Júlia Oliveira Dias de Souza

Associação entre as características da mãe e do recém-nascido e a prevalência de prematuridade no Estado do Rio de Janeiro: um estudo utilizando modelo de regressão log-linear de Poisson

Niterói - RJ, Brasil

13 de setembro de 2021

Universidade Federal Fluminense

Júlia Oliveira Dias de Souza

Associação entre as características da mãe e do recém-nascido e a prevalência de prematuridade no Estado do Rio de Janeiro: um estudo utilizando modelo de regressão log-linear de Poisson

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador: Dr. José Rodrigo de Moraes

Co-Orientadora: Dra. Patrícia Viana Guimarães Flores

Niterói - RJ, Brasil

13 de setembro de 2021

Universidade Federal Fluminense

Júlia Oliveira Dias de Souza

Associação entre as características da mãe e do recém-nascido e a prevalência de prematuridade no Estado do Rio de Janeiro: um estudo utilizando modelo de regressão log-linear de Poisson

Monografia de Projeto Final de Graduação sob o título *"Associação entre as características da mãe e do recém-nascido e a prevalência de prematuridade no Estado do Rio de Janeiro: um estudo utilizando modelo de regressão log-linear de Poisson"*, defendida por Júlia Oliveira Dias de Souza e aprovada em 13 de setembro de 2021, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Prof. Dr. José Rodrigo de Moraes
Orientador
Departamento de Estatística - UFF

Dra. Patrícia Viana Guimarães Flores
Co-orientadora
Hospital Federal de Bonsucesso - HFB

Profa. Dra. Jéssica Pronestino de Lima Moreira
Instituto de Estudos em Saúde Coletiva - UFRJ

Prof. Dr. Carlos Augusto Faria
Departamento Materno-Infantil - UFF

Niterói, 13 de setembro de 2021

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

S719a Souza, Júlia Oliveira Dias de
Associação entre as características da mãe e do recém-nascido e a prevalência de prematuridade no Estado do Rio de Janeiro: um estudo utilizando modelo de regressão log-linear de Poisson / Júlia Oliveira Dias de Souza ; José Rodrigo de Moraes, orientador ; Patrícia Viana Guimarães Flores, coorientadora. Niterói, 2021.
55 f. : il.

Trabalho de Conclusão de Curso (Graduação em Estatística)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2021.

1. Prematuridade. 2. Sistemas de Informação em Saúde. 3. Modelo Log-Linear de Poisson. 4. Aprendizado de Máquinas. 5. Produção intelectual. I. Moraes, José Rodrigo de, orientador. II. Flores, Patrícia Viana Guimarães, coorientadora. III. Universidade Federal Fluminense. Instituto de Matemática e Estatística. IV. Título.

CDD -

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

A prematuridade é um problema alarmante de saúde pública e a principal causa de óbito neonatal no Brasil. A complexidade relacionada ao cuidado com o prematuro está associada com a imaturidade geral, que pode levar qualquer órgão à disfunção, podendo o recém-nascido sofrer comprometimento ao longo do seu desenvolvimento. O presente trabalho teve como objetivo avaliar a associação entre as características maternas e do recém-nascido e a prevalência de prematuridade no Estado do Rio de Janeiro em 2019. Foi realizado um estudo transversal, com os dados do Sistema de Informações sobre Nascidos Vivos (SINASC 2019) utilizando o modelo de regressão log-linear de Poisson com variância robusta, numa abordagem de aprendizado de máquina (machine learning). A prevalência de prematuridade foi de 9,5%; e na modelagem estatística observou-se maior prevalência de prematuridade entre mães de 35 anos ou mais versus 19 anos ou menos (RP=1,136; p-valor<0,001), casadas ou com união estável (RP=1,105; p-valor<0,001), viúvas ou separadas/divorciadas (RP=1,196; p-valor<0,001), que realizaram parto cesáreo (RP=1,207; p-valor< 0,001) e que tiveram bebês com apresentação pélvica ou podálica/transversa (RP=1,260; p-valor<0,001). Adicionalmente, uma menor prevalência de prematuridade foi observada entre mães com 20 a 34 anos versus 19 anos ou menos (RP=0,949; p-valor=0,022), não brancas (RP=0,942; p-valor<0,001), entre mães primíparas versus nulíparas (RP=0,944; p-valor = 0,002), que realizaram seis ou mais consultas pré-natal (RP=0,642; p-valor<0,001), bem como entre bebês do sexo feminino (RP=0,858; p-valor<0,001), que receberam escore de Apgar com a avaliação "normal" versus "muito baixo" (RP=0,722; p-valor<0,001) e que não tiveram baixo peso ao nascer (RP=0,101; p-valor<0,001). Conclui-se que tanto características demográficas da mãe, como idade, situação conjugal e raça/cor, quanto características clínicas do bebê e às relativas ao parto estão associadas com a prematuridade.

