Análise de Dados Clínicos e Sociodemográficos sobre Casos de Sífilis Congênita no Brasil (2013-2021)

Victor P P de Melo[1], **Eduarda S Figueredo**[2], **Arthur P Padilha**[3]

[1]CESAR School - Ciência da Computação

[2]CESAR School - Ciência da Computação

[3]CESAR School - Ciência da Computação

vppm@cesar.school, esf@cesar.school, app2@cesar.school

Abstract. *This study presents an analysis of clinical and sociodemographic data related to congenital syphilis cases in Brazil between 2013 and 2021. Using machine learning techniques, we performed data preprocessing, classification models to predict the VDRL test result, and regression models to analyze the relationship with patients' age. The results highlight the main variables influencing the occurrence of congenital syphilis, providing insights for public health interventions.*

Keywords*: Congenital syphilis, machine learning, data analysis, public health, classification, regression,age, VDRL.*

Resumo. *Este estudo apresenta uma análise de dados clínicos e sociodemográficos relacionados a casos de sífilis congênita no Brasil entre 2013 e 2021. Utilizando técnicas de aprendizado de máquina, realizamos pré-processamento dos dados, modelos de classificação para prever o resultado do exame VDRL e modelos de regressão para analisar a relação com a idade dos pacientes. Os resultados destacam as principais variáveis que influenciam a ocorrência de sífilis congênita, fornecendo insights para intervenções em saúde pública.*

Palavras-chave*: Sífilis congênita, aprendizado de máquina, análise de dados, saúde pública, classificação, regressão, idade, VDRL.*

1. Introdução

A sífilis congênita permanece como um desafio significativo de saúde pública no Brasil, afetando milhares de recém-nascidos anualmente. A detecção precoce e a intervenção adequada são cruciais para prevenir complicações graves. Este projeto visa aplicar técnicas de análise de dados e aprendizado de máquina para investigar fatores clínicos e sociodemográficos associados aos casos de sífilis congênita, utilizando o dataset "Clinical and sociodemographic data on congenital syphilis cases, Brazil, 2013-2021".

Ao aplicar modelos preditivos, buscamos identificar padrões e fatores de risco que possam auxiliar na formulação de políticas públicas e estratégias de intervenção. A utilização de técnicas de classificação e regressão permitirá compreender melhor as variáveis que influenciam os desfechos clínicos, promovendo ações direcionadas na prevenção e controle da sífilis congênita.

2. Objetivos

Este estudo tem como objetivos principais:

1. **Compreender e preparar dados de saúde pública:**
    - Realizar análise exploratória dos dados.
    - Identificar e tratar problemas de qualidade, como valores ausentes e inconsistentes.
    - Aplicar técnicas de pré-processamento adequadas.
2. **Desenvolver e avaliar modelos de classificação:**
    - Construir modelos para prever o resultado do exame VDRL.
    - Justificar a escolha dos modelos e parâmetros.
    - Avaliar o desempenho utilizando métricas apropriadas.
3. **Desenvolver e avaliar modelos de regressão:**
    - Analisar a relação entre variáveis clínicas/sociodemográficas e a idade (*AGE*).
    - Utilizar modelos de regressão adequados e avaliar seu desempenho.
4. **Interpretar resultados e relacionar com práticas de saúde pública:**
    - Identificar fatores de risco relevantes.
    - Sugerir intervenções preventivas baseadas nos insights obtidos.

3. Descrição do Dataset

O dataset utilizado neste estudo é intitulado ***"Clinical and sociodemographic data on congenital syphilis cases, Brazil, 2013-2021"***. Este conjunto de dados compreende informações clínicas e sociodemográficas relacionadas a casos de sífilis congênita notificados no Brasil entre 2013 e 2021.

1. **Características do Dataset:**
    - Contém 41.762 registros.
    - Inclui variáveis como resultado do exame VDRL (*VDRL_RESULT*), idade da paciente (*AGE*), fatores clínicos e condições socioeconômicas.
    - As variáveis possuem diferentes tipos de dados: numéricos contínuos, categóricos e binários.
2. **Relevância:**
    - Permite a análise de fatores que contribuem para a incidência da sífilis congênita.
    - Auxilia na identificação de grupos de risco e na formulação de políticas públicas.

