integração: Atributos nas quais davam informações repetidas foram unidas nas bases ou escolhida a que possuía mais campos completos. Códigos que representam dados ignorados e os incompletos foram unidos a fim de facilitar a análise pelo algoritmo. Para a melhor análise, foram criados grupos de dados para atributos que possuíam uma grande variedade de dados, como idades, datas e horários.
- Transformação: A fim de obter uma análise mais precisa dos dados, em específico na redes neurais, os dados foram convertidos para tipo numérico, divididos (treino classe e previsores) e submetidos para uma normalização utilizando cálculos de desvio padrão utilizando a a biblioteca *sklearn*, como pode ser observado na Figura 4;

Seguindo as regras, dos 92 atributos iniciais, foi criada a primeira base para análise com 28 atributos relativos a: Tipo Óbito, Data Óbito, Hora Óbito, Município Residencia, Data Nascimento, Idade, Sexo, Raça, Local Ocorrência, Estabelecimento, Idade da Mãe, Escolaridade da Mãe, Escolaridade da Mãe em 2010, Ocupação Mãe, Quantidade de filhos mortos e vivos, Tipo de Gravidez, Semana de Gestação, Gestação, Tipo de Parto, Óbitos durante a gravidez, Causa Base das mortes, Acidente de trabalho, Escolaridade 2010 agregada e Morte no Parto.

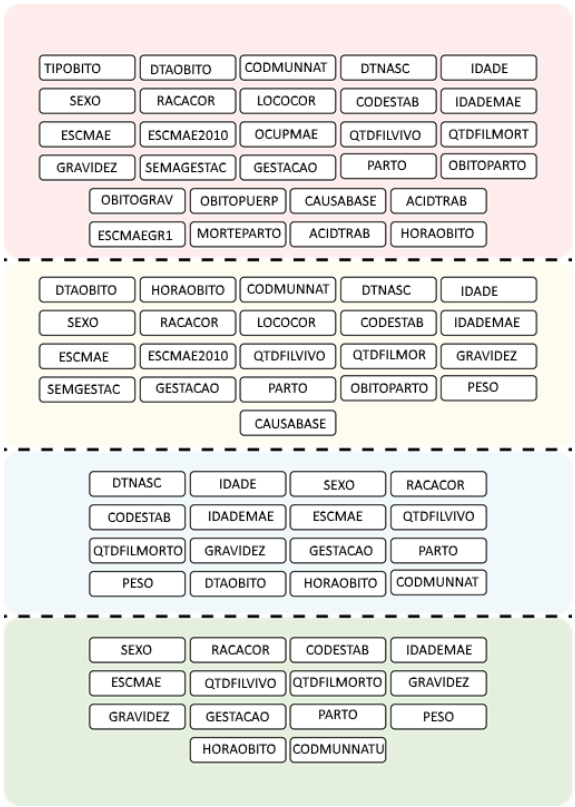

Figura 3. Normalização a partir das regras

Avaliando a base de dados inicial no WEKA, foi-se aplicando as regras citadas analisando a variabilidade dos atributos na ferramenta, gerando as versões mostradas na

Figura 3, até produzir a versão final contendo os atributos: Hora Óbito, Município de natureza, idade, sexo, raça, idade mãe, quantidade de filhos vivos e mortos, Gestação, Parto, Peso e Causas Bases.

4.3. Mineração de Dados

Na realização da mineração de dados, para evidenciar os atributos mais relevantes na base de dados para o propósito desta pesquisa, foram aplicadas as árvores de decisão J48 e REPtree na ferramenta WEKA. Com foco no atributo ”causa base de doenças”, juntamente com a regra *cross validation*.

4.4. Redes Neurais

Utilizando a biblioteca Keras, foi instanciado um modelo sequencial que permite inserir diferentes camadas de uma estrutura de rede neural. Foram implementadas três camadas, a primeira camada, possuiu como parâmetros 14 neurônios de entrada, definidos a partir da média aritmética entre o número de classes e os previsores. O tipo de camada*Dense* permitiu calcular a função de ativação junto com os dados de entrada e pesos. A função *Rectified Linear Unit*, foi escolhida na primeira e segunda camadas com 14 neurônios como parâmetros definidos por testes.

