Análise de Dados de Sífilis Congênita no Brasil (2013-2021)

1. Introdução

A sífilis congênita é uma doença infecciosa grave que pode ser transmitida da mãe para o bebê durante a gestação ou parto, resultando em complicações sérias para a saúde do recém-nascido. No Brasil, a incidência de sífilis congênita tem sido uma preocupação crescente para as autoridades de saúde pública, exigindo estratégias eficazes de prevenção e controle.

Este projeto visa proporcionar uma experiência prática de análise de dados aplicada à saúde pública, focando nos casos de sífilis congênita no Brasil entre os anos de 2013 e 2021. Utilizando o dataset "Clinical and sociodemographic data on congenital syphilis cases, Brazil, 2013-2021", aplicaremos técnicas de pré-processamento, classificação e regressão para investigar fatores clínicos e sociodemográficos associados aos desfechos de sífilis congênita.

Objetivos do Projeto

1. **Compreensão e Preparação dos Dados de Saúde Pública:**
    - Realizar análise exploratória dos dados.
    - Identificar e tratar problemas de qualidade nos dados, como valores ausentes e outliers.
    - Aplicar técnicas de pré-processamento, incluindo one-hot encoding e balanceamento de classes.
2. **Desenvolvimento e Avaliação de Modelos de Classificação:**
    - Prever o resultado do exame VDRL (`VDRL_RESULT`) como indicador de sífilis congênita.
    - Implementar múltiplos modelos de classificação.
    - Avaliar o desempenho dos modelos utilizando métricas apropriadas.
3. **Desenvolvimento e Avaliação de Modelos de Regressão:**
    - Modelar a relação entre variáveis clínicas e sociodemográficas e a idade (`AGE`) dos pacientes.
    - Implementar múltiplos modelos de regressão.
    - Avaliar o desempenho dos modelos utilizando métricas de erro.
4. **Interpretação dos Resultados e Relação com Práticas de Saúde Pública:**
    - Identificar fatores de risco e associações relevantes.
    - Sugerir intervenções preventivas com base nos insights obtidos.

2. Descrição do Dataset

- **Nome do Dataset:** Clinical and sociodemographic data on congenital syphilis cases, Brazil, 2013-2021
- **Fonte:** Dataset no Mendeley
- **Artigo Base:** Predicting congenital syphilis cases: A performance evaluation of different machine learning models

Características do Dataset

- **Dados Clínicos:** Informações sobre exames médicos, histórico clínico e tratamento.
- **Dados Sociodemográficos:** Informações sobre idade, educação, renda e acesso a serviços de saúde.
- **Variáveis-Alvo:**
  - **Classificação:** `VDRL_RESULT` (resultado do exame VDRL).
  - **Regressão:** `AGE` (idade do paciente).

Estrutura do Dataset

O dataset contém diversas colunas que representam características clínicas e sociodemográficas dos casos de sífilis congênita. A seguir, uma visão geral das principais variáveis:

- **VDRL_RESULT:** Resultado do exame VDRL (Positivo/Negativo).
- **AGE:** Idade da gestante.
- **EDUCATION_LEVEL:** Nível de educação da gestante.
- **INCOME:** Renda familiar.
- **ACCESS_TO_HEALTH_SERVICES:** Acesso a serviços de saúde.
- **CLINICAL_HISTORY:** Histórico clínico da gestante.
- **TRIMESTER:** Trimestre de gestação no momento do diagnóstico.
- **LOCATION:** Região geográfica do caso.

3. Metodologia

3.1 Análise Exploratória e Pré-processamento

3.1.1 Importação de Bibliotecas

Utilizamos bibliotecas essenciais para manipulação de dados, visualização e modelagem estatística e de machine learning, como pandas, sklearn etc...