4. Metodologia

4.1. Análise Exploratória e Pré-processamento

1. **Carregamento dos dados:**
    - Importação das bibliotecas necessárias: `pandas`, `numpy`, `seaborn`, `matplotlib`, `sklearn` e `imbleran`.
    - Leitura dos arquivos `attributes.csv` e `data_set.csv`.
2. **Pré-visualização e Limpeza:**

- Utilização de `data.head()` e `data.info()` para inspeção inicial.
- Remoção de valores inconsistentes, como idades negativas.
  - `data = data[data['AGE'] ≥ 0]`
- Verificação de valores ausentes e tipos de dados.

3. **Análise Estatística:**
   - Cálculo de estatísticas descritivas com `data.describe()`.
   - Análise da distribuição da variável alvo `VDRL_RESULT`.
4. **Matriz de Correlação:**
   - Geração da matriz de correlação para identificar relações entre variáveis.
     - `correlation_matrix = data.corr()`
   - Visualização com mapa de calor utilizando `seaborn`.
5. **Codificação de Variáveis Categóricas:**
   - Aplicação de One-Hot Encoding para variáveis categóricas.
     - `from sklearn.preprocessing import OneHotEncoder`
     - `encoder = OneHotEncoder(sparse_output=False)`
   - Concatenação dos dados codificados com as variáveis numéricas.

4.2. Modelos de Classificação

1. **Divisão dos Dados:**
   - Separação das variáveis independentes (X) e da variável alvo (y: `VDRL_RESULT`).
   - Divisão em conjuntos de treino e teste (80% treino, 20% teste).
2. **Tratamento de Desbalanceamento:**
   - Aplicação da técnica SMOTEENN para balancear as classes, usando a biblioteca 'imblearn'.
     - `from imblearn.combine import SMOTEENN`
     - `smote_enn = SMOTEENN(random_state=42)`
     - `X_balanced, y_balanced = smote_enn.fit_resample(X, y)`
3. **Modelos Utilizados:**
   - **Random Forest Classifier:**
     - **Justificativa:** Alta capacidade de generalização e robustez contra overfitting.
     - **Hiperparâmetros:** Número de árvores (n_estimators), profundidade máxima (max_depth), e amostras mínimas para divisão (min_samples_split).
     - **Implementação:** 'RandomForestClassifier' da biblioteca 'sklearn'.
     - **Validação:** Utilização de GridSearchCV com validação cruzada de 5 folds.
   - **Decision Tree Classifier:**
     - **Justificativa:** Simplicidade e facilidade de interpretação.
     - **Hiperparâmetros:** Profundidade máxima (max_depth), amostras mínimas para divisão (min_samples_split).
     - **Implementação:** 'DecisionTreeClassifier' da biblioteca 'sklearn'.
4. **Validação e Avaliação:**
   - Validação cruzada k-Fold e ajuste de hiperparâmetros com GridSearchCV.
   - Avaliação utilizando métricas: precisão, recall, f1-score.
   - Relatório de classificação usando `classification_report`.

4.3. Modelos de Regressão

1. **Variável Alvo:**
    - Utilização da variável contínua `AGE`.
2. **Modelo Utilizado:**
    - **Random Forest Regressor:**
        - **Justificativa:** Bom para capturar relações complexas.
        - **Hiperparâmetros:** Número de árvores (n_estimators), profundidade máxima (max_depth).
        - **Implementação:** 'RandomForestRegressor' da biblioteca 'sklearn'.
        - **Validação:** GridSearchCV com validação cruzada de 5 folds.
3. **Avaliação dos Modelos:**
    - Métricas: MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error).
    - Comparação entre modelos para identificar o melhor ajuste.

4.4. Interpretação e Análise de Fatores

1. **Importância das Variáveis:**
    - Extração das principais variáveis que influenciam os modelos.
        - `feature_importance = rf_classifier.feature_importances_`
    - Visualização das 10 variáveis mais importantes em gráficos de barras.
2. **Discussão dos Resultados:**
    - Interpretação dos fatores de risco identificados.
    - Relação com práticas e políticas de saúde pública.

5. Resultados e Discussão

5.1. Análise Exploratória

1. **Distribuição das Classes:**
    - Classe positiva (`VDRL_RESULT = 1`): 40.935 casos.
    - Classe negativa (`VDRL_RESULT = 0`): 826 casos.
    - Observou-se um desbalanceamento significativo entre as classes.
2. **Correlação entre Variáveis:**
    - A variável `AGE` apresentou correlações positivas com `NUM_PREGNANCIES` e `NUM_RES_HOUSEHOLD`, embora baixas.
    - A matriz de correlação indicou baixa multicolinearidade entre as variáveis independentes.

