Imunidade na Gravidez: Diferenças entre Populações Tsimane e Estadounidense

Jonatan Andres Gomez Parada

1. Introdução

Entre os mamíferos placentários, a reprodução feminina exige que o organismo da mãe tolere a presença de um feto geneticamente diferente durante toda a gestação, sem rejeitá-lo. Em humanos, esse processo é particularmente desafiador devido à natureza invasiva da placenta e exige mudanças no sistema imunológico da mãe, o que pode alterar sua suscetibilidade a infecções e doenças autoimunes.

Até hoje, a maior parte dos estudos sobre tolerância fetal foi feita em países industrializados e de alta renda, onde a exposição a microrganismos ao longo da vida é muito menor. Essa privação microbiana reduz a mortalidade por doenças infecciosas, mas também compromete o desenvolvimento equilibrado do sistema imunológico, podendo levar a uma maior sensibilidade a estímulos inofensivos e ao surgimento de doenças inflamatórias crônicas, como alergias e artrite reumatoide.

Este estudo compara os efeitos da gravidez na imunidade materna em dois grupos muito diferentes: mulheres da comunidade Tsimane, que vivem em um ambiente com alta carga de patógenos e têm fertilidade natural, e mulheres nos EUA.

Os Tsimane habitam os neotrópicos da Bolívia, um ambiente rico em biodiversidade e repleto de inúmeros agentes infecciosos. Devido à exposição variada e crônica, as infecções (principalmente respiratórias e gastrointestinais) são as principais causas de morbidade e mortalidade entre os Tsimane, enquanto alergias, atopia e doenças autoimunes são raras.

Serão considerados os seguintes tipos de células:

- Leucócitos totais: Também conhecido como contagem de glóbulos brancos (WBC).
- Linfócitos: Responsáveis pela imunidade específica contra antígenos.
- Neutrófilos: Combatem infecções extracelulares e fúngicas, ativam a resposta adaptativa e promovem inflamação.
- Monócitos: Semelhantes aos neutrófilos, porém menos numerosos e com vida mais curta. Migram para locais de infecção e se transformam em macrófagos especializados nos tecidos.
- Eosinófilos: Granulócitos envolvidos na resposta a parasitas e alergias.
- Basófilos: Tipo raro de granulócito envolvido em inflamações.

2. Informações gerais do conjunto de dados e análise exploratória dos dados.

O conjunto de dados é formado pela união de duas fontes. A primeira delas referente ao *Tsimane Health and Life History Project* (THLHP) correspondente a dados de mulheres Tsimane. A segunda fonte de dados corresponde a dados de mulheres morando nos Estados Unidos, e são referentes ao National Health and Nutrition Examination Survey (NHANES, 2003–2016). A idade das mulheres das duas fontes de dados varia entre 18 e 45 anos.

2.1 Descrição dos dados

O dataset inicial contém 6 grupos de variáveis no conjunto de dados distribuídos da seguinte forma:

- Contagem Leucocitária (células/μL)
    - WBC: Leucócitos totais
    - NEU: Neutrófilos
    - LYM: Linfócitos
    - MON: Monócitos
    - EOS: Eosinófilos
    - BAS: Basófilos
- Proporções Leucocitárias (%)
    - neu_pct: % de neutrófilos
    - lym_pct: % de linfócitos
    - mon_pct: % de monócitos
    - eos_pct: % de eosinófilos
    - bas_pct: % de basófilos
- Marcadores Inflamatórios
    - crp: Proteína C-reativa (mg/L)
- Dados Antropométricos
    - BMI: Índice de massa corporal (kg/m²)
    - Age: Idade em anos
- Dados Reprodutivos
    - NumPartos: Número de partos
    - RepStatus: Status reprodutivo (Cycling: Não grávida. T1, T2, T3 : Trimestre da gravidez)
- Metadados e Controle
    - pid: Identificador único do participante
    - Population: Origem populacional (THLHP:Tsimane, NHANES: Mulheres EUA)
    - Repeats: Número de repetições
    - REF: Indicador de medidas repetidas (0=não, 1=sim)

Inicialmente o dataset contém dados de 2330 mulheres, sendo 935 mulheres Tsimane e 1395 mulheres dos EUA, das quais 256 Tsimane são grávidas e dos EUA temos 277 grávidas.