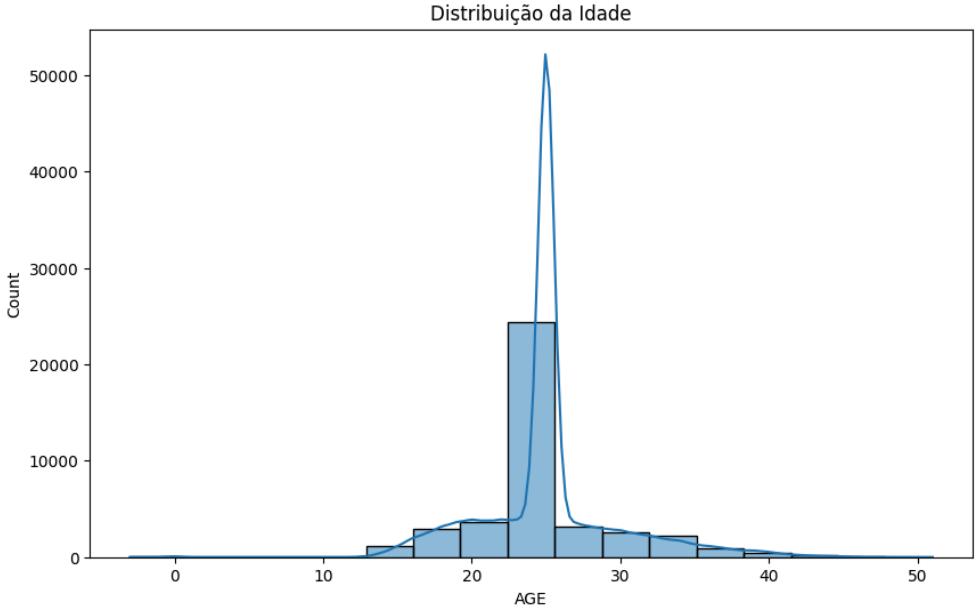
3.1.2 Carregamento do Dataset

O dataset foi carregado utilizando a biblioteca pandas, permitindo uma inspeção inicial das características dos dados.

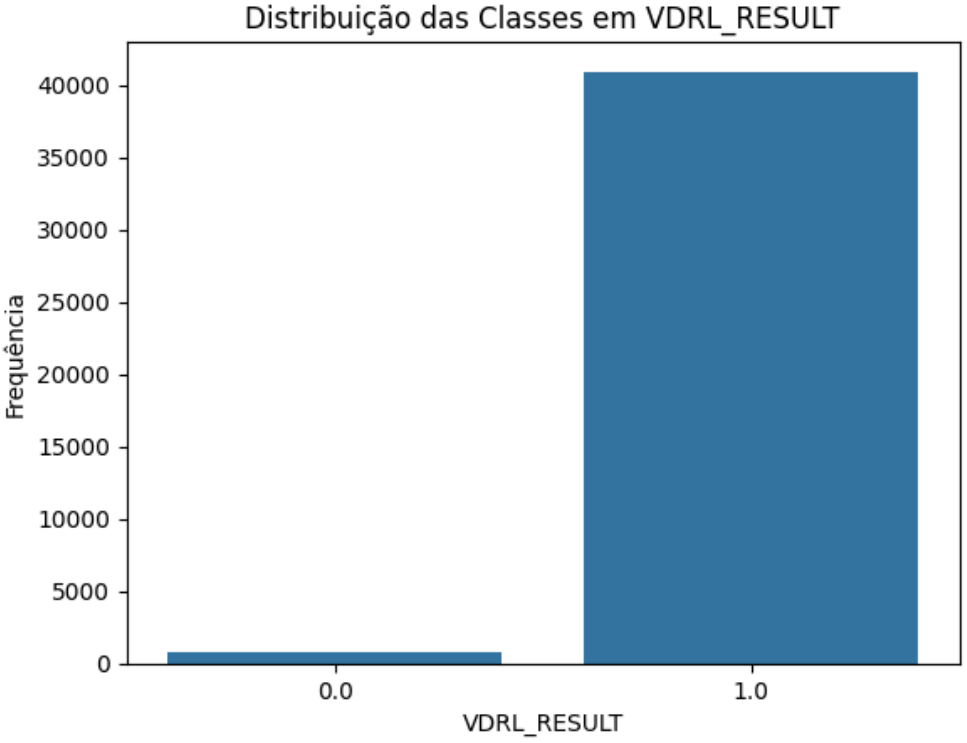
3.1.3 Análise Exploratória dos Dados

Realizamos uma análise exploratória para entender a distribuição das variáveis, identificar correlações e visualizar a distribuição das classes da variável alvo.

