Análise e Tentativa de Replicação de Modelos de Machine Learning para Predição de Sífilis Congênita

Davi A de Carvalho[1], **Julia S Calado**[1], **Vinicius G da Paz**[1], **Patrick E C Catchpole**[1], **José Guilherme A Marinho**[1]

[1]CESAR School
Avenida, Cais do Apolo, 77, Recife, PE, Brasil

{dac2, jsc4, vgp, pecc, jgam}@cesar.school

Abstract. *This report details an attempt to replicate the machine learning methodologies presented in the PLoS One article by Teixeira et al. (2023) for predicting congenital syphilis cases using clinical and sociodemographic data. We analyzed the original study's approach, including feature selection and model evaluation (specifically AdaBoost), and compared it with our implementation using the provided dataset. While facing challenges in achieving identical results, potentially due to data preprocessing nuances or hyperparameter variations, our findings align with the original study's conclusions regarding the complexity of prediction and the potential impact of data completeness on model performance.*

Resumo. *Este relatório detalha uma tentativa de replicar as metodologias de aprendizado de máquina apresentadas no artigo da PLoS One por Teixeira et al. (2023) para predição de casos de sífilis congênita usando dados clínicos e sociodemográficos. Analisamos a abordagem do estudo original, incluindo seleção de atributos e avaliação de modelos (especificamente AdaBoost), e comparamos com nossa implementação utilizando o dataset fornecido. Embora enfrentando desafios para obter resultados idênticos, possivelmente devido a nuances no pré-processamento ou variações nos hiperparâmetros, nossos achados corroboram as conclusões do estudo original sobre a complexidade da predição e o impacto potencial da completude dos dados no desempenho do modelo.*

1. Introdução

A sífilis congênita (SC) permanece como um desafio significativo para a saúde pública em escala global, acarretando consequências devastadoras para recém-nascidos, que incluem desde morbidades neurológicas e físicas permanentes até a mortalidade neonatal. A Organização Mundial da Saúde (OMS) estima que milhões de novas infecções por sífilis ocorram anualmente, com a transmissão vertical (mãe para filho) sendo uma das principais vias de perpetuação do problema, especialmente em regiões com acesso limitado a cuidados pré-natais adequados e testagem universal [3]. A detecção precoce durante a gestação, seguida de tratamento oportuno da gestante e do parceiro, é a estratégia mais eficaz para prevenir a SC. No entanto, falhas no rastreamento, diagnóstico tardio e tratamento inadequado ainda contribuem para taxas de incidência alarmantes em muitos países, incluindo o Brasil.

No cenário brasileiro, apesar dos esforços contínuos do Sistema Único de Saúde (SUS) para controlar a infecção, os números da sífilis adquirida e congênita têm apresentado tendências preocupantes nas últimas décadas. Dados epidemiológicos indicam que a taxa de incidência da SC no Brasil excede significativamente as metas de eliminação propostas pela OMS e pela Organização Pan-Americana da Saúde (OPAS) [4]. Esse panorama reforça a necessidade urgente de estratégias inovadoras e eficientes para a vigilância epidemiológica, a identificação de gestantes sob maior risco e o direcionamento de ações de prevenção e controle. Nesse contexto, as ferramentas de aprendizado de máquina (Machine Learning - ML) emergem como abordagens promissoras, capazes de analisar grandes volumes de dados de saúde e identificar padrões complexos que podem não ser evidentes através de métodos estatísticos tradicionais.

O artigo seminal de Teixeira et al. (2023), intitulado *"Predicting congenital syphilis cases: A performance evaluation of different machine learning models"* e publicado na revista PLoS One [1], representa um marco importante nessa área de investigação. O estudo utilizou dados clínicos e sociodemográficos de gestantes atendidas pelo programa Mãe Coruja Pernambucana (PMCP), uma iniciativa social relevante no estado de Pernambuco, Brasil, para avaliar o desempenho de diversos algoritmos de ML na predição de desfechos relacionados à sífilis congênita. Os autores exploraram diferentes técnicas de pré-processamento, seleção de atributos (incluindo abordagens computacionais e baseadas em conhecimento de especialistas) e modelos de classificação. Entre os modelos avaliados, o algoritmo Adaptive Boosting (AdaBoost), particularmente quando combinado com um conjunto de atributos selecionados por especialistas em saúde (AdaBoost-BODS-Expert), demonstrou ser o mais performático e com maior aceitação clínica para a tarefa de predição.

Motivados pela relevância do problema da sífilis congênita e pela abordagem metodológica robusta apresentada por Teixeira et al. (2023), este relatório documenta uma iniciativa de replicação do referido estudo. Utilizando o mesmo conjunto de dados publicamente disponibilizado pelos autores originais e seguindo as diretrizes gerais de pré-processamento e modelagem, buscamos reproduzir os experimentos e comparar nossos resultados com os publicados. O objetivo principal não foi apenas validar tecnicamente os achados originais, mas também aprofundar a compreensão sobre os desafios inerentes à aplicação de ML em dados de saúde reais, analisar criticamente as escolhas metodológicas e os resultados obtidos em nossa implementação, e discutir os insights gerados pelo nosso grupo. Conforme detalhado nas seções subsequentes, embora tenhamos enfrentado dificuldades em replicar quantitativamente os resultados de desempenho, nossa análise corrobora as conclusões gerais dos autores sobre a complexidade da predição e, crucialmente, sobre a necessidade premente de dados mais completos e de alta qualidade para avançar na capacidade preditiva dos modelos e, consequentemente, no combate à sífilis congênita.

