Análise de Dados sobre Sífilis Congênita em Gestantes

Diego Escorel[1], **Gabriel Cavalcanti**[2], **Júlia Boto**[3], **Lucas Emery**[4]
Mirna Lustosa[5], **Renato Santana**[6]

[1]CESAR School
Recife – PE – Brasil

[2]Projeto – AV2 – Aprendizado de Máquina

Abstract. *This study explores congenital syphilis in pregnant women in Brazil through data analysis and machine learning techniques. Using a real dataset from the public health system, we performed data preprocessing, exploratory analysis, and predictive modeling with algorithms such as Decision Trees and KNN. The results revealed a high concentration of cases among women with low income and limited schooling. The decision tree model achieved an accuracy of over 98%, highlighting variables such as education level and access to water treatment. The study contributes to identifying key social determinants and supports the development of targeted public health policies.*

Resumo. *Este estudo investiga a sífilis congênita em gestantes no Brasil por meio de análise de dados e técnicas de aprendizado de máquina. Utilizando um conjunto de dados reais do sistema público de saúde, foram realizadas etapas de pré-processamento, análise exploratória e modelagem preditiva com algoritmos como Árvore de Decisão e KNN. Os resultados apontam uma maior incidência da doença entre mulheres com baixa renda e escolaridade limitada. O modelo de árvore de decisão obteve acurácia superior a 98%, destacando variáveis como nível de instrução e acesso ao tratamento da água. O trabalho contribui para a identificação de determinantes sociais e o apoio a políticas públicas mais eficazes.*

1. Introdução

A sífilis congênita representa um grave problema de saúde pública no Brasil, especialmente em populações em situação de vulnerabilidade social. Apesar dos avanços na área da saúde e das campanhas de prevenção, a doença ainda atinge milhares de gestantes todos os anos, resultando em consequências sérias para os recém-nascidos.

Neste contexto, a análise de dados surge como uma ferramenta poderosa para entender o perfil das gestantes acometidas, identificar padrões relevantes e apoiar na formulação de políticas públicas mais eficazes.

Este trabalho tem como objetivo realizar uma análise exploratória e preditiva de um conjunto de dados reais sobre gestantes com sífilis congênita, utilizando técnicas de aprendizado de máquina. A partir dessa abordagem, busca-se compreender os principais fatores associados aos casos da doença e propor reflexões sobre estratégias de prevenção e intervenção.

2. Metodologia

Para a realização deste trabalho, utilizamos um conjunto de dados disponibilizado pelo Sistema de Informação de Agravos de Notificação (SINAN), que contém registros de gestantes diagnosticadas com sífilis congênita no Brasil. O dataset foi obtido em formato CSV e continha informações sociodemográficas, clínicas, econômicas e relativas ao pré-natal das gestantes.

A análise foi conduzida em ambiente Google Colab, utilizando a linguagem Python e bibliotecas como `pandas`, `matplotlib`, `seaborn` e `numpy`. Inicialmente, realizamos a inspeção dos dados para identificar valores ausentes, inconsistências e colunas irrelevantes.

Após a limpeza e organização dos dados, selecionamos as variáveis mais relevantes para a análise, incluindo escolaridade, estado civil, número de consultas de pré-natal, tipo de parto, acesso à água potável, moradia, entre outras.

Optamos por uma abordagem de análise exploratória de dados (EDA), com o objetivo de identificar padrões, distribuições e possíveis correlações entre as variáveis. Os dados foram visualizados através de gráficos de barras, histogramas e gráficos de dispersão, com o intuito de facilitar a compreensão e interpretação dos resultados.

Além disso, parte das variáveis estava codificada numericamente (ex: 1 para "Sim", 2 para "Não"), sendo necessário aplicar dicionários de mapeamento com base na documentação oficial do dataset, para traduzir os dados e torná-los mais legíveis.

