O Uso de Ciência de Dados Para Identificar o Impacto da Sífilis na Gravidez na Morte de Crianças de Até 1 Ano de Idade

Arthur Volpe[1], **Giovanna M. Piccolo**[1], **Rafael F. Facco**[1], **Prof. Dr. Ismar F. Silveira**[1]

[1]Faculdade de Computação e Informática - Universidade Presbiteriana Mackenzie
Rua da Consolação, 930 Consolação, São Paulo, SP, 01302-907

arthur.volpe@hotmail.com, giovanna.mpiccolo@gmail.com,
rafaelferrari99@gmail.com, ismarfrango@gmail.com

Abstract. *This research aims to analyze the impact of the syphilis disease on deaths of children up to 1 year of age after the delivery period, taking into account that the pregnant woman was infected. Therefore, the Datasus database will be used to analyze the history of the pregnant woman's diagnosis, which may be early or late and relate to the infant mortality rate. To achieve this goal, classification algorithms will be used to predict the cases of congenital syphilis that resulted in the death of children up to 1 year of age.*

Resumo. *Esta pesquisa tem como intuito analisar o impacto da doença sífilis em mortes de crianças com até 1 ano de idade após o período de parto, tendo em consideração que a gestante estava infectada. Para isso, a base do Datasus foi utilizada para analisar o histórico de diagnóstico da gestante, podendo esse ser precoce ou tardio e relacionar com a taxa de mortalidade infantil. Para atingir tal objetivo, foram utilizados algoritmos de classificação para predizer os casos de sífilis congênita que resultaram na morte de crianças até 1 ano de idade.*

1. Introdução

No mundo, cerca de 2 milhões de gestantes são infectadas pela sífilis a cada ano. A maioria das gestantes não realiza o teste para sífilis, e as que o fazem não são tratadas adequadamente ou sequer recebem tratamento. Aproximadamente 50% das gestantes não tratadas ou inadequadamente tratadas podem transmitir a doença ao concepto, levando a resultados adversos como morte fetal, morte neonatal, prematuridade, baixo peso ao nascer ou infecção congênita [WHO 2011].

1.1. Contextualização e Relevância do Tema

Nos dias atuais, a taxa de mortalidade infantil vem sendo elevada por inúmeros fatores, dentre eles causas propositais, acidentes ou por doenças. Em meio desta situação, o Hospital Santa Casa planeja um estudo com o tema "Doenças na gestação que causam a morte de crianças já nascidas com até 1 ano de idade". Desta forma, no contexto que a população vive e com inúmeras doenças, a sífilis ainda afeta um número elevado de gestantes.

Estima-se que, em 2008, cerca de 1,36 milhão (IC95%: 1,16-1,56) de gestantes apresentavam sífilis ativa, com mais de meio milhão de desfechos negativos, representados por perdas fetais com 22 ou mais semanas gestacionais, óbitos neonatais, recém-natos prematuros ou com baixo peso ao nascer e recém-natos infectados [Domingues and Leal 2016].

Segundo o Ministério da Saúde, a sífilis é uma doença infecciosa sistêmica causada pela espiroqueta Treponema pallidum, de evolução crônica e muitas vezes assintomática, que tem como principais formas de transmissão as vias sexual e vertical.

Sabe-se que a sífilis no período de gestação vem sendo um problema desde o século passado, mas que vem se agravando desde o início do século. Segundo o Ministério da Saúde, 56,5% das gestantes com sífilis receberam tratamento inadequado, 27,3% não receberam tratamento, 12,1% dos casos foram ignorados e apenas 4,1% receberam a terapêutica adequada, através desses motivos a taxa só vem aumentando.

Nos dias de hoje, falar sobre doenças sexualmente transmissíveis (DST) ainda é um tabu, e segundo Domingues, isso acaba ocasionando na falta de conhecimento e despreparo da sociedade em relação aos protocolos nacionais da sífilis e, por fim, sua contaminação.

Em 2007, foi criado um projeto pela OMS (Organização Mundial da Saúde) com o objetivo de eliminar a transmissão de Sífilis, porém esse projeto não deu muito certo já que houve falta de proatividade dos pacientes em irem atrás dos exames e das respectivas soluções. Dessa forma, o estudo não investigou áreas geográficas importantes e consequentemente não foi efetivo.

Com isso, grande parte das mulheres grávidas não realizam o teste para sífilis, e as que realizam e recebem o diagnóstico não são tratadas adequadamente ou sequer recebem tratamento, o que vem agravando os casos de morte das crianças com até 1 ano de idade.

1.2. Objeto de Pesquisa

1.2.1. Contextualização do Problema de Pesquisa

Em um estudo feito pela Secretaria de Estado de Saúde do Distrito Federal sobre a Sífilis, dentre 67 mulheres, 15 eram gestantes e entre elas, três tiveram como desfecho da gestação o abortamento, enquanto 52 eram puérperas, entre as quais duas tiveram natimorto. Quando questionados os antecedentes obstétricos, 54 (80,6%) gestantes/puérperas referiram mais de uma gestação e 46 (68,7%) referiram ter realizado pré-natal nas gestações anteriores [Magalhães et al. 2013].

