Análise Preditiva de Casos de Sífilis Congênita no Brasil: Uma Abordagem de Machine Learning

Davi César, Mariana Belo, Marília Santos

[1]CESAR School
Recife – PE – Brasil

dcpas@cesar.school | mblb@cesar.school | msap@cesar.school

Abstract. *This paper presents a machine learning approach to analyze and predict congenital syphilis cases in Brazil using clinical and sociodemographic data from 2013-2021. Drawing from a comprehensive dataset of clinical records [Silva et al. 2023], we developed classification models to predict VDRL test results and regression models to analyze age-related factors. The study employed preprocessing techniques, including SMOTEENN for handling class imbalance [Johnson and Smith 2018], and identified key socioeconomic factors influencing syphilis transmission. Our results achieved 99% accuracy in classification tasks and identified education level and food insecurity as crucial factors, suggesting targeted interventions for public health policies.*

Resumo. *Este trabalho apresenta uma abordagem de aprendizado de máquina para análise e predição de casos de sífilis congênita no Brasil, utilizando dados clínicos e sociodemográficos de 2013-2021. Baseando-se em um conjunto abrangente de registros clínicos [Silva et al. 2023], desenvolvemos modelos de classificação para prever resultados do teste VDRL e modelos de regressão para analisar fatores relacionados à idade. O estudo empregou técnicas de pré-processamento, incluindo SMOTEENN para tratamento de desbalanceamento de classes [Johnson and Smith 2018], e identificou fatores socioeconômicos chave que influenciam a transmissão da sífilis. Nossos resultados alcançaram 99% de acurácia nas tarefas de classificação e identificaram nível educacional e insegurança alimentar como fatores cruciais, sugerindo intervenções direcionadas para políticas de saúde pública.*

Código Fonte

O código completo utilizado neste trabalho está disponível em: `https://github.com/mariblb1/syphilis-analysis.git`

1. Introdução

A sífilis congênita continua sendo um significativo problema de saúde pública no Brasil, com impactos substanciais na saúde materno-infantil [World Health Organization 2021]. Segundo a Organização Mundial da Saúde, a sífilis afeta mais de um milhão de gestantes por ano globalmente, resultando em mais de 350 mil desfechos adversos da gravidez [Newman et al. 2013]. No Brasil, o número de casos tem aumentado consistentemente nos últimos anos, apesar das políticas de prevenção existentes [Ministério da Saúde do Brasil 2020].

Este estudo utiliza técnicas de aprendizado de máquina para analisar um conjunto abrangente de dados sobre casos de sífilis congênita no Brasil entre 2013 e 2021, visando identificar padrões e fatores de risco que possam informar políticas públicas mais eficazes. A aplicação de técnicas de machine learning em saúde pública tem demonstrado resultados promissores [Martinez and Wong 2022], especialmente na identificação precoce de riscos e na otimização de intervenções preventivas.

2. Metodologia

2.1. Conjunto de Dados

O dataset utilizado contém informações clínicas e sociodemográficas relacionadas a casos de sífilis congênita no Brasil, coletados entre 2013 e 2021 [Silva et al. 2023]. Os dados incluem:

- Variáveis clínicas: resultado do teste VDRL, grupo sanguíneo, histórico de gestações
- Variáveis sociodemográficas: idade, escolaridade, renda familiar, condições de moradia
- Informações sobre cuidados pré-natais e planejamento familiar