Após análise das colunas correspondentes à contagem Leucocitária, considerando distribuição e outliers, foram mantidos os dados satisfazendo as seguintes condições:

- WBC: Leucócitos totais < 22000 células/μL
- NEU: Neutrófilos < 15000 células/μL
- LYM: Linfócitos < 8000 células/μL
- MON: Monócitos < 1250 células/μL
- EOS: Eosinófilos < 5000 células/μL
- BAS: Basófilos < 400 células/μL

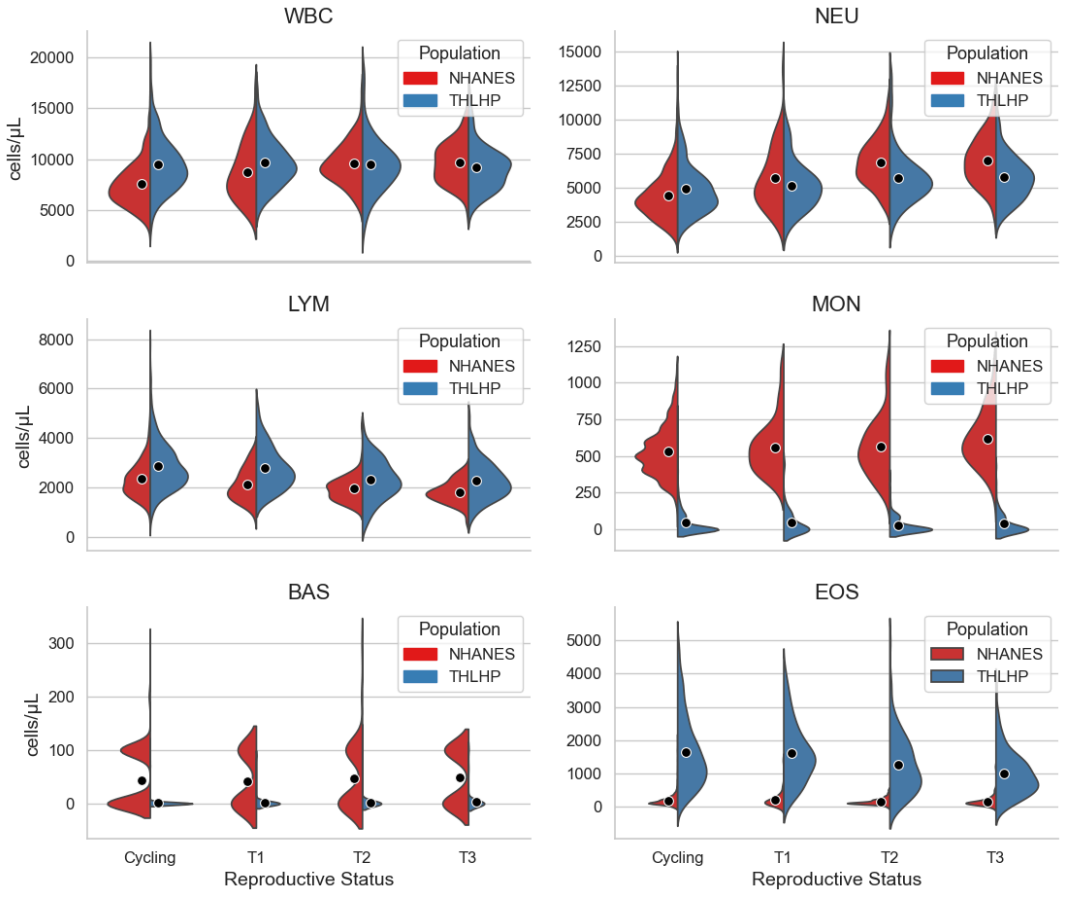
No total foram removidas 106 linhas, onde 86 linhas correspondem a valores NaN em todas as colunas de contagem leucocitária. Logo, foram removidas 20 linhas de dados que não satisfaziam as condições acima mencionadas. Além disso, a coluna referente a variável crp apresentava 64% de valores NaN, essa coluna foi desconsiderada, assim como as colunas pid, Repeats e REF. As colunas sobre Proporções Leucocitárias foram desconsideradas, uma vez que a sua informação é obtida a partir das respectivas colunas de contagem leucocitárias e a contagem leucocitária total,

além do fato de, neste estudo, termos principal interesse na contagem total e não nas porcentagem das células leucocitárias.

Também foi realizado o seguinte procedimento: Foram criadas duas colunas a partir da coluna RepStatus chamadas RepStatus_cat e RepStatus_bin. RepStatus_bin assume valores 0 e 1, onde 0 corresponde a mulheres não grávidas e 1 a mulheres grávidas. Enquanto RepStatus_cat assume valores 0, 1, 2 e 3, sendo 0 para não grávidas, e para grávidas assume valores 1, 2 ou 3 a depender do trimestre da gravidez.

	Conjunto inicial de dados		Relação de dados após tratamento	
	Tsimane	EUA	Tsimane	EUA
Grávidas 1 trimestre	71	53	70	52
Grávidas 2 trimestre	94	113	93	104
Grávidas 3 trimestre	91	111	91	106
Não Grávidas	679	1118	653	1055
Total	935	1395	907	1317

A seguinte imagem mostra a distribuição das diferentes contagem leucocitária no conjunto de dados final, agrupando pelo tipo de população e estado da gravidez (ou não gradivez).



Note-se que temos maior diferenciais nos valores correspondentes a os Monócitos, Basófilos e Eosinófilos.

3. Análises Estatísticas

Temos especial interesse em conhecer se existe uma diferença na média das contagens das células leucocitárias, quando consideramos o tipo de população e estado da gravidez. Para isto, consideramos Testes Mann-Whitney U, para testar hipóteses de um dos grupos populacionais ter menores valores em cada uma das contagens. Na seguinte tabela são apresentadas as interpretações dos resultados dos testes Mann-Whitney U, indicando qual dos grupos populacionais têm menor contagem por tipo de célula.

	Leucócitos	Neutrófilos	Linfócitos	Monócitos	Basófilos	Eosinófilos
Não Grávida	NHANES	NHANES	NHANES	THLHP	THLHP	NHANES
T1	NHANES	*Iguais*	NHANES	THLHP	THLHP	NHANES
T2	*Iguais*	THLHP	NHANES	THLHP	THLHP	NHANES
T3	*Iguais*	THLHP	NHANES	THLHP	THLHP	NHANES

Chamamos principalmente a atenção à contagem total de Leucócitos, a qual não tem resultado menor para o grupo dos EUA em não grávidas e grávidas no primeiro trimestre, mas que não representa diferença significativa nos dois últimos trimestres, assim como para a classe dos Neutrófilos que também apresenta variação entre não grávidas até grávidas no terceiro trimestre.

4. Agrupamento K-Means

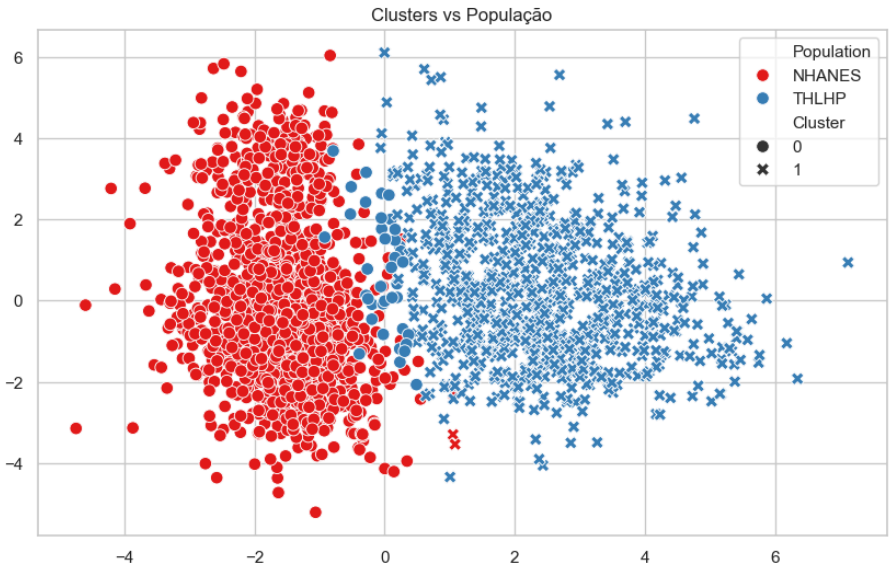
Na Análise estatística pode ser vista a diferença existente na contagem dos diferentes tipos de células entre os dois tipos de população nas diferentes classificações da gravides.
A seguir, com o propósito de determinar uma possível diferença considerando um conjunto maior das variáveis, consideramos a implementação de um modelo K-Means, com dois clusters, excluindo a variável Population, e fazemos uma comparação da distribuição de cada um dos dois tipos de população nos agrupamentos fornecidos pelo modelo K-Means.

As variáveis consideradas no modelo foram: WBC, NEU, LYM, MON, BAS, EOS, lym_pct, neu_pct, eos_pct, mon_pct, bas_pct, BMI, Age, NumPartos, RepStatus_bin, RepStatus_cat.

A distribuição obtida é apresentada na seguinte tabela:

População	NHANES	THLHP
Cluster		
0	1314	39
1	3	868

Pode-se considerar que existe sim a nível geral das variáveis uma diferença entre as duas diferentes populações no estudo. Agora, para apreciação gráfica consideramos uma Análise de Componentes Principais (PCA) para redução da dimensionalidade, e obtemos o seguinte gráfico apresentando a distribuição dos tipos de populações nos agrupamentos do K-Means.



5. Modelo de regressão para estimativa da contagem de neutrófilos.

Nas análises prévias, foi identificado que há diferenças estatisticamente significativas na contagem de células sanguíneas entre duas populações distintas. Dentre essas, a temos interesse na contagem de neutrófilos (NEU) por seu papel central na resposta imune inata e sua relevância clínica no contexto de infecções, inflamação e alterações fisiológicas como a gestação.

A proposta de utilizar um modelo de regressão para estimar os níveis de NEU a partir de outras variáveis hematológicas, clínicas e demográficas é fundamentada tanto em aspectos biológicos quanto práticos:

- Os neutrófilos constituem normalmente a fração mais abundante dos leucócitos. Variações em WBC tendem a refletir proporcionalmente alterações nos níveis de NEU.
- O Índice de Massa Corporal e a Idade são variáveis que podem influenciar na distribuição de células imunológicas. Por exemplo, estados de sobrepeso/obesidade estão associados a inflamação crônica de baixo grau, o que pode elevar contagens de neutrófilos.
- O Número de Partos e as variáveis RepStatus_cat (estado gestacional) e RepStatus_bin (gestante/não gestante) permitem capturar o histórico reprodutivo e o estado hormonal atual da mulhe.
- Population_bin, representando o grupo populacional de origem (NHANES ou THLHP), é fundamental para considerar possíveis diferenças genéticas, ambientais ou nutricionais que possam influenciar os parâmetros hematológicos.

Do ponto de vista operacional, muitas dessas variáveis podem ser obtidas de forma simples e acessível: idade, número de partos e estado gestacional são dados autodeclarados ou extraídos de prontuário, enquanto WBC e BMI são medidas de baixo custo e rotineiramente disponíveis em exames laboratoriais e avaliações clínicas básicas. Já a contagem

diferencial de leucócitos (incluindo NEU), embora mais informativa, pode ser inviável em contextos com infraestrutura laboratorial limitada.

Portanto, desenvolver um modelo que permita estimar NEU com base em WBC e outras variáveis clínicas acessíveis representa uma solução custo-efetiva, com potencial aplicação em triagens rápidas, monitoramento de populações vulneráveis e ampliação do acesso à informação clínica, mesmo em cenários de recursos restritos.

5.1 Escolha do modelo

Para estimar a contagem de neutrófilos (NEU) com base em variáveis clínicas e demográficas, diferentes modelos de regressão foram testados, incluindo modelos lineares e não lineares, a fim de identificar aqueles com melhor desempenho preditivo.

O conjunto de dados foi dividido em subconjuntos de treino (80%) e teste (20%), com padronização das variáveis explicativas para garantir comparabilidade entre modelos sensíveis à escala.

Os modelos avaliados (nos hiperparâmetros default) foram:

- LinearRegression (regressão linear)
- Ridge (regressão linear com regularização L2)
- Lasso (regressão linear com regularização L1)
- SVR (máquina de vetor de suporte)
- RandomForestRegressor
- GradientBoostingRegressor

Os critérios utilizados para a comparação foram: O MAE (Erro médio absoluto), o R^2 (Coeficiente de Determinação), e o R^2 ajustado.

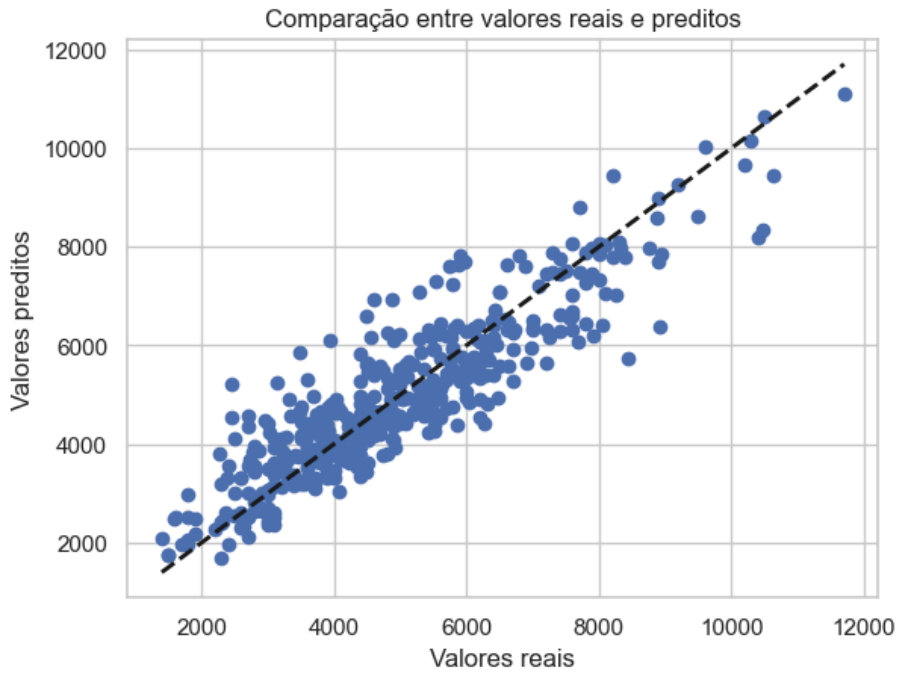
Entre os modelos testados, o Gradient Boosting Regressor apresentou o melhor desempenho global, com: MAE: 616.03, R^2: 0.802, e R^2 ajustado: 0.799.

O Gradient Boosting combina múltiplos modelos (árvores de decisão) de forma sequencial, corrigindo os erros anteriores a cada etapa.

