Predição de casos de sífilis congênita: uma avaliação de desempenho de diferentes modelos de aprendizado de máquina

Igor Vitor Teixeira[1], **Patricia Takako Endo**[1]

[1]Programa de Pós-Graduação em Engenharia da Computação,
Universidade de Pernambuco, Recife, Brasil

ivt@ecomp.poli.br, patricia.endo@upe.br

Resumo. *As Infecções Sexualmente Transmissíveis (ISTs) são um problema preocupante, especialmente em países em desenvolvimento e subdesenvolvidos, nos quais fatores ambientais e outros determinantes da saúde contribuem para sua rápida disseminação. Diante desta situação, o principal objetivo deste trabalho é avaliar o desempenho de diferentes modelos de aprendizado de máquina na predição de resultados indesejáveis da sífilis congênita, a fim de auxiliar na alocação de recursos e otimizar as ações na área de saúde, especialmente em um ambiente de saúde com poucos recursos. Utilizamos dados clínicos e sociodemográficos de gestantes atendidas em um programa social do estado de Pernambuco, Brasil, denominado Programa Mãe Coruja Pernambucana (PMCP). Os modelos SVM e AdaBoost apresentaram os melhores resultados, utilizando 13 e 11 atributos como entrada, respectivamente.*

1. Caracterização do problema e motivação da pesquisa

As Infecções Sexualmente Transmissíveis (ISTs) estão entre as doenças transmissíveis mais comuns [da Saúde do Brasil 2021], afetando negativamente a qualidade de vida e a saúde das pessoas. Dentre elas, a sífilis é uma infecção sistêmica causada pela bactéria *Treponema pallidum*, transmitida de três formas: sexual, congênita ou por transfusão sanguínea. Os resultados adversos da sífilis gestacional não tratada são aborto precoce (40%), morte fetal (11%) e prematuridade ou baixo peso (12% a 13%). Além disso, pelo menos 20% dos recém-nascidos apresentam sinais sugestivos de sífilis congênita precoce [Domingues et al. 2021]. Em recém-nascidos, a sífilis congênita pode se manifestar de maneira precoce ou tardia, tendo como principais consequências a prematuridade e baixo peso ao nascer, lesões de pele, periostite, alterações radiográficas, tíbia em lâmina de sabre e articulações de Clutton [da Saúde do Brasil 2020a]. A sífilis congênita deve ser precisamente combatida por causa dessas graves consequências em recém-nascidos, sendo objetivamente erradicável. O aumento do número de testes rápidos, principalmente na Atenção Primária à Saúde (APS), a partir de 2017, permitiu uma melhor compreensão do cenário da sífilis gestacional e congênita no Brasil.

A falta de interesse no tratamento da sífilis por parte dos parceiros sexuais da gestante pode estar impactando nas taxas de incidência, possibilitando a descontinuidade da cadeia de infecção da sífilis e evitando possíveis reinfecções após o sucesso do tratamento [da Saúde do Brasil 2020b]. A sífilis é uma condição evitável desde que a gestante e seus parceiros sexuais sejam identificados e o tratamento adequado seja realizado, com intervenções simples e orientadas para a gestante, parceiros sexuais e recém-nascidos.

Neste contexto, o Programa Mãe Coruja Pernambucana (PMCP) é um programa social brasileiro de referência na área materno-infantil, implantado em 2007, reconhecido e premiado pela Organização das Nações Unidas (ONU) e Organização dos Estados Americanos (OEA). Tem como objetivo garantir a atenção integral às gestantes usuárias do Sistema Único de Saúde (SUS) e seus filhos até 5 anos de idade, formando uma rede solidária para reduzir a mortalidade materno-infantil e melhorar os indicadores sociais. No entanto, a escassez de canais de comunicação e informação e a existência de estigmatização pela sociedade sobre as ISTs, principalmente no contexto de populações vulneráveis, fragilizam a conscientização e o acesso às medidas preventivas e, consequentemente, ao tratamento da sífilis.