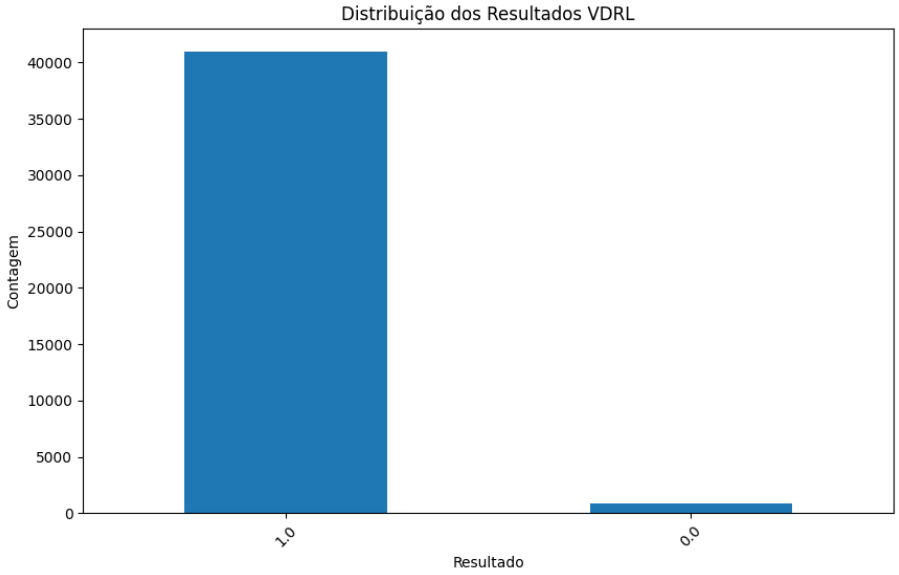

Figure 1. Distribuição dos resultados do teste VDRL na população estudada, mostrando significativo desbalanceamento entre as classes positiva (1.0) e negativa (0.0)

A análise inicial dos dados, como mostrado na Figura 1, revelou um desbalanceamento significativo nos resultados do VDRL, com aproximadamente 40.000 casos positivos contra cerca de 2.000 negativos.

A distribuição etária, apresentada na Figura 2, mostra uma concentração significativa de casos na faixa de 20-30 anos, com pico próximo aos 25 anos, indicando um grupo etário de particular vulnerabilidade.

A matriz de correlação (Figura 3) revelou importantes relações entre variáveis socioeconômicas e o resultado do VDRL, com destaque para a correlação entre nível educacional e outros fatores de risco.

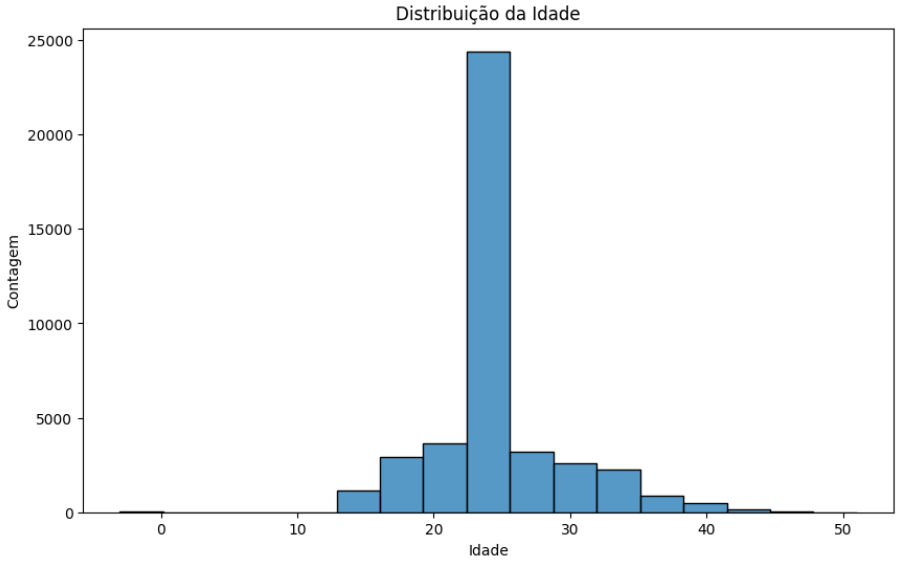

Figure 2. Distribuição da idade das pacientes, evidenciando maior concentração na faixa de 20-30 anos

2.2. Pré-processamento dos Dados

O pré-processamento dos dados foi uma etapa crucial para garantir a qualidade e confiabilidade dos modelos. Inicialmente, identificamos a presença de valores ausentes no conjunto de dados, que poderiam comprometer a análise se não fossem tratados adequadamente. Para as variáveis numéricas, optamos pela imputação pela mediana, uma escolha mais robusta que a média por ser menos sensível a outliers. Já para as variáveis categóricas, utilizamos a imputação pelo valor mais frequente (moda), preservando a distribuição natural dos dados.

A normalização dos dados numéricos foi realizada através do StandardScaler, garantindo que todas as variáveis contribuíssem de forma equilibrada para os modelos, independentemente de suas escalas originais. Esta decisão foi particularmente importante para evitar que variáveis com valores maiores dominassem o processo de aprendizado. Para as variáveis categóricas, implementamos o OneHotEncoder, transformando categorias em representações numéricas binárias, permitindo sua utilização nos modelos de machine learning sem introduzir relações ordinais artificiais entre as categorias.