Palavras-chave: Estudos Transversais. Prematuridade. Sistemas de Informação em Saúde. Modelo log-linear de Poisson. Aprendizado de Máquina.

Agradecimentos

Agradeço primeiramente à Deus, que me abençoou com a oportunidade de viver essa experiência incrível que foi a minha primeira graduação, na universidade dos meus sonhos - a Universidade Federal Fluminense.

Em especial, agradeço a minha mãe, Michelle, minha maior inspiração, que me apoiou e vibrou comigo em todas as minhas conquistas ao longo da graduação. Agradeço também os meus avós Altemízia e Manoel, aos meus tios Paula e Vínicius, ao meu irmão Pietro e minhas primas Manuella e Luiza por sempre estarem do meu lado e confiarem em mim.

Agradeço ao meu namorado Anderson, por ser meu maior suporte durante esses semestres finais, por me ouvir apresentar o trabalho incontáveis vezes, por acreditar em mim e por me acalmar durante todas as crises de ansiedade.

Agradeço a todos os amigos que fiz na Estatística, em especial Beatriz e Gabriel, que não só tornaram a caminhada mais leve, como a fizeram inesquecível.

Agradeço ao meu orientador Dr. José Rodrigo de Moraes, o qual tive a honra em conhecer e o privilégio de poder aprender com toda sua didática. Um profissional que inspira pela sua paixão no que faz. Por sua confiança em mim e sua dedicação e paciência à cada etapa deste trabalho, serei para sempre grata.

À minha co-orientadora Dra. Patrícia Guimarães Flores, que esteve disposta a contribuir para este trabalho mesmo com toda a correria que o hospital demanda, minha eterna gratidão por toda sua ajuda, é uma honra poder contar com seus conhecimentos.

Agradeço aos professores Dr. Bruno Francisco Teixeira Simões, Dr. Carlos Augusto Faria e Dra. Jéssica Pronestino de Lima Moreira, por terem aceitado fazer parte deste trabalho, é gratificante tê-los na minha banca.

Aos amigos, familiares e professores do Departamento de Estatística que me ajudaram ao longo dessa jornada, minha gratidão.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Quadros

1 Introdução

1.1 Prematuridade

A prematuridade é um problema alarmante de saúde pública e a principal causa de óbito neonatal no Brasil (FRANÇA et al., 2017). De acordo com a Organização Mundial de Saúde (OMS), é definido como prematuro ou pré-termo os bebês que nascem antes das 37 semanas de gestação. São classificados em prematuros extremos os que nasceram antes das 28 semanas, circunstância que exige muito cuidado visto que os órgãos, apesar de formados, ainda estão imaturos. Existem também os muitos pré-termos, nascidos entre 28 e 31 semanas, e os prematuros moderados ou tardios, onde o parto ocorre entre 32 até as 36 semanas (BLENCOWE et al., 2012). Com 37 semanas ou mais, a gestação é considerada a termo.

A complexidade relacionada ao cuidado com o prematuro está associada com a imaturidade geral, que pode levar qualquer órgão à disfunção, podendo o recém-nascido sofrer comprometimento ao longo do seu desenvolvimento (RAMOS e CUMAN, 2009). O bebê pré-termo tem riscos aumentados de adoecer e falecer em função do incompleto desenvolvimento fetal e maior suscetibilidade às infecções devido ao período de internação nas unidades neonatais (GUIMARÃES et al., 2017).

Dentre as complicações dos bebês prematuros, particularmente aqueles nascidos de gestações de menos de 32 semanas, pode-se citar maior probabilidade de apresentar problemas auditivos (RECHIA et al., 2016) e paralisia cerebral (CASTRO et al., 2012). A prematuridade também pode aumentar o risco de alterações visuais, entre elas a retinopatia da prematuridade, que é uma doença que pode causar graves problemas oculares, como, por exemplo, miopia, descolamento de retina e até cegueira infantil (MELLO e MEIO, 2003; AGUIAR et al., 2007). Blencowe et al. (2012) apontam ainda outras complicações do nascimento prematuro específicas da imaturidade, como a síndrome do desconforto respiratório, hemorragia intracraniana e enterocolite necrosante. Além disso, a hipotermia é uma outra complicação que pode ocorrer em bebês prematuros, sendo

considerado fator de risco para um pior prognóstico, aumentando a chance de morbidade e mortalidade neonatal. Apesar de rara, a hipertermia também pode ocorrer em bebês prematuros (BRASIL, 2012).