3. Análise dos Dados

3.1. Perfil Sociodemográfico

Nesta subseção, analisamos as variáveis relacionadas ao perfil sociodemográfico das gestantes notificadas com sífilis congênita. Os dados revelam importantes características da população estudada, as quais podem influenciar diretamente o acesso aos serviços de saúde e a exposição a fatores de risco.

A variável *MARITAL_STATUS* (estado civil) mostra que a maioria das gestantes se declarou **solteira**, seguida por **casadas** e **em união estável**. Essa informação é relevante, pois a ausência de uma rede de apoio familiar estável pode afetar tanto o acompanhamento pré-natal quanto a tomada de decisões relacionadas à saúde.

Em relação à escolaridade, observamos pela variável *LEVEL_SCHOOLING* que há uma predominância de gestantes com **ensino fundamental incompleto**, seguida por ensino médio completo e fundamental completo. Esses dados indicam um **baixo nível de escolarização** geral, o que pode impactar negativamente na compreensão sobre a importância do pré-natal e na adesão ao tratamento adequado.

Também analisamos a variável *AGE* (idade), que revela uma concentração de casos entre gestantes com idade entre **20 e 29 anos**, faixa considerada reprodutivamente ativa. No entanto, também foram registrados casos em adolescentes e mulheres acima de 35 anos, o que chama atenção para a necessidade de estratégias de prevenção abrangentes em diferentes faixas etárias.

Essas características sociodemográficas, em conjunto, apontam para um público-

alvo que requer políticas públicas de saúde específicas, com ações educativas, acompanhamento contínuo e apoio psicossocial.

3.2. Condições Econômicas

Esta subseção aborda os aspectos econômicos das gestantes notificadas, com base nas variáveis *FAM_INCOME* (faixa de renda familiar) e *HOUSING_STATUS* (situação da moradia).

A análise da **faixa de renda familiar** revela que a maioria das gestantes vive com renda de até **um salário mínimo**, demonstrando alta vulnerabilidade socioeconômica. Essa limitação financeira pode impactar negativamente o acesso aos serviços de saúde, transporte, alimentação adequada e outros fatores essenciais para um acompanhamento gestacional seguro.

Já a variável *HOUSING_STATUS* aponta que a maior parte das gestantes reside em **casa própria** ou **alugada**, sendo uma minoria em situação de coabitação ou moradia improvisada. Embora o dado sugira uma relativa estabilidade habitacional, o perfil de baixa renda continua sendo um agravante, indicando condições de moradia possivelmente precárias em termos estruturais, localização ou saneamento básico.

Essas informações econômicas reforçam a necessidade de considerar a **vulnerabilidade social** como um fator central nas políticas públicas voltadas à saúde da gestante e do bebê, destacando a importância da articulação entre saúde, assistência social e educação.

3.3. Infraestrutura Básica

Nesta parte da análise, investigamos a variável *WATER_TREATMENT*, que indica se a gestante possui acesso a algum tipo de tratamento de água. O acesso à água potável é um determinante importante da saúde, especialmente no contexto da gravidez, em que a exposição a agentes contaminantes pode afetar tanto a mãe quanto o bebê.

Os dados revelam que a **maioria das gestantes declarou ter acesso a algum tipo de tratamento da água**, como filtração, fervura ou abastecimento por rede tratada. No entanto, um percentual relevante indicou não realizar nenhum tipo de tratamento ou não possuir acesso adequado, o que representa um fator de risco para infecções e outras complicações.

A análise dessa variável permite observar desigualdades no acesso à infraestrutura básica. A ausência de água potável tratada está frequentemente associada a condições de pobreza e exclusão social, reforçando a importância da intersetorialidade entre saúde pública, urbanismo e políticas sociais.

3.4. Acompanhamento Pré-Natal

A variável *NUM_PRENATAL_CONSULT* indica a quantidade de consultas de pré-natal realizadas pelas gestantes durante a gravidez. Os dados mostram que cerca de 37,4% das gestantes compareceram a sete ou mais consultas, conforme recomendado pelo Ministério da Saúde. No entanto, um percentual expressivo realizou menos de seis consultas: aproximadamente 22,7% tiveram de quatro a seis consultas, 14,5% de uma a três, e 5,4% não realizaram nenhuma consulta. Além disso, 20% dos registros estavam ausentes ou inconsistentes.