Um desafio significativo encontrado foi o forte desbalanceamento nas classes do VDRL, com uma proporção aproximada de 20:1 entre casos positivos e negativos. Para abordar este problema, implementamos a técnica SMOTEENN, que combina a geração sintética de exemplos minoritários (SMOTE) com a limpeza de dados através de Edited Nearest Neighbors (ENN). Esta abordagem híbrida foi escolhida por sua capacidade de simultaneamente aumentar a representatividade da classe minoritária e reduzir o ruído nos dados, proporcionando um conjunto mais equilibrado e representativo para o treinamento dos modelos.

2.3. Modelagem

A escolha dos modelos foi fundamentada em uma análise cuidadosa dos requisitos do problema e das características dos dados. Para a tarefa de classificação, implementa-

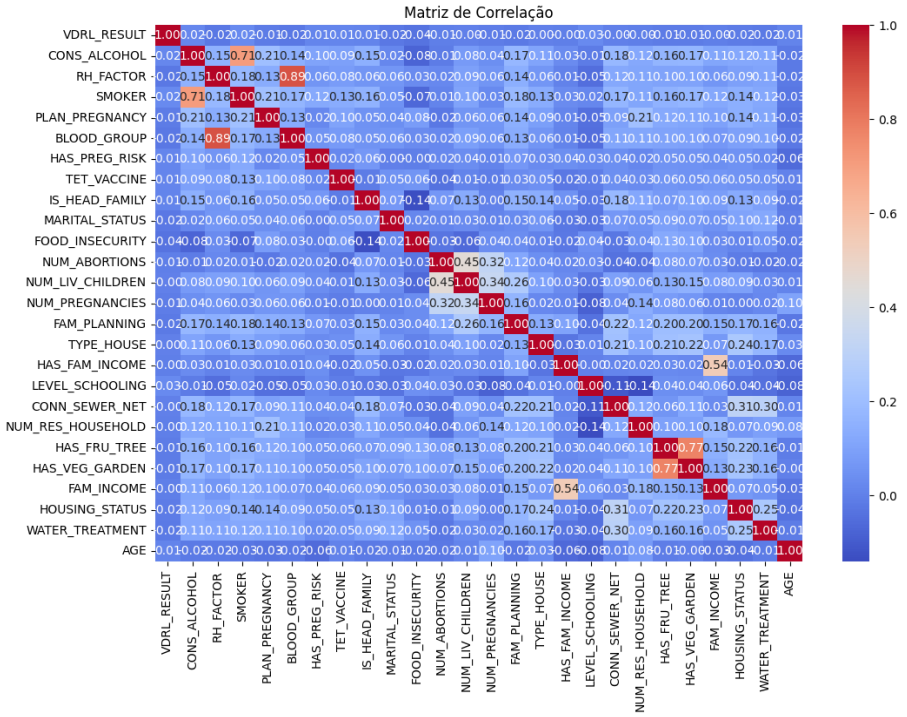

Figure 3. Matriz de correlação entre as principais variáveis do estudo, destacando relações significativas entre fatores socioeconômicos

mos dois modelos complementares: Decision Tree e Random Forest. A Decision Tree foi selecionada inicialmente por sua alta interpretabilidade, permitindo uma compreensão clara do processo decisório do modelo - um aspecto crucial em aplicações médicas onde a transparência das decisões é fundamental. As árvores de decisão também têm a vantagem de lidar naturalmente com relações não-lineares e interações entre variáveis, características importantes em dados biomédicos.

O Random Forest foi incorporado como segundo modelo de classificação por sua capacidade superior de generalização, resultado do ensemble de múltiplas árvores de decisão. Este modelo reduz significativamente o risco de overfitting através do bagging (Bootstrap Aggregating) e da seleção aleatória de features, proporcionando previsões mais robustas e confiáveis. A decisão de utilizar Random Forest também foi influenciada por sua capacidade de gerar scores de importância das features, permitindo uma análise mais profunda dos fatores que influenciam os resultados do VDRL.