Dessa forma, pode-se perceber que a doença é mais grave para o bebê quando a mulher está em período de gestação. Sendo assim, a pergunta que será respondida nesta pesquisa é: qual é o impacto da sífilis durante a gravidez para o bebê que morreu até 1 ano de idade considerando características da mãe como idade, escolaridade e estado de residência, caraterísticas da gravidez como tipo (simples, dupla, tripla, etc.), semanas de gestação, se houve assistência médica e o tipo do parto.

1.2.2. Hipótese

A sífilis materna constitui uma importante causa potencialmente evitável de óbito fetal e de outros resultados perinatais adversos ocorrendo principalmente nas regiões menos desenvolvidas do mundo.[Nascimento et al. 2012].

Quando a mãe tem acompanhamento regular com o médico no período da gravidez, possui maior grau de escolaridade e têm uma maior idade gestacional a chance da

criança morrer até 1 ano de idade é menor do que aquelas que não tiveram todos estes acompanhamentos e características.

1.3. Objetivos do Estudo

1.3.1. Objetivo Geral

O presente trabalho tem como objetivo final ou geral realizar uma análise do impacto da doença sífilis durante a gravidez em relação à morte de crianças de até 1 ano de idade, e através dessa análise, realizar um estudo usando algoritmos de *Machine Learning* para prever se a criança, cuja mãe tinha sífilis, vai morrer com menos de 1 ano de idade ou mais.

Este estudo possui como objetivo fazer uma análise a respeito da doença sífilis na gravidez, com o intuito de analisar a grandeza deste problema que vem se tornando cada vez mais comum, principalmente em jovens, onde muitas vezes eles não utilizam métodos contraceptivos para evitar esta DST, e acabam tendo graves consequências.

Esta análise foi realizada por meio de consultas feitas em diversas bases do Datasus que contém informações sobre o assunto apresentado, além de pesquisas executadas em diversos artigos científicos cujo intuito foi se aprofundar no assunto. Outro aspecto relevante a ser citado é referente a utilização de algoritmos já existentes com o objetivo de relacionar as informações a fim de chegar em conclusões que dizem respeito ao tema abordado neste projeto.

Aprofundando um pouco mais na parte prática desta pesquisa, tem-se como principal objetivo o uso de algoritmos de *Machine Learning* buscando fazer uma classificação de qual é o grupo de mortalidade que o bebê se encaixa (abaixo ou acima de 1 ano) baseado em seus históricos de dados que foram obtidos através do estudo de máquina.

1.3.2. Objetivos Específicos

Tem-se como objetivo específico deste projeto realizar uma análise de casos de mães que foram diagnosticadas com sífilis durante o período de gestação, onde tiveram ou não o acompanhamento e tratamento da doença considerando sua idade, escolaridade, estado de residência, tipo de gravidez, semanas de gestação predizendo por meio de algoritmos de classificação o impacto dessa doença na vida de seus bebês.

1.4. Justificativa

O presente trabalho se justifica dada a relevância social do impacto da sífilis durante a gravidez que acabam causando a morte de crianças de até 1 ano de idade. É de extrema importância o teste, o acompanhamento e o tratamento da sífilis em mulheres grávidas para evitar a transmissão para os bebês e, consequentemente, evitar sua morte pela doença. Além disso, a sífilis materna pode ser potencialmente evitável, o que consequentemente diminuiria a taxa de óbito fetal.

No mundo, cerca de 2 milhões de gestantes são infectadas pela sífilis a cada ano. A maioria das gestantes não realiza o teste para sífilis, e as que o fazem não são tratadas adequadamente ou sequer recebem tratamento. Aproximadamente 50% das gestantes não

tratadas ou inadequadamente tratadas podem transmitir a doença ao concepto, levando a resultados adversos como morte fetal, morte neonatal, prematuridade, baixo peso ao nascer ou infecção congênita [Nonato et al. 2015].

Desta forma, a análise feita por meios tecnológicos é necessária para acompanhar, compreender tentar eliminar a sífilis congênita que impacta diretamente três dos oito Objetivos do Milênio: ODM 4 - Reduzir a mortalidade infantil; ODM 5 - Melhorar a saúde materna e ODM 6 - Combater a HIV/AIDS, malária e outras enfermidades, segundo a Organização Pan-Americana de Saúde.

1.5. Delimitação do Estudo

O estudo tem como delimitação a análise de dados, a partir de 2011, do óbito de crianças causadas por sífilis congênita considerando características da mãe como idade, raça/cor, escolaridade, estado de residência, além de características como tipo de gravidez e semanas de gestação.

Os dados utilizados estão disponíveis e registrados no Sistema de Informações sobre Mortalidade (SIM) localizado na base de dados de saúde pública do governo, o Datasus.

A adoção da Mineração de Dados permitiu identificar fatores e evidenciar fragilidades, criando hipóteses que ajudem os gestores a entender melhor o problema e traçar estratégias para seu enfrentamento.[Nakamura et al. 2016].

1.6. Organização do Estudo

Este estudo tem como intuito ajudar o programa do Hospital Santa Casa com o tema morte de crianças com até 1 ano causadas por doenças que a mãe teve durante a gravidez. Neste trabalho, a doença a ser analisada será a sífilis.