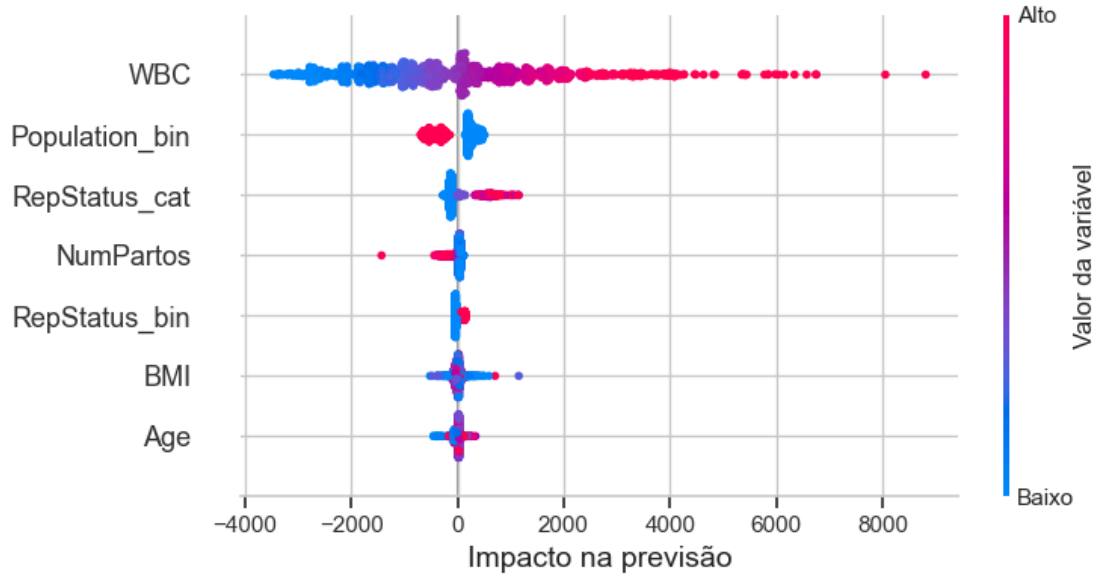
5.1.1 Comentários sobre o desempenho do modelo

Embora o modelo Gradient Boosting tenha apresentado um erro absoluto médio (MAE) de aproximadamente 616 unidades, esse valor deve ser interpretado à luz do contexto clínico e da distribuição dos valores de neutrófilos na base de dados. Esse erro representa o 12.56% da média dos valores da variável alvo usados no teste (média NEU_teste: 4905.4). Vale ressaltar que: O R^2 ajustado de 0,799 indica que o modelo consegue explicar cerca de 80% da variância total da variável NEU, o que é um desempenho considerado elevado para dados clínicos reais, geralmente marcados por variabilidade biológica e ruído. Além disso, o objetivo do modelo não é substituir a medição direta com precisão exata, mas sim fornecer uma estimativa confiável e acessível em contextos onde a contagem diferencial não está disponível. Podendo-se considerar como uma ferramenta de triagem ou apoio à decisão, e não como diagnóstico definitivo.

A seguinte imagem mostra o comportamento do modelo no conjunto de teste.



O impacto de cada variável no modelo pode ser apreciado no seguinte gráfico.



6. Conclusões.

A partir dos conjunto de dados usado neste trabalho, pode se concluir que:

- Os dois tipos de populações apresentam diferentes contagens de células leucocitárias, o qual é esperado considerando os entornos sociais e ambientais nos quais vivem.
- Alguns tipos de células leucocitárias podem apresentar variações de na contagem ao longo da gravidez.
- Modelos de aprendizado de máquina considerando variáveis imunológicas e variáveis antropométricas, reprodutivas e populacionais, podem ajudar a entender a interação entre comportamentos biológicos e entornos de moradia.
- Uma melhor descrição da população na amostra dos EUA pode ser útil para entender melhor a diferença entre as duas populações.

Recomendações para Pesquisas Futuras

- Ampliação da amostra: Incluir outras populações para análise comparativa
- Dados adicionais: Coletar informações sobre dieta, atividade física e exposição ambiental
- Modelagem avançada: Explorar técnicas de deep learning para padrões mais complexos
- Estudo longitudinal: Acompanhar as variações nos parâmetros ao longo do tempo

7. Documentação.

O estudo foi realizado usando a linguagem Python 3.12.3

Bibliotecas usadas com a respetiva versão:
pandas 2.2.3
matplotlib 3.10.3
numpy 2.2.6
seaborn 0.13.2
scikit-learn 1.7.0
scipy 1.15.3

O conjunto de dados corresponde aos dados usados como referência no artigo
Immune function during pregnancy varies between ecologically distinct populations,
(https://doi.org/10.1093/emph/eoaa022)

O conjunto de dados considerados no trabalho pode ser achado em:
https://datadryad.org/dataset/doi:10.25349/D94C77#citations

Dados gerais referentes ao National Health and Nutrition Examination Survey se encontram em:
https://wwwn.cdc.gov/nchs/nhanes/

Informações sobre o The Tsimane Health and Life History Project sem encontram em:
https://tsimane.anth.ucsb.edu/index.html