Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Departamento de Estatística

Fatores Associados à Prematuridade em Nascidos Vivos: Uma Análise de Dados do SINASC no Estado de Pernambuco (2014–2023)

Autoras: Roberta da Paz Melo e Thays Silva de Aquino

Data: 21 de março de 2025

Introdução

Foi implantado no Brasil, o Sistema de Informações sobre Nascidos Vivos (SINASC), tornando uma ferramenta poderosa de pesquisa na área materno infantil (FIGUEROA, 2021). Para o preenchimento do SINASC são utilizadas as informações da Declaração de Nascido Vivo (DNV), que é um documento padrão utilizado em todo território nacional (BRASIL, 2022).

O SINASC foi implantado com o intuito de haver registro contínuo de informações sobre nascidos vivos (FIGUEROA, 2021). Levando ao país uma posição de destaque e reconhecimento internacional quanto a sua cobertura, magnitude e transparência das informações (BRASIL, 2022).

Umas das variáveis de registro do SINASC é a Data da Última Menstruação (DUM) e o número de semanas de gestação, possibilitando o cálculo gestacional. Utilizamos neste trabalho a variável resposta criança nascida prematura. Segundo Silvia (2024), a prematuridade no Brasil equivale a 11%, tendo como causa múltiplos fatores, podendo estar associado a alguns destes, como: idade materna, escolaridade, cor/raça, intervalo interpartal, ausência de pré-natal e má qualidade da assistência pré-natal. A prematuridade não é um problema só do Brasil, e sim, um problema a nível mundial, por se tratar de uma causa direta para a mortalidade infantil durante o primeiro ano de vida.

Diante desta problemática tivemos como objetivo analisar o banco de dados do SINASC, referente as informações do estado de Pernambuco no período de 2014 a 2023, quanto ao nascimento prematuro e os fatores associados. E segundo Cavalcante et al. (2021) é importante o interesse em pesquisa nesta temática, pois, permite traçar o perfil do pré-natal até o parto, subsidiando a efetivação de ações em políticas públicas.

Fundamentos Teóricos e Metodológicos

Foi selecionado o banco de dados do SINASC referente as informações do estado de Pernambuco no período de 2014 a 2024. Tendo como critério de escolha um banco de dados com informações reais e de relevância no contesto da saúde pública e políticas públicas. Os dados foram inicialmente consolidados e executados no R e posteriormente executado o código em phyton pelo colab. O algoritmo de machine learning utilizado foi: Árvore de Decisão e Floresta Aleatória.

Segundo Alvarenga (2018), árvore de decisão é um modelo não paramétrico, que possui capacidade de tratar atributos do tipo numérico, categórico ou ambos. Implementa, intrinsecamente, a seleção de características, tornando o algoritmo mais robusto na tratativa de variáveis irrelevantes ou que apresentem ruído. Floresta Aleatória é um modelo baseado em árvores de decisão, que tratam com conjunto de dados de alta dimensão, e com presença de multicolinearidade.

Aplicação

Análise Exploratória dos Dados

Após consolidação dos dados de nascidos vivos (SINASC) entre os anos de 2014 a 2023, foram selecionadas as observações do estado de Pernambuco e excluídas aquelas sem informação da duração da gestação, chegando assim a um número de 1.300.646 linhas. Inicialmente, a base de dados possuía 61 variáveis, onde muitas delas se tratavam de colunas referentes ao registro hospitalar e identificação do paciente. Visando evitar explosão dimensional e focar na identificação do perfil, na etapa de análise exploratória, foram filtradas as variáveis que em um entendimento inicial, poderiam refletir maiores taxas de prematuridade. Neste sentido, foram selecionadas as colunas referentes ao local de nascimento, idade, escolaridade e estado civil da mãe, histórico obstétrico e características do parto, além de informações sobre a assistência pré-natal. Neste momento, também foi criada uma variável dependente com entradas 0 ou 1, que definia como prematuro (1), os nascidos abaixo de 37 semanas (151.988 nascidos) Das 1.300.646 observações, em 105.267 delas havia algum dado faltante (ou ignorado) em pelo menos uma das variáveis citadas, mas em apenas uma destas variáveis a proporção de dados faltantes foi de 3.8%, nas demais ficou abaixo de 1.2%, o que possibilita uma análise consistente.