O segundo capítulo apresenta a revisão bibliográfica dos artigos utilizados na pesquisa. No terceiro capítulo, está descrita a metodologia de pesquisa utilizada no projeto, detalhando as etapas em ordem cronológica. O cronograma das atividades descritas no terceiro capítulo está presente no quarto capítulo.

2. Referencial Teórico

Nonato, Melo e Guimarães (2015) justificam a sífilis congênita devido aos erros na assistência pré-natal e, desta forma, os autores quiseram acompanhar quais são os resultados finais quando se tem um diagnóstico e um acompanhamento ou quando se tem um diagnóstico sem qualquer tratamento durante a gestação de mulheres em Belo Horizonte pela rede pública de Saúde. Além disso, tiveram como objetivo mostrar dentre os casos de gestantes com Sífilis quais eram as condições sociais e comportamentais de cada uma. Por fim, o artigo discorre por esse tema apresentando dados e chegando à conclusão de que o atual cenário para acompanhamento das gestantes e os exames que são feitos precisam ser revistos e melhorados.

Domingues e Leal (2016) afirmam que o número de casos de sífilis na gestação vem aumentando a cada ano e associaram a elevada transmissão vertical da sífilis com a ausência do tratamento durante a gravidez, podendo ter sua taxa de transmissão próxima a 100% e consequentemente, apresentando uma alta taxa de mortalidade entre os bebês.

Sendo assim o objetivo deste artigo foi realizar um estudo com mulheres durante o período da gestação e após o parto para mostrar as possibilidades de diagnósticos em relação à Sífilis que a criança terá após seu nascimento.

A pesquisa feita por Magalhães, Kawaguchi, Dias e Calderon (2015) indica que, para garantir o controle da sífilis congênita, é necessário tomar medidas mais efetivas de prevenção do que apenas oferecer uma boa qualidade no pré-natal, sendo imprescindível que o parceiro da gestante também passe por um tratamento para garantir maior chances de cura para a parceira. Além disso, os autores relacionam a sífilis durante a gestação sendo mais frequente em mulheres jovens de baixa renda e baixa escolaridade, logo a assistência e exames precoces para detectar a sífilis devem ser de fácil acesso para elas.

Fernandes e Filho (2019) fizeram um estudo sobre o uso da mineração de dados e aprendizado de máquina em saúde e segurança do trabalho e afirmam que, na área da saúde, é cada vez mais frequente o uso de *Data Mining* e *Machine Learning* na detecção de doenças, predição de riscos, entre outros. Aprendizado de Máquina é uma área de IA cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática [Monard and Baranauskas 2003]. Um dos algoritmos utilizados para fins de agrupamento de perfis foi o *k-means*, algoritmo não supervisionado que utiliza medidas de distância para agrupar os pontos em cada centróide, identificando grupos parecidos. Porém, atualmente, a metodologia de *Machine Learning* mais utilizada em análise de dados é a árvore de decisão, sendo implementada principalmente em *R* ou *Python*, segundo Alexandre Dias Porto Chiavegatto Filho (2015).

Nakamura, Otero e Carvalho (2016) frisam a importância do tratamento do parceiro da mãe infectada utilizando o algoritmo de classificação *J48* disponível no software *WEKA*. No caso, elas utilizaram dados das bases SINAN, SINASC e SCNES para extrair informações úteis que contribuem na tomada de decisão para enfrentar a transmissão da sífilis mãe-filho. Segundo as autoras, esse algoritmo foi escolhido pois permite identificar quais são as variáveis que mais se relacionam com a variável de tratamento adequado da gestante e, como resultado, a variável que mais exerceu influência sobre o tratamento adequado da gestante é a realização do tratamento do parceiro.

Silveira e Moreira (2020) também fizeram uso de algoritmos de aprendizagem de máquina para predizer o número de casos de doenças transmitidas pelo mosquito *Aedes Aegyoti* a partir de características inerentes ao paciente como idade, sexo, sintomas, etc. Essa predição pode ser utilizada no momento das tomadas de decisões de contenções das doenças pelos especialistas. No caso, os algoritmos *J48*, *Random Forest* e Redes Neurais foram utilizados sendo que o algoritmo *Random Forest* apresentou maior acurácia.

Há 3 técnicas mais conhecidas para aplicação da análise de dados, Azevedo e Santos estabelecem uma comparação entre os processos *KDD*, *SEMMA* e *CRISP-DM*. Como resultado da pesquisa, os autores chegaram a conclusão que os processos *SEMMA* e *CRISP-DM* são uma implementação do processo *KDD*. Examinando-o minuciosamente, podemos afirmar que as cinco etapas do processo da *SEMMA* podem ser vistas como uma implementação prática das cinco etapas do processo *KDD* [Azevedo and Santos 2008]. Comparar os estágios *KDD* com os estágios *CRISP-DM* não é tão simples como no *SEMMA*. No entanto, podemos, em primeiro lugar, observar que a metodologia CRISP-

DM incorpora as etapas que, conforme referido acima, deve preceder e seguir o processo *KDD* [Azevedo and Santos 2008].