5.2. Modelos de Classificação

1. **Desempenho do Random Forest Classifier:**
    - **Melhores Hiperparâmetros:**
        - `n_estimators`: 50
        - `max_depth`: `None`
        - `min_samples_split`: 5
        - `min_samples_leaf`: 1
    - **Desempenho do conjunto de teste:**
        - Precisão média: 99%.
        - Recall médio: 99%.
        - F1-score médio: 99%.
    - **Interpretação:**
        - O modelo apresentou excelente desempenho após o balanceamento das classes utilizando SMOTEENN.
        - As principais variáveis influentes foram `FOOD_INSECURITY`, `PLAN_PREGNANCY`, `FAM_PLANNING`, `HOUSING_STATUS` e `CONN_SEWER_NET`.
2. **Desempenho do Decision Tree Classifier:**
    - **Melhores Hiperparâmetros:**
        - `max_depth`: `None`
        - `min_samples_split`: 10
        - `min_samples_leaf`: 4
    - **Desempenho do conjunto de teste:**
        - Precisão média: 98%.
        - Recall médio: 98%.
        - F1-score médio: 98%.
    - **Interpretação:**
        - O modelo de árvore de decisão apresentou desempenho ligeiramente inferior ao Random Forest.
        - Houve menor generalização e maior suscetibilidade a overfitting.
        - Variáveis mais importantes: `AGE`, `LEVEL_SCHOOLING`, `NUM_PREGNANCIES`, `NUM_RES_HOUSEHOLD`, `FAM_INCOME`.

5.3. Modelos de Regressão

1. **Random Forest Regressor:**
    - **Melhores Hiperparâmetros:**
        - `n_estimators`: 200
        - `max_depth`: 10
        - `min_samples_split`: 2
        - `min_samples_leaf`: 4
    - **Desempenho do conjunto de teste:**
        - MAE: 2,85.
        - MSE: 17,28.
        - RMSE: 4,16.
    - **Interpretação:**
        - O erro médio absoluto indica que as previsões, em média, diferem em aproximadamente 2,85 anos da idade real.

- O modelo capturou bem as relações não lineares entre as variáveis.
- Variáveis mais impactantes: NUM_PREGNANCIES, LEVEL_SCHOOLING, NUM_LIV_CHILDREN, MARITAL_STATUS, NUM_RES_HOUSEHOLD, HAS_PREG_RISK, WATER_TREATMENT, FAM_INCOME, FOOD_INSECURITY, HOUSING_STATUS.

5.4. Análise dos Fatores de Risco

1. **Idade (AGE):**
   - Variável mais influente na predição do resultado do exame VDRL.
   - Indica que a idade da mãe está diretamente relacionada ao risco de sífilis congênita.
2. **Escolaridade (LEVEL_SCHOOLING):**
   - Alto impacto nos modelos de classificação e regressão.
   - Sugere que níveis mais baixos de escolaridade estão associados a maiores riscos.
3. **Condições Socioeconômicas:**
   - Variáveis como FAM_INCOME, HOUSING_STATUS e WATER_TREATMENT mostraram influência significativa.
   - Reflete a importância de fatores socioeconômicos na saúde materna e neonatal.
4. **Histórico Reprodutivo:**
   - NUM_PREGNANCIES, NUM_LIV_CHILDREN e NUM_ABORTIONS foram relevantes.
   - Indica a necessidade de acompanhamento mais próximo de mulheres com múltiplas gestações.

6. Conclusão

Este estudo aplicou técnicas de aprendizado de máquina para analisar dados clínicos e sociodemográficos relacionados à sífilis congênita no Brasil. Os modelos desenvolvidos permitiram identificar os principais fatores de risco, destacando a idade materna, escolaridade e condições socioeconômicas como influências significativas.

Os resultados sugerem que intervenções focadas em educação e melhoria das condições socioeconômicas podem ser eficazes na prevenção da sífilis congênita. Além disso, o acompanhamento especializado de gestantes em grupos de risco é fundamental.

Para trabalhos futuros, recomenda-se a inclusão de dados adicionais, como acesso a serviços de saúde e informações sobre tratamentos anteriores, para aprimorar os modelos preditivos e a compreensão dos fatores envolvidos.