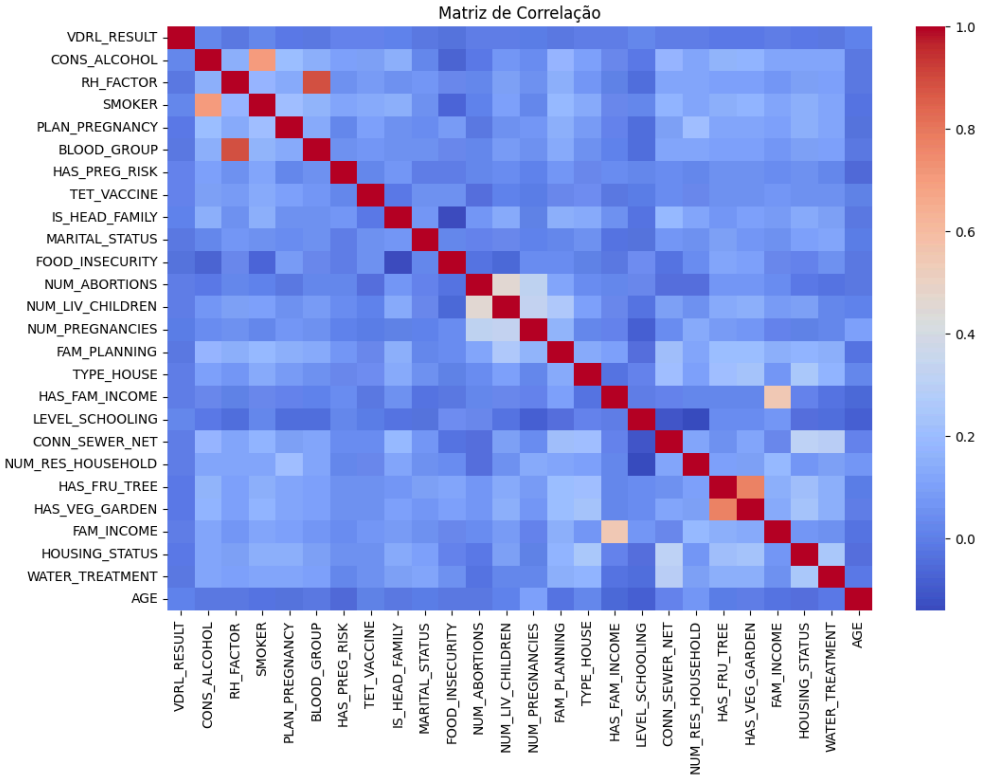
- **Distribuição da idade :**



- **Distribuição de 'VDRL_RESULT' :**



- **Matriz de Correlação:**



3.1.4 Tratamento de Valores Ausentes

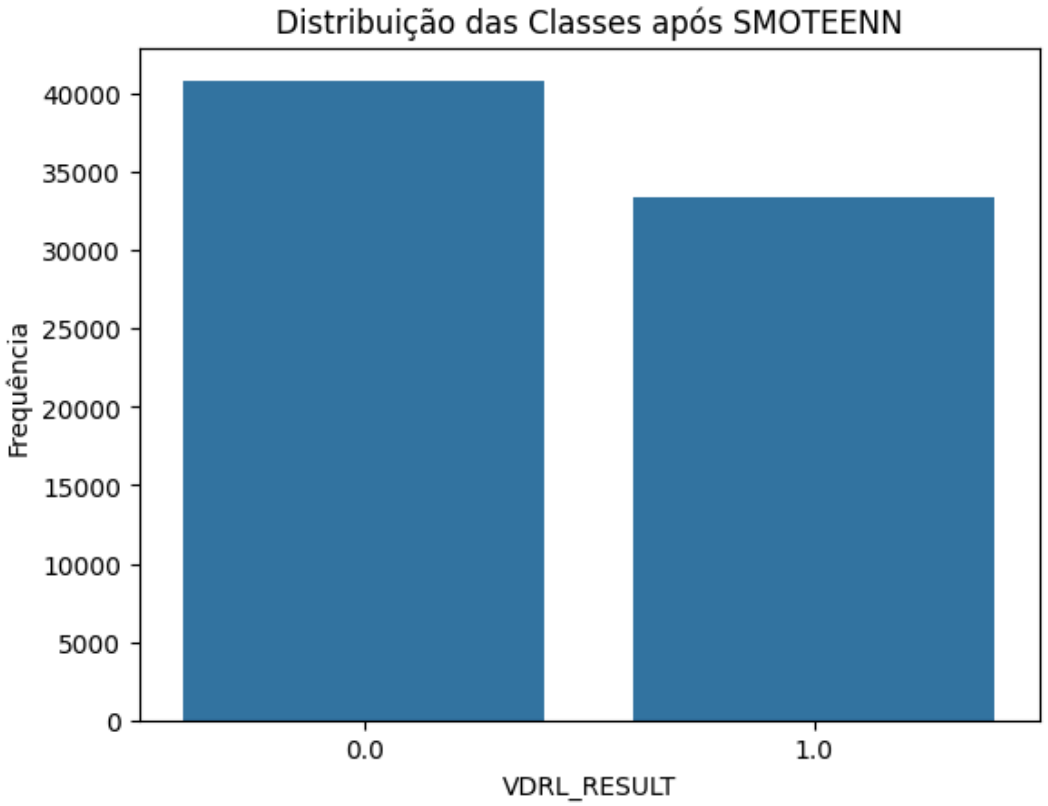
Não há nenhum valor ausente no dataset, entretanto, algumas features possuem a opção "Não informado" nas variáveis categóricas. Portanto, nesses casos, as linhas quem possuiam a maior parte dos valores não informados foram descartadas.