Utilizando o algoritmo de aprendizado de máquinas, Noemi, Amaral e Melo (2019), desenvolveram métodos que através do uso de dados obtidos por meio de exames que foram realizados no laboratório de instrumentação biomédica da UERJ, foi realizado 5 experimentos com duas possíveis saídas (não portador e portador), avaliando separadamente a classificação da fibrose cística por meio de técnicas que envolvem o uso do algoritmo *K-NN*, *Adaboost*, *Random Forest*, *K-pasta* e *Wrapper*, estas avaliações comprovaram que quando utilizamos dados oferecidos pelas técnicas de oscilações forçadas nos algoritmos de aprendizado de máquinas, os mesmos demonstram ser eficientes na identificação de portadores de fibrose cística.

Turlapati e Prusty enfrentaram um conjunto de dados desequilibrado referente ao COVID-19 onde apenas 9% das pessoas testaram positivo para o coronavírus. Os autores propuseram um método de sobreamostragem baseado na técnica *SMOTE* que cria novos dados sintéticos com base nos dados existentes. O *SMOTE* cria amostras sintéticas da classe minoritária calculando a distância euclidiana entre quaisquer dois k-vizinhos mais próximos escolhidos aleatoriamente e introduzindo novas amostras sintéticas ao longo da linha que une as duas amostras minoritárias [Turlapati and Prusty 2020].

Nascimento, Cunha, Guimarães, Alvarez, Oliveira e Bôas (2012) dizem sobre os óbitos fetais causados pela sífilis durante a gravidez, mostrando uma série de características que podem agravar a situação, tais como o estágio da infecção materna, idade gestacional, o país em que a gestante vive, etc. Além disso, demonstra as diversas possibilidades de resultados que podem ocorrer caso a infecção por sífilis recente não seja tratada, dentre elas, destacam-se o aborto tardio, óbito fetal, óbito neonatal, parto prematuro, e por fim, sífilis congênita. Por fim, é abordado também o grau de importância dos óbitos fetais gerados por conta da sífilis, visto que este é um tema que não está tendo seu devido reconhecimento, e por conta disso, o número de gestantes com sífilis vem crescendo cada vez mais, sendo necessária a realização de uma análise de saúde e das principais causas do óbito fetal, em diferentes regiões.

Koul, Becchio e Cavallo (2018) propuseram a incorporação de técnicas de validação cruzada em estudos de pesquisas, ou seja, dividir aleatóriamente um conjunto de dados em subconjunto de dados com o objetivo de evitar *overfitting* e consequentemente aumentar a confiabilidade do modelo. A validação cruzada evita esse risco avaliando o desempenho do modelo em um conjunto de dados independente (conjunto de teste) [Koul et al. 2018].

A Organização Mundial da Saúde (OMS) (2011) afirma que, para eliminar a sífilis congênita, os países devem ter uma política de monitoramento com indicadores e metas, inclusive naqueles cujas mulheres grávidas possuem baixo risco de infecção. As principais ações para erradicá-la são o rastreamento da sífilis através de atendimento pré-natal precoce, o tratamento dessas infectadas e seus parceiros e o tratamento dos bebês cuja sorologia da mãe tenha dado positivo. Além disso, algumas ações preventivas também podem ser tomadas, como o uso de preservativos e a disseminação de informações sobre o tema. A sífilis materna pode causar, entre outros problemas, mortes fetais e infantis. Sendo assim, a OMS aborda sua iniciativa para a Eliminação Global da Sífilis Congênita,

que afeta diretamente três Objetivos de Desenvolvimento do Milênio que dizem sobre a redução da mortalidade infantil, a melhora da saúde materna e o combate de doenças como HIV, AIDS, entre outras. Em vários trechos, a OMS frisa a importância da coleta, análise e disseminação desses dados para que países que tiverem êxito sirvam como modelos para outros.

3. Metodologia de Pesquisa

A metodologia de pesquisa aplicada neste trabalho foi baseada no processo *Knowledge Discovery in Databases* (*KDD*). O processo *KDD*, conforme apresentado em (Fayyad et al, 1996) é o processo de usar métodos de Data Mining para extrair o que é considerado conhecimento de acordo com a especificação de medidas e limites, usando um banco de dados ao longo com qualquer pré-processamento, subamostragem e transformação do banco de dados necessários [Azevedo and Santos 2008]. O *KDD* possui 5 etapas: Seleção, Processamento, Transformação, Mineração de Dados e Interpretação/Avaliação.

A etapa de seleção consiste em criar um conjunto de dados no qual será analisado. A etapa de Processamento diz respeito a validação da qualidade dos dados selecionados, ou seja, ocorre a verificação de dados inconsistentes podendo ser feitas limpezas, exclusões de registros, entre outras ações com o objetivo de ter dados mais consistentes. A etapa de Transformação consiste na transformação dos dados, ou seja, selecionamos apenas os campos que fazem sentido para a análise e que podem passar por agregações, padronizações, além da criação de novos campos. A etapa de *Data Mining* consiste na aplicação de técnicas de mineração de dados e criação de modelos preditivos alterando seus hiperparâmetros. Por fim, a etapa de Interpretação/Avaliação diz sobre a interpretação e avaliação dos resultados obtidos dos modelos, ou seja, nessa etapa pode-se avaliar e comparar o desempenho dos modelos testados anteriormente.