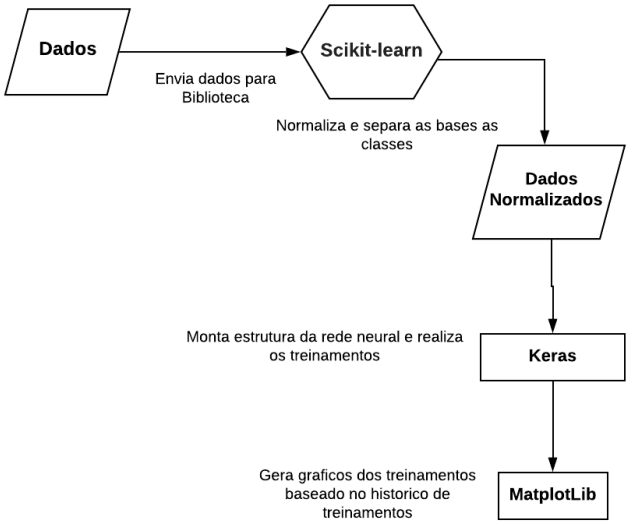

Figura 4. Representação da metodologia utilizada

A última camada consta, novamente 14 neurônios como parâmetros e a função de ativação *softmax*, por auxiliar gerar as probabilidades para definir bases com mais de uma classe de entrada e saída. Montada a estrutura das camadas, o método *compile* foi empregado no intuído de otimizar o processo de aprendizagem de máquina, possuindo como parâmetros respectivamente: funções para cálculo dos pesos, cálculos de validação e métricas escolhidas. Com o método *fit* definiu-se por meio de testes o número de iterações (*epochs*) igual a 150 e número de dados de treinamento lidos durante a interação (*batchs*) de 32, para respectivamente serem inseridas as variáveis dos valores a serem treinados.

Com o método *evaluate*, foi verificada a veracidade do resultado do treino junto com a base treinada e predições, para assim, analisar os resultados da aplicação com auxílio da biblioteca *matplot*(figura 3).

5. Resultados

De acordo com o gráfico gerado na figura 5, a partir do treinamento, o algoritmo conseguiu aprender e compreender 70% das causas das doenças relacionadas com as classes testes restantes.

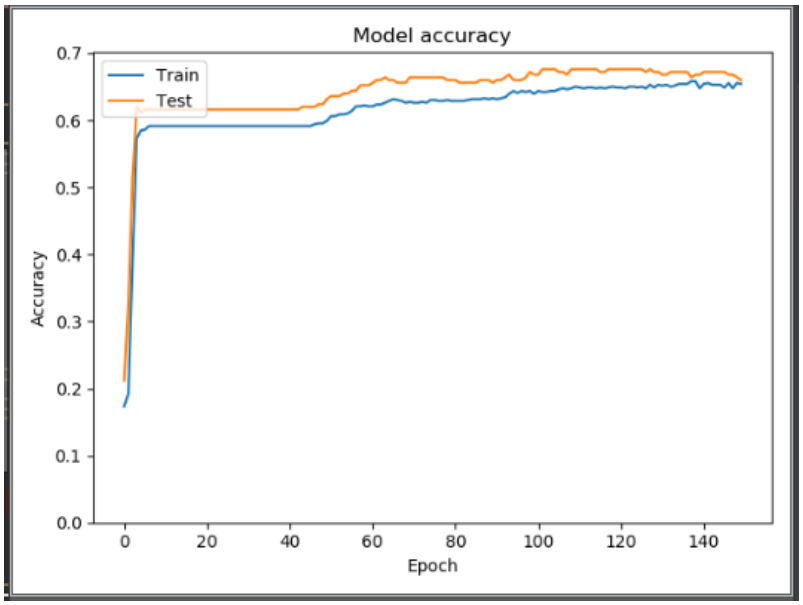

Figura 5. Gráfico acurácia por número de iterações

Aplicando os algoritmos de árvore de decisão J48 e REPtree obteram-se uma acurácia de 60% a 61%. Das 89 regras geradas pelo algoritmo J48 em relação a causa base das doenças, se destacaram:

- Se o tempo de vida do nascido for de menos de um dia, o falecimento da criança foi decorrente de complicações no período perinatal [3];
- Se o tempo de vida do nascido for de menos de uma hora e a criança possuir menos de um quilo ao nascer, sua morte foi decorrente de complicações na gravidez, parto e puerpério[4];
- Caso o tempo de vida do nascido tenha sido de até 28 dias e a gestação da mãe tenha durado menos de 28 semanas, o falecimento da criança foi decorrente de complicações no período perinatal;
- Partos do tipo vaginal possuíram um maior risco de mortes devido a doenças infecciosas e parasitárias;
- Partos do tipo cesáreo possuíram um maior risco de mortes devido a doenças no período perinatal;
- Filhos de mães de Santa Maria com menos de 5 a 12 anos de estudo, faleceram na base devido a causas externas;
- Filhos de mães com idade acima de 35 anos faleceram devido a complicações na gravidez, parto e puerpério;

[3]Período da gravidez humana que decorre entre as 22 semanas completas e os 7 dias completos após o nascimento

[4]Período que decorre desde o parto até que o corpo da mulher voltem às condições anteriores à gestação.

- Nascidos no estabelecimento DES4 (verificar anexo para mais detalhes),de raça amarela, parda e indígena, tiveram maior incidência a doenças respiratórias ou pulmonares;
- Nascidos no estabelecimento DES4,de raça preta ou branca nos quais as mães possuíam mais de dois filhos, foram afetados por doenças no sistema nervoso;
- Crianças que faleceram com até 6 meses de vida, cujas mães possuíam mais de 25 anos, faleceram devido a problemas respiratórios ou pulmonares;
- Assim como o caso anterior, mas para mães com menos de 25 anos, houve casos não definidos;
- No local HUSM, em casos de crianças que faleceram com até 6 meses de vida foram formadas 3 regras: 1. Caso a mãe possuía gravidez do tipo dupla, tripla ou mais, foram detectadas mortes por problemas respiratórios ou pulmonares; 2. Relacionados a gestações de mais de 32 semanas e tempo de escolaridade menor ou igual a 4 anos de idade para mães fora de Santa Maria, foram detectados casos de morte por doenças no sistema nervoso; 3. Caso as mães forem de Santa Maria e o tempo de gestação for de mais de 42 semanas, o algoritmo vinculou essas causas a doenças infecciosas e parasitárias ou a problemas na gravidez, parto e puerpério;

Das regras geradas pelo algoritmo REPtree podemos destacar:

- Se o tempo de vida do nascido for de até 28 dias e o tempo de gestação da mãe for de menos de 22 semanas, o algoritmo associa a morte do tipo perinatal;
- Crianças que nasceram com menos de 1 quilo tendem a falecer em menos de uma hora;
- Crianças que faleceram com até 6 meses de vida são ligadas aos estabelecimentos Caridade e HUSM, com mortes ocorridas por complicações na gravidez, parto e puerpério e perinatal.
- Ainda no HUSM em mortes de até 6 meses de vida, caso o peso da criança tenha sido maior de 2kg e a semana de gestação maior que 37 semanas, crianças Pardas e Indígenas tenderam a falecer por doenças respiratórias ou pulmonares e com complicações na gravidez, parto e puerpério.
- Em relatórios do HUSM para crianças falecidas em até 11 meses foram relatados 2 regras:
  - Relação de mães com menos de 5 anos de estudo e com mais de um filho.Nesses casos as tendencias são de mortes por neoplasias[5] e problemas circulatórios, menos veias e linfáticos
  - Mães com menos de um filho tiveram mortes com complicações na gravidez, parto e puerpério;
- No local HUSM, para crianças falecidas com até 6 meses foram encontradas relações com doenças respiratórias ou pulmonares quando a gestação da mãe foi maior de 28 semanas e para partos do tipo cesáreo;
- No mesmo período ainda, foram vinculados tipo de parto vaginal com doenças infecciosas e parasitárias;
- No local DES4, mortes por doenças nutricionais e metabólicas foram relacionadas a mães que residem fora da cidade de Santa Maria com anos de estudo menor que 4 anos;
- Filhos de mães que possuíam mais de um filho morto e com mais de 40 anos foram vinculados com doenças do tipo infecciosas e parasitárias;