3.1.5 Codificação de Variáveis Categóricas

Aplicamos `Label Encoding` e `One-Hot Encoding` para transformar variáveis categóricas em formatos numéricos adequados para modelagem.

3.1.6 Balanceamento das Classes

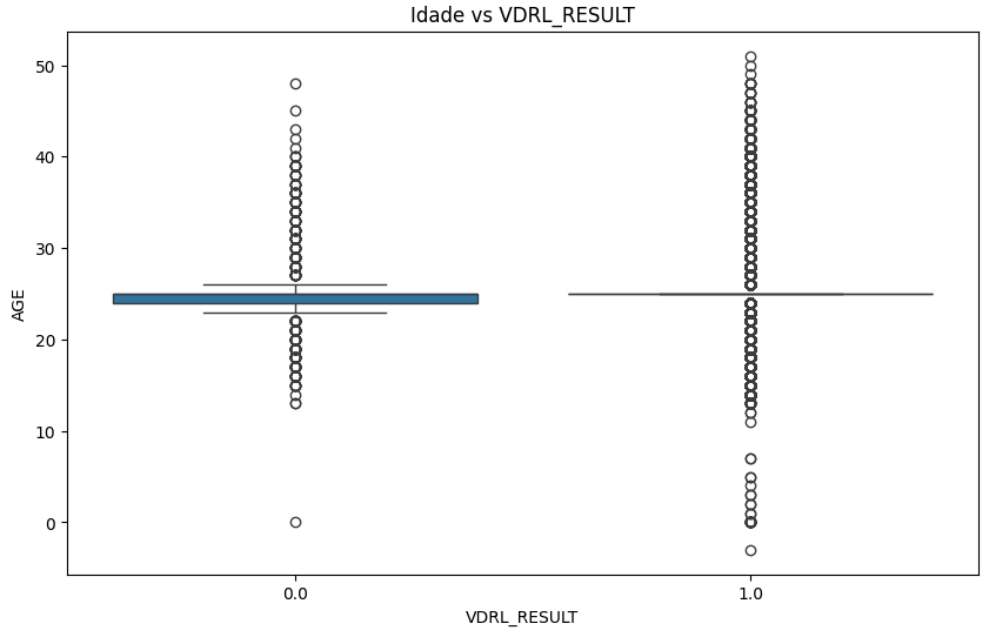
A variável alvo `VDRL_RESULT` apresentava desbalanceamento. Utilizamos a técnica de **SMOTEENN** para balancear as classes, melhorando a performance dos modelos de classificação.