4. Desenvolvimento e Resultados

O presente trabalho utilizou a base de dados SIM (Sistema de Informações sobre Mortalidade) do DATASUS (Departamento de Informática do Sistema Único de Saúde do Brasil). Os dados são referentes às mortalidades da população brasileira desde 1996.

Para realizar as etapas computacionais do trabalho foi utilizada a plataforma *Google Colab* devido a sua praticidade e poder de processamento. A primeira etapa do trabalho foi realizar a extração da base SIM e SINASC (Sistema de Informações sobre Nascidos Vivos) através da biblioteca *Python PySUS* [Coelho et al. 2021] e armazenar os arquivos *CSVs* referentes as bases no *Google Drive*. Duas linguagens de programação têm conquistado o apoio crescente dos cientistas na última década: *R* e *Python*. A expectativa é que essas duas linguagens passem a ser dominantes também entre epidemiologistas. Ambas são *open source*, gratuitas e têm uma comunidade de programadores e cientistas extremamente ativa, o que significa que novas metodologias estatísticas são rapidamente incorporadas pelos usuários por meio de pacotes e bibliotecas [Filho 2015]. Inicialmente, o objetivo era cruzar informações das duas bases, porém a base SINASC não possui um identificador único de registro, os indivíduos não são identificáveis (não possui nome, nome da mãe, etc.) e não possui uma coluna referente a doença sífilis para os dados serem filtrados. Sendo assim, o uso dessa base foi descartado.

Após a extração e armazenamento dos dados, utilizamos o módulo *drive* da biblioteca *google.colab* para acessar os arquivos *CSVs* do *Google Drive* e espelhá-los no sistema de arquivos do *Google Colab*. Os dados passaram por tratamentos como a padronização do tipo *string* para o tipo *datetime* dos campos DTOBITO (data de óbito da criança) e DTNASC (data de nascimento da criança), do tipo *string* para o tipo *int* do campo ESTADO (estado do óbito da criança), transformações como a coluna IDADE que passou a medir a idade da criança em meses e não mais em anos, exclusão de registros nulos e a criação de uma variável binária onde o valor 1 corresponde à morte da criança com menos de 1 ano e 0 à morte da criança com mais de 1 ano. Além disso, muitas colunas não são relacionadas ao contexto da sífilis congênita, logo apenas os dados relevantes foram filtrados, ou seja, apenas os registros de mortes a partir de 2011 causadas por sífilis congênita foram filtrados. O aprendizado de máquina induzido do tipo supervisionado deve iniciar pela coleta do conjunto de dados, o que pode ocorrer com a indicação de que atributos são os mais significativos [Silveira et al. 2020]. Após o filtro, as seguintes colunas foram utilizadas:

Tabela 1. Colunas selecionadas para a modelagem

Coluna	**Descrição**
IDADE	Idade da criança no momento do óbito
RACACOR	Raça/cor da criança
IDADEMAE	Idade da mãe
ESCMAE2010	Nível de escolaridade da mãe
QTDFILVIVO	Quantidade de filhos vivos da mãe
QTDFILMORT	Quantidade de filhos mortos da mãe
SEMAGESTAC	Idade gestacional em semanas
GRAVIDEZ	Tipo de gravidez (simples, dupla, etc.)
PARTO	Tipo de parto (normal, cesariano, etc.)
ASSISTMED	Flag indicando se houve assistência médica durante a gestação
ESTADO	Estado que ocorreu o óbito da criança

Concluída a fase de tratamento, limpeza e padronização dos dados, foi realizada uma análise exploratória a partir de agrupamentos da quantidade de óbitos de crianças até 1 ano de idade causada por sífilis congênita por escolaridade da mãe, por tempo gestacional em semanas e por assistência médica. Pode-se observar que a maior incidência de mortes de crianças está entre as mães de baixa escolaridade e em nascimentos prematuros. A prematuridade é uma das morbimortalidades causada pela sífilis congênita. Além disso, pode-se observar que houve assistência médica na maioria dos óbitos como mostram as figuras abaixo. Para a geração dos gráficos em barra, foi utilizada a biblioteca *Seaborn* do *Python*.

Figura 1. Número de Mortes de Crianças Até 1 Ano por Sífilis Congênita por Escolaridade da Mãe