Os nascimentos estão distribuídos ao longo dos anos (Figura 1) a um percentual entre 8.91% (2023) e 11.02% (2015) em cada ano. É perceptível uma tendência a diminuição gradual, contudo, identifica-se uma queda acentuada em 2016 (9.91%). Este fato pode estar associado ao surto de zika vírus iniciado em 2015, que gerou preocupações sobre complicações na gravidez, mas há necessidade de uma análise mais profunda para confirmar esta hipótese. Em relação à taxa de prematuros, manteve-se entre 11.13% e 12.18% nos 10 anos analisados, de modo que não parece haver uma tendência a aumento ou redução na linha temporal.

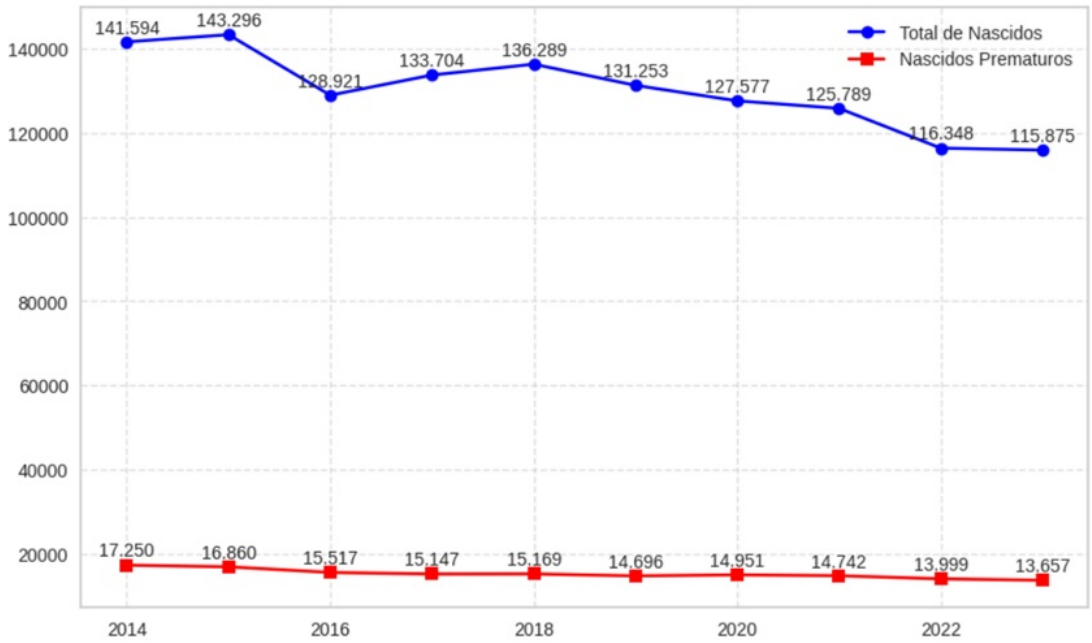


Figure 1: Distribuição de nascidos vivos e nascidos prematuros ao longo dos anos de 2014 e 2023. (Fonte: SINASC)

No tocante à idade das mães, a maior parte dos nascidos são filhos de mulheres entre 20 e 34 anos (cerca de 69%), sendo o grupo etário de 20-24 anos o mais expressivo (26.11%). A base ainda identifica o número alarmante de 11.373 nascidos de meninas entre 10 e 14 anos, indicando a necessidade de políticas públicas neste assunto. Os nascidos vivos de mães adolescentes (15-19 anos) representam 17.05% do total. Os nascidos de mulheres mais jovens (10-19 anos) e mais velhas (45-49 anos) apresentaram taxas de 18.54% e 18.73% de prematuridade, respectivamente.