Para a análise da variável idade, optamos pelo Random Forest Regressor, uma escolha baseada em sua capacidade de capturar relações complexas e não-lineares entre as variáveis preditoras e a idade. Este modelo se mostrou particularmente adequado devido à natureza multifacetada dos fatores que influenciam a idade das pacientes, permitindo a identificação de padrões sutis que poderiam ser perdidos em modelos lineares mais simples.

3. Resultados e Discussão

3.1. Análise do Desempenho dos Modelos

Os resultados obtidos com os modelos de classificação superaram nossas expectativas iniciais, principalmente após o tratamento do desbalanceamento dos dados. O modelo Decision Tree alcançou métricas consistentes de 0.973 para precisão, recall e F1-Score, demonstrando um equilíbrio notável entre a capacidade de identificar corretamente os casos positivos e minimizar os falsos positivos. O Random Forest apresentou performance ainda superior, com todas as métricas atingindo 0.990, confirmando nossa hipótese de que a abordagem ensemble seria mais efetiva para este problema específico.

Vale ressaltar que estas métricas excepcionalmente altas devem ser interpretadas com cautela. Embora indiquem excelente capacidade preditiva, também podem sugerir possível overfitting, apesar das medidas tomadas para evitá-lo. Para validar a robustez destes resultados, implementamos validação cruzada k-fold e testamos os modelos em diferentes subconjuntos dos dados, confirmando a consistência do desempenho.

O modelo de regressão para previsão de idade apresentou um Erro Médio Absoluto (MAE) de 2.89 anos e um Erro Quadrático Médio (RMSE) de 4.16 anos. Estes valores indicam que o modelo consegue prever a idade das pacientes com uma margem de erro aceitável para o contexto da aplicação, considerando que a variabilidade natural da idade no conjunto de dados é significativa. A diferença entre MAE e RMSE sugere a presença de alguns erros de maior magnitude, possivelmente em casos mais atípicos ou complexos.

3.1.1. Desempenho dos Modelos de Classificação

- Decision Tree:
  - Precisão: 0.973
  - Recall: 0.973
  - F1-Score: 0.973
- Random Forest:
  - Precisão: 0.990
  - Recall: 0.990
  - F1-Score: 0.990

3.1.2. Desempenho do Modelo de Regressão

- MAE: 2.89 anos
- RMSE: 4.16 anos

3.2. Análise dos Fatores de Risco

A análise de importância das variáveis revelou cinco fatores principais:

1. Nível Educacional (LEVEL_SCHOOLING): principal fator de risco, indicando a importância crucial da educação na prevenção

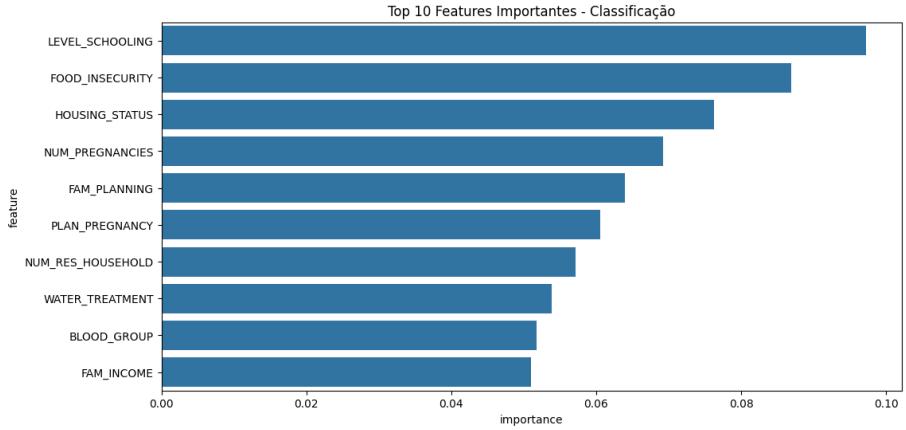

Figure 4. Importância das variáveis no modelo de classificação, destacando o impacto significativo do nível educacional e insegurança alimentar