3.2 Análise Exploratória

Realizamos visualizações adicionais para compreender melhor a distribuição da idade e sua relação com o resultado do exame VDRL.

- **Boxplot de Idade por `VDRL_RESULT`:**



4. Modelos de Classificação

4.1 Divisão dos Dados

Dividimos os dados balanceados em conjuntos de treino e teste com uma proporção de 80/20, garantindo a representatividade das classes através da estratificação.

4.2 Treinamento e Avaliação de Múltiplos Modelos

Implementamos uma gama de modelos de classificação para prever o resultado do exame VDRL:

- **Logistic Regression**
- **Decision Tree**
- **Random Forest**
- **Gradient Boosting**
- **AdaBoost**
- **XGBoost**
- **LightGBM**
- **SVM (Support Vector Machine)**
- **K-Nearest Neighbors (KNN)**

Cada modelo foi treinado utilizando os dados de treino e avaliado no conjunto de teste utilizando métricas como **Accuracy**, **Precision**, **Recall**, **F1-Score** e **ROC AUC**.

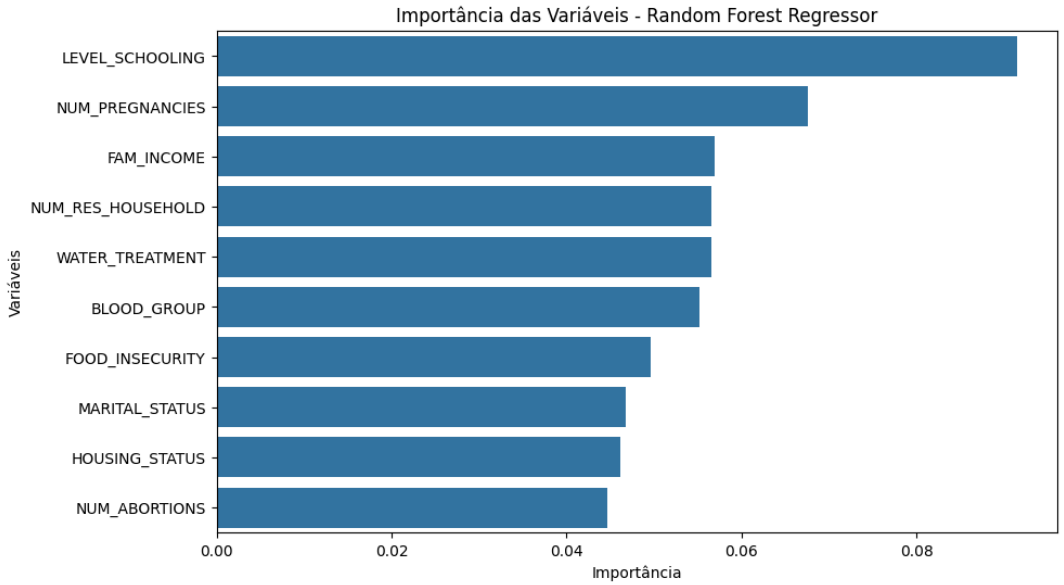
4.3 Comparação dos Modelos

Os resultados de desempenho dos modelos foram compilados em um DataFrame para facilitar a comparação:

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	0.644205	0.642143	0.644205	0.638265	0.682045
Decision Tree	0.974191	0.974207	0.974191	0.974179	0.973465
Random Forest	0.989151	0.989407	0.989151	0.989162	0.996968
Gradient Boosting	0.967655	0.969094	0.967655	0.967718	0.989024
XGBoost	0.987803	0.988125	0.987803	0.987817	0.994609
SVM (Linear Kernel)	0.489218	0.520597	0.489218	0.464088	0.524296
KNN	0.981806	0.982370	0.981806	0.981767	0.996109

4.4 Análise de Importância das Variáveis

Utilizamos o modelo **RandomForest** para extrair e visualizar a importância das variáveis mais influentes na predição do resultado do exame VDRL.