A prematuridade pode ser classificada em espontânea, devido a trabalho de parto espontâneo ou ruptura prematura das membranas, e eletiva, quando a interrupção da gestação é indicada por complicações maternas (como exemplo, doença hipertensiva, placenta prévia) e/ou fetais (restrição do crescimento ou sofrimento do feto) (RADES et al., 2004; BITTAR e ZUGAIB, 2009). A prematuridade espontânea corresponde a 75% dos nascimentos prematuros e sua etiologia é complexa e multifatorial ou desconhecida, sendo muitas vezes difícil inferir sobre suas causas (BITTAR e ZUGAIB, 2009).

Em termos mundiais, no ano de 2005, nasceram cerca de 13 milhões de prematuros, o que correspondeu a 10% do total de nascimentos (BECK et al., 2010). A taxa de prematuridade vem aumentando nas últimas décadas, variando entre 5% (países europeus) e 18% (países africanos) (BLENCOWE et al., 2012). Em recente levantamento realizado pela Escola Nacional de Saúde Pública (ENSP/Fiocruz) para a pesquisa Nascer no Brasil, a taxa de prematuridade brasileira ficou em 11,5% dos nascimentos (LEAL et al., 2016a).

A prematuridade acarreta custos econômicos elevados aos setores de saúde públicos e privados, visto que o tempo de permanência hospitalar do bebê pode ser prolongado e demandar internação em unidades de cuidados intermediários ou intensivos, com equipamentos avançados e a utilização de medicamentos de alto custo (BLENCOWE et al., 2013). De acordo com um levantamento do Centro Paulista de Economia da Saúde da Universidade Federal de São Paulo (UNIFESP) entre 2009 e 2011, o custo médio diário de internação de um recém-nascido prematuro no Brasil varia de U$ 97,00 a U$ 157,00 (MWAMAKAMBA e ZUCCHI, 2014). Além dos custos econômicos, a prematuridade acarreta um elevado custo social às famílias, seja pela longa permanência do recém-nascido em hospital, por seus impactos ao longo da vida da criança e, sobretudo, pela morte do recém-nascido (BLENCOWE et al., 2013).

As possibilidades de sobrevivência do prematuro estão condicionadas pela idade gestacional, o peso ao nascer e pelas complicações que o bebê apresenta (NOMURA et al., 2001; FERRAZ e NEVES, 2011). De todos estes fatores, o mais importante é a idade gestacional, uma vez que esta determina a maturidade dos órgãos (NOMURA et al., 2001). As taxas de sobrevivência aumentam gradativamente de acordo com a idade gestacional, assim como diminuem as chances de sequelas graves (SOUSA et el., 2017).

Com relação a mensuração da idade gestacional, métodos de estimação como exame

físico, data da última menstruação (DUM) e ultrassonografia são comumente utilizados, porém a DUM tem sido considerada a primeira escolha, apesar da limitação pela dependência da memória da gestante (ALEXANDER e ALEN, 1996). Nos casos em que a qualidade da informação da DUM seja duvidosa e a gestante não tenha realizado a ultrassonografia no início da gestação, a idade gestacional pode ser estimada pela observação das características físicas e neurológicas do recém-nascido ao nascimento (MORAES e REICHENHEIM, 2000).

Alguns fatores de risco para o parto prematuro são conhecidos, como: tabagismo, consumo de álcool, drogas ilícitas, diabetes mellitus, gestação gemelar e infecções de trato urinário (CASCAES et al., 2008; BRANDI et al., 2020; TEIXEIRA et al.,2018). Além desses, são apontados ainda como fatores de risco para a prematuridade a quantidade insuficiente de consultas de pré-natal e a realização de parto cesáreo (GUIMARÃES et al., 2017).

A idade materna também pode ser um fator de risco para a prematuridade. As complicações da gravidez na adolescência estão relacionadas a múltiplas condições, como o número reduzido de consultas de pré-natal e o baixo nível de escolaridade (MARTINS et al., 2011). Além disso, a mãe adolescente possui maior risco de ter um parto prematuro por sua imaturidade biológica, pois a gestação ocorre em um organismo que ainda está em formação física e emocional, e poderá desencadear problemas de crescimento e desenvolvimento, devido a insuficiência uteroplacentária e ao comprometimento da transferência de nutrientes para o bebê (BULHÕES et al., 2018).