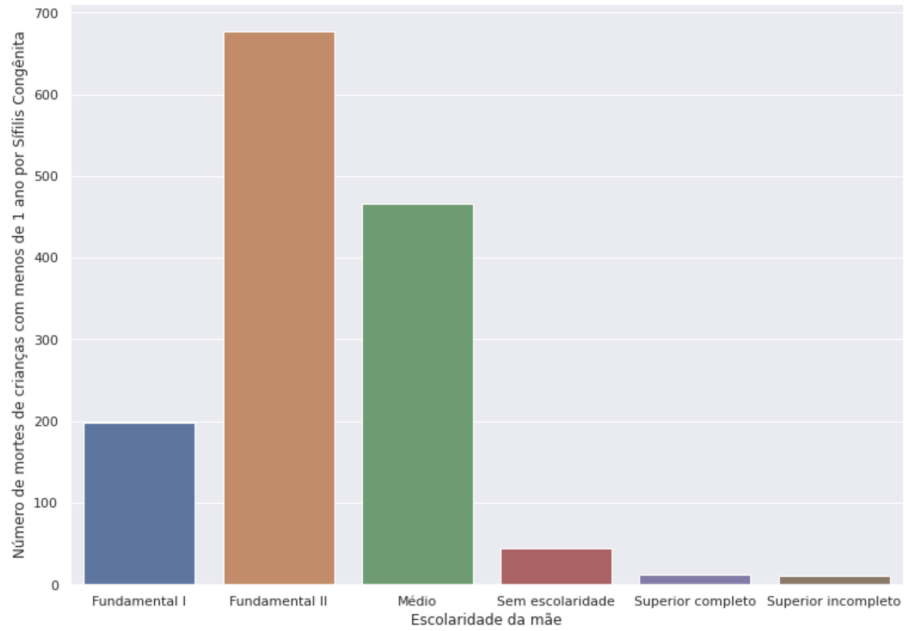


Figura 2. Número de Mortes de Crianças Até 1 Ano por Sífilis Congênita por Assistência Médica

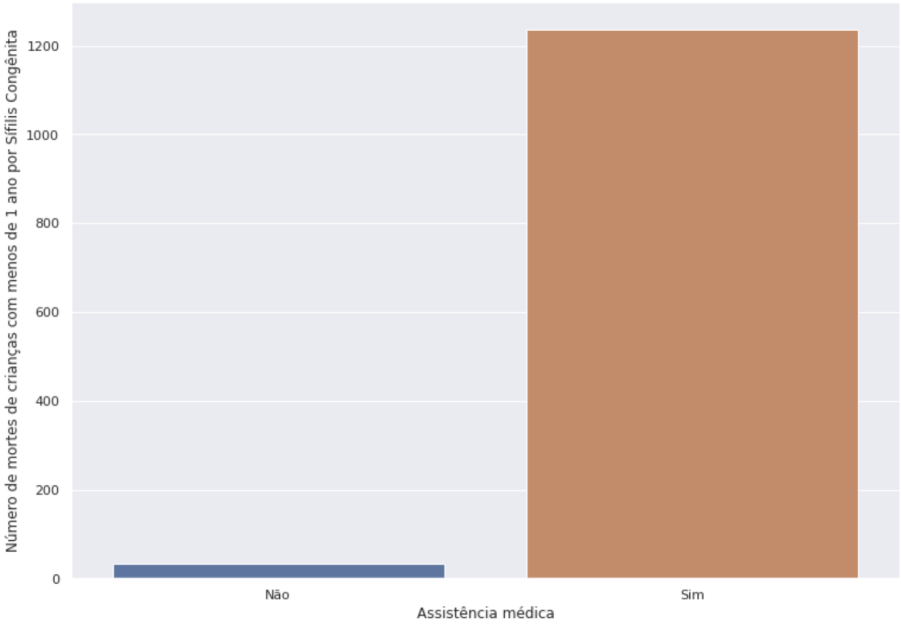
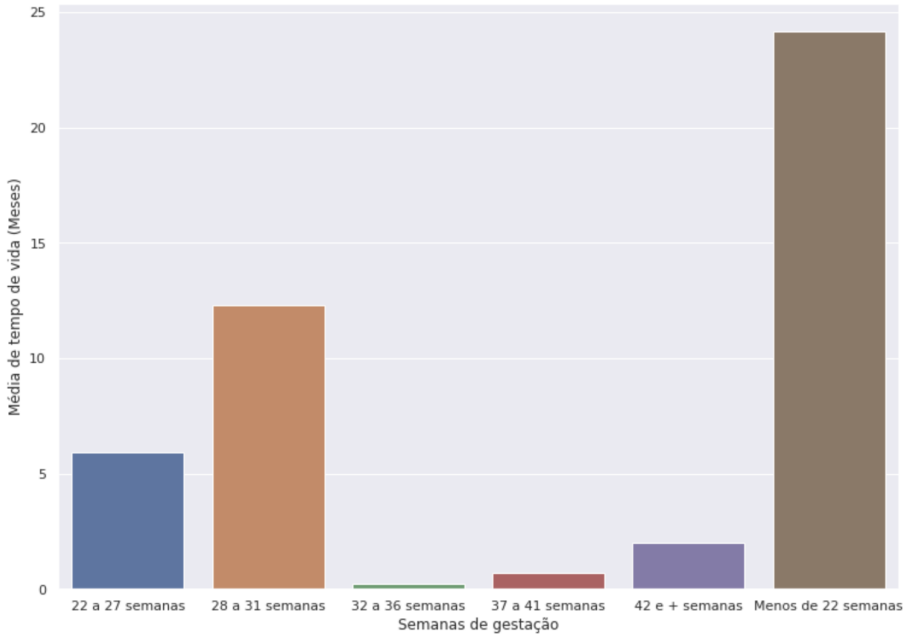


Figura 3. Número de Mortes de Crianças Até 1 Ano por Sífilis Congênita por Semanas de Gestação