[5]Massa anormal de tecido, cujo crescimento é excessivo, descontrolado e persistente

6. Considerações finais

Este trabalho teve como objetivo entender padrões de mortalidades ocorridas nos anos de 2000 até 2017 na cidade de Santa Maria. Os dados utilizados foram extraídos do Sistemas de Informação de Mortalidade no DATASUS referentes a declarações de óbitos infantis. As bases de MI de cada ano possuíam características diferentes e atributos inclusos, o que demonstra a preocupação do Ministério da Saúde na coleta de informações mais precisas em relação à investigação das origens dos óbitos.

Apesar da renovação das bases nos anos decorrentes, falta por parte do Ministério da Saúde um incentivo para que enfermeiros e médicos entendam a importância da qualidade das informações inseridas nos seus atendimentos. Informações como códigos de estabelecimentos, código de ocupação da mãe, presença de anomalias, acidente de trabalho e morte no parto, encontravam-se desatualizados nas bases (não possuindo mais registros nos códigos brasileiros de ocupação ou de estabelecimentos de saúde) ou incompletas.

Utilizando como apoio bibliotecas da linguagem Python, foi possível normalizar, criar uma estrutura de redes neurais, treinar as bases e predizer com uma acurácia aceitável as causas dos óbitos infantis de Santa Maria. No entanto, o algoritmo não proporcionou os detalhes desejados na mesma maneira que a aplicação de árvores utilizadas na ferramenta WEKA. Recomenda-se o algoritmo de Redes Neurais para a aplicação de predições.

A facilidade de prever as causas da morte se deve pela normalização do atributo Causa Base de Morte pelos códigos do CID - Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde. Entretanto, as classificações gerais do CID se tornam gerais principalmente nas complicações em períodos perinatais, onde são incluídas mortes referentes a hemorragias, septicemias[6] e asfixias. Na geração de regras de causas de doenças mais precisas, será necessário o acompanhamento de um profissional na saúde para a caracterização correta deste atributo.

As regras de associação geradas pelas árvores J48 e REPtree foram relativas principalmente aos atributos: idade (tempo de vida de criança), peso e estabelecimento. As causas de mortes que relacionaram com estes atributos com maior frequência foram: mortes relacionadas a complicações na gravidez, parto, puerpério e no período perinatal. Tais regras, foram de acordo com o padrão da literatura como Sartorelli(2017) e Ferreira(2012), o que reforça a ideia de que algoritmos computacionais podem ser utilizados para análise de dados da saúde, podendo auxiliar na tomada de decisões relacionadas a de políticas públicas.

Referências

Boente, A. and Braga, G. (2004). Metodologia científica contemporânea para universitários e pesquisadores. *Rio de Janeiro: Brasport*, pages 79–98.

Chollet, F. (2018). *Deep Learning with Python*. Manning Publications Co., 1th edition.

Datasus B (2019). Site Oficial:`http://www2.datasus.gov.br/DATASUS/index.php?area=060701`,Acesso Setembro 2019.

Datasus C (2019). Site Oficial:`http://datasus.saude.gov.br/datasus`,Acesso Setembro 2019.

[6]Infecção no corpo, seja por bactérias, fungos ou vírus

Dias, B. A. S., dos Santos, E. T., and Andrade, M. A. C. (2017). Classificações de evitabilidade dos óbitos infantis: diferentes métodos, diferentes repercussões? *Caderno de Saúde Pública*, 33(5):1–15.

Ferreira, D., Oliveira, A., and Freitas, A. (2012). Applying data mining techniques to improve diagnosis in neonatal jaundice. *Medical Informatics and Decision Making*, 12(1):143.

Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I. (2004). Data mining in bioinformatics using weka. *Bioinformatics*, 20(15):2479–2481.