O levantamento estatístico e o monitoramento dos nascimentos são de extrema importância na área epidemiológica para fins de prevenção da prematuridade. Neste contexto, a partir dos dados do Sistema de Informações sobre Nascidos Vivos (SINASC) do Ministério da Saúde, é possível identificar quais são as prioridades de intervenção relacionadas à saúde da mãe e do bebê, além de fornecer indicadores de saúde sobre pré-natal e assistência ao parto imprescindíveis para o combate da prematuridade (RAMOS e CUMAN, 2009).

1.2 Revisão Bibliográfica

Nesta revisão, encontrou-se artigos a respeito da prematuridade os quais possuem diferentes áreas de pesquisa e diferentes fatores de riscos apresentados com formas distintas de delineamento do estudo. Nesta seção, serão apresentados os resultados obtidos

por alguns estudos que relacionam diferentes variáveis com o desfecho de prematuridade apresentado.

Guimarães et al. (2017) utilizaram a mesma base de dados do presente trabalho (SINASC) empregando o modelo de regressão logística múltipla, no período de 2008 a 2011, na região de Divinópolis, Minas Gerais. Excluindo os casos de partos gemelares e com menos de 22 semanas gestacionais, concluíram que a prevalência de prematuridade foi de 8,0% e que maior chance de prematuridade está associada ao parto cesáreo, à realização de menos de sete consultas de pré-natal e à menor idade materna.

Cascaes et al. (2008) buscaram estimar a prevalência de prematuridade, no Estado de Santa Catarina em 2005, e, por meio do ajuste do modelo de regressão logística, avaliar os fatores associados à chance de prematuridade utilizando a base de dados do SINASC. Verificou-se que a prevalência de prematuridade foi igual a 6,1% e a chance de nascer antes da 37ª semana de gestação foi maior quanto menor o número de consultas pré-natal, idade materna superior a 40 ou inferior a 20 anos e parto cesáreo.

Bezerra et al. (2006), realizaram um estudo transversal com dados de prontuários de gestantes atendidas no Hospital Universitário da Universidade de São Paulo entre os anos 1995 e 2000 e submetidas a tratamento para inibição de parto prematuro, no qual procurou-se identificar a prevalência e fatores associados à prematuridade por meio de ajustes de modelos logísticos univariados e multivariados. Constatou-se a ocorrência de parto prematuro em 66,3% das gestantes e observou-se associação estatisticamente significativa entre parto prematuro e a gestante ser nulípara e ter realizado uma baixa quantidade de consultas pré-natal.

O artigo de Oliveira et al. (2016) teve como objetivo identificar características maternas e neonatais associados à prematuridade no município de Porto Alegre, por meio de um estudo do tipo caso-controle de base populacional, no qual os casos eram compostos por recém-nascidos prematuros (com menos de 37 semanas de gestação) e os controles eram formados pelos recém-nascidos não prematuros (37 semanas ou mais de gestação). Por meio da análise de regressão logística segundo modelo hierárquico, os autores verificaram associação significativa entre o desfecho de prematuridade e um conjunto de características sociodemográficas e do nascimento, tais como idade materna baixa (menor do que 19 anos) ou elevada (35 anos ou mais), escolaridade materna inadequada, gestação múltipla, pré-natal inadequado, realização de parto cesáreo, baixo peso ao nascer (<2.500g) e índice de Apgar no 5º minuto muito baixo (pontuação de 0 a 3).

O artigo de Teixeira et al. (2018) teve como finalidade traçar um perfil de mães de

bebês prematuros e a termo em partos ocorridos entre os meses de abril e setembro de 2015, em uma maternidade pública da região Nordeste do país, utilizando um estudo analítico-descritivo. Constataram que o sedentarismo, sobrepeso e obesidade, pressão arterial alta e baixo nível de escolaridade e de renda caracterizaram o perfil das mães de recém-nascidos prematuros e são fatores de risco para a saúde do bebê.

Souza et al. (2019) tiveram como objetivo estudar a prevalência de prematuridade no estado do Rio Grande do Sul em 2014, utilizando um estudo de delineamento do tipo transversal com base nos dados do SINASC. Neste estudo, os autores observaram que a prevalência de prematuridade no estado foi cerca de 11,5% e, usando modelo de regressão de Poisson, concluíram que a maior prevalência de prematuridade estava estatisticamente associada a mães com 40 anos ou mais de idade, mães de cor/raça amarela, parda e indígena e que realizaram parto cesáreo. O número baixo de consultas pré-natal e o baixo nível de escolaridade também foram variáveis que estiveram associadas à uma maior prevalência de prematuridade.