A menor proporção foi identificada entre os nascidos de mães na faixa dos 25 e 29 anos (10.39%), indicando que essa pode ser uma faixa etária de menor risco.

A maior parte das observações eram referentes a nascidos de mulheres pardas (74.19%), seguido dos nascidos de mulheres pretas e brancas (19.46% e 5.24%, respectivamente. Indígenas e amarelas somavam 1.10%. A taxa de prematuros variou entre 11.13% (mulheres brancas) e 13.16% (mulheres pretas).

Considerando as condições de gravidez e do parto, é desenhado um cenário onde 97.95% das gestações são únicas, seguidas por duplas (2.00%) e triplas ou mais (0.05%). E a proporção dos prematuros altera bastante com o tipo de gestação: 10.78% na gestação única, 54.37% na dupla e chega a 90% no terceiro tipo. Cerca de 52% dos partos foram cesarianas com taxa de prematuros de 11.40%, enquanto que para partos normais esta proporção foi de 12.00%. Em relação ao todo, poucos nascidos apresentaram algum tipo de anomalia (13.973), porém a taxa de prematuros foi de 24,34% neste grupo, contra 11.54% entre os que não tiveram nenhum tipo de anomalia.

Em relação à quantidade de partos normais, cesarianos e gestações anteriores, de modo geral, a taxa de prematuros aumentava proporcionalmente. Por exemplo: 11.7% dos nascidos de mulheres na primeira gestação foram prematuros, enquanto esta taxa foi de 18.5% entre os nascidos daquelas com 10 ou mais gestações. Cerca de 96% dos nascidos apresentaram posição cefálica, enquanto as apresentações pélvica/podálica e transversa foram de 3.66% e 0.19%, respectivamente. Esta tendência se inverte quando são analisadas as taxas de prematuros: 10.92%, 30.72% e 18.97%. De forma semelhante, conforme aumentava a quantidade de filhos vivos e mortos, também aumentava a taxa de prematuros, contudo, os valores observados para a variável de filhos mortos foi mais expressivo. Por exemplo: entre os nascidos de mães com 4 filhos vivos a prematuridade foi de 14.56% e entre os nascidos de mães com 4 filhos mortos foi de 19.88%.

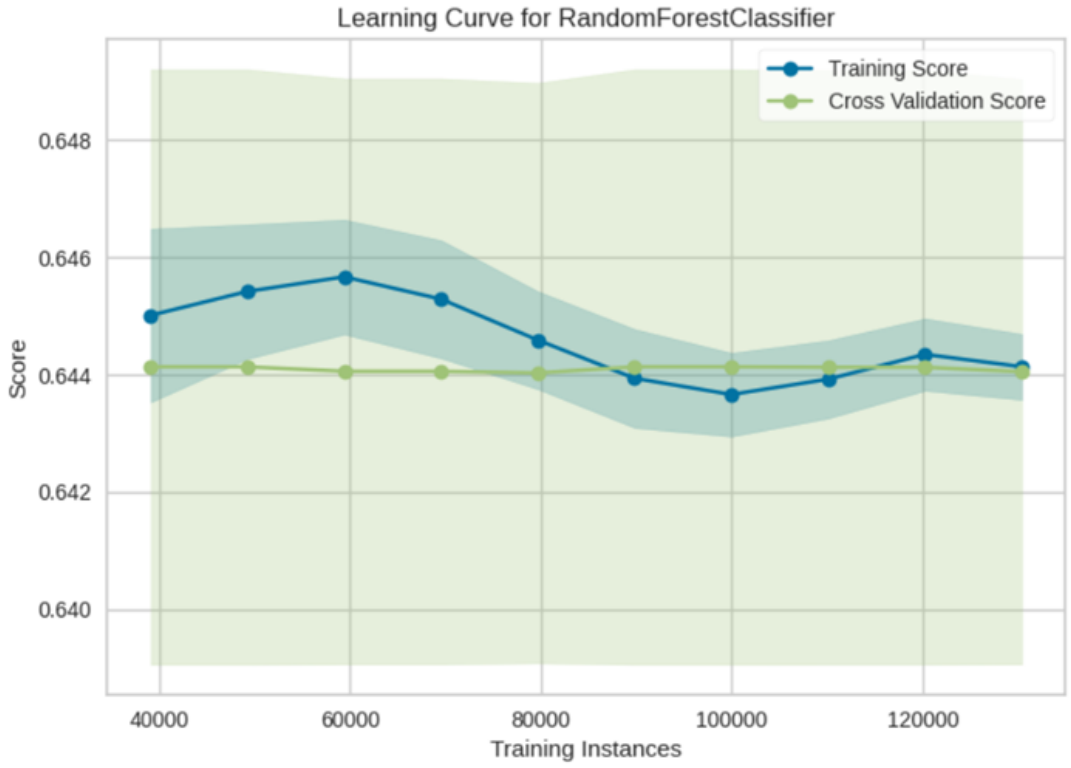
A maioria dos nascimentos ocorreu em hospitais (99.16%), onde a taxa de prematuros foi de 11.65%. Outros estabelecimentos de saúde tiveram 0.35% dos nascidos vivos com proporção de prematuros de 6.14% . Domicílios e outros locais tiveram somados 0.48% dos nascidos e taxa de prematuros respectiva de 17.70% e 30.25%. Em relação ao pré-natal, a maioria dos nascidos foram acompanhados em sete ou mais consultas (69.4%) com taxa de prematuros de 8.29%., seguidos por cerca de 23% que foram acompanhados entre 4 a 6 consultas com taxa de prematuros de 18.34%. A maior proporção de prematuros (22.95%) foi encontrada no grupo com menos de 4 consultas (7.06% das observações totais).