Principais Variáveis:

1. **Nível de Escolaridade**
2. **Número de Gravidezes**
3. **Renda Familiar**
4. **Número de Residentes no Ambiente Doméstico**

5. Modelos de Regressão

5.1 Preparação dos Dados

Para a tarefa de regressão, focamos na variável AGE (idade dos pacientes). Removemos registros com valores inválidos e dividimos os dados em conjuntos de treino e teste.

5.2 Treinamento e Avaliação de Múltiplos Modelos

Implementamos diversos modelos de regressão para modelar a idade dos pacientes:

- **Linear Regression**
- **Ridge Regression**
- **Random Forest Regressor**
- **Gradient Boosting Regressor**
- **AdaBoost Regressor**
- **XGBoost Regressor**
- **LightGBM Regressor**
- **SVR (Support Vector Regression)**
- **KNN Regressor**

Cada modelo foi treinado e avaliado utilizando métricas de erro como **MAE (Mean Absolute Error)**, **RMSE (Root Mean Squared Error)** e **R^2 (Coeficiente de Determinação)**.

5.3 Comparação dos Modelos

Os resultados de desempenho dos modelos de regressão foram compilados em um DataFrame:

Model	MAE	RMSE	R^2
Linear Regression	2.869040	4.413516	0.035497
Ridge Regression	2.869026	4.413516	0.035497
Random Forest Regressor	2.868961	4.062000	0.183015
Gradient Boosting Regressor	2.829128	4.144237	0.149600
AdaBoost Regressor	3.272035	4.488923	0.002257
XGBoost Regressor	2.872881	4.058112	0.184578
LightGBM Regressor	2.825826	4.080238	0.175662
SVR	2.605583	4.426254	0.029921
KNN Regressor	3.181324	4.607804	-0.051289

6. Análise de Fatores e Discussão

6.1 Identificação de Fatores de Risco

A análise dos modelos de classificação e regressão revelou que certos fatores socioeconômicos e demográficos desempenham papéis cruciais nos desfechos de sífilis congênita. Especificamente, os seguintes atributos mostraram-se os mais influentes na classificação do resultado do exame VDRL (`VDRL_RESULT`):

1. **Nível de Escolaridade (`LEVEL_SCHOOLING`):**
    - **Descrição:** Representa o nível de escolaridade da gestante, categorizado desde educação básica até superior.
    - **Impacto:** Gestantes com níveis mais baixos de escolaridade tendem a ter menor acesso a informações sobre saúde, o que pode contribuir para diagnósticos tardios e tratamento inadequado da sífilis.
2. **Número de Gravidezes (`NUM_PREGNANCIES`):**
    - **Descrição:** Indica o número de gravidezes que a gestante já teve, categorizado de "Nenhuma" a "Mais de duas".
    - **Impacto:** Gestantes com múltiplas gravidezes podem enfrentar maior estresse e menos tempo para cuidados pré-natais adequados, aumentando o risco de transmissão da sífilis para o bebê.
3. **Renda Familiar (`FAM_INCOME`):**

- **Descrição:** Representa a renda familiar informada durante os cuidados pré-natais, categorizada em faixas de renda.
- **Impacto:** Baixa renda familiar está associada a limitações no acesso a serviços de saúde de qualidade, tornando difícil para as gestantes receberem o tratamento necessário para prevenir a transmissão da sífilis.

4. **Número de Residentes no Ambiente Doméstico (`NUM_RES_HOUSEHOLD`):**
    - **Descrição:** Indica o número de residentes na residência da gestante, categorizado de "Nenhum" a "Mais de três".
    - **Impacto:** Ambientes domésticos com mais residentes podem apresentar desafios em termos de privacidade e acesso a cuidados de saúde, além de potenciais condições de higiene inadequadas que facilitam a propagação de doenças infecciosas.