No estudo de Brandi et al. (2020) procurou-se evidenciar fatores de risco materno-fetais pra o nascimento pré-termo em um hospital de referência de Minas Gerais, através de um estudo de corte transversal, retrospectivo e analítico, baseado em dados de prontuários de todos os nascimentos em 2017. Neste artigo, verificou-se uma prevalência de prematuridade de 13,8% e concluiu-se, por meio da aplicação de testes bivariados (teste Qui-Quadrado ou Exato de Fisher) que os principais fatores de risco para a mãe e seu bebê associados ao parto prematuro foram diabetes mellitus, síndromes hipertensivas na gravidez, sífilis materna, gestação gemelar e parto cesáreo, além de malformações fetais e recém-nascido com peso pequeno para idade gestacional.

2 Objetivos

2.1 Objetivo geral

O objetivo geral do presente trabalho é avaliar a associação entre as características maternas e do recém-nascido e o desfecho de prematuridade no Estado do Rio de Janeiro, no ano de 2019.

2.2 Objetivos específicos

1. Analisar as distribuições das características maternas e do recém-nascido;
2. Analisar a distribuição da prematuridade segundo as características maternas e dos recém-nascidos;
3. Identificar as características da mãe e do bebê que estão estatisticamente associadas à prevalência de prematuridade;
4. Avaliar o sentido e o grau da associação das características maternas e do recém-nascido que estão associadas à prevalência da prematuridade;
5. Analisar a capacidade do modelo de predizer o desfecho de prematuridade, considerando as características maternas e dos recém-nascidos.

3 Materiais e Métodos

3.1 Sistema de Informações sobre Nascidos Vivos (SINASC)

O Sistema de Informação sobre Nascidos Vivos (SINASC) foi implantado oficialmente em 1990, com o objetivo de coletar dados sobre os nascimentos informados em todo território nacional e fornecer dados sobre a natalidade para todos os níveis do Sistema de Saúde, como os dados da mãe, da gestação e do recém-nascido, permitindo conhecer o perfil dos nascidos vivos tais como peso ao nascer, condições de vitalidade e prematuridade (MELLO JORGE et al., 1993).

A implantação do SINASC ocorreu de forma gradual em todas as unidades da Federação e já vem apresentando, em muitos municípios, um número maior de registros do que o publicado pelo IBGE com base nos dados de Cartório de Registro Civil (BRASIL, 2001).

O documento padrão de uso obrigatório em todo o território nacional e imprescindível à coleta de dados é a Declaração de Nascidos Vivos (DN), considerado como documento hábil para os fins do Art. 51 da Lei nº 6.015/1973, para lavratura da Certidão de Nascimento pelo Cartório de Registro Civil e do inciso IV do Art. 10 da Lei nº 8.069/1990.

O Ministério da Saúde, através da Secretaria de Vigilância à Saúde (SVS)/Departamento de Análise de Saúde e Vigilância de Doenças Não Transmissíveis (DASNT)/ Coordenação Geral de Informações e Análises Epidemiológicas (CGIAE), tem incentivado aos gestores municipais e estaduais a fazerem uso dos dados contidos no SINASC, para a formulação de indicadores epidemiológicos como instrumentos de suporte ao planejamento de ações, atividades e programas voltados à gestão em saúde.

3.1.1 Formulário da Declaração de Nascido Vivo (DN)

O formulário da Declaração de Nascido Vivo (DN) deve ser preenchido por uma pessoa previamente treinada para esse fim, como por exemplo o enfermeiro, médico ou profissional da área administrativa. A DN é impressa, pré-numerada sequencialmente e preenchida em três vias. Sua distribuição para os estados é da competência exclusiva do Ministério da Saúde, e cada uma das vias tem o seguinte destino (BRASIL, 2011):

- Via branca – Estabelecimentos de Saúde que realizam partos, devem enviar para Supervisão Técnica de Saúde de sua região/ Cartórios de Registro Civil da capital e profissionais cadastrados que prestam assistência nos partos domiciliares, devem encaminhar para Gerência do SINASC;
- Via amarela - Entregar ao pai ou responsável legal para assentamento do nascimento em cartório e obtenção da certidão de nascimento;
- Via rosa – Arquivar no prontuário da gestante ou do recém-nascido.

A DN deve ser preenchida, preferencialmente, após a segunda avaliação do recém-nascido realizada pelo neonatologista, que ocorre geralmente 6 horas após o nascimento, ou próximo à alta da mãe. O formulário (ver ANEXO 1) é composto por oito blocos de informações: Recém-Nascido, Local da ocorrência, Mãe, Pai, Gestação e parto, Anomalia congênita, Identificação do responsável pelo preenchimento e Cartório.

Bloco I - Recém-Nascido

- Nome do Recém-nascido;
- Data e hora do Nascimento;
- Sexo;
- Peso ao nascer (em gramas);
- Índice de Apgar;
- Detectada alguma anomalia ou defeito congênito?

