Análise de casos de diagnósticos de sífilis em gestantes utilizando arvores de decisão, KNN e SVM

Getúlio Santos Mendes, João Gustavo Silva Guimarães, João Pedro Freitas de Paula Dias

15 de fevereiro de 2025

1 Introdução e Problemática

A sífilis em gestantes permanece um desafio crítico para a saúde pública global, com repercussões graves para a saúde materno-infantil. Segundo (ORGANIZAÇÃO MUNDIAL DA SAÚDE, 2024), aproximadamente 7,1 milhões de novos casos de sífilis foram registrados em 2022, sendo a transmissão vertical responsável por mais de muitas mortes fetais e neonatais anuais, além de complicações como prematuridade, baixo peso ao nascer e sequelas neurológicas irreversíveis. A persistência desses agravos, mesmo diante de um tratamento eficaz e acessível — a penicilina —, revela lacunas estruturais nos sistemas de saúde, desde a oferta de pré-natal de qualidade até a vigilância epidemiológica.

No contexto latino-americano, a situação é ainda mais preocupante. Dados da Organização Pan-Americana da Saúde (OPAS, 2024) apontam que a região das Américas concentra uma das maiores taxas de sífilis congênita do mundo, com 3,37 milhões de casos, refletindo desigualdades no acesso a diagnósticos precoces e tratamentos adequados. A subnotificação de casos, o estigma social associado à infecção e a fragmentação das redes de cuidado ampliam a complexidade do cenário, exigindo análises que transcendam a mera descrição estatística.

Este artigo propõe-se a analisar dados epidemiológicos de diagnósticos de sífilis em gestantes, explorando padrões temporais, distribuição geográfica e fatores de risco associados, por meio de técnicas de inteligência artificial (IA) como árvores de decisão, KNN (K-Nearest Neighbors) e SVM (Support Vector Machines). Esses algoritmos permitem modelar relações complexas entre variáveis clínicas, socioeconômicas e geográficas, identificando não apenas correlações estatísticas, mas padrões preditivos que orientem intervenções direcionadas.

Ao analisar dados de gestantes notificadas com sífilis por Unidades Básicas de Saúde(UBS) em Montes Claros e alguns de seus municípios, pode-se entender que é possível inferir a zona de residência das gestantes com base em seu tipo de tratamento e do tipo de tratamento do parceiro.

Além de tornar possível a criação de um modelo isso possibiliza traçar um perfil específico, de que pessoas com um tipo específico de tratamento de sífilis tendem à ter uma parceira gestante residindo em uma zona específica, ou que as gestantes que estão aderindo um tratamento específico tendem à residir em uma zona específica.

2 Metodologia

Para a implementação dos algoritmos de inteligência artificial, foi usado a biblioteca para *python sklearn*, devido á facilidade de implementação dos modelos de aprendizado e também pelo fator de visualização. O código esta disponível em: Disponível em: Link do repositorio.

Uma árvore de decisão binária é uma estrutura hierárquica utilizada para tomada de decisões, onde cada nó interno representa um teste em um atributo, cada ramo corresponde a um resultado do teste e cada folha indica uma decisão ou classificação. Ao utilizar essa abordagem, foram consideradas várias profundidades diferentes para obter a melhor acurácia de predição. A árvore final foi concebida com a menor profundidade responsável pela melhor acurácia alcançada.

O **K-Nearest Neighbors (KNN)** é um algoritmo de aprendizado baseado em instâncias que classifica um novo ponto com base na proximidade dele em relação aos pontos de treinamento. A classificação é determinada pela maioria dos vizinhos mais próximos. Para avaliar seu desempenho, foram testados diferentes valores de **K**, buscando encontrar aquele que proporcionasse a melhor acurácia.

O **Support Vector Machine (SVM)** é um método de aprendizado supervisionado que busca encontrar um hiperplano ótimo para separar os dados em diferentes classes. Dependendo da distribuição dos dados, diferentes *kernels* podem ser utilizados para transformar o espaço de entrada e melhorar a separação. Neste estudo, foram testados os *kernels Radial Basis Function* (RBF), polinomial, sigmoide e linear, selecionando aquele que resultasse

na melhor acurácia.

2.1 Resultados e Discussão

A acurácia obtida através dos três métodos não se deferiram, isso pode ser causado pelo nosso *fine tunning* de cada método para o seu melhor resultado, o pequeno tamanho da base de dados ou uma demonstração da utilidade e convergência de diferentes métodos. Para uma análise mais completa de cada método também se deve considerar a performance e tempo de treinamento de cada um, mas como nossa base é pequena, essa análise se torna difícil de se efetuar.

Para a Arvore de decisão binária utilizamos uma variedade de profundidades para conseguirmos ver a variança de acurácia que esse parâmetro proporciona para o modelo:

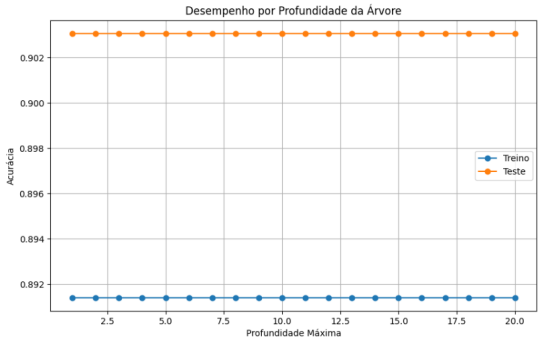

Figura 1

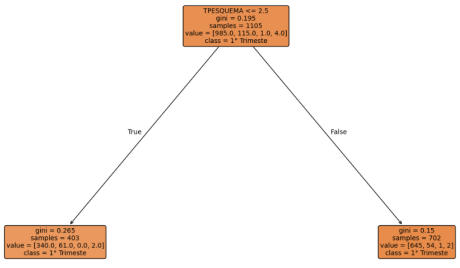

Figura 2

Temos que para o algoritmo da arvore de decisão a melhor profundidade foi de $\mathbf{K = 1}$ com $\mathbf{90}\%$ de acurácia.

No Algoritmo do **KNN (K-Nearest Neighbors)** o valor de **K** é correspondente á quantidade de vizinhos a serem analisados, a determinação do rótulo é feita pela análise de qual é a classificação majoritária dentre os Vizinhos. Para os valores de **K** e acurácia obtidos temos a tabela e o Gráfico:

k	Acurácia
1	0.51
3	0.87
5	0.87
7	0.87
9	0.90
11	0.90
13	0.90
15	0.90
17	0.90
19	0.90

Tabela 1: Valores de k no KNN e suas respectivas acurácias

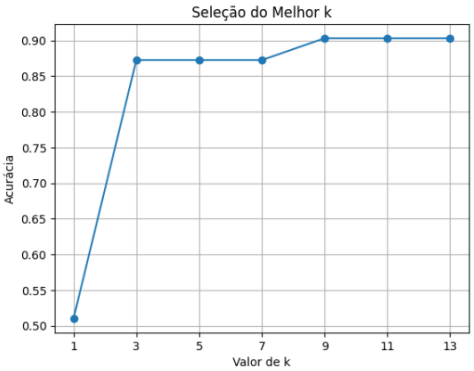

Figura 3: Acurácia por K no KNN

Temos que o melhor **K** encontrado foi o $\mathbf{K = 9}$ com a acurácia de 90%.

No modelo com base em **SVM (Support Vector Machine)**. O parâmetro **C** é o fator de regularização que controla o equilíbrio entre maximizar a margem de separação e minimizar os erros de classificação no conjunto de treinamento.

Valores menores de **C** permitem uma margem mais ampla, tolerando mais erros e promovendo melhor generalização. Em contrapartida, valores maiores de **C** penalizam fortemente os erros, buscando classificar corretamente todos os pontos de treinamento, o que pode levar ao *overfitting*.

Apresenta-se uma tabela com os resultados obtidos para diferentes configurações do modelo implementado de **SVM**:

Configuração	Acurácia
SVM Linear com diferentes valores de C	
C = 0.001	0.90
C = 0.010	0.90
C = 0.100	0.90
C = 1.000	0.90
C = 10.000	0.90
C = 100.000	0.90
SVM com diferentes kernels	
Kernel: linear	0.90
Kernel: rbf	0.90
Kernel: poly	0.90
Kernel: sigmoid	0.89

Tabela 2: Acurácias para diferentes configurações de SVM.

A melhor configuração foi alcançada com o kernel linear e C = 0.001.

3 Conclusão

A análise dos dados de gestantes notificadas com sífilis, provenientes das UBS de Montes Claros e municípios adjacentes, permitiu identificar padrões relevantes entre o tipo de tratamento adotado pelas gestantes e seus parceiros e a zona de residência. Os modelos preditivos desenvolvidos – por meio de Árvore de Decisão, KNN e SVM – apresentaram desempenho consistente, alcançando acurácias idênticas á 90%.

Esse resultado pode ser interpretado de duas maneiras: por um lado, os diferentes métodos, mesmo com abordagens teóricas diversas, convergiram para resultados similares, sugerindo que os padrões presentes na base de dados são robustos e bem definidos; por outro, o alto desempenho pode estar relacionado ao ajuste fino de cada algoritmo e ao tamanho relativamente pequeno do *dataset*, o que pode limitar a generalização dos modelos para cenários mais complexos.

Em síntese, os resultados indicam a viabilidade de utilizar modelos preditivos para traçar perfis específicos de gestantes e seus parceiros, possibilitando intervenções mais direcionadas na área da saúde pública. Contudo, para uma validação mais abrangente dos modelos e a consolidação dos padrões identificados, futuras análises deverão considerar bases de dados mais amplas e a inclusão de métricas adicionais, como tempo de treinamento e complexidade computacional.

4 Referências

CORMEN, T. H. et al. Introduction to Algorithms, third edition. [s.l.] MIT Press, 2009. Acessado em 20 de Março de 2023.

RUSSEL, STUART, and PETER Norvig. Inteligência Artificial. 3rd ed., Elsevier Editora Ltda., 2013.

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE. Casos de sífilis aumentam nas Américas. Washington, DC: OPAS, 22 de maio de 2024. Disponível em: `https://www.paho.org/pt/noticias/22-5-2024-casos-sifilis-aumentam-nas-americas`. Acessado em 14 de fevereiro de 2025.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. Implementando as estratégias globais do setor de saúde para HIV, hepatites virais e infecções sexualmente transmissíveis, 2022–2030: relatório sobre progressos e lacunas. Genebra: OMS, 2024. Disponível em: `https://www.who.int/publications/i/item/9789240094925`. Acessado em 14 de fevereiro de 2025.

DISCENTES DA UNIFIPMOC. Expansão da Base de Dados de Gestantes Notificadas com Sífilis. [s.l.]: UNIFIPMOC, 2025. Base de dados original disponível em: `https://portalsinan.saude.gov.br/sifilis-em-gestante`. Expansão realizada por alunos da UNIFIPMOC. Acessado em 14 de fevereiro de 2025.

Silva, J. G. G. *KNN_SVM_on_pregnant-women-with-syphilis.* GitHub repository. Disponível em: `https://github.com/jAzz-hub/KNN_SVM_on_pregnant-women-with-syphilis`. Acessado em: 14 de fevereiro de 2025.