França, E., Lansky, S., Rego, M. A., Malta, D., França, J., Teixeira, R., Porto, D., de Almeida, M., de Souza, M. d. F., Szwarcwald, C., Mooney, M., Naghavi, M., and Vasconcelos, A. M. (2017). Principais causas da mortalidade na infância no brasil, em 1990 e 2015: Estimativas do estudo de carga global de doença. *Revista Brasileira de Epidemiologia*, 20(15):46–60.

Jothi, N. and Husain, W. (2015). Data mining in healthcare–a review. *Procedia Computer Science*, 72(7):306–313.

Keras (2019). Site Oficial:`https://keras.io/`,Acesso Outubro 2019.

Librelotto, S. R. and Mozzaquatro, P. M. (2014). Análise dos algoritmos de mineração j48 e apriori aplicados na detecção de indicadores da qualidade de vida e saúde. *Revista Interdisciplinar de Ensino, Pesquisa e Extensão*, 1(1).

Martins, P. C. R. e. a. (2018). Convergência entre as taxas de mortalidade infantil e os Índices de desenvolvimento humano no brasil no período de 2000 a 2010. *Interações*, 19:291–313.

Matplotlib (2019). Site Oficial:`https://matplotlib.org/devel/index.html`,Acesso Outubro 2019.

Menezes, M. B. A., Barros, C. F., Horta, L. B., Matijasevich, A., Bertodi, D. A., Oliveira, D. P., and Victora, G. C. (2019). Stillbirth, newborn and infant mortality: trends and inequalities in four population-based birth cohorts in pelotas, brazil, 1982–2015. *International Journal of Epidemiology*, 48(9):54–62.

Ministério da Saúde (2000). `http://bvsms.saude.gov.br/bvs/publicacoes/manual_obito_infantil_fetal_2ed.pdf`, Acessado em: Outubro 2019.

Miotto, R. e. a. (2017). Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19:1236–1246.

Monard, M. C. and José, A. B. (2003). Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, 1(1):32.

Nascimento, E. O. e. a. (2017). Redes neurais artificiais aplicadas na correlação entre óbitos de dengue, automedicação e fatores abióticos em goiânia-goiás. *Scientia Plena*, 13(3).

Numpy (2019). Site Oficial:`https://numpy.org/`,Acesso Outubro 2019.

Paiz, J. C., Bigolin, M., dos Santos, R. R., and Bordina, R. (2018). Mortalidade infantil e serviços de atenção primária à saúde em porto alegre (rs). *Revista Brasileira de Medicina de Família e Comunidade*, 13(40):1–13.

Pandas (2019). Site Oficial:https://pandas.pydata.org/,Acesso Outubro 2019.

Rede Integrada de Informações para a Saúde (2000). Site Oficial:http://www.ripsa.org.br/fichasIDB/pdf/ficha_C.1.pdf,Acessado em: Julho 2019.

Rede Integrada de Informações para a Saúde (2019). Site Oficial:http://www.ripsa.org.br/vhl/rede-de-instituicoes/ms/departamento-de-informatica-do-sus,Acessado em: Outubro 2019.

Sartorelli, A. P. e. a. (2017). Fatores que contribuem para a mortalidade infantil utilizando a mineração de dados. *Saúde e Pesquisa*, 10(1):33–41.

Scikit-learn (2019). Site Oficial:https://scikit-learn.org/stable//,Acesso Outubro 2019.

Silva, L. A. e. a. (2017). *Introdução à mineração de dados: com aplicações em R*. Elsevier Brasil, 1th edition.

Tan, P., M., S., and V., K. (2009). *Introdução ao Data Mining*. Editora Ciência Moderna, 1th edition.

United Nations Children's Fund (2017). Site Oficial: https://data.unicef.org/resources/state-worlds-children-2017-statistical-tables/,Acessado em: Outubro 2019.

Weka A (2019). Site Oficial:https://www.cs.waikato.ac.nz/ml/weka/index.html,Acessado em: Outubro 2019.

Weka B (2019). Site Oficial:https://www.cs.waikato.ac.nz/ml/index.html,Acessado em: Outubro 2019.

Weka C (2019). Site Oficial:http://weka.sourceforge.net/doc.stable/allclasses-noframe.html,Acessado em: Outubro 2019.

World Health Organization (2019). Site Oficial:https://www.who.int/gho/child_health/mortality/neonatal_infant_text/en,Acessado em: Outubro 2019.

Tabelas dos Atributos Utilizados

IDADE	DESCRIÇÃO	COD
300 - 305	Até 5 meses de vida	5
306 - 311	Até 11 meses de vida	11
200 - 229	Até 30 dias de vida	30
100 - 123	Até 23 horas de Vida	23
0 - 60	Minutos de vida	60
99/9/0	Ignorado	I

Figura 6. Tabela Idade

PESO	DESCRIÇÃO	COD
de 300 a 999	- de 1kg	0
1000 - 1999	Faixa dos 1kg	1
2000 - 2999	Faixa dos 2kg	2
3000 - 3999	Faixa dos 3kg	3
4000 - 4999	Faixa dos 4kg	4

Figura 7. Tabela Peso

IDADE MÃE	DESCRIÇÃO	COD
15	0 - 15 anos	15
25	Até 25 anos	25
35	Até 35 anos	35
48	Até 48 anos	48
9	Ignorado	I

Figura 8. Tabela Idade Mãe

HORARIO	COD
NOITE	1
MANHA	2
TARDE	3
MADRUGADA	4
IGNORADO	99

Figura 9. Tabela Horário

CODESTAB	DESCRIÇÃO	COD
5066	Deslocado ou Fechado	4
18598	Deslocado ou Fechado	3
18614	Deslocado ou Fechado	2
23366	Deslocado ou Fechado	1
2244268	HGUSM - Hospital Geral de Santa Maria	RETIRADO
2243474	PAFMS - PRONTO ATENDIMENTO MEDICO MUNICIPAL FLAVIO MIGUEL SCHNEIDER	5
2244276	CARIDADE- HOSPITAL DE CARIDADE ASTROGILDO DE AZEVEDO	6
2244284	CSAUDE - HOSPITAL MUNICIPAL CASA DE SAUDE	7
2244292	HOSPITAL SAO FRANCISCO DE ASSIS	RETIRADO
2244306	HUSM - HOSPITAL UNIVERSITARIO DE SANTA MARIA	8
6359167	CAUZZO MEIRELLES POLICLINICA	RETIRADO
7015887	UPA 24 H SANTA MARIA RS	RETIRADO

Figura 10. Tabela Estabelecimentos

CID	CLASSIFICAÇÃO CID	CLASSIFICAÇÃO
A09-B377	Doenças infecciosas e parasitárias	1
C400-D434	Neoplasias	2
D065 -D762	Sangue, órgãos hematopoiéticos e transtornos imunitários	3
E046 - E889	Doenças nutricionais e metabólicas	4
G001 - G938	Sistema nervoso	5
I038 - I629	Circulatório, menos veias e linfáticos	6
J09 - J988	Infecções respiratórias agudas e Doenças Respiratórias	7
K550 - K631	Doenças dos intestinos e peritôneo	8
K729 - K758	Doenças do fígado	9
M009	Osteomuscular e tecido conjuntivo	RETIRADO
N171 - N390	Doenças urinárias	10
P000 - P696	Afecções originadas no período perinatal	11
Q08 - Q999	Gravidez, parto e puerpério	12
R092 - R099	Sintomas, sinais e afecções mal definidas	99
R092 - R099	Lesões, envenenamentos e causas externas	13

Figura 11. Tabela Causas Bases

TIPO PARTO	COD
CESARIANA	2
VAGINAL	1
IGNORADO/VAZIO	99

Figura 12. Tabela Tipo Parto

GRAVIDEZ (TIPO GRAVIDEZ)	COD
UNICA	1
DUPLA	2
TRIPLA OU MAIS	3
IGNORADO/VAZIO	99

Figura 13. Tabela Tipo Gravidez

RACACOR	COD
BRANCA	1
PRETA	2
AMARELA	3
PARDA	4
INDIGENA	5
IGNORADO	9

Figura 14. Tabela Raça/Cor