Bloco II – Local da Ocorrência

- Local da Ocorrência: Hospital, outro estabelecimento de Saúde, Domicílio, Outros ou Ignorado;

- Estabelecimento de ocorrência do parto: nome do estabelecimento de saúde, número do Cadastro Nacional de Estabelecimentos de Saúde – CNES e endereço completo de onde ocorreu o parto (logradouro, n°, CEP, município).

Bloco III - Mãe

- Nome da mãe;
- Cartão SUS;
- Escolaridade;
- Ocupação Habitual e Ramo de Atividade;
- Data de nascimento da mãe;
- Idade da mãe;
- Naturalidade da mãe;
- Situação conjugal;
- Raça/Cor da mãe;
- Residência da mãe.

Bloco IV – Pai

- Nome do pai;
- Idade do pai.

Bloco V - Gestação e Parto

- Histórico gestacional;
- Data da última menstruação (DUM);
- Nº de semanas de gestação, se DUM ignorada;
- Número de consultas de pré-natal;
- Mês da gestação que iniciou o pré-natal;
- Tipo de gravidez: única, dupla, tripla ou mais;
- Apresentação: cefálica, pélvica ou transversa;
- O trabalho de parto foi induzido?
- Tipo de parto;
- Cesárea ocorreu antes do trabalho de parto iniciar?
- Nascimento assistido por.

Bloco VI – Anomalia Congênita

- Descrever todas as anomalias ou defeitos congênitos se observados.

Bloco VII - Identificação do responsável pelo preenchimento

- Data do preenchimento;
- Nome do responsável pelo preenchimento;
- Função;
- Tipo documento;
- Nº do documento;
- Órgão emissor.

Bloco VIII - Cartório

- Nome do cartório, município, registro.

3.2 População de estudo

A população de estudo foi composta por bebês nascidos vivos a termo, pré-termo e pós-termo filhos de mães residentes no estado do Rio de Janeiro, no ano de 2019. Foram excluídos do presente estudo, os recém-nascidos com algum tipo de anomalia congênita, os partos de gestações com menos de 22 semanas (abortos) e partos gemelares.

3.3 Variáveis de estudo

3.3.1 Desfecho de prematuridade

O desfecho de prematuridade foi baseado nas seguintes informações requeridas no formulário da DN, contida no bloco V - Gestação e parto: *"Data da última menstruação (DUM)"* e *"Nº de semanas de gestação, se DUM ignorada"* estimado por meio de exame físico ou outro método. A partir das respostas de ambos os quesitos, foi originada a variável categórica "idade gestacional (semanas de gestação)" disponibilizada no banco de dados do SINASC. A partir desta variável foi construída no presente trabalho o desfecho de prematuridade (Y) com duas categorias possíveis, onde Y=1 indica que o recém-nascido é *prematuro*, no caso da idade gestacional ser menor que 37 semanas, e Y=0 indica que o recém-nascido é *a termo ou pós-termo*, isto é, quando a gestação durou 37 semanas ou mais.

3.3.2 Variáveis relativas a mãe e ao recém-nascido

As variáveis consideradas no presente estudo que se referem às características sociodemográficas e de saúde relativas a mãe e ao recém-nascido, estão no Quadro 1.

Quadro 1: Variáveis referentes às características sociodemográficas e de saúde relativas a mãe e ao recém-nascido.

Variáveis	**Características**
1. Idade da mãe	19 anos ou menos 20 a 34 anos 35 anos ou mais
2. Raça/Cor da mãe	Branca Não Branca
3. Escolaridade da mãe	Até ensino fundamental Ensino médio ou superior incompleto Ensino superior completo
4. Situação conjugal da mãe	Solteira Casada/ União estável Viúva/Separada judicialmente/divorciada
5. Paridade (Número de gestações anteriores, excluídas as gestações com menos de 22 semanas)	Núlipara Primípara Multípara
6. Número de consultas durante o pré-natal	6 ou mais consultas pré-natal Até 5 consultas pré-natal
7. Tipo de parto	Vaginal Cesáreo
8. Prematuridade (desfecho)	Prematuro (pré-termo) Não prematuro (a termo ou pós-termo)
9. Sexo do recém-nascido	Masculino Feminino
10. Índice Apgar no quinto minuto	Muito baixo (0 a 3) Baixo (4 a 6) Normal (7 a 10)
11. Baixo peso ao nascer (< 2500g)	Sim Não
12. Tipo de apresentação do recém-nascido	Cefálico Pélvica ou podálica/Transversa

3.4 Modelos Lineares Generalizados (MLG)

Modelos lineares generalizados (MLG) são modelos estatísticos caracterizados por uma distribuição de probabilidade para a variável resposta Y, um conjunto de variáveis explicativas (estrutura linear) e uma função de ligação. Um MLG apresenta três compo-

nentes chamadas de componente aleatória, componente sistemática e função de ligação. Estas componentes são descritas a seguir:

i. Componente aleatória

A componente aleatória é composta por uma variável aleatória Y_i com parâmetro θ_i e uma distribuição de probabilidade pertencente à família exponencial. Supondo que Y é uma variável aleatória discreta, é dito que sua distrubuição pertence a família exponencial se a função de probabilidade puder ser expressa da seguinte forma:

$$P(Y_i = y_i, \theta) = \exp\{a(y_i)b(\theta_i) + c(\theta_i) + d(y_i)\}$$

onde $a(y_i), b(\theta_i), c(\theta_i)$ e $d(y_i)$ são funções conhecidas.

Algumas distribuições discretas que pertencem a família exponencial são a Bernoulli, a Binomial e a Poisson. Com relação as contínuas, pode-se citar as distribuições Normal e Gama (RENCHER, 2008).

ii. Componente sistemática

A componente sistemática é composta pelo preditor linear que segue a forma $\boldsymbol{\eta_i} = \boldsymbol{x_i'\beta}$, onde $\boldsymbol{x_i'}$ é o vetor das variáveis explicativas (covariáveis ou variáveis dummy para os níveis dos fatores) e $\boldsymbol{\beta}$ é o vetor dos parâmetros desconhecidos do modelo. Isto é,

$$\boldsymbol{\eta_i} = \boldsymbol{x_i'\beta} = \beta_0 + \beta_1 x_{i1} + ... + \beta_{k-1} x_{ik-1}$$

iii. Função de ligação

A função de ligação é uma função monótona e diferenciável, denotada por $g(\mu_i)$, tal que:

$$g(\mu_i) = \boldsymbol{x_i'\beta} \ ; \ i = 1, 2, ..., n.$$

A função de ligação relaciona a média da variável resposta Y com o preditor linear do modelo.

3.5 Modelo de regressão log-linear de Poisson

O modelo de regressão log-linear de Poisson é um modelo linear generalizado que é usualmente utilizado para análise de dados de contagem (números inteiros positivos) ou taxas (FARAWAY, 2006). Segundo Coutinho et al. (2008), na área epidemiológica, este

modelo é geralmente adotado em estudos longitudinais, onde a variável resposta representa o número de sucessos ocorridos num determinado intervalo de tempo. Porém, o modelo log-linear de Poisson vem sendo adotado na análise de variáveis respostas binárias levantadas em estudos seccionais, os quais são considerados que todos os indivíduos foram observados no mesmo período de tempo (COUTINHO et al., 2008). Quando o modelo de regressão de Poisson é aplicado a desfechos binários, o erro padrão para a razão de prevalência é superestimado, pois a variância da distribuição de Poisson aumenta gradativamente, enquanto a variância da distribuição Binomial atinge o seu valor máximo quando a prevalência é igual a 0,5 (COUTINHO et.al, 2008).

Segundo Cummings (2009), para desfechos raros, o erro padrão da distribuição de Poisson convergirá para o erro padrão da distribuição Binomial, entretanto no caso de desfechos frequentes a utilização do modelo de regressão de Poisson para estimar razões de prevalências em estudos transversais, pode gerar valores elevados para os erros padrão dos estimadores, p-valores dos testes de Wald de significância e para os intervalos de confiança dos parâmetros. Este problema pode ser resolvido ajustando o modelo de regressão log-linear de Poisson com o procedimento de variância robusta, isto é, empregando o estimador robusto de variância (ZOU, 2004), conhecido também como estimador Huber-White ou "estimador sanduíche"(CUMMINGS, 2009). Utilizando o estimador de variância robusta, o modelo de regressão log-linear de Poisson é o modelo estatístico mais adequado por produzir estimativas mais consistentes para os parâmetros de interesse com dados provenientes de estudos seccionais (BARROS e HIRAKATA, 2003)

3.5.1 Especificação do modelo

O modelo de regressão log-linear de Poisson é um MLG, onde Y tem distribuição de probabilidade de Poisson e a função de ligação é a função logarítmica $\ln(p_i)$. Quanto a sua especificação, o modelo é representado por:

$$\ln(p_i) = \boldsymbol{x_i'\beta}$$

$$\ln(p_i) = \begin{bmatrix} 1 & x_{i1} & x_{i2} & ... & x_{ik-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ ... \\ \beta_{k-1} \end{bmatrix}$$

$$\ln(p_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_{k-1} x_{ik-1}$$

onde:

$\ln(p_i)$ é o logaritmo da média da variável resposta Y referente a i-ésima unidade da amostra, sendo denotada por $E(y_i) = p_i$, por representar a prevalência de sucesso no caso de desfecho binário.

$\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos do modelo de dimensão kx1.

$\boldsymbol{x_i}$ é o vetor das variáveis explicativas (variáveis *dummy* para os níveis de fatores) da i-ésima unidade da amostra, de dimensão kx1.

3.5.2 Método de máxima verossimilhança (MV)

A estimação dos parâmetros dos modelos lineares generalizados é realizada pelo método de máxima verossimilhança, nas situações em que as observações são independentes e identicamente distribuídas (IID). Sejam $y_1, y_2, ..., y_n$ observações das variáveis $Y_1, Y_2, ..., Y_n$, independentes, todas com distribuição de Poisson com parâmetro p_i, a função de probabilidade de Y_i é:

$$P(Y_i = y_i) = \frac{e^{-p_i} p_i^{y_i}}{y_i!};\ y_i = 0, 1, 2, ...\ i = 1, 2, ..., n \tag{3.1}$$

A função de verossimilhança da amostra é dada por:

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{n} P(Y_i = y_i) = \prod_{i=1}^{n} \frac{e^{-p_i} p_i^{y_i}}{y_i!} \tag{3.2}$$

Para estimação de $\boldsymbol{\beta}$ é necessário maximizar a função de verossimilhança da amostra 3.2 em relação ao vetor de parâmetros $\boldsymbol{\beta}$. O valor que maximiza a função de verossimilhança é também o que maximiza o logaritmo da função de verossimilhança (função log-verossimilhança), uma vez que a função logarítmica é uma função monotônica. Logo,

em geral, aplica-se logaritmo à função de verossimilhança da amostra, como mostrado a seguir:

$$lnL(\boldsymbol{\beta}|\boldsymbol{y}) = \ln \prod_{i=1}^{n} P(Y_i = y_i) = \sum_{i=1}^{n} \ln P(Y_i = y_i)$$

$$= \sum_{i=1}^{n} \ln \left(\frac{e^{-p_i} p_i^{y_i}}{y_i!} \right) = \sum_{i=1}^{n} [y_i \; ln(p_i) - p_i - ln(y_i!)]$$

Substituindo p_i por $e^{\boldsymbol{x_i'\beta}}$, é possível obter:

$$lnL(\boldsymbol{\beta}|\boldsymbol{y}) = \sum_{i=1}^{n} [y_i \boldsymbol{x_i'\beta} - e^{\boldsymbol{x_i'\beta}} - \ln(y_i!)] \tag{3.3}$$

Derivando a equação 3.3 em relação a $\boldsymbol{\beta}$, tem-se:

$$\begin{aligned}
\frac{\partial \ln L(\boldsymbol{\beta}|\boldsymbol{y})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}} [y_i \boldsymbol{x_i'\beta} - e^{\boldsymbol{x_i'\beta}} - ln(y_i!)] \\
&= \sum_{i=1}^{n} \boldsymbol{x_i'} y_i - \boldsymbol{x_i'} e^{\boldsymbol{x_i'\beta}} \\
&= \sum_{i=1}^{n} \boldsymbol{x_i'} [y_i - e^{\boldsymbol{x_i'\beta}}] \\
&= \sum_{i=1}^{n} \boldsymbol{u_i(\beta)}
\end{aligned}$$

onde $\sum_{i=1}^{n} \boldsymbol{u_i(\beta)} = \sum_{i=1}^{n} \boldsymbol{x_i'}[y_i - e^{\boldsymbol{x_i'\beta}}]$ é o vetor dos escores da i-ésima unidade da amostra, de dimensão kx1.

Igualando a derivada parcial a $\mathbf{0}$, é obtido o seguinte sistema de equações de verossimilhança amostrais:

$$\frac{\partial \ln L(\boldsymbol{\beta}|\boldsymbol{y})}{\partial \boldsymbol{\beta}} = \mathbf{0}. \tag{3.4}$$

A solução desse sistema de equações é o estimador de $\boldsymbol{\beta}$ obtido pelo método de máxima verossimilhança, denotado por $\hat{\boldsymbol{\beta}}_{\boldsymbol{MV}}$, dado por:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{MV}} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_{k-1}]$$

Em geral, o sistema 3.4 contém equações não-lineares e precisam ser resolvidas numericamente por processos iterativos. Para obter numericamente o estimador de máxima

verossimilhança de $\boldsymbol{\beta}$, utiliza-se o método do escore de Fisher.

Para grandes amostras, a distribuