Sífilis Congênita em Pernambuco: Uma Análise Preditiva de Dados Clínicos e Sociodemográficos (2013-2021)

Ana Beatriz Ximenes Alves, Caio Barreto de Albuquerque, Isabela Spinelli, Maria Luisa Arruda, Victor Hora

Graduação em Ciência da Computação - Centro de Estudos e Sistemas Avançados do Recife (CESAR School) - Recife, PE - Brasil

abxa@cesar.school, cba2@cesar.school, isfsca@cesar.school, mlva2@cesar.school, vht@cesat.school

Abstract. *This paper focuses on developing predictive models using clinical and sociodemographic data from the Mãe Coruja Pernambucana Program (PMCP), Pernambuco, Brazil (2013-2021), encompassing 41,762 records. The primary goals were: 1) to predict age using regression techniques and 2) to classify Venereal Disease Research Laboratory (VDRL) test results for congenital syphilis, where positive cases (826 instances, ~1.98%) represent the minority class. The methodology included data preprocessing, application of RandomForestRegressor for age prediction, and a DecisionTreeClassifier with SMOTEENN for VDRL classification. Age prediction yielded an RMSE of approximately 4.16. The VDRL classification model achieved an AUC-ROC of ~0.56, with a recall of ~0.166 and precision of ~0.028 for the positive (syphilis) class. Features like LEVEL_SCHOOLING were identified as important. The study underscores the significant challenge in detecting the minority class (congenital syphilis cases) effectively with the current models and features.*

Resumo. *Este artigo foca no desenvolvimento de modelos preditivos utilizando dados clínicos e sociodemográficos do Programa Mãe Coruja Pernambucana (PMCP), Pernambuco, Brasil (2013-2021), abrangendo 41.762 registros. Os objetivos principais foram: 1) prever a idade utilizando técnicas de regressão e 2) classificar os resultados do teste Venereal Disease Research Laboratory (VDRL) para sífilis congênita, onde os casos positivos (826 instâncias, ~1,98%) representam a classe minoritária. A metodologia incluiu pré-processamento de dados, aplicação do RandomForestRegressor para previsão de idade, e um DecisionTreeClassifier com SMOTEENN para classificação VDRL. A previsão de idade resultou em um RMSE de aproximadamente 4.16. O modelo de classificação VDRL alcançou um AUC-ROC de ~0,56, com um recall de ~0,166 e precisão de ~0,028 para a classe positiva (sífilis). Atributos como LEVEL_SCHOOLING foram identificados como importantes. O estudo sublinha o desafio significativo na detecção eficaz da classe minoritária (casos de sífilis congênita) com os modelos e atributos atuais.*

1. Introdução

A sífilis congênita persiste como um grave problema de saúde pública, e a detecção precoce através de testes como o VDRL (Venereal Disease Research Laboratory) durante o pré-natal e no parto é fundamental. O Programa Mãe Coruja Pernambucana (PMCP) visa oferecer suporte a gestantes e crianças, monitorando seu desenvolvimento. Este estudo utiliza o conjunto de dados "*Clinical and sociodemographic data on congenital syphilis cases, Brazil, 2013-2021*", proveniente de cidades atendidas pelo PMCP no estado de Pernambuco, Brasil, para desenvolver modelos de aprendizado de máquina. Os objetivos são: (1) prever a idade, um indicador demográfico relevante, e (2) classificar o resultado do teste VDRL para sífilis congênita ao nascer, focando na identificação dos casos positivos.

A aplicação de técnicas de aprendizado de máquina pode auxiliar na identificação de fatores de risco associados e na melhoria dos processos de triagem. Este trabalho investiga a performance do algoritmo RandomForest para a tarefa de regressão da idade e do *DecisionTreeClassifier*, com a técnica de reamostragem *Synthetic Minority Oversampling Technique*, combinado com *Edited Nearest Neighbors*, culminando no SMOTEENN para lidar com o desbalanceamento de classes, para a classificação dos resultados do VDRL.

2. Descrição do Conjunto de Dados e do Dataset

O conjunto de dados deste estudo, intitulado "*Clinical and sociodemographic data on congenital syphilis cases, Brazil, 2013-2021*", contém 41.762 registros e 26 atributos. Estes dados abrangem informações clínicas e sociodemográficas sobre cuidados pré-natais, desfechos de gestantes e os resultados do teste VDRL de seus filhos, coletados entre 2013 e 2021 em municípios cobertos pelo PMCP em Pernambuco.