A etapa de data mining acontece após as etapas de seleção e transformação de dados para tentar descobrir padrões nos dados. O uso de data mining, combinado com algoritmos de *Machine Learning*, pode auxiliar o especialista da saúde em momentos críticos que demandem decisões rápida [Fernandes and Filho 2019]. O *K-NN* é considerado um dos mais simples algoritmos de AM, com o aprendizado feito por instâncias e o conjunto de treinamento armazenado durante a aprendizagem (FACELI et al., 2011). As florestas aleatórias (*Random Forests*) são comitês de árvores de decisão e se baseiam em dois conceitos principais: seleção aleatória dos atributos de entrada e o *bagging* (*Bootstrap Aggregation*) (LIAW et al., 2002). [Pinto et al. 2019]. Sendo assim, o primeiro passo foi separar a base em dados de treino e teste para utilizar 4 algoritmos de classificação alterando seus hiperparâmetros: o *K-NN* foi testado alterando o número de vizinhos mais próximos de 1 a 5, para o *Naive Bayes* foram testados os tipos *GaussianNB*, *MultinominalNB*, *ComplementNB* e *BernoulliNB*, o modelo de Árvore de Decisão foi testado alterando o número de nós de 2 a 5 e o modelo *Random Forest* alterando a profundidade máxima de 2 a 6. Os modelos foram testados com o objetivo de prever se os casos de óbito por sífilis congênita ocorreram com mais ou com menos de 1 ano de idade e então comparar a acurácia de cada modelo.

Devido ao fato da base ter poucos registros, foi utilizada a técnica de validação cruzada via biblioteca *Python sklearn* para aumentar a capacidade de generalização do modelo. A validação cruzada envolve um conjunto de técnicas que particionam o conjunto de dados e geram modelos repetidamente e testam seu futuro poder preditivo [Koul et al. 2018]. A validação cruzada tem a vantagem computacional de evitar o ajuste de um modelo muito próximo às peculiaridades de um conjunto de dados (*overfitting*) [Koul et al. 2018]. Além disso, como há mais registros de mortes com menos de 1 ano do que com mais de 1 ano, além da acurácia geral do modelo foi calculada a acurácia de cada um dos modelos prevendo mortes com mais de 1 ano para identificar se o mo-

delo ficou enviesado. Sendo assim, após o cálculo das acurácias mencionadas, foi utilizada uma das técnicas de *Oversampling* chamada *SMOTE* via biblioteca *Python imblearn* para balancear a base e calcular a acurácia de cada um dos modelos novamente. O *SMOTE* gera amostras sintéticas da classe minoritária por meio da sobreamostragem de cada ponto de dados, considerando combinações lineares de vizinhos de classe minoritária existentes. Cada amostra de dados minoritários gera um número igual de dados sintéticos.[Turlapati and Prusty 2020]. As tabelas abaixo mostram as acurácias antes e depois do balanceamento respectivamente.

Tabela 2. Acurácias antes do balanceamento da base

Modelo	Acurácia Geral	Acurácia Maior 1 Ano
KNN k=1	0.948235	0.428571
KNN k=2	0.938824	0.607143
KNN k=3	0.960000	0.357143
KNN k=4	0.957647	0.589286
KNN k=5	0.962941	0.446429
GaussianNB	0.957647	0.839286
MultinomialNB	0.942941	0.857143
ComplementNB	0.936471	0.875000
BernoulliNB	0.962941	0.000000
Árvore de Decisão 2	0.962941	0.000000
Árvore de Decisão 3	0.962941	0.000000
Árvore de Decisão 4	0.962941	0.000000
Árvore de Decisão 5	0.962941	0.000000
Random Forest 2	0.962941	0.000000
Random Forest 3	0.962941	0.000000
Random Forest 4	0.962941	0.000000
Random Forest 5	0.962941	0.000000

Tabela 3. Acurácias depois do balanceamento da base

Modelo	Acurácia Geral	Acurácia Maior 1 Ano
KNN k=1	0.910584	0.877129
KNN k=2	0.943127	0.971411
KNN k=3	0.932482	0.939781
KNN k=4	0.942518	0.975061
KNN k=5	0.921533	0.923358
GaussianNB	0.917275	0.878345
MultinomialNB	0.913321	0.888078
ComplementNB	0.913321	0.888078
BernoulliNB	0.823905	0.952555
Árvore de Decisão 2	0.823905	0.952555
Árvore de Decisão 3	0.823905	0.952555
Árvore de Decisão 4	0.823905	0.952555
Árvore de Decisão 5	0.823905	0.952555
Random Forest 2	0.823905	0.952555
Random Forest 3	0.823905	0.952555
Random Forest 4	0.823905	0.952555
Random Forest 5	0.823905	0.952555

A tabela 2 contém as acurácias antes do balanceamento da base, é possível notar que todos os modelos apresentaram resultados satisfatórios quando se tem como objetivo a acurácia geral, porém apenas os modelos *KNN* e *Naive Bayes*, com exceção do tipo *BernoulliNB*, apresentaram resultados medianos e satisfatórios respectivamente quando

se tem como objetivo a acurácia prevendo mortes por sífilis congênita acima de 1 ano. Pode-se observar que os modelos Árvore de Decisão e *Random Forest* tiveram a acurácia prevendo mortes acima de 1 ano de 0%, ou seja, devido ao desbalanceamento da base o modelo ficou enviesado. Já a tabela 3 contém as acurácias após o balanceamento da base via técnica de *Oversampling*. É evidente que, assim como na tabela 2, todos os modelos apresentaram resultados satisfatórios quando se tem como objetivo a acurácia geral. Pode-se notar também que, após o balanceamento, todos os modelos também apresentaram resultados satisfatórios quando se tem como objetivo a acurácia prevendo mortes por sífilis congênita acima de 1 ano.

5. Conclusões e Recomendações

Como resultado, pode-se observar a partir dos gráficos que a maior incidência de mortes de crianças com menos de 1 ano por sífilis congênita ocorre quando a mãe possui baixa escolaridade, sendo que a maior concentração está na escolaridade "Fundamental II". Além disso, a sífilis congênita possui como morbimortalidade a prematuridade, justificando a maior incidência de mortes quando a idade gestacional é inferior a 22 semanas, ou seja, em nascimentos prematuros. Uma das hipóteses era que a maior incidência de mortes estaria em filhos de mães que não tiveram assistência médica. Entretanto, conforme resultados na Figura 2 na seção **Desenvolvimento e Resultados**, a maior incidência está em filhos de mães que tiveram assistência médica. Contudo, a documentação da base SIM não esclarece se a assistência médica é algum tratamento específico para a sífilis durante a gestação ou se a mãe apenas teve contato com algum hospital ou clínica. Sendo assim, não podemos afirmar que a criança virá a óbito com menos de 1 ano apenas com essa variável.

De qualquer forma, podemos concluir a partir do modelo que as variáveis idade da mãe, raça/cor da criança, escolaridade da mãe, quantidade de filhos vivos da mãe, quantidade de filhos mortos da mãe, idade gestacional, tipo de gravidez, tipo do parto, assistência médica e estado de residência da mãe quando analisados juntas apresentam resultados satisfatórios para os modelos preditivos utilizados. É importante ressaltar que, apesar dos modelos apresentarem valores de acurácia altos, o que poderia indicar *overfitting*, foram aplicados os métodos de validação cruzada e *oversampling* para avaliar a capacidade de generalização do modelo e para evitar que o modelo fique enviesado, respectivamente. Além disso, é relevante ressaltar a importância da descoberta e tratamento da sífilis antes da mulher engravidar devido sua alta taxa de morbimortalidade, evidenciando o impacto da sífilis congênita na morte de bebês de até 1 ano de idade.

Para trabalhos futuros, recomenda-se utilizar variáveis que especificam quais tratamentos e acompanhamentos a gestante realizou como consultas com especialistas, exames laboratoriais e de imagem, identificando também a qualidade do atendimento que a gestante recebeu para demonstrar o impacto que o tratamento da sífilis durante a gestação traz para os bebês.

Referências

Azevedo, A. and Santos, M. F. (2008). Kdd, semma and crisp-dm: A parallel overview.

Coelho, F. C., Baron, B. C., de Castro Fonseca, G. M., Reck, P., and Palumbo, D. (2021). Alertadengue/pysus: Vaccine.

Domingues, R. and Leal, M. (2016). Incidência de sífilis congênita e fatores associados à transmissão vertical da sífilis: dados do estudo nascer no brasil. 32(6):e00082415.

Fernandes, F. and Filho, A. (2019). Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho. 44:e13.

Filho, A. (2015). Uso de big data em saúde no brasil: perspectivas para um futuro próximo. 24(2):325–332.

Koul, A., Becchio, C., and Cavallo, A. (2018). Cross-validation approaches for replicability in psychology. 9(1117).

Magalhães, D., Kawaguchi, I., Dias, A., and Calderon, I. (2013). Sífilis materna e congênita: ainda um desafio. 29:1109–1120.

Monard, M. and Baranauskas, J. (2003). Conceitos sobre aprendizado de máquina. sistemas inteligentes-fundamentos e aplicações. 1:31.

Nakamura, C., Otero, C., and Carvalho, D. (2016). Mineração de dados no enfrentamento da transmissão vertical de sífilis. pages 171–180.

Nascimento, M., Cunha, A., Guimarães, E., Alvarez, F., Oliveira, S., and Bôas, B. (2012). Gestações complicadas por sífilis materna e óbito fetal. 34:56–62.

Nonato, S., Melo, A., and Guimarães, M. (2015). Sífilis na gestação e fatores associados à sífilis congênita em belo horizonte-mg 2010-2013. 24(4):681–694.

Pinto, N., Amaral, J., and Melo, P. (2019). Detecção de alterações respiratórias na fibrose cística com o uso de algoritmos de aprendizado de máquinas. pages 5–8.

Silveira, F., Moreira, L., and Carvalho, D. (2020). Utilização de algoritmos de aprendizagem de máquina na predição de arboviroses transmitidas pelo aedes aegypti. 14(1):1824.

Turlapati, V. P. K. and Prusty, M. R. (2020). Outlier-smote: A refined oversampling technique for improved detection of covid-19.

WHO (2011). *Methods for surveillance and monitoring of Congenital syphilis elimination within existing systems*. WHO Library Cataloguing-in-Publication Data.