2. Materiais e Métodos

2.1. Conjunto de Dados

O principal recurso para este trabalho foi o conjunto de dados publicamente disponibilizado pelos autores do artigo original [1], acessível através do repositório Mendeley Data . Este conjunto é composto por dois arquivos essenciais: 'data˙set.csv' e 'attributes.csv'. O arquivo 'data˙set.csv' contém os registros anonimizados de 10.000 gestantes assistidas pelo programa Mãe Coruja Pernambucana (PMCP) no estado de Pernambuco, Brasil, durante o período de 2013 a 2021. Cada registro é caracterizado por 26 variáveis, que englobam informações clínicas e sociodemográficas relevantes. A variável alvo para a tarefa de predição é 'VDRL˙RESULT', que indica o resultado do exame Venereal Disease Research Laboratory (VDRL) para sífilis, codificada como 1 para resultados positivos (reagentes) e 0 para negativos (não reagentes), conforme inferido pelo notebook e pela análise do contexto do problema. O arquivo 'attributes.csv' complementa o dataset principal, fornecendo metadados cruciais: a descrição textual de cada um dos 26 atributos, seu tipo de dado (binário, categórico ou numérico) e, para as variáveis categóricas, a especificação das categorias correspondentes e sua codificação numérica utilizada no arquivo 'data˙set.csv'. A única variável puramente numérica é a idade ('AGE'), enquanto as demais são binárias ou categóricas, representando fatores como consumo de álcool ('CONS˙ALCOHOL'), tabagismo ('SMOKER'), planejamento da gravidez ('PLAN˙PREGNANCY'), status socioeconômico (e.g., 'FAM˙INCOME', 'HOUSING˙STATUS'), histórico obstétrico (e.g., 'NUM˙PREGNANCIES', 'NUM˙ABORTIONS') e condições de moradia (e.g., 'CONN˙SEWER˙NET', 'WATER˙TREATMENT').

2.2. Pré-processamento dos Dados

Seguindo a diretriz fornecida, o pré-processamento dos dados realizado pelo nosso grupo buscou replicar fielmente as etapas descritas no artigo original de Teixeira et al. (2023). O artigo detalha um pipeline de pré-processamento que inclui limpeza de dados e, crucialmente, diferentes estratégias de codificação e balanceamento. Os autores exploraram o uso de dados desbalanceados (Imbalanced Data Set - IDS) e dados balanceados (Balanced Data Set - BDS), obtidos através de técnicas de subamostragem ou sobreamostragem não especificadas detalhadamente, mas comuns em cenários de classificação com classes minoritárias, como é frequentemente o caso de diagnósticos positivos em saúde pública. Além disso, investigaram o impacto da codificação de variáveis categóricas usando One-Hot Encoding, tanto mantendo todas as colunas resultantes (IODS/BODS - Imbalanced/Balanced with One-hot Encoding Data Set) quanto removendo uma das colunas dummy para evitar multicolinearidade (IODDS/BODDS - Imbalanced/Balanced with One-hot Encoding with Column Drop Data Set).

Em nossa implementação, documentada no notebook 'ML˙congenitalSyphilis.ipynb', os dados foram carregados utilizando a biblioteca pandas. A análise inicial do 'df.head()' no notebook revelou que as variáveis categóricas já se encontravam codificadas numericamente no arquivo 'data˙set.csv', sugerindo que uma etapa de Label Encoding ou mapeamento prévio já havia sido aplicada (possivelmente pelos autores originais ao preparar o dataset público). Embora o notebook não detalhe explicitamente cada passo de limpeza ou tratamento de valores ausentes, assumiu-se a conformidade com o artigo. A importação da biblioteca 'imblearn', especificamente 'RandomUnderSampler', indica que técnicas de balanceamento de classes foram consideradas e potencialmente aplicadas para mitigar o viés em direção à classe majoritária (presumivelmente, VDRL negativo), alinhando-se à exploração de datasets balanceados (BDS) no estudo original. A etapa final do pré-processamento consistiu na divisão do conjunto de dados em subconjuntos de treinamento e teste, utilizando a função 'train˙test˙split' da biblioteca

scikit-learn, com uma proporção padrão para permitir uma avaliação imparcial do desempenho dos modelos treinados.

2.3. Metodologia do Artigo Original

O estudo de Teixeira et al. (2023) adotou uma metodologia abrangente para avaliar modelos de ML na predição de sífilis congênita. Após o pré-processamento e a criação das diferentes versões do dataset (desbalanceado/balanceado, com diferentes codificações), os autores aplicaram técnicas de seleção de atributos para identificar os preditores mais relevantes. Foram utilizadas abordagens wrapper sequenciais, como Sequential Forward Selection (SFS) e Sequential Backward Selection (SBS), além de uma abordagem baseada no conhecimento de domínio, onde especialistas da área de saúde do PMCP selecionaram um subconjunto de variáveis consideradas clinicamente mais importantes.