Alguns trabalhos da literatura apresentaram estudos sobre a incidência e os fatores de risco para sífilis, como Santos et al. [Santos et al. 2021] e Lima et al. [Lima et al. 2013]. Porém, este trabalho se diferencia de outros da literatura por conduzir e discutir uma avaliação de desempenho de modelos de aprendizado de máquina usando dados clínicos e sociodemográficos em um sistema integrado que registra dados de pré-natal, parto e acompanhamento do desenvolvimento infantil. Esses dados permitem focar na classificação de possíveis casos de sífilis congênita em diversos cenários e avaliar quais modelos de aprendizado de máquina são mais eficientes, auxiliando os profissionais de saúde no acompanhamento das gestantes.

2. Objetivos e contribuições

Este trabalho tem como objetivo apresentar uma avaliação de desempenho de diferentes modelos de aprendizado de máquina para classificar possíveis desfechos indesejáveis da sífilis congênita, utilizando dados de gestantes atendidas pelo PMCP. Os modelos utilizam dados clínicos e sociodemográficos para possibilitar um melhor acompanhamento e cuidado durante a gestação.

Foram utilizadas dados anônimos fornecidos pelo PMCP, extraídos de seu sistema de informações, denominado SIS-MC. Esses dados são compostos por dados clínicos e sociodemográficos referentes ao pré-natal, desfechos das gestantes e seus filhos, dos municípios atendidos pelo PMCP no Estado de Pernambuco, Brasil, entre os anos de 2013 e 2021. A utilização dos dados do SIS-MC foi autorizada pelo Comitê de Ética em Pesquisa (CEP) da Universidade Federal de Pernambuco (UFPE) com o número de Certificado de Apresentação de Apreciação Ética (CAAE) 12438019.2.0000.5208 e autorizada pela instituição parceira, o PMCP. As principais contribuições deste trabalho são:

- Conjunto de dados pré-processado disponível publicamente[1], contendo dados clínicos e sociodemográficos;
- Registro de software no Instituto Nacional da Propriedade Industrial (INPI): BR512022002788-7;
- Artigo em revisão na PLOS ONE: Predicting congenital syphilis cases: a performance evaluation of different machine learning models [Teixeira et al. 2022];
- Avaliação de desempenho de diferentes modelos de aprendizado de máquina para predição de casos de sífilis congênita utilizando dados clínicos e sociodemográficos.

[1] `https://data.mendeley.com/datasets/3zkcvybvkz/1`

3. Metodologia

A Figura 1 apresenta a metodologia de pré-processamento dos dados aplicada para unificar os data sets fornecidas pelo PMCP e para realizar a seleção manual de atributos, lidar com dados faltantes, remover *outliers* e criar novos atributos. Devido a restrição de páginas, mais detalhes podem ser vistos em: `https://github.com/dotlab-brazil/congenital-syphilis/blob/main/dissertacao_ppgec_igor_vitor_teixeira.pdf`.

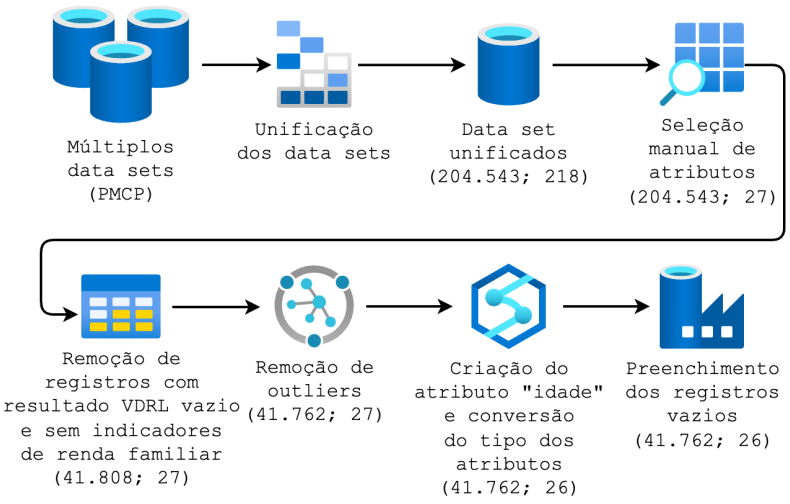

Figura 1. Metodologia de pré-processamento dos dados.