A variável alvo para a tarefa de classificação é o resultado do teste VDRL (VDRL_RESUL), que é um teste de triagem para sífilis congênita. No dataset, existem 826 casos positivos para sífilis congênita (aproximadamente 1.98% do total) e 40.936 casos negativos (aproximadamente 98.02%). Esta distribuição evidencia um severo desbalanceamento de classes, onde a classe de interesse (positiva para sífilis congênita) é minoritária. A variável alvo para regressão é a idade (AGE). A análise exploratória inicial dos dados (conforme visualizações em seu notebook, como histograma de idade e matriz de correlação) fornece insights sobre as distribuições e relações entre os atributos.

2.1 Estrutura do dataset

- **Número de variáveis:** 26.
- **Tipo das Variáveis:** Atributos numéricos, categóricos e binários.
- **Número de registros:** 41.762

- **Período de coleta:** Dados coletados entre 2013 e 2021.

2.2 Atributos do Dataset:

Tabela 1. Atributos do dataset

CATEGORIA	NOME	TIPO	DESCRIÇÃO
Clínico	VDRL RESULT	Binary	Resultado do exame de sífilis VDRL. Valores: Positivo (0), Negativo (1).
	HAS PREG RISK	Categorical	Indica risco na gravidez.
	TET VACCINE	Categorical	Vacinação contra tétano (Positivo, Negativo ou Não informado).
	NUM ABORTIONS	Categorical	Numero de abortos (Nenhum, Um, Dois, Mais de dois).
Sociodemográfico	LEVEL SCHOOLING	Categorical	Escolaridade (Ensino fundamental, médio, superior completo, etc.).
	HOUSING STATUS	Categorical	Condição de moradia (Própria, Alugada, Cedida, etc.).
	FAM INCOME	Categorical	Renda familiar (faixas de valor). Infraestrutura e Ambiente Familiar:
Infraestrutura e Ambiente Familiar	CONN SEWER NET	Categorical	Casa conectada a rede de esgoto.
	TYPE HOUSE	Categorical	Tipo de construção (palha, madeira, alvenaria, etc.).
Regressão	AGE	Numerical	Idade registrada no dataset, utilizada para análise de regressão.

3. Processo de Pré-Processamento:

- **Carregamento e análise inicial:** Verificação da qualidade dos dados, incluindo valores ausentes e outliers.
- **Tratamento de outliers:** Foram removidos valores anômalos na variável AGE, como idades negativas.

- **Tratamento de valores ausentes:** Atributos com valores ausentes foram tratados com categorias como “Não informado”.

4. Análise Exploratória dos Dados (EDA)

Após realizar a etapa de pré-processamento, utilizamos a análise exploratória para preparar o conjunto de dados para nossa análise, que incluiu verificações de distribuição dos dados de estatísticas descritivas como na análise da variável AGE:

4.1. Distribuição da Idade

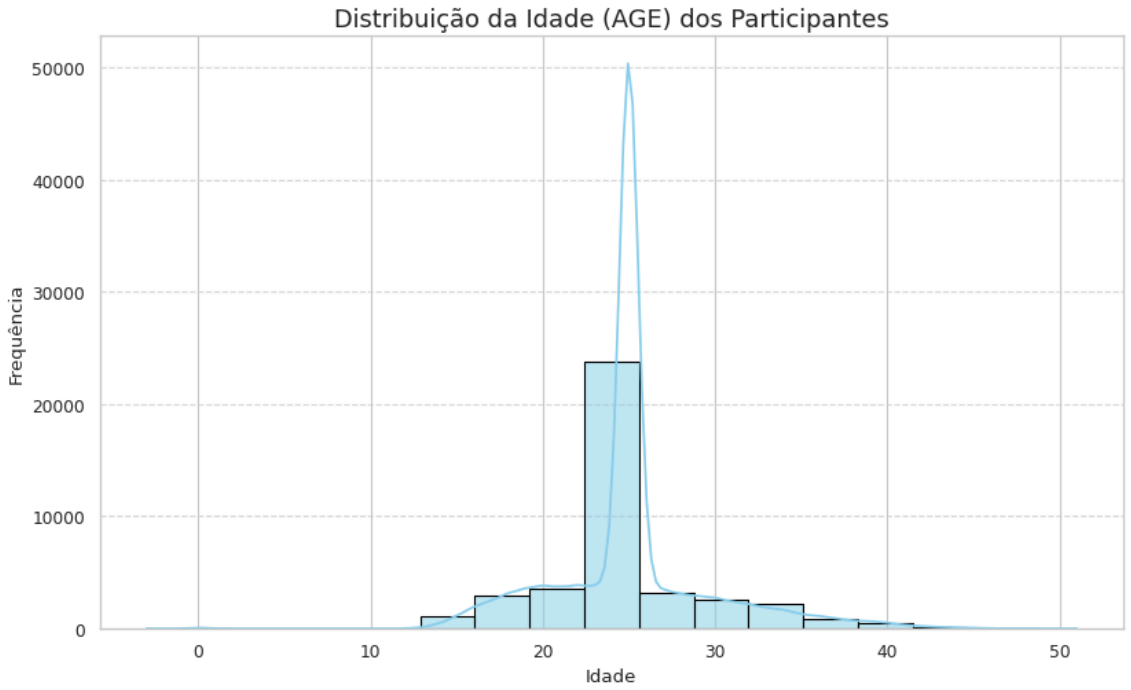

Figura 1. Distribuição da idade das pacientes

A Figura 1 mostra a distribuição das idades. A maioria das pacientes se encontra na faixa entre 20 e 30 anos, com média próxima de 25 anos. Essa concentração é compatível com a faixa etária reprodutiva, sendo um ponto relevante para políticas públicas de prevenção.

4.2. Distribuição dos Resultados do VDRL

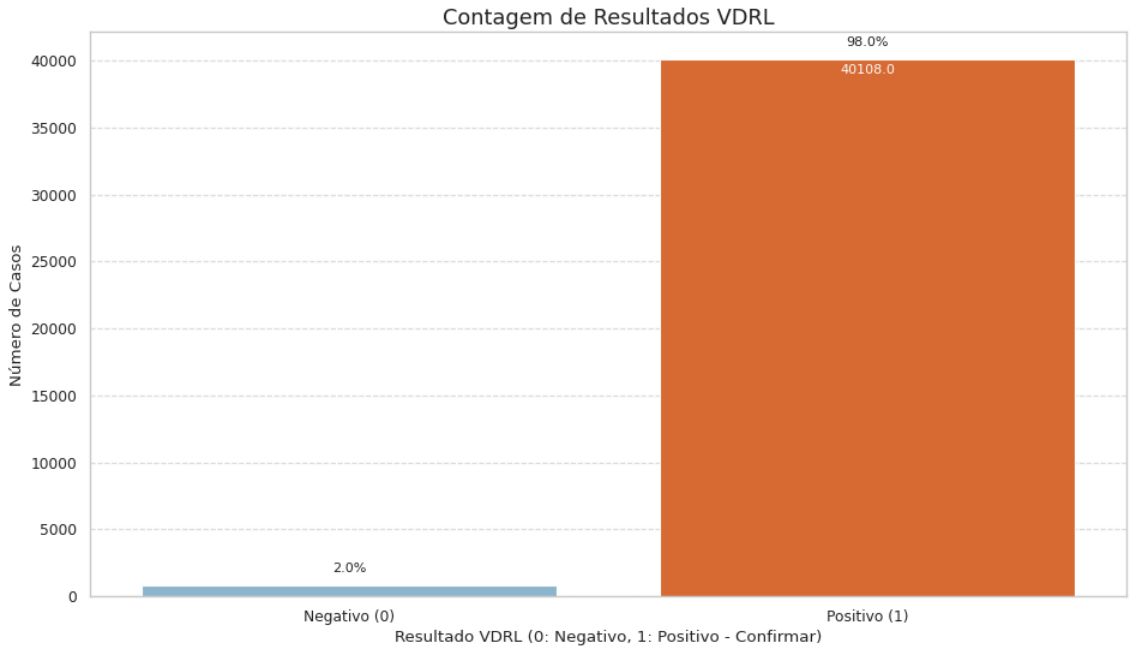

Figura 2. Contagem de resultados do teste VDRL

A Figura 2 evidencia o severo desbalanceamento de classes, com 98% dos resultados positivos. Isso exige cuidados na modelagem, como uso de técnicas de reamostragem para evitar que o modelo simplesmente aprenda a prever sempre "positivo".

4.3. Correlação entre Variáveis Numéricas

A análise de correlação entre as variáveis categóricas e as variáveis-alvo (VDRL RESULT e AGE) ajudou a identificar potenciais preditores relevantes. Por exemplo, variáveis como "*Level of Schooling*" e "*Family Income*" podem ter correlação com os resultados de saúde observados e deverão ser exploradas em análises mais aprofundadas.

A análise de correlação entre as variáveis categóricas e as variáveis-alvo (VDRL RESULT e AGE) não ajudou a identificar potenciais preditores relevantes. Por exemplo, variáveis como "*Level of Schooling*" e "*Family Income*" indica apenas que mais pessoas que possuem maior renda familiar são mais propensas a ir para a escola e SMOKER apresenta uma correlação de cerca de 0.55 com CONS ALCOHOL (Consumo de álcool). Essa correlação positiva indica que indica que os que fumam tendem a ter uma maior probabilidade de consumir álcool. Essa relação faz sentido, pois geralmente é este comportamento, mas não implica fortemente em indicadores para guiar nosso estudo.

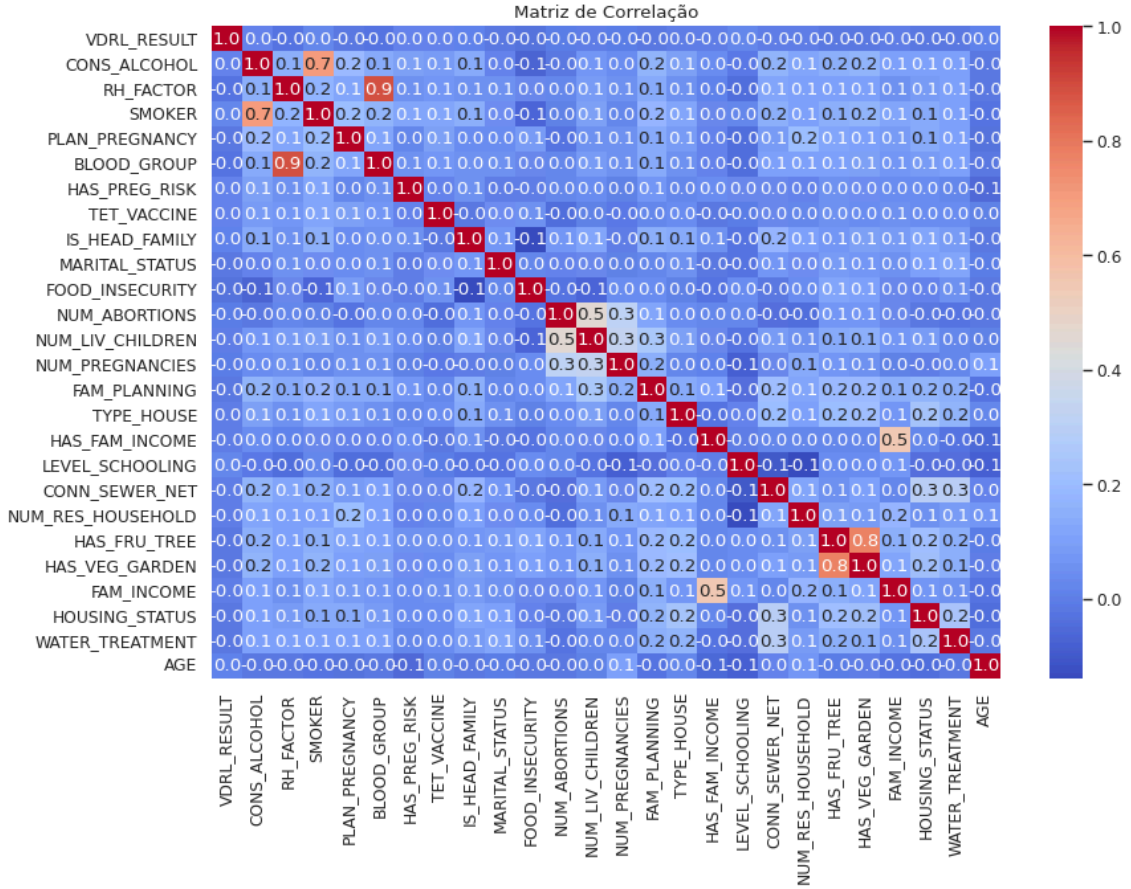

Figura 3. Matriz de Correlação

A Figura 3 apresenta a matriz de correlação. Não há correlações fortes entre variáveis numéricas e os alvos (VDRL_RESULT e AGE). Isso reforça a importância de variáveis categóricas e da aplicação de modelos capazes de capturar relações não lineares.

5. Metodologia

A abordagem metodológica compreendeu o pré-processamento dos dados aplicando engenharia de atributos, a divisão em conjuntos de treino e teste, e o treinamento e avaliação de modelos de regressão e classificação.

5.1. Pré-processamento dos Dados

As variáveis alvo (AGE para regressão e VDRL_RESUL para classificação) foram definidas. Os atributos preditores (X) foram identificados como numéricos ou categóricos. Um *ColumnTransformer* foi empregado para aplicar *StandardScaler* aos atributos numéricos e OneHotEncoder (com handle_unknown='ignore') aos atributos categóricos.

5.2. Modelo de Regressão para Previsão de Idade(AGE)

Para prever a idade, um modelo *RandomForestRegressor* (random_state=42) foi utilizado dentro de um pipeline contendo o pré-processador. Os dados foram divididos em 80% para treino e 20% para teste. A avaliação do modelo foi feita com base nas métricas MAE, MSE e RMSE no conjunto de teste.

5.3. Modelo de Classificação para Resultado do VDRL (VDRL_RESUL)

Para classificar o resultado do VDRL, focando na detecção da classe minoritária (positiva para sífilis congênita), um DecisionTreeClassifier (max_depth=10, min_samples_leaf=5, random_state=123, criterion='gini') foi implementado. Devido ao desbalanceamento de classes, a técnica de reamostragem SMOTEENN foi aplicada aos dados de treinamento dentro do pipeline. A performance do modelo foi avaliada usando AUC-ROC, Matriz de Confusão, Precisão, Recall e F1-Score, com ênfase nas métricas para a classe positiva.

5.4. Análise de Importância dos Atributos

A importância dos atributos para o modelo *DecisionTreeClassifier* na predição do VDRL_RESUL foi analisada para identificar os fatores mais influentes segundo o modelo.

6. Resultados e Discussão

6.1. Resultados da Regressão (Previsão de AGE)

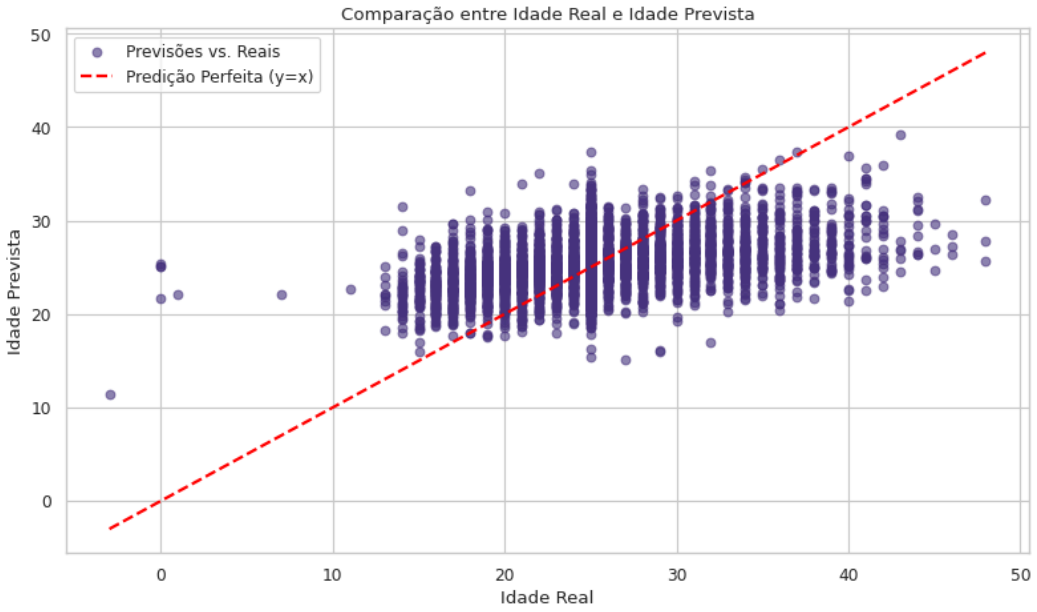

Figura 4. Comparação entre idade real e idade prevista

O modelo *RandomForestRegressor* para previsão da idade (AGE) obteve os seguintes resultados no conjunto de teste:

- Erro Médio Absoluto (MAE): ~2.93
- Erro Médio Quadrático (MSE): ~17.33
- Raiz do Erro Médio Quadrático (RMSE): ~4.16

Estes resultados, juntamente observando os resultados da Figura 1 e Figura 4, sugerem que o modelo consegue estimar a idade com um erro médio de aproximadamente 4 anos.

6.2. Resultados da Classificação (VDRL_RESUL)

A performance do *DecisionTreeClassifier* com SMOTEENN na predição da classe positiva para sífilis congênita foi:

- AUC-ROC (com SMOTEENN): ~0.5641

Já referente ao comparativo AUC-ROC, temos:

- Sem reamostragem (Decision Tree): 0.5709
- Com SMOTEENN (Decision Tree): 0.5641

Ao olharmos para as métricas para a classe positiva (VDRL Positivo = 1), temos:

- Precisão: ~0.0279 (Quando o modelo prevê sífilis, está correto em apenas ~2.8% das vezes)
- Recall (Sensibilidade): ~0.1659 (O modelo identifica apenas ~16.6% dos casos reais de sífilis)
- F1-Score: ~0.0478

Sendo assim, e sabendo que a matriz de Confusão detalha os erros de classificação, os valores de AUC-ROC próximos a 0.5 indicam um desempenho similar a um classificador aleatório. O baixo recall para a classe positiva é particularmente preocupante, pois significa que uma grande proporção de casos de sífilis congênita não seria detectada pelo modelo. A precisão extremamente baixa também é um problema, levando a muitos falsos alarmes se o modelo fosse usado para triagem. A aplicação do SMOTEENN, neste caso, não melhorou a AUC-ROC em relação ao modelo sem reamostragem, e o desafio de prever corretamente a classe minoritária persiste.

6.3. Análise de Importância dos Atributos para VDRL_RESUL

Os atributos identificados como os mais importantes pelo *DecisionTreeClassifier* para a predição do VDRL_RESUL foram:

- LEVEL_SCHOOLING

- HOUSING_STATUS_1.0
- NUM_ABORTIONS_3.0
- BLOOD_GROUP_0.0
- NUM_LIV_CHILDREN_4.0

A influência desses fatores, especialmente LEVEL_SCHOOLING, no contexto da sífilis congênita, poderia ser objeto de investigações mais aprofundadas.

6.4. Matriz de Confusão e Análise do Classificador

A seguir, a Figura 5 apresenta a matriz de confusão do modelo com SMOTEENN. Embora o modelo tenha identificado corretamente 6.792 casos positivos, ainda houve 1.230 falsos negativos, ou seja, pacientes com sífilis congênita que não foram detectados. Além disso, observam-se 115 falsos positivos, indicando previsões incorretas de sífilis em gestantes saudáveis.

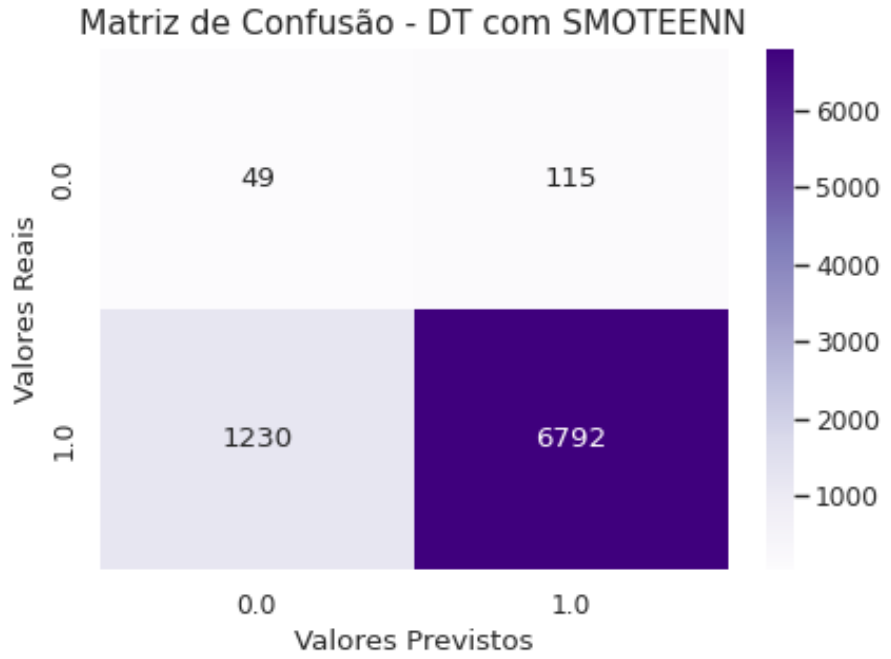

Figura 5 . Matriz de confusão (reamostragem)

Apesar da aplicação de técnicas de reamostragem, a quantidade significativa de falsos negativos torna o modelo inadequado para uso em triagem clínica sensível, já que pode deixar de identificar muitos casos reais. Esse resultado reforça a necessidade de explorar modelos mais sofisticados, estratégias de engenharia de atributos e, principalmente, variáveis mais informativas no conjunto de dados.

6.5. Importância das Features para Previsão de Sífilis Congênita

A análise de importância das variáveis pelo modelo *DecisionTreeClassifier* (com SMOTEENN) permite entender quais atributos mais contribuíram para a decisão do modelo. A Figura 6 apresenta as 10 features com maior impacto, medidas com base na redução de Gini.

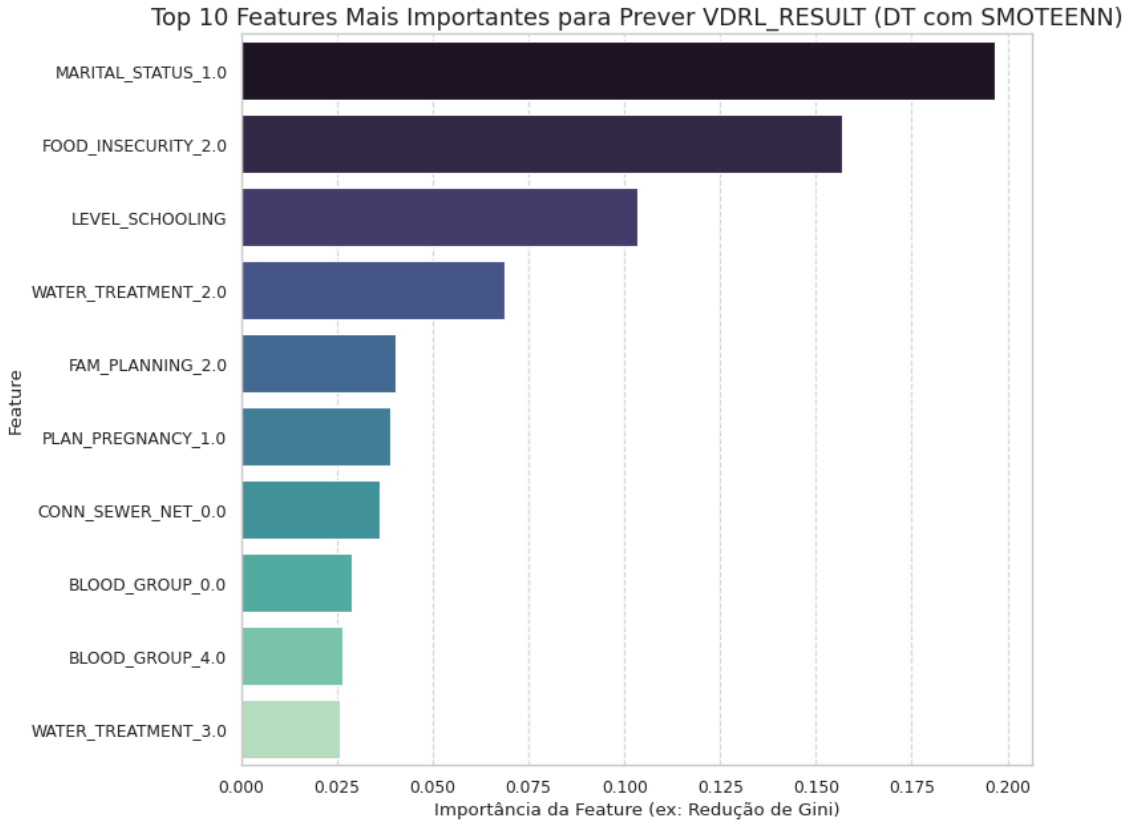

Figura 6 . Top 10 features mais importantes para previsão de VDRL_RESULT

As três variáveis com maior peso foram:

- MARITAL_STATUS_1.0 (estado civil)
- FOOD_INSECURITY_2.0 (grau de insegurança alimentar)
- LEVEL_SCHOOLING (escolaridade)

Esses resultados indicam que fatores socioeconômicos têm papel relevante na identificação de risco para sífilis congênita. Adicionalmente, variáveis como WATER_TREATMENT, FAM_PLANNING e BLOOD_GROUP também aparecem como importantes, sugerindo influência de aspectos estruturais e de saúde reprodutiva.

6.6. Visualização da Árvore de Decisão

A Figura 7 apresenta os três primeiros níveis da árvore de decisão treinada para prever o resultado do teste VDRL. Ela mostra como as regras de decisão foram estruturadas com base nas features mais importantes. Os nós são coloridos de acordo com a predominância da classe (0 = negativo, 1 = positivo), e os critérios de divisão envolvem principalmente escolaridade, planejamento familiar e acesso à infraestrutura básica.

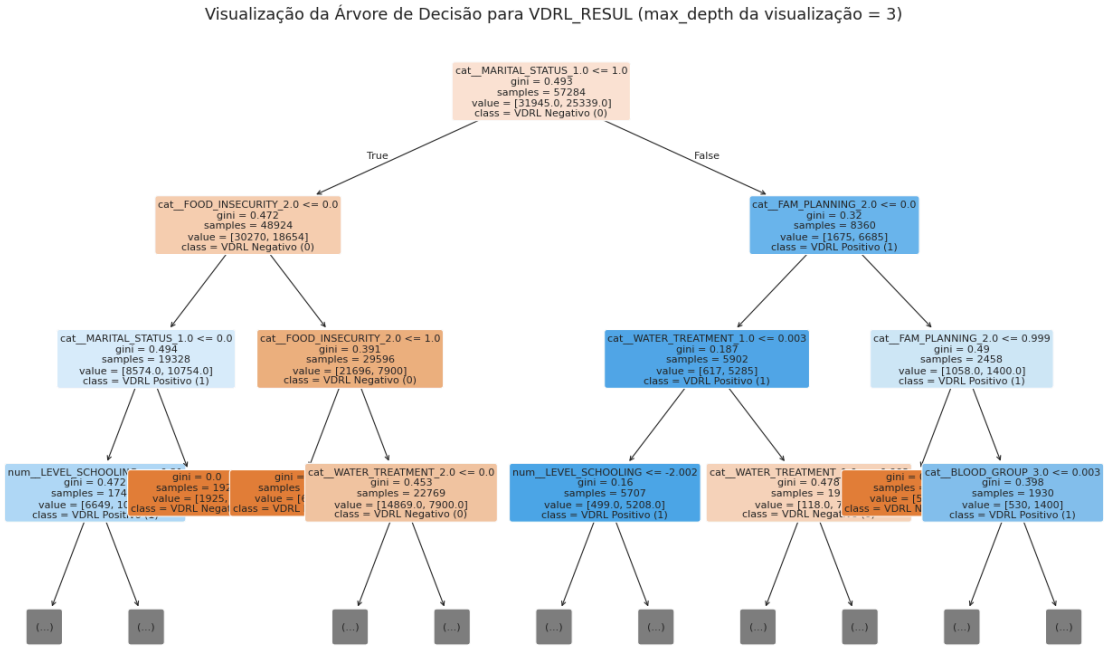

Figura 7. Árvore de Decisão para VDRL_RESULT

Observa-se que, por exemplo, a combinação de MARITAL_STATUS, FOOD_INSECURITY e LEVEL_SCHOOLING já separa a maior parte das amostras com boa pureza. O uso dessa visualização pode contribuir para o esclarecimento do modelo, algo essencial em contextos de saúde pública.

7. Conclusão

Este estudo buscou desenvolver modelos preditivos para idade e resultados do teste VDRL para sífilis congênita utilizando dados clínicos e sociodemográficos do Programa Mãe Coruja Pernambucana (PMCP). Para a tarefa de regressão, o modelo *RandomForestRegressor* apresentou desempenho satisfatório, com um RMSE de aproximadamente 4.16, o que indica uma boa capacidade de estimar a idade das pacientes com base nos atributos disponíveis.

No entanto, para a tarefa de classificação da sífilis congênita, os resultados foram limitados. O modelo *DecisionTreeClassifier*, mesmo com o uso da técnica de

reamostragem SMOTEENN para tratar o severo desbalanceamento de classes (apenas ~1,98% dos casos eram positivos), obteve um AUC-ROC de ~0.56, com um recall de ~0.166 e uma precisão de apenas ~0.028. Isso evidencia a dificuldade do modelo em detectar corretamente a classe minoritária, resultando em uma grande quantidade de falsos negativos, o que compromete sua aplicação em triagens sensíveis na área da saúde pública. Além disso, observou-se que variáveis sociodemográficas como MARITAL_STATUS, FOOD_INSECURITY e LEVEL_SCHOOLING foram as que mais contribuíram para a predição do teste VDRL positivo, o que ressalta a importância de fatores estruturais e sociais no contexto da sífilis congênita.

Esses resultados enfatizam o desafio de prever a sífilis congênita com os atributos e modelos utilizados, mesmo com abordagens clássicas de balanceamento de dados e algoritmos interpretáveis. A baixa capacidade de detecção da classe positiva é uma limitação crítica, especialmente em cenários que exigem sensibilidade elevada.

Referências

[Pedregosa et al. 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

[Chawla et al. 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.