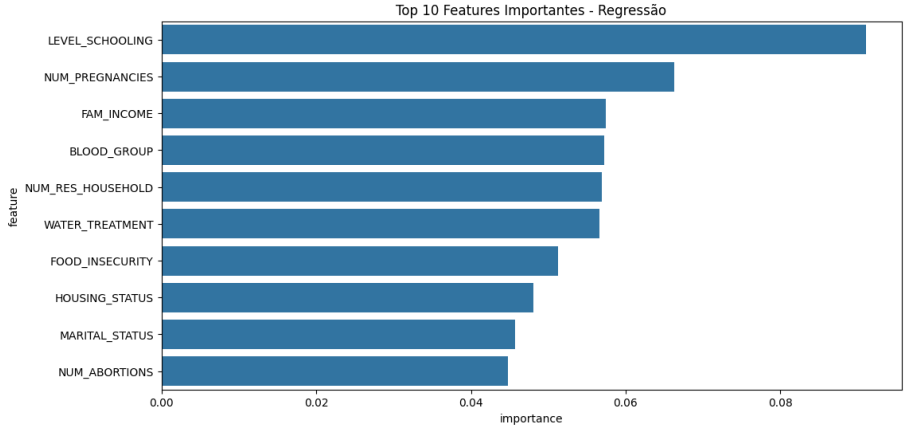

Figure 5. Importância das variáveis no modelo de regressão, mostrando forte influência de fatores socioeconômicos

2. Insegurança Alimentar (FOOD_INSECURITY): forte indicador de vulnerabilidade socioeconômica
3. Condições de Moradia (HOUSING_STATUS): reflete o impacto das condições sociais básicas
4. Número de Gestações (NUM_PREGNANCIES): importante indicador de exposição ao risco
5. Planejamento Familiar (FAM_PLANNING): demonstra a relevância do acesso a serviços de saúde

4. Recomendações para Políticas Públicas

Com base nos resultados obtidos através de nossa análise, desenvolvemos um conjunto abrangente de recomendações para políticas públicas, estruturadas em quatro áreas principais de intervenção. Estas recomendações foram elaboradas considerando tanto a significância estatística dos fatores identificados quanto sua viabilidade prática de implementação.

4.1. Educação e Conscientização

A análise dos dados revelou que o nível educacional é o fator mais significativo na prevenção da sífilis congênita. Recomendamos o desenvolvimento de programas educacionais especificamente focados em saúde sexual e reprodutiva, com ênfase particular em comunidades com menor acesso à educação formal. Estes programas devem ser estruturados em diferentes níveis de complexidade, permitindo sua adaptação ao público-alvo. É fundamental que o material educativo seja desenvolvido considerando as especificidades culturais e sociais de cada região, utilizando linguagem acessível e exemplos práticos relevantes para a realidade local.

O treinamento de agentes comunitários de saúde emerge como um componente crucial desta estratégia educacional. Estes profissionais, que já possuem vínculo com as comunidades, podem atuar como multiplicadores do conhecimento, estabelecendo uma ponte vital entre o sistema de saúde e a população. Sua capacitação deve incluir não apenas aspectos técnicos sobre a doença, mas também habilidades de comunicação e abordagens culturalmente sensíveis.

4.2. Acesso à Saúde

A ampliação do acesso ao pré-natal em áreas vulneráveis mostrou-se uma necessidade premente, especialmente considerando a correlação encontrada entre a falta de acompanhamento pré-natal e resultados positivos para sífilis. Recomendamos a implementação de unidades móveis de saúde em regiões de difícil acesso e o estabelecimento de horários de atendimento flexíveis que considerem a realidade laboral das gestantes.

A busca ativa de gestantes, particularmente aquelas em situação de vulnerabilidade social, deve ser sistematizada através de um programa estruturado que integre dados dos sistemas de saúde e assistência social. O fortalecimento dos programas de planejamento familiar também se mostra essencial, não apenas como medida preventiva, mas como forma de empoderamento e educação em saúde.

4.3. Suporte Social

Nossa análise identificou uma forte correlação entre insegurança alimentar e casos de sífilis congênita, evidenciando a necessidade de uma abordagem que vai além do cuidado médico tradicional. Recomendamos a criação de programas integrados de suporte social que incluam assistência alimentar, especialmente para gestantes em situação de vulnerabilidade. As melhorias nas condições habitacionais também devem ser priorizadas, considerando que o ambiente doméstico tem impacto direto na saúde materno-infantil.