Neste trabalho, foram definidos experimentos com o objetivo de comparar o desempenho do aprendizado dos modelos ao lidar com diferentes configurações dos data sets, utilizando: (*i*) dados desbalanceados, (*ii*) dados balanceados e (*iii*) dados aplicando a técnica de one-hot encoding. Com isso, foram totalizados seis experimentos, sendo avaliadas as seguintes técnicas de aprendizado de máquina: Decision Tree, Random Forest, AdaBoost, Gradient Boosting Machines (GBM), eXtreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN) e Support Vector Machine (SVM). Para cada experimento, foi construído um data set de acordo com as seguintes características:

1. ***Imbalanced Data Set* (IDS)**: data set desbalanceado com 2.327 registros (826 casos positivos e 1.501 casos negativos) e 26 atributos. Como o data set original continha muitos casos negativos quando comparado ao número de casos positivos (40.936 casos negativos e 826 casos positivos), a técnica de random undersampling foi utilizada para reduzir a diferença entre essa classes, estabelecendo uma proporção de 55% do número de amostras na classe minoritária (casos positivos) sobre o número de amostras na classe majoritária (casos negativos) após a nova amostragem.
2. ***Balanced Data Set* (BDS)**: data set balanceado usando a técnica de random undersampling com 1.652 registros (826 casos positivos e 826 casos negativos).
3. ***Imbalanced with One-hot Encoding Data Set* (IODS)**: data set desbalanceado com técnica de one-hot encoding aplicada para transformar dados categóricos em dados binários. Neste caso, o número de atributos aumentou para 97, pois o one-hot encoding cria um novo atributo para cada classe de um determinado atributo categórico. Por exemplo, o atributo categórico que indica se a gestante é fumante

(SMOKER) possui três categorias possíveis: positivo, negativo e não informado. Mediante a aplicação da técnica de one-hot encoding, esse atributo único será dividido em três atributos binários: (*i*) positivo para fumantes; (*ii*) negativo para fumantes; e (*iii*) não informado.

4. ***Balanced with One-hot Encoding Data Set*** **(BODS)**: data set balanceado com a técnica de one-hot encoding.
5. ***Imbalanced with One-hot Encoding with Column Drop Data Set*** **(IODDS)**: data set desbalanceado com técnica de one-hot encoding aplicada, porém com a coluna relacionada a não informada pela paciente removida. Alguns atributos possuem uma classe que representa os dados faltantes. Neste experimento, após aplicar o one-hot encoding, a coluna relacionada a ele foi removida, diminuindo o número de atributos para 75.
6. ***Balanced with One-hot Encoding with Column Drop Data Set*** **(BODDS)**: data set balanceado e com a técnica de one-hot aplicada, porém a coluna referente a não informado pela paciente foi removida.

Após a criação dos data sets para cada experimento, eles foram divididos em conjuntos de treinamento (80%) e de teste (20%). A fim de avaliar a melhor abordagem para selecionar o subconjunto de atributos mais eficaz, aplicamos dois processos de treinamento diferentes em todos os experimentos, usando a técnica de seleção automática de atributos *Sequential Feature Algorithm* (SFA), onde foi aplicado os tipos *Sequential Forward Selection* (SFS) e *Sequential Backward Selection* (SBS), e atributos selecionados pelas especialistas em saúde do PMCP, resultando em 126 modelos para serem avaliados.

4. Resultados e discussão

A métrica F1-Score foi utilizada para comparar o desempenho dos modelos de cada experimento, devido ela ser calculada a partir de uma média harmônica entre precisão e sensibilidade, duas métricas relevantes para análise de problemas de saúde. Os experimentos levaram mais de 5 meses para serem finalizados, e foram executados em quatro servidores do tipo c5a.4xlarge com um processador AMD EPYC series 2ª geração, 16 vCPUs, 32 GB de RAM e 30 GB de armazenamento, providos pelo serviço Amazon Elastic Compute Cloud (EC2) da Amazon Web Services (AWS). Esses recursos foram obtidos através do Edital CNPq/AWS 032/2019.

Dentre os modelos que utilizaram técnicas de SFA para fazer a seleção de atributos, os que apresentaram os melhores desempenhos de cada experimento foram selecionados e comparados entre si, onde foi possível notar resultados diferentes, principalmente para as métricas F1-Score, sensibilidade e especificidade. Esses modelos apresentaram subconjuntos de atributos que variaram entre 13 e 42 atributos, sendo que os atributos mais comuns foram: escolaridade, estado civil, insegurança alimentar, tipo de tratamento de água residencial e se a gestante é fumante.

O modelo que apresentou o melhor valor de F1-Score foi o SVM quando executado no experimento BDS, sendo chamado de SVM-BDS-SFA. Este modelo utilizou 13 atributos. Dentre os modelos que utilizaram os atributos selecionados pelas especialistas em saúde do PMCP, o modelo que apresentou o melhor F1-Score foi o AdaBoost quando executado no experimento BODS, sendo chamado de AdaBoost-BODS-Expert.

Os modelos SVM-BDS-SFA e AdaBoost-BODS-Expert atingiram resultados similares, apresentando maiores diferenças na sensibilidade, na especificidade e nos atributos utilizados. A Figura 2 apresenta uma comparação entre eles e a Tabela 1 lista o resultado do *grid search* e os atributos utilizados por ambos os modelos.

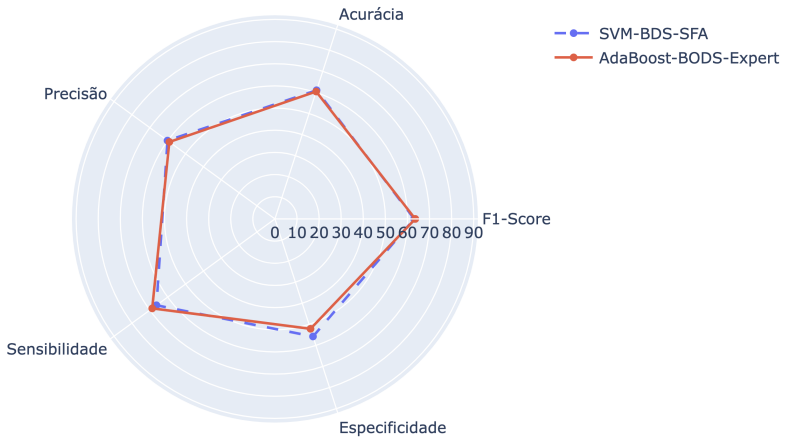

Figura 2. Comparação dos modelos SVM-BDS-SFA e AdaBoost-BODS-Expert.

Tabela 1. Resultados do *grid search* e da seleção de atributos.

Modelo	Hiperparâmetros	Qtd. Atr.	Atributos
SVM-BDS-SFA	gamma: scale kernel: rbf	13	EDUC_LEVEL, HAS_FRU_TREE, WATER_TREATMENT, RH_FACTOR, PLAN_PREGNANCY, HAS_PREG_RISK, TET_VACCINE, IS_HEAD_FAMILY, MARITAL_STATUS, FOOD_INSECURITY, NUM_ABORTIONS, NUM_PREGNANCIES e HAS_FAM_INCOME
AdaBoost-BODS-Expert	learning_rate: 0,5 n_estimators: 150	11	AGE, LEVEL_SCHOOLING, FAM_INCOME, PLAN_PREGNANCY, HAS_PREG_RISK, MARITAL_STATUS, FOOD_INSECURITY, NUM_ABORTIONS, NUM_LIV_CHILDREN, NUM_PREGNANCIES e FAM_PLANNING

Nenhum dos modelos apresentou acurácia superior a 70%, variando de 56,50% a 64,38%, demonstrando a dificuldade de classificar possíveis desfechos da sífilis congênita utilizando apenas dados clínicos e sociodemográficos. A razão que pode explicar esse fato é a abundância de dados faltantes, que reduzem a qualidade dos dados e o aprendizado do modelo [Ehsani-Moghaddam et al. 2019]. Para lidar com essa questão, criamos um valor categórico para representar que não foi informado pela paciente. No entanto, isso pode ter causado: (*i*) dificuldade em encontrar padrões nos dados que permitissem uma classificação mais precisa e (*ii*) relevância para categorias relacionadas a dados não informados, o que pode dificultar a classificação de dados mais realistas, quando os atributos são preenchidos corretamente.

É importante ressaltar que segundo as especialistas em saúde do PMCP, este trabalho é aplicável no acompanhamento diário das gestantes, sendo relevante para o desenvolvimento de novas estratégias no PMCP, permitindo um acompanhamento mais qualificado e possibilitando a criação de protocolos que hoje não existem no PMCP. Além disso, a utilização de modelos como o AdaBoost-BODS-Expert aumentará o nível de alerta dos profissionais de saúde do PMCP. Isso permitirá um melhor acompanhamento de todas as gestações, não apenas daquelas consideradas possíveis casos de sífilis

congênita, o que melhorará a qualidade do atendimento. O feedback dos especialistas em saúde do PMCP sobre os resultados obtidos neste trabalho está disponível em `https://youtu.be/a80XlyrTH0M`.

5. Conclusões e trabalhos futuros

Não foram encontrados na literatura estudos utilizando técnicas de aprendizado de máquina para classificar casos de sífilis congênita. Este trabalho explorou essa limitação por meio da aplicação de técnicas de otimização no data set, bem como nos modelos e experimentos propostos. Nossos resultados mostraram que, apesar de ser um desafio, é possível predizer a sífilis congênita durante a gravidez apenas por meio de dados clínicos e sociodemográficos. Ao mesmo tempo, também identificamos as limitações geradas pela grande quantidade de dados faltantes, que podem ter impossibilitado resultados mais acurados, indicando a necessidade do PMCP melhorar a qualidade da aquisição dos dados.

Como trabalho futuro, pretendemos aplicar diferentes técnicas de balanceamento e imputação de dados para investigar métodos alternativos para lidar com o desequilíbrio do data set e dados faltantes. Também planejamos expandir nosso trabalho para predição de sífilis gestacional, uma vez que pode ajudar a reduzir a incidência de sífilis congênita.

Referências

da Saúde do Brasil, M. (2020a). Guia de vigilância em saúde. Acessado em 12 de dezembro de 2020.

da Saúde do Brasil, M. (2020b). Protocolo clínico e diretrizes terapêuticas para atenção integral às pessoas com infecções sexualmente transmissíveis (ist). Acessado em 7 de dezembro de 2022.

da Saúde do Brasil, M. (2021). Boletim epidemiológico sífilis.

Domingues, C. S. B., Duarte, G., Passos, M. R. L., Sztajnbok, D. C. d. N., and Menezes, M. L. B. (2021). Brazilian protocol for sexually transmitted infections, 2020: congenital syphilis and child exposed to syphilis. *Revista da Sociedade Brasileira de Medicina Tropical*, 54.

Ehsani-Moghaddam, B., Martin, K., and Queenan, J. A. (2019). Data quality in healthcare: A report of practical experience with the canadian primary care sentinel surveillance network data. *Health Information Management Journal*, 50(1-2):88–92.

Lima, M. G., Santos, R. F. R. d., Barbosa, G. J. A., and Ribeiro, G. d. S. (2013). Incidência e fatores de risco para sífilis congênita em belo horizonte, minas gerais, 2001-2008. *Ciência & Saúde Coletiva*, 18:499–506.

Santos, M. M. d., Rosendo, T. M. S. d. S., Lopes, A. K. B., Roncalli, A. G., and Lima, K. C. d. (2021). Weaknesses in primary health care favor the growth of acquired syphilis. *PLoS neglected tropical diseases*, 15(2):e0009085.

Teixeira, I. V., Leite, M. T. d. S., Melo, F. L. d. M., Rocha, E. d. S. R., Sadok, S., Carrarine, A. S. P. d. C., Santana, Marilia, R. C. P., Oliveira, A. M. d. L., Gadelha, K. V., Morais, C. M. d., Kelner, J., and Endo, P. T. (2022). Predicting congenital syphilis cases: a performance evaluation of different machine learning models. *PLOS ONE*.