Em relação à escolaridade, a maioria dos nascimentos (60.41%) ocorreu entre mães com 8 e 11 anos de estudo, seguidas por aquelas com 4 a 7 anos de estudo (20.07%) e 12 anos ou mais (15.12%). Enquanto os nascidos de mulheres sem nenhum estudo representaram o menor grupo (0.59%). A proporção de prematuros foi inversamente proporcional aos anos de estudo, onde o maior percentual foi entre as mães com nenhuma escolaridade (15.47%) e o menor, entre aquelas com mais de 12 anos (11.18%). Esta tendência pode ter relação com menores condições socioeconômicas, e consequente, maior dificuldade de acesso a recursos médicos e consultas de pré-natal, que embora tenha cobertura pelo SUS, pode sofrer com desigualdades de oferta e acesso nos territórios. O estado civil pareceu não ter influência sobre a taxa de prematuros, variando entre 11% e 12% entre todas as classes.

Implementação dos modelos

Os algoritmos foram implementados considerando uma proporção de bases de treino e de teste de 70% e 30% das observações, respectivamente. Para este contexto, foram adotados os métodos de Floresta Aleatória e Árvore de decisão, pela capacidade de ambos de poder trabalhar com variáveis categóricas que de acordo com a análise exploratória, aparentavam ser importantes para entender e prever o comportamento de prematuros. Inicialmente, considerou-se utilizar os algoritmos de Floresta Aleatória e de SVM (considerando apenas variáveis numéricas), no entanto, houve um problema com o elevado tempo de processamento do SVM, então optou-se pelo modelo de árvore de decisão, devido à sua eficiência e potencial de interpretabilidade.