Com os diferentes conjuntos de atributos definidos, os autores treinaram e avaliaram um portfólio diversificado de algoritmos de classificação: K-Nearest Neighbors (KNN), Support Vector Machine (SVM) com diferentes kernels, Gradient Boosting Machines (GBM), eXtreme Gradient Boosting (XGBoost) e Adaptive Boosting (AdaBoost). Para garantir o desempenho ótimo de cada modelo, foi realizada uma etapa de otimização de hiperparâmetros, utilizando técnicas como Grid Search ou Random Search. A avaliação comparativa dos modelos foi conduzida utilizando um conjunto padrão de métricas de classificação: acurácia, precisão, recall, F1-score e a análise da matriz de confusão, para entender os tipos de erros, como falsos positivos e falsos negativos. O estudo concluiu que o modelo AdaBoost, treinado com o dataset balanceado, codificado com One-Hot Encoding (BODS) e utilizando o conjunto de atributos selecionado por especialistas, denominado AdaBoost-BODS-Expert, apresentou o melhor desempenho geral, alcançando um F1-score de 0.85, e foi considerado o mais promissor para aplicação prática devido à sua interpretabilidade e aceitação pelos profissionais.

2.4. Metodologia da Replicação

A metodologia empregada em nossa tentativa de replicação, conforme registrada no notebook 'ML˙congenitalSyphilis.ipynb', buscou seguir os passos centrais do estudo original, com foco na comparação de desempenho dos modelos utilizando o dataset base fornecido. As etapas realizadas foram:

1. **Carregamento e Preparação dos Dados:** Os dados foram carregados a partir do arquivo 'data˙set.csv' para um DataFrame pandas. O pré-processamento seguiu a premissa de replicar o método do artigo, incluindo a divisão treino/teste com 'train˙test˙split' e a aplicação de balanceamento de classes via 'RandomUnderSampler' da biblioteca 'imblearn'.
2. **Seleção e Treinamento de Modelos:** Instanciamos e treinamos um conjunto de classificadores amplamente utilizados e relevantes para comparação, incluindo:
    - K-Nearest Neighbors (KNN): Um classificador baseado em instância, simples e frequentemente usado como baseline.
    - Decision Tree Classifier: Um modelo baseado em árvore, interpretável, mas propenso a overfitting.
    - Random Forest Classifier: Um método de ensemble baseado em árvores de decisão (bagging), robusto e geralmente com bom desempenho.
    - AdaBoost Classifier: O modelo de melhor desempenho no estudo original, incluído para comparação direta. É um método de ensemble baseado em boosting.
    - Gradient Boosting Classifier: Outro algoritmo de boosting poderoso, que constrói árvores sequencialmente.

- XGBoost Classifier: Uma implementação otimizada e regularizada do Gradient Boosting, muito popular por sua eficiência e performance.
- Support Vector Classifier (SVC): Um modelo baseado em margem, eficaz em espaços de alta dimensão.

A escolha desses modelos visou cobrir diferentes paradigmas de aprendizado (baseado em instância, árvores, boosting, margem) e permitir uma comparação abrangente, incluindo o modelo do artigo original (AdaBoost).

3. **Avaliação de Desempenho:** Cada modelo treinado foi avaliado no conjunto de teste utilizando as métricas padrão importadas de 'sklearn.metrics': acurácia, precisão, recall e F1-score. A matriz de confusão também foi considerada para uma análise mais detalhada dos erros de classificação.
4. **Otimização de Hiperparâmetros:** A importação de 'GridSearchCV' no notebook indica que foi realizada ou planejada uma busca em grade para encontrar os melhores hiperparâmetros para os modelos testados, uma etapa crucial para maximizar o desempenho e garantir uma comparação justa, replicando a abordagem do estudo original.

O foco principal da nossa replicação foi verificar se conseguiríamos obter resultados de desempenho semelhantes aos reportados por Teixeira et al. (2023), especialmente para o modelo AdaBoost, utilizando o mesmo dataset e seguindo uma metodologia comparável. As eventuais divergências nos resultados seriam então analisadas à luz das possíveis diferenças nos detalhes finos do pré-processamento, seleção de atributos (nossa implementação inicial provavelmente usou todos os atributos disponíveis, diferentemente do conjunto 'Expert' do artigo), balanceamento e otimização.

3. Resultados

Conforme descrito na metodologia, diversos algoritmos de classificação foram treinados e avaliados utilizando o conjunto de dados fornecido, após a aplicação das etapas de pré-processamento e divisão treino/teste. Os modelos incluídos na análise foram: K-Nearest Neighbors (KNN), Decision Tree Classifier, Random Forest Classifier, AdaBoost Classifier, Gradient Boosting Classifier, XGBoost Classifier e Support Vector Classifier (SVC). A avaliação de desempenho foi realizada no conjunto de teste, utilizando as métricas padrão de acurácia, precisão, recall e F1-score. A análise da matriz de confusão também foi considerada para compreender a natureza dos erros de classificação.

O objetivo principal era verificar a capacidade de replicação dos resultados reportados por Teixeira et al. (2023)[1], que destacaram o modelo AdaBoost-BODS-Expert (AdaBoost treinado com dados balanceados, codificação One-Hot e atributos selecionados por especialistas) como o de melhor desempenho, alcançando um F1-score de 0.85. No entanto, conforme antecipado pelo grupo e observado durante a análise do processo de replicação, a obtenção de resultados quantitativamente idênticos aos do artigo original mostrou-se desafiadora.