6.2 Sugestões para Intervenções Preventivas

Com base nos fatores de risco identificados, propomos as seguintes intervenções para a prevenção e controle da sífilis congênita:

1. **Educação e Capacitação:**
    - **Programas Educacionais:** Desenvolver programas educativos voltados para gestantes com baixos níveis de escolaridade, enfatizando a importância do pré-natal e a prevenção de doenças infecciosas como a sífilis.
    - **Capacitação de Profissionais de Saúde:** Treinar profissionais de saúde para identificar gestantes em risco e fornecer informações claras e acessíveis sobre a prevenção e tratamento da sífilis.
2. **Melhoria do Acesso a Serviços de Saúde:**
    - **Ampliação dos Serviços Pré-Natais:** Garantir que todas as gestantes, especialmente aquelas com múltiplas gravidezes, tenham acesso fácil e regular a serviços de pré-natal de qualidade.
    - **Facilitação do Acesso a Tratamentos:** Implementar políticas que reduzam barreiras financeiras e geográficas ao acesso a tratamentos para sífilis, garantindo que todas as gestantes recebam o tratamento adequado.
3. **Apoio Socioeconômico:**
    - **Programas de Assistência Financeira:** Desenvolver programas que apoiem famílias de baixa renda, melhorando as condições socioeconômicas das gestantes e aumentando sua capacidade de acessar cuidados de saúde.
    - **Iniciativas de Redução de Pobreza:** Implementar políticas abrangentes de redução de pobreza que abordem fatores subjacentes que contribuem para a vulnerabilidade das gestantes.

4. **Gestão de Espaços Domésticos:**
    - **Melhoria das Condições de Habitação:** Investir em melhorias das condições de habitação, garantindo ambientes mais saudáveis e adequados para gestantes e recém-nascidos.
    - **Programas de Higiene e Saneamento:** Promover programas que incentivem práticas de higiene e saneamento adequadas nas residências, reduzindo a propagação de doenças infecciosas.

6.3 Relação com Políticas de Saúde Pública

Os resultados deste projeto fornecem uma base sólida para a formulação de políticas de saúde pública direcionadas à prevenção da sífilis congênita. A identificação de fatores de risco específicos permite que as autoridades de saúde desenvolvam estratégias mais eficazes e direcionadas, tais como:

1. **Monitoramento e Vigilância:**
    - **Sistemas de Vigilância Preditiva:** Implementar sistemas que utilizem modelos preditivos para identificar áreas e populações de alto risco, permitindo uma alocação mais eficiente de recursos.
    - **Integração de Dados:** Combinar dados de diferentes fontes para aprimorar a precisão dos modelos e a eficácia das intervenções.
2. **Alocação de Recursos:**
    - **Direcionamento de Recursos para Áreas Prioritárias:** Utilizar os insights dos modelos para direcionar recursos financeiros e humanos para regiões e grupos demográficos identificados como de alto risco.
    - **Planejamento Estratégico:** Incorporar as descobertas dos modelos no planejamento estratégico das ações de saúde pública, garantindo que as intervenções sejam baseadas em evidências robustas.

6.4 Considerações sobre os Atributos do Dataset

É importante notar que os atributos analisados neste estudo refletem uma variedade de fatores clínicos e socioeconômicos que podem influenciar os desfechos de saúde das gestantes e de seus bebês. A categorização detalhada de cada atributo permitiu uma análise mais granular e precisa dos fatores de risco associados à sífilis congênita.

Atributos Chave:

- **VDRL_RESULT:** Fundamental para a tarefa de classificação, indicando a presença ou ausência da sífilis.
- **CONS_ALCOHOL, SMOKER, PLAN_PREGNANCY, etc.:** Variáveis categóricas que capturam comportamentos e condições de saúde que podem influenciar os resultados.