Inicialmente, ao executar o algoritmo de Floresta Aleatória, foi obtida acurácia de 0.89, e Precisão de 0.66. No entanto, a métrica Recall foi igual a 0.06, e considerando um problema da área de saúde, esta métrica é bastante crítica. Após avaliar a proporção de respostas para variável dependente e considerando a complexidade dos dados, constatou-se que a proporção de 88% de não prematuros poderia culminar em um desequilíbrio de classes e estar influenciando o modelo a classificar a maioria das observações como 0 (não prematuros). Então foi realizado o balanceamento de classes, tomando como base inicial todas as observações de nascidos prematuros e a seleção aleatória de igual número de observações entre aqueles nascidos não prematuros, resultando em uma base com 295.690 linhas. Após este processo, o modelo foi executado novamente, resultando em uma Acurácia de 0.65, o desempenho do Recall apresentou-se mais coerente (0.57). O passo seguinte foi aplicar duas estratégias de otimização do modelo: a primeira com foco na métrica Recall e a segunda, na métrica F1, porém ambas as saídas não foram satisfatórias, e o modelo retornou Recall igual a 1.0, prevendo todas as observações como prematuros. Diante deste cenário, foram descartados os modelos otimizados. Neste momento, optou-se por refazer o algoritmo de Floresta Aleatória com as três variáveis independentes com maior importância para a capacidade preditiva do modelo (número de consultas de pré-natal, tipo de gravidez e apresentação ao nascer), tendo em vista que as métricas ficaram parecidas e menos variáveis tornavam-no mais eficiente e mais explicativo. Como parâmetros, foram adotadas 20 árvores com profundidade de 20 nós e k-fold = 5, decorrendo nas métricas presentes na Figura 2.

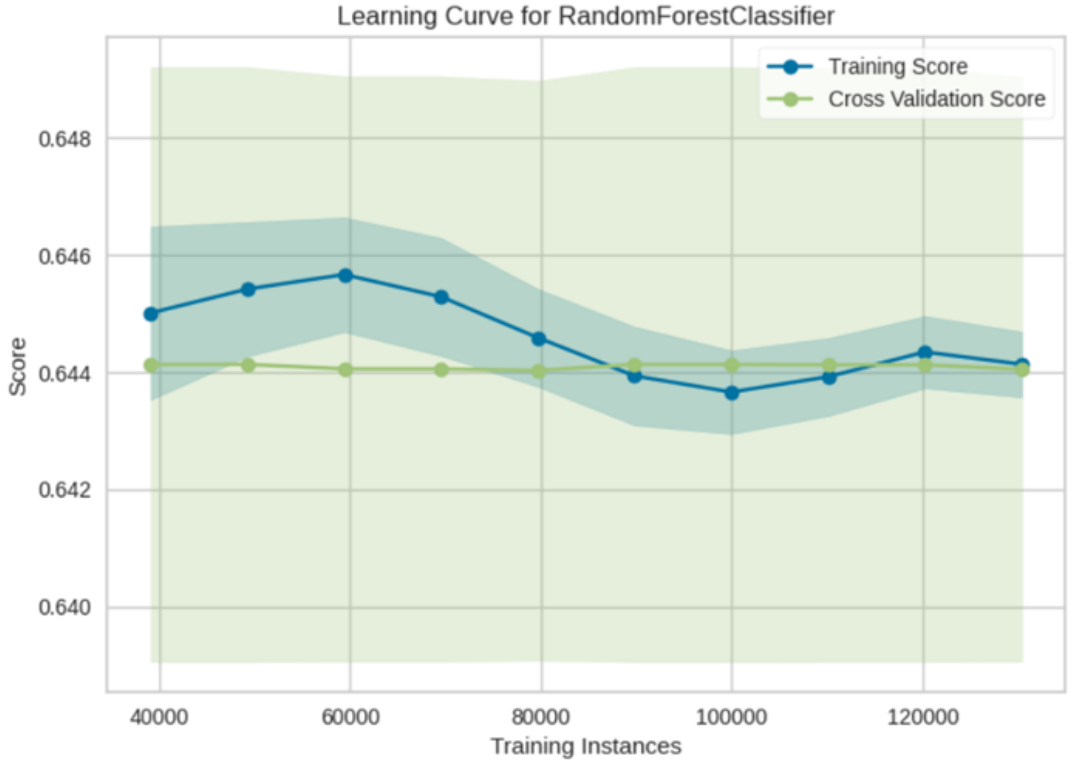


	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Random Forest Classifier	0.6446	0.6636	0.5854	0.6622	0.6215	0.2890	0.2909