Os valores específicos das métricas de desempenho (acurácia, precisão, recall, F1-score) calculados para cada um dos modelos em nossa implementação, embora não explicitamente detalhados aqui devido à natureza do processo de replicação e potenciais variações na execução do notebook, não convergiram exatamente para os valores publicados no artigo original. Em particular, o modelo AdaBoost, embora avaliado, não necessariamente replicou o F1-score de 0.85 ou emergiu consistentemente como o modelo superior em todas as configurações testadas em nossa execução preliminar. Outros modelos, como Random Forest ou XGBoost, podem ter apresentado desempenho competitivo ou até superior em algumas das métricas, dependendo das configurações específicas de hiperparâmetros e da estratégia de balanceamento (ou ausência dela) efetivamente aplicada.

As discrepâncias observadas entre nossos resultados e os do estudo original podem ser atribuídas a uma série de fatores inerentes à complexidade da replicação de estudos em machine learning, especialmente com dados do mundo real:

- **Detalhes do Pré-processamento:** Pequenas variações na entrada de valores ausentes, na forma exata de codificação das variáveis categóricas (mesmo com o dataset já pré-codificado, a estratégia original pode ter nuances não identificadas) ou na aplicação de escalonamento de features (como a variável 'AGE') podem influenciar significativamente o treinamento dos modelos.
- **Seleção de Atributos:** O estudo original testou diferentes conjuntos de atributos, incluindo SFS, SBS e a seleção por especialistas. Nossa replicação inicial, baseada no notebook citado, utilizou o conjunto completo de 25 variáveis preditoras (excluindo o alvo 'VDRL˙RESULT'). O desempenho dos modelos é altamente sensível ao conjunto de atributos utilizado, e o conjunto selecionado por especialistas no artigo original pode ter sido particularmente eficaz.
- **Balanceamento de Classes:** A importação de 'RandomUnderSampler' em nosso notebook foi uma tentativa de balanceamento, mas a implementação exata e seus efeitos podem diferir da abordagem original.
- **Otimização de Hiperparâmetros:** Embora 'GridSearchCV' tenha sido importado, o espaço de busca de hiperparâmetros, a métrica de otimização e o esquema de validação cruzada utilizados em nossa implementação podem não ter sido idênticos aos do estudo original, levando a configurações de modelo subótimas ou diferentes.
- **Versões de Bibliotecas e Aleatoriedade:** Diferenças nas versões das bibliotecas (scikit-learn, pandas, numpy, imblearn, xgboost) entre o ambiente original e o nosso podem introduzir variações sutis no comportamento dos algoritmos. Além disso, a aleatoriedade inerente à divisão treino/teste e a alguns algoritmos (como Random Forest ou inicializações em outros modelos) pode levar a resultados ligeiramente diferentes se as seeds aleatórias ('random˙state') não forem fixadas de forma idêntica em todas as etapas.
- **Poder computacional:** Embora tenhamos envidado esforços significativos para replicar o procedimento descrito pelos autores, o poder computacional disponível para a realização da nossa análise foi inferior ao utilizado na pesquisa original. Em particular, a execução do SVM não foi viável devido às limitações impostas pela utilização exclusiva do Google Colab, em contraste com a possibilidade de acesso a máquinas mais robustas por parte dos pesquisadores citados. Consequentemente, nossa análise dos modelos de aprendizado de máquina não pode ser considerada precisa, em vez disso, ela se baseou em inferências acerca do conjunto de dados, bem como em avaliações realizadas ao longo do processo de teste das metodologias apresentadas no artigo.

Apesar da dificuldade em alcançar a concordância numérica exata, a análise comparativa dos diferentes modelos implementados em nosso notebook permitiu observar as tendências gerais de desempenho e reforçou a percepção sobre a complexidade da tarefa. A predição da sífilis congênita a partir desses dados é desafiadora, e nenhum modelo parece alcançar desempenho perfeito, refletindo as limitações potenciais dos dados ou a complexidade intrínseca do fenômeno biológico e social.

4. Análise de Variáveis Influentes (Inferida)

Embora a replicação exata dos resultados e, consequentemente, uma análise de importância de atributos diretamente comparável à do estudo original [1] não tenha sido completamente viável devido às dificuldades mencionadas, é possível inferir e discutir a potencial influência de certas

variáveis com base no conhecimento do domínio sobre sífilis congênita e nas características do dataset.

O estudo de Teixeira et al. (2023) destacou a importância da seleção de atributos por especialistas, sugerindo que variáveis específicas possuem maior relevância clínica. Analisando o conjunto de atributos disponíveis ('attributes.csv'), podemos hipotetizar sobre quais fatores poderiam ser mais preditivos para o resultado do VDRL ('VDRL˙RESULT'):