- **LEVEL_SCHOOLING, FAM_INCOME, NUM_PREGNANCIES, NUM_RES_HOUSEHOLD:** Variáveis socioeconômicas que emergiram como os principais fatores de risco, destacando a interseção entre condições socioeconômicas e saúde pública.
- **AGE:** Utilizada tanto em tarefas de classificação quanto de regressão, proporcionando insights sobre a demografia das gestantes afetadas.

7. Conclusão

Este projeto demonstrou a aplicação eficaz de técnicas de aprendizado de máquina na análise de dados de saúde pública, especificamente nos casos de sífilis congênita no Brasil. Através da implementação de múltiplos modelos de classificação e regressão, identificamos fatores socioeconômicos e demográficos como determinantes críticos para os desfechos da doença.

Principais Achados:

1. **Desempenho dos Modelos de Classificação:**
    - **Modelos de Ensemble: Random Forest**, **XGBoost** e **K-Nearest Neighbors (KNN)** apresentaram os melhores desempenhos em termos de **Accuracy**, **Precision**, **Recall**, **F1-Score** e **ROC AUC**, com **Random Forest** alcançando uma **Accuracy** de **98.92%** e **ROC AUC** de **99.70%**.
    - **Modelos Lineares: Logistic Regression** apresentou desempenho moderado, com uma **Accuracy** de **64.42%** e **ROC AUC** de **68.20%**, indicando limitações na captura de relações complexas entre as variáveis.
    - **Modelos Não Lineares: Support Vector Machine (SVM)** com kernel linear apresentou desempenho inferior, com uma **Accuracy** de **48.92%**, sugerindo inadequação para este conjunto de dados específico.
2. **Desempenho dos Modelos de Regressão:**
    - **Modelos de Ensemble: Random Forest Regressor** e **XGBoost Regressor** se destacaram, alcançando **R^2** de **0.183015** e **0.184578** respectivamente, indicando uma capacidade moderada de explicar a variância na variável AGE.
    - **Modelos Lineares: Linear Regression** e **Ridge Regression** apresentaram **R^2** de aproximadamente **0.0355**, refletindo uma baixa capacidade de modelar a relação entre as variáveis independentes e a idade das gestantes.
    - **Modelos Não Lineares: Support Vector Regression (SVR)** e **KNN Regressor** mostraram desempenho insatisfatório, com **R^2** negativos e próximos de zero, sugerindo que não conseguiram capturar adequadamente as relações presentes nos dados.

3. **Importância das Variáveis:**
    - **Classificação:** As variáveis mais influentes na predição do resultado do exame VDRL foram:
        1. **Nível de Escolaridade (`LEVEL_SCHOOLING`)**
        2. **Número de Gravidezes (`NUM_PREGNANCIES`)**
        3. **Renda Familiar (`FAM_INCOME`)**
        4. **Número de Residentes no Ambiente Doméstico (`NUM_RES_HOUSEHOLD`)**

Interpretação dos Resultados:

Os modelos de ensemble, especialmente **Random Forest** e **XGBoost**, demonstraram uma alta capacidade preditiva, capturando relações complexas entre as variáveis. A identificação de **Nível de Escolaridade**, **Número de Gravidezes**, **Renda Familiar** e **Número de Residentes no Ambiente Doméstico** como os principais fatores de risco reforça a importância de condições socioeconômicas e demográficas na ocorrência e detecção da sífilis congênita.

8. Referências

- **Dataset:** Clinical and sociodemographic data on congenital syphilis cases, Brazil, 2013-2021
- **Artigo Base:** Predicting congenital syphilis cases: A performance evaluation of different machine learning models
- **Scikit-learn Documentation:** https://scikit-learn.org/stable/
- **Imbalanced-learn Documentation:** https://imbalanced-learn.org/stable/
- **XGBoost Documentation:** https://xgboost.readthedocs.io/en/latest/
- **LightGBM Documentation:** https://lightgbm.readthedocs.io/en/latest/

9. Autores

- **Henrique Leal, Gabriel Galdino e Pedro Tojal**
- **Disciplina:** Aprendizado de Máquina - 2024.2
- **Instituição:** CESAR School