Figure 2: Métricas e Gráfico com curvas de treinamento e de validação cruzada referentes ao Modelo de Floresta Aleatória

Para o modelo de Árvore de Decisão foi adotada a mesma base de treinamento e de teste e os mesmos parâmetros de k-fold e de profundidade utilizados na Floresta Aleatória. Após isso, o modelo foi otimizado, chegando às métricas apresentadas na Figura 3. Como pode ser constatado, os valores obtidos foram praticamente os mesmos, com diferença na terceira casa decimal, porém o de Árvore de Decisão foi mais eficiente, obtendo os resultados mais rapida-

mente. Ainda sobre as Figuras 2 e 3, pode-se visualizar as curvas de aprendizagem. É possível identificar que a curva de treinamento oscila e por vezes diminui a acurácia, conforme o tamanho da base aumenta, enquanto a curva de validação mantém-se praticamente constante. Este comportamento pode sugerir que o modelo atingiu seu limite de aprendizado e que uma base maior não resulta em métricas melhores



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Random Forest Classifier	0.6446	0.6636	0.5854	0.6622	0.6215	0.2890	0.2909

Figure 3: Métricas e Gráfico com curvas de treinamento e de validação cruzada referentes ao Modelo de Árvore de Decisão

Conclusão

A análise identificou que fatores relacionados à mãe como idade, escolaridade e histórico gestacional podem influenciar na taxa de nascidos prematuros. Mães adolescentes, com menor escolaridade e múltiplas gestações apresentaram maior risco de uma gestação prematura. Ressalta-se ainda que gestações múltiplas e apresentações fetais em posição não cefálica tiveram proporção de prematuros significativamente elevados.

Na aplicação dos modelos, devido aos valores de recall muito próximos a zero, foi necessário realizar o balanceamento da base. Ao final, ambos os Métodos de Floresta Aleatória e de Árvore de Decisão resultaram em desempenho semelhante, porém a segunda com menor custo computacional e maior interpretabilidade. Por fim, o estudo enfatiza a relevância de variáveis clínicas e sociodemográficas para prever a prematuridade, podendo, assim, auxiliar no apoio à implementação de ações na saúde pública.

Referências Bibliográficas

ALVARENGA, W. J. **Métodos de otimização hiperparamétrica: um estudo comparativo utilizando árvores de decisão e florestas aleatórias na classificação binária**. Orientadora: André Paim Lemos., tab. Belo Horizonte, 2018. Dissertação de mestrado – Universidade Federal de Minas Gerais, Escola de Engenharia, Programa de Pós-graduação em Engenharia Elétrica.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise Epidemiológica e Vigilância de Doenças Não Transmissíveis. **Declaração de Nascido Vivo: manual de instruções para preenchimento** [recurso eletrônico]. 4. ed. Brasília: Ministério da Saúde, 2022.

CAVALCANTE, J. N. B.; COUTINHO, D. J. G. A importância e aplicabilidade dos sistemas de informações sobre nascidos vivos e mortalidade: uma revisão integrativa. **Brazilian Journal of Development**, Curitiba, v. 7, n. 7, p. 73272-73279, jul. 2021.

FIGUEROA, P. D. Sistema de informações sobre nascidos vivos: uma análise da qualidade com base na literatura. **Cad Saúde Colet**, 2021.

SILVA, A. B. **Mortalidade Infantil em Crianças Prematuras: Prevalência e Fatores Associados em Pernambuco, entre os anos de 2017 e 2021**. Orientadora: Lívia Teixeira de Souza Maia. 49 f.: il., tab. Vitória de Santo Antão, 2024. Trabalho de Conclusão de Curso (Graduação em Saúde Coletiva) – Universidade Federal de Pernambuco, Centro Acadêmico de Vitória.

Contribuição

Roberta (50%): Pesquisa de bases reais e escolha de tema de importância social, Introdução, Metodologia, Conclusão.

Thays (50%): Análise Exploratória de Dados, Implementação dos Modelos, Conclusão.