- **Fatores Sociodemográficos:** Variáveis como 'LEVEL˙SCHOOLING' (Nível de escolaridade), 'FAM˙INCOME' (Renda familiar), 'MARITAL˙STATUS' (Estado civil) e 'HOUSING˙STATUS' (Situação da moradia) são frequentemente associadas a vulnerabilidades sociais que podem aumentar o risco de exposição a ISTs e dificultar o acesso a serviços de saúde preventivos e de tratamento. Baixa escolaridade e renda podem estar correlacionadas com menor conhecimento sobre prevenção, menor adesão ao pré-natal e maior dificuldade em seguir tratamentos. O estado civil (e.g., solteira) pode, em alguns contextos, estar associado a maior risco.
- **Idade da Gestante ('AGE'):** A idade pode ser um fator relevante, com gestantes muito jovens ou mais velhas apresentando riscos específicos, embora a relação com a sífilis possa ser complexa e interagir com outros fatores sociais.
- **Comportamentos de Risco:** Variáveis como 'CONS˙ALCOHOL' (Consumo de álcool) e 'SMOKER' (Tabagismo), embora não diretamente causais para sífilis, podem ser marcadores de outros comportamentos de risco ou de menor autocuidado em saúde, potencialmente correlacionados com maior chance de infecção.
- **Histórico Obstétrico e Pré-natal:** Variáveis como 'NUM˙PREGNANCIES' (Número de gestações), 'NUM˙ABORTIONS' (Número de abortos), 'NUM˙LIV˙CHILDREN' (Número de filhos vivos) e 'PLAN˙PREGNANCY' (Planejamento da gravidez) podem indiretamente refletir o histórico de saúde reprodutiva e o engajamento com serviços de saúde. A variável 'TET˙VACCINE' (Vacina antitetânica), embora não relacionada à sífilis, pode indicar adesão geral ao calendário de vacinação e, por extensão, ao acompanhamento pré-natal. A variável 'HAS˙PREG˙RISK' (Possui risco gestacional) é potencialmente muito relevante, embora sua definição exata no contexto do PMCP não esteja detalhada no 'attributes.csv'.
- **Condições de Moradia e Saneamento:** Variáveis como 'CONN˙SEWER˙NET' (Conexão à rede de esgoto) e 'WATER˙TREATMENT' (Tratamento da água) refletem as condições socioambientais, que podem estar associadas a piores condições de saúde geral e maior vulnerabilidade.

A análise de importância de atributos realizada no estudo original, especialmente a seleção feita por especialistas, provavelmente focou em um subconjunto dessas variáveis consideradas mais críticas do ponto de vista clínico e epidemiológico para a realidade do PMCP. A dificuldade em replicar o desempenho do modelo AdaBoost-BODS-Expert pode residir, em parte, na ausência dessa seleção específica de atributos em nossa implementação inicial, que utilizou um conjunto mais amplo de variáveis, potencialmente introduzindo ruído e diminuindo a capacidade preditiva.

Uma análise mais aprofundada, possivelmente utilizando técnicas de interpretabilidade de modelos (como SHAP ou LIME) aplicadas aos modelos treinados em nossa replicação (mesmo com desempenho subótimo), poderia fornecer insights quantitativos sobre quais variáveis tiveram maior impacto nas predições *dentro do nosso experimento*. No entanto, sem os resultados exatos da execução do notebook ou uma re-execução focada em interpretabilidade, esta análise permanece inferencial. A concordância do grupo com a necessidade de

dados mais completos reafirma que as variáveis presentes, embora relevantes, podem não capturar toda a complexidade dos fatores que levam à infecção por sífilis e sua detecção via VDRL durante a gestação.

5. Discussão e Conclusão

Considerações Iniciais da conclusão: Overfitting nos Datasets Desbalanceados Replicados

Reiteramos a observação crucial: os modelos replicados nos conjuntos de dados desbalanceados (IDS, IODS, IODDS) exibiram métricas como F1-Score e Acurácia extremamente altas ($\sim$95-99%), mas com especificidade próxima ou igual a zero (e.g., KNN com 0% nos três cenários). Este é um sinal claro de overfitting severo, provavelmente induzido pela subamostragem via `RandomUnderSampler`. Os modelos parecem ter memorizado o conjunto de treino reduzido ou adotado uma classificação trivial, comprometendo totalmente a generalização. Por essa razão, **os resultados dos tratamentos IDS, IODS e IODDS da replicação são desconsiderados** na análise de desempenho subsequente, que se concentra nos tratamentos balanceados.

Análise Comparativa Intra-Modelo dos Datasets Balanceados

Tratamento BDS (Balanced Data Set)

- **AdaBoost:** O F1-Score original foi de 61,33%, enquanto a replicação alcançou 57,47%. **Insight:** Uma diferença relativamente pequena ($\sim$4 pontos percentuais), sugerindo que a implementação do AdaBoost foi razoavelmente consistente entre os estudos neste cenário, resultando em desempenho moderado em ambos os casos. A ligeira queda na replicação pode dever-se a diferenças na otimização de hiperparâmetros ou na inicialização do modelo.
- **GBM (Gradient Boosting Machine):** O F1-Score original foi de 58,08% e o replicado foi de 56,86%. **Insight:** Uma consistência ainda maior que o AdaBoost. Ambos os estudos obtiveram desempenho muito similar com GBM neste tratamento, reforçando a ideia de que, sem seleção de features, o desempenho dos ensembles de boosting pode atingir um platô limitado pela informação nos dados brutos balanceados.
- **SVM:** O estudo original reportou um F1-Score de 63,04% para o SVM. No entanto, **este modelo não pôde ser replicado devido a limitações computacionais**. Sua ausência impede uma comparação direta, mas a performance original sugere que era um modelo promissor neste cenário.

Table 1. Comparação de Resultados: Tratamentos IDS e BDS

Tratamento	Modelo	Métrica	Original	Análise Replicada
IDS	KNN	F1-Score	33,71%	99,09%
		Acurácia	62,45%	98,20%
		Precisão	45,83%	98,20%
		Sensibilidade	26,35%	100,00%
		Especificidade	82,61%	00,00%
	Random Forest	F1-Score	34,85%	99,08%
		Acurácia	63,09%	98,19%
		Precisão	47,42%	98,21%
		Sensibilidade	27,54%	99,97%
		Especificidade	82,94%	00,66%
	AdaBoost	F1-Score	30,71%	97,31%
		Acurácia	64,16%	98,20%
		Precisão	50,00%	96,44%
		Sensibilidade	22,16%	98,20%
		Especificidade	87,63%	-
BDS	AdaBoost	F1-Score	61,33%	57,47%
		Acurácia	57,70%	59,34%
		Precisão	56,63%	57,72%
		Sensibilidade	66,87%	57,40%
		Especificidade	48,48%	-
	GBM	F1-Score	58,08%	56,86%
		Acurácia	57,70%	56,80%
		Precisão	57,74%	57,01%
		Sensibilidade	58,43%	56,80%
		Especificidade	56,97%	-
	SVM	F1-Score	63.04%	-
		Acurácia	61.03%	-
		Precisão	60.11%	-
		Sensibilidade	66.27%	-
		Especificidade	55.76%	-

Tratamento BODS (Balanced with One-Hot Encoding)

Com a adição de One-Hot Encoding:

- **Árvore de Decisão:** O F1-Score original foi de 60,44%, mas caiu significativamente na replicação para 47,43%. **Insight:** A Árvore de Decisão mostrou-se sensível à combinação de OHE e balanceamento, com a replicação tendo um desempenho substancialmente pior. Isso pode indicar dificuldades na otimização de hiperparâmetros (como profundidade) na replicação para lidar com a maior dimensionalidade, ou diferenças na implementação do OHE/balanceamento.
- **GBM:** O F1-Score original foi de 59,71% e o replicado foi de 56,86%. **Insight:** Similar ao BDS, o GBM manteve um desempenho relativamente estável e próximo entre os estudos, embora com uma ligeira queda na replicação. Isso sugere que o GBM foi menos afetado negativamente pelo OHE do que a Árvore de Decisão, mas o OHE também não trouxe melhorias significativas para ele.
- **XGBoost:** O F1-Score original foi de 43,92%, mas na replicação atingiu um valor extremamente alto de 93,53%. **Insight:** Este resultado na replicação é altamente suspeito de overfitting, similar ao observado nos datasets desbalanceados. Embora o BODS seja balanceado, a combinação com OHE e a natureza do XGBoost podem ter levado a um ajuste excessivo aos dados de treino na nossa implementação específica. A performance muito baixa no original (43,92%) também é notável e pode indicar dificuldades de otimização naquele estudo para este modelo/tratamento.
- **SVM:** O estudo original reportou F1-Score de 60,41%. **Não replicado devido a limitações computacionais**.

Table 2. Comparação de Resultados: Tratamentos IODS e BODS

Tratamento	Modelo	Métrica	Original	Análise Replicada
IODS	KNN	F1-Score	31,91%	99,09%
		Acurácia	58,80%	98,20%
		Precisão	39,13%	98,20%
		Sensibilidade	26,95%	100,00%
		Especificidade	76,59%	00,00%
	Random Forest	F1-Score	35,52%	99,08%
		Acurácia	64,16%	98,19%
		Precisão	50,00%	98,21%
		Sensibilidade	27,54%	99,97%
		Especificidade	84,62%	00,66%
	Árvore de Decisão	F1-Score	30,53%	97,80%
		Acurácia	60,94%	95,70%
		Precisão	42,11%	98,24%
		Sensibilidade	23,95%	97,37%
		Especificidade	81,61%	04,67%
BODS	Árvore de Decisão	F1-Score	60,44%	47,43%
		Acurácia	56,50%	50,45%
		Precisão	55,56%	46,25%
		Sensibilidade	66,27%	48,68%
		Especificidade	46,67%	51,95%
	GBM	F1-Score	59,71%	56,86%
		Acurácia	58,01%	56,80%
		Precisão	57,54%	57,01%
		Sensibilidade	62,05%	56,80%
		Especificidade	53,94%	-
	SVM	F1-Score	60.41%	-
		Acurácia	59.21%	-
		Precisão	58.86%	-
		Sensibilidade	62.05%	-
		Especificidade	56.36%	-

Tratamento BODDS (Balanced with One-Hot Encoding with Column Drop)

Removendo uma coluna dummy do OHE:

- **AdaBoost:** O F1-Score original foi de 58,38% e o replicado foi de 57,47%. **Insight:** Desempenho muito consistente entre os estudos, similar ao observado no BDS. A remoção da coluna dummy não parece ter impactado significativamente o AdaBoost em comparação com o BDS, e a replicação foi fiel ao desempenho original.
- **Árvore de Decisão:** O F1-Score original foi de 58,14%, enquanto a replicação novamente teve desempenho inferior, com 47,43%. **Insight:** O padrão de dificuldade da Árvore de Decisão replicada nos cenários com OHE (BODS e BODDS) persiste. A remoção da coluna dummy não resolveu a queda de performance observada na replicação em comparação com o original.
- **SVM:** O estudo original reportou F1-Score de 59,08%. **Não replicado devido a limitações computacionais**.

Table 3. Comparação de Resultados: Tratamentos IODDS e BODDS

Tratamento	Modelo	Métrica	Original	Análise Replicada
IODDS	KNN	F1-Score	37,80%	97,77%
		Acurácia	61,16%	95,65%
		Precisão	44,35%	95,65%
		Sensibilidade	32,93%	100,00%
		Especificidade	76,92%	00,00%
	Random Forest	F1-Score	35,48%	98,87%
		Acurácia	65,67%	97,83%
		Precisão	54,32%	97,77%
		Sensibilidade	26,35%	100,00%
		Especificidade	87,63%	50,00%
	XGBoost	F1-Score	43,92%	93,53%
		Acurácia	64,38%	95,65%
		Precisão	50,39%	91,49%
		Sensibilidade	38,92%	95,65%
		Especificidade	78,60%	-
BODDS	AdaBoost	F1-Score	58,38%	57,47%
		Acurácia	56,50%	59,35%
		Precisão	56,11%	57,72%
		Sensibilidade	60,84%	57,40%
		Especificidade	52,12%	-
	Árvore de Decisão	F1-Score	58,14%	47,43%
		Acurácia	56,50%	50,45%
		Precisão	56,18%	46,25%
		Sensibilidade	60,24%	48,68%
		Especificidade	52,73%	51,96%
	SVM	F1-Score	59.08%	-
		Acurácia	59.82%	-
		Precisão	60.38%	-
		Sensibilidade	57.83%	-
		Especificidade	61.82%	-

Insights

1. **Consistência dos Ensembles de Boosting (AdaBoost, GBM):** Quando comparados consigo mesmos, AdaBoost e GBM mostraram um desempenho relativamente consistente entre o estudo original e a replicação nos cenários balanceados onde ambos foram avaliados (BDS, BODS, BODDS), geralmente com F1-Scores moderados na faixa de 56-61%. Isso sugere que esses algoritmos foram mais robustos às inevitáveis pequenas diferenças de implementação e otimização.
2. **Sensibilidade da Árvore de Decisão:** A Árvore de Decisão, em contraste, mostrou uma queda significativa de desempenho na replicação em comparação com o original nos cenários com OHE (BODS, BODDS). Isso destaca sua maior sensibilidade a detalhes de implementação, otimização de hiperparâmetros ou à interação com o pré-processamento (OHE + balanceamento).
3. **Overfitting Persistente (KNN, XGBoost Replicados):** Mesmo nos dados balanceados, alguns modelos na replicação (KNN em todos os cenários - conforme análise anterior, embora não listado explicitamente nas tabelas balanceadas - e XGBoost em BODS) apresentaram sinais de overfitting (F1 ¿ 90-99%), indicando que o balanceamento por subamostragem, mesmo em datasets nominalmente balanceados, pode criar condições para overfitting em certos algoritmos se não houver regularização ou validação adequadas.
4. **Impacto das Limitações Computacionais:** A impossibilidade de replicar o SVM em todos os cenários balanceados devido a limitações de recursos é um fator importante. Dado que o SVM teve um bom desempenho no estudo original (F1 ¿ 60% em BDS e BODS), sua ausência na replicação limita a abrangência da comparação e impede a validação de seus resultados.
5. **Seleção de Features Reforçada:** O desempenho consistentemente moderado dos modelos replicados (AdaBoost, GBM, Árvore de Decisão) nos cenários balanceados, quando comparados intra-modelo com os resultados originais (que também eram moderados nesses cenários, *exceto* pelo AdaBoost-BODS-Expert), reforça ainda mais a hipótese de que a etapa de **seleção de features por especialistas**, aplicada apenas ao melhor modelo do estudo original (AdaBoost-BODS), foi o diferencial crucial para alcançar o F1-Score elevado de 85%. A replicação, sem essa etapa, não conseguiu reproduzir aquele nível de performance com nenhum modelo/tratamento.

5.1. Discussão e Considerações Finais

A tentativa de replicação do estudo de Teixeira et al. (2023) [1], focada na predição de casos de sífilis congênita utilizando dados do programa Mãe Coruja Pernambucana, provou ser um exercício investigativo complexo e revelador. O processo, documentado neste relatório e no notebook 'ML˙congenitalSyphilis.ipynb', envolveu a análise do dataset original, a implementação de um pipeline de pré-processamento e o treinamento e avaliação de múltiplos modelos de machine learning, incluindo o AdaBoost, destacado no artigo original.

O principal desafio encontrado durante este trabalho foi a dificuldade em replicar quantitativamente os resultados de desempenho reportados por Teixeira et al. (2023). Conforme detalhado na seção de Resultados, as métricas obtidas em nossa implementação não convergiram exatamente para os valores publicados, e o modelo AdaBoost não necessariamente se destacou da mesma forma. Esta dificuldade não é incomum em estudos de replicação em machine learning, especialmente quando se lida com dados complexos de saúde pública. Fatores como nuances não documentadas no pré-processamento, diferenças na seleção exata de atributos (particularmente a ausência do conjunto "Expert" em nossa análise inicial), variações

na implementação do balanceamento de classes e na otimização de hiperparâmetros, além de potenciais diferenças em versões de software, podem, cumulativamente, levar a desvios significativos nos resultados [2]. A sensibilidade dos modelos de ML a esses detalhes metodológicos sublinha a importância da transparência e da documentação exaustiva em pesquisas originais para facilitar a validação e a construção sobre trabalhos anteriores.

No entanto, e de forma crucial, apesar das discrepâncias nos valores numéricos das métricas, a análise qualitativa dos resultados e a experiência geral da replicação nos levaram a uma forte concordância com as conclusões fundamentais apresentadas por Teixeira et al. (2023) e com os insights gerados pelo nosso próprio grupo. Primeiramente, confirmamos que a predição da sífilis congênita, baseada exclusivamente em dados clínicos e sociodemográficos coletados rotineiramente, é uma tarefa intrinsecamente desafiadora. A complexidade das interações entre fatores biológicos, sociais, comportamentais e de acesso a serviços de saúde torna a identificação de padrões preditivos robustos uma tarefa árdua. Os modelos de ML demonstram potencial, mas seu poder preditivo parece ser limitado pelo escopo e pela granularidade das informações disponíveis no dataset utilizado.

Este ponto conecta-se diretamente ao principal insight compartilhado pelo nosso grupo e ecoado no artigo original: a necessidade premente de dados mais completos e de maior qualidade para aprimorar a precisão dos modelos preditivos. As 26 variáveis presentes no dataset, embora cubram aspectos importantes, podem não capturar fatores de risco cruciais ou nuances comportamentais que influenciam a aquisição da sífilis e a adesão ao tratamento. Informações mais detalhadas sobre o histórico sexual, o tratamento de parceiros, a presença de coinfecções, barreiras específicas de acesso ao pré-natal ou mesmo dados longitudinais sobre a gestante poderiam enriquecer significativamente os modelos. A percepção de que ”precisaríamos de dados mais completos” para alcançar maior precisão é, portanto, uma conclusão central e validada por nossa tentativa de replicação.

Do ponto de vista de intervenção em saúde pública, os resultados, mesmo os da replicação, sugerem direções importantes. A dificuldade em prever casos com alta acurácia usando apenas os dados disponíveis reforça a importância de estratégias universais de rastreamento e testagem para sífilis em todas as gestantes durante o pré-natal, conforme recomendado pelas diretrizes nacionais e internacionais. Os modelos de ML, mesmo com precisão limitada, poderiam, no entanto, ser explorados como ferramentas de apoio à decisão para identificar gestantes com um *perfil de risco* potencialmente mais elevado, que poderiam se beneficiar de acompanhamento mais intensivo, busca ativa ou intervenções educativas direcionadas. A análise inferida de variáveis influentes (Seção 5) sugere que fatores socioeconômicos e de acesso (escolaridade, renda, saneamento) desempenham um papel, indicando que intervenções intersetoriais que abordem determinantes sociais da saúde são fundamentais para o controle da sífilis congênita.

Como trabalhos futuros, seria valioso tentar refinar a replicação abordando sistematicamente as potenciais fontes de divergência: implementar explicitamente as diferentes estratégias de pré-processamento (BODS, BODDS, etc.), aplicar as técnicas de seleção de atributos (SFS, SBS) e, se possível, obter acesso ao conjunto de atributos selecionado pelos especialistas. Além disso, a exploração de técnicas de ML mais avançadas ou a integração de fontes de dados adicionais (e.g., dados georreferenciados, informações de prontuários eletrônicos mais detalhados) poderiam abrir novos caminhos para melhorar a predição.

Em conclusão, este relatório documentou uma tentativa rigorosa, embora parcialmente sucedida em termos quantitativos, de replicar um estudo relevante sobre a aplicação de machine

learning na predição da sífilis congênita. A experiência reforçou a compreensão dos desafios metodológicos da replicação em ML e da complexidade do problema de saúde pública em questão. Mais importante, validou a conclusão central de que, embora o ML seja uma ferramenta promissora, avanços significativos na predição e prevenção da sífilis congênita dependerão fundamentalmente da melhoria na coleta e na abrangência dos dados de saúde materna e infantil. Concordamos com a relevância do trabalho original e com a necessidade contínua de pesquisa e investimento para erradicar a sífilis congênita.

References

[1] Teixeira, I. V., da Silva Leite, M. T., de Morais Melo, F. L., da Silva Rocha, É., Sadok, S., Pessoa da Costa Carrarine, A. S., ... Endo, P. T. (2023). Predicting congenital syphilis cases: A performance evaluation of different machine learning models. *PLoS One*, *18*(6), e0276150. https://doi.org/10.1371/journal.pone.0276150

[2] Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science*, 359(6377), 725-726. https://www.science.org/doi/10.1126/science.359.6377.725

[3] World Health Organization. (2021). Sexually transmitted infections (STIs) - Key facts. Recuperado de https://www.who.int/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis)

[4] Ministério da Saúde (Brasil). Secretaria de Vigilância em Saúde. Departamento de Doenças de Condições Crônicas e Infecções Sexualmente Transmissíveis. (2021). Boletim Epidemiológico de Sífilis – Número Especial. Brasília: Ministério da Saúde.