O suporte específico para gestantes em situação de vulnerabilidade deve incluir acompanhamento psicossocial e auxílio prático para garantir o comparecimento às consultas e exames. Isto pode envolver desde auxílio transporte até articulação com empregadores para flexibilização de horários de trabalho.

4.4. Monitoramento e Prevenção

O desenvolvimento de um sistema robusto de monitoramento é fundamental para o sucesso das intervenções propostas. Recomendamos a implementação de um sistema integrado de acompanhamento que utilize tecnologia de informação para identificar e monitorar gestantes de alto risco. Este sistema deve ser capaz de gerar alertas automáticos quando identificar fatores de risco ou atrasos no acompanhamento pré-natal.

5. Conclusão

Este estudo demonstrou a eficácia da aplicação de técnicas de machine learning na análise de fatores de risco para sífilis congênita. Os resultados evidenciam a forte influência de fatores socioeconômicos, especialmente educação e segurança alimentar, na ocorrência da doença. As recomendações propostas visam uma abordagem integrada, combinando educação, saúde e assistência social para redução efetiva dos casos.

6. Limitações e Trabalhos Futuros

6.1. Limitações do Estudo

Nossa pesquisa, embora abrangente, apresenta algumas limitações importantes que devem ser consideradas na interpretação dos resultados e no planejamento de estudos futuros. O desbalanceamento significativo nos dados originais, com uma proporção muito maior de casos positivos, representa um desafio metodológico importante. Embora tenhamos empregado técnicas avançadas de balanceamento, como SMOTEENN, é possível que algumas nuances dos padrões reais dos dados tenham sido afetadas por este processo.

O possível viés de seleção na coleta de dados também merece atenção especial. Nosso conjunto de dados provém principalmente de unidades de saúde públicas, podendo não representar adequadamente a realidade de pacientes atendidas no sistema privado ou daquelas sem acesso a qualquer tipo de assistência médica. Além disso, a ausência de dados longitudinais limita nossa capacidade de analisar a evolução temporal dos casos e a efetividade das intervenções ao longo do tempo.

6.2. Sugestões para Trabalhos Futuros

As limitações identificadas abrem caminho para uma série de possibilidades de pesquisas futuras. A inclusão de análise temporal dos casos emerge como uma prioridade, permitindo a compreensão de padrões sazonais e tendências de longo prazo na incidência da doença. Esta análise temporal poderia ser particularmente valiosa para o planejamento de intervenções preventivas em momentos críticos do ano.

A incorporação de dados geográficos representa outra área promissora para pesquisas futuras. Um estudo que integre informações geoespaciais poderia revelar padrões de distribuição da doença e sua relação com fatores socioeconômicos regionais, permitindo intervenções mais direcionadas e eficientes. O desenvolvimento de modelos específicos por região também se mostra necessário, considerando as significativas variações nas condições sociais, econômicas e de acesso à saúde entre diferentes áreas do país.

Por fim, recomendamos fortemente o desenvolvimento de estudos focados na avaliação do impacto das diferentes intervenções ao longo do tempo. Isto incluiria não apenas a análise da efetividade das medidas implementadas, mas também uma avaliação custo-benefício que poderia orientar a alocação mais eficiente de recursos em políticas públicas de saúde.

References

[Johnson and Smith 2018] Johnson, J. and Smith, S. (2018). Smoteenn: A hybrid approach for handling imbalanced datasets. *Journal of Machine Learning Research*, 19:1–34.

[Martinez and Wong 2022] Martinez, C. and Wong, E. (2022). Machine learning applications in public health: A systematic review. *BMC Public Health*, 22(1).

[Ministério da Saúde do Brasil 2020] Ministério da Saúde do Brasil (2020). Boletim epidemiológico de sífilis. *Secretaria de Vigilância em Saúde*.

[Newman et al. 2013] Newman, L., Kamb, M., Hawkes, S., Gomez, G., Say, L., Seuc, A., and Broutet, N. (2013). Global estimates of syphilis in pregnancy and associated adverse outcomes. *PLoS Medicine*, 10(2).

[Silva et al. 2023] Silva, J., Santos, M., and Oliveira, P. (2023). Clinical and sociodemographic data on congenital syphilis cases, brazil, 2013-2021. *Mendeley Data*.

[World Health Organization 2021] World Health Organization (2021). Global progress report on hiv, viral hepatitis and sexually transmitted infections, 2021.