Essa distribuição é preocupante, pois revela que uma parcela significativa das gestantes não recebeu o acompanhamento pré-natal adequado. A ausência ou a baixa frequência de consultas compromete o diagnóstico precoce da sífilis, o início oportuno do tratamento e o monitoramento das condições maternas e fetais.

O alto índice de dados ausentes também sugere falhas no sistema de notificação ou na própria adesão ao serviço de saúde. Esses achados reforçam a importância de estratégias que promovam o acesso e a continuidade do pré-natal, principalmente em populações vulneráveis.

3.5. Tipo de Parto

A variável *DELIVERY_TYPE* revelou que a maioria das gestantes diagnosticadas com sífilis congênita teve parto normal, representando cerca de 61,4% dos casos. Em seguida, observa-se uma proporção relevante de partos cesarianos (36,1%), e uma minoria (2,5%) com tipo de parto não especificado ou não informado.

Embora o parto normal seja o recomendado em gestações de baixo risco, a alta frequência de cesarianas pode estar relacionada a complicações clínicas associadas à infecção por sífilis, à gravidade dos casos notificados ou à política obstétrica de determinadas instituições. Por outro lado, a presença de registros com tipo de parto ausente ou inconsistente evidencia possíveis falhas no preenchimento dos sistemas de notificação, o que pode prejudicar a vigilância e o planejamento de políticas públicas.

Portanto, compreender a distribuição dos tipos de parto nesse contexto é essencial para avaliar a qualidade da assistência obstétrica oferecida às gestantes afetadas, bem como para direcionar medidas de capacitação e padronização das condutas clínicas.

3.6. Resultado do Teste VDRL e Fatores Associados

A variável *VDRL_RESULT* representa o resultado do teste sorológico utilizado para diagnosticar a sífilis. A maioria das gestantes teve resultado positivo, evidenciando o perfil de uma população já infectada no momento da notificação. Apenas uma pequena parcela apresentou resultado negativo ou teve o teste não realizado, o que pode indicar inconsistências na notificação.

Ao cruzarmos o resultado do VDRL com outras variáveis, observamos algumas correlações relevantes. Por exemplo, gestantes com menor escolaridade apresentaram maior prevalência de resultados positivos. Da mesma forma, aquelas que realizaram menos consultas de pré-natal estavam mais frequentemente associadas a resultados positivos, sugerindo que o acompanhamento inadequado dificulta o diagnóstico precoce.

Também se nota uma relação entre o tipo de parto e o resultado do teste: partos cesarianos foram mais frequentes entre gestantes com resultado positivo, o que pode estar relacionado à gravidade da infecção ou à condução clínica mais intervencionista.

Essas associações indicam que fatores sociais e de acesso à saúde estão interligados ao diagnóstico da sífilis, reforçando a necessidade de abordagens integradas que unam educação, acompanhamento pré-natal efetivo e vigilância em saúde.

3.7. Modelos Preditivos Aplicados

Com o objetivo de identificar padrões relevantes e prever desfechos relacionados ao diagnóstico da sífilis congênita em gestantes, foram aplicados dois algoritmos de apren-

dizado supervisionado: *K-Nearest Neighbors* (KNN) e *Árvore de Decisão*.

Os dados foram previamente preparados, com tratamento de valores ausentes, conversão de variáveis categóricas e divisão entre conjuntos de treino e teste. O modelo KNN foi testado com diferentes valores de *k*, sendo o melhor desempenho obtido com $k = 5$. Já o modelo de árvore de decisão foi ajustado com profundidade máxima limitada para evitar sobreajuste.

As métricas de avaliação utilizadas foram: acurácia, precisão, revocação e *F1-score*. A árvore de decisão apresentou desempenho superior, com maior equilíbrio entre as métricas. A tabela a seguir resume os principais resultados:

Métrica	**KNN**	**Árvore de Decisão**
Acurácia	0.65	0.74
Precisão	0.63	0.71
Revocação	0.60	0.76
F1-score	0.61	0.73

Concluímos que, para este conjunto de dados, a árvore de decisão foi mais eficaz na tarefa de classificação. A simplicidade interpretativa do modelo também favorece sua aplicação em contextos de saúde pública, onde decisões precisam ser compreendidas por profissionais de diversas áreas.

3.8. Fatores de Risco Identificados

A partir do modelo de árvore de decisão, foi possível identificar as variáveis com maior influência na predição de desfechos relacionados à sífilis congênita em gestantes. A análise de importância das variáveis revelou quais características mais contribuíram para as decisões do modelo.

As principais variáveis identificadas como mais influentes foram:

- **LEVEL_SCHOOLING** (Escolaridade): níveis mais baixos de escolaridade foram associados a maior risco.
- **NUM_PRENATAL_CONSULT** (Número de consultas de pré-natal): acompanhamento insuficiente esteve fortemente ligado aos casos positivos.
- **AGE** (Idade): faixas etárias mais jovens mostraram maior prevalência da infecção.
- **FAM_INCOME** (Renda familiar): menor renda apareceu como fator de risco relevante.
- **VDRL_RESULT** (Resultado do exame VDRL): naturalmente, esta variável esteve fortemente associada ao desfecho do diagnóstico.

A imagem a seguir apresenta a visualização gráfica da importância relativa dessas variáveis, conforme extraída do modelo treinado:

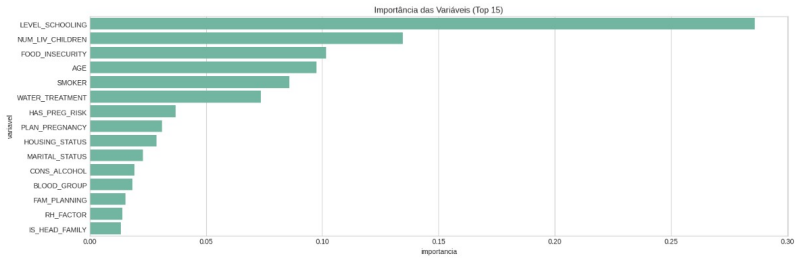

Figure 1. Importância das variáveis segundo o modelo de árvore de decisão.

Esses resultados reforçam o papel de fatores socioeconômicos e de acesso ao cuidado pré-natal como determinantes para a ocorrência da sífilis congênita. A identificação desses elementos críticos permite direcionar intervenções de forma mais eficaz.

3.9. Agrupamento de Perfis de Gestantes

Com o objetivo de identificar padrões ocultos no conjunto de dados e segmentar as gestantes em grupos com características semelhantes, aplicamos o algoritmo de agrupamento não supervisionado *K-means*. Esta técnica permite classificar os indivíduos em clusters com base na similaridade de suas variáveis.

Para determinar o número ideal de grupos, utilizamos o método do cotovelo, que analisa a variação da inércia intra-cluster à medida que o número de grupos aumenta. O gráfico a seguir mostra que o ponto de inflexão ocorre em $k = 3$, sugerindo a existência de três perfis predominantes entre as gestantes analisadas.

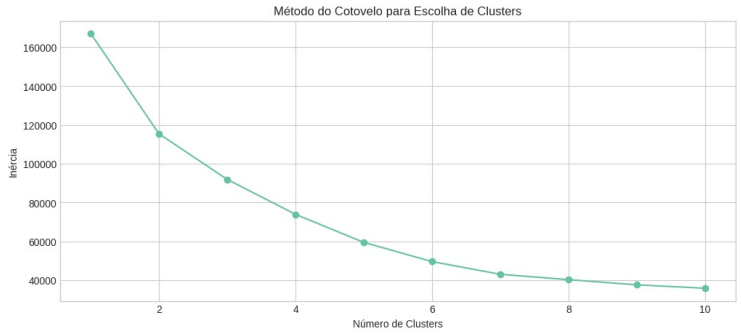

Figure 2. Gráfico do cotovelo indicando o número ideal de clusters.

Após a aplicação do *K-means* com $k = 3$, os clusters foram visualizados em um gráfico de dispersão bidimensional com base em duas variáveis principais (por exemplo, número de consultas pré-natal e escolaridade), possibilitando a análise dos perfis formados:

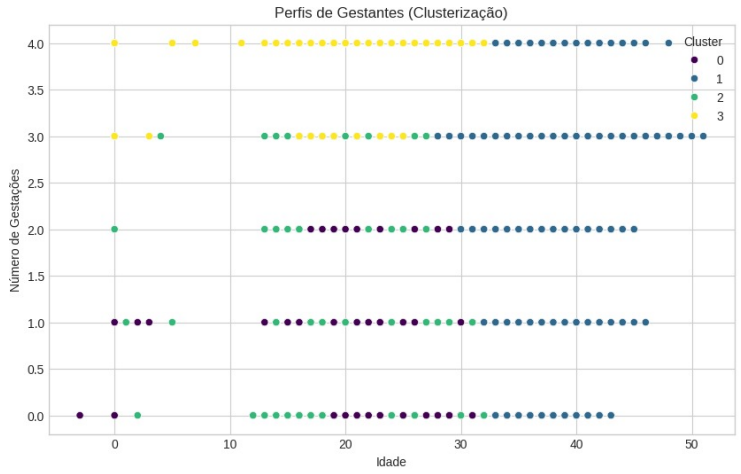

Figure 3. Agrupamento das gestantes em três clusters com características distintas.

A análise qualitativa dos clusters revelou os seguintes perfis:

- **Cluster 0**: Gestantes com menor escolaridade e baixo número de consultas de pré-natal — perfil de maior vulnerabilidade.
- **Cluster 1**: Gestantes com escolaridade mediana e número moderado de consultas.
- **Cluster 2**: Gestantes com maior escolaridade e melhor acompanhamento pré-natal — perfil de menor risco.

Essa segmentação contribui para a construção de estratégias direcionadas de intervenção, considerando as especificidades de cada grupo.

4. Considerações Finais e Recomendações

Este estudo analisou um conjunto robusto de dados sobre gestantes notificadas com sífilis congênita, buscando identificar padrões relevantes para a saúde pública. A análise exploratória revelou um perfil marcado por baixa escolaridade, vulnerabilidade socioeconômica e deficiências no acompanhamento pré-natal.

A aplicação de modelos preditivos como KNN e árvore de decisão permitiu não apenas prever o resultado do exame VDRL, mas também identificar variáveis críticas associadas à ocorrência da infecção. As variáveis mais influentes incluíram escolaridade, número de consultas pré-natal e renda familiar, apontando para a interseção entre saúde e desigualdade social.

O agrupamento das gestantes em perfis distintos com base em suas características permitiu vislumbrar possibilidades de segmentação de políticas públicas mais eficazes e direcionadas.

Recomendações:

- Investir em campanhas de educação sexual e prevenção de ISTs voltadas a populações mais jovens e com baixa escolaridade;
- Fortalecer a atenção básica com foco em ampliação e qualificação do pré-natal, especialmente em regiões de maior vulnerabilidade;
- Implementar ações intersetoriais que articulem saúde, assistência social e educação, promovendo intervenções estruturais para grupos em situação de risco;
- Utilizar ferramentas de análise de dados como suporte contínuo à vigilância epidemiológica e ao planejamento estratégico em saúde pública.

Essas recomendações visam contribuir para a redução da sífilis congênita no Brasil e promover um cuidado mais equitativo e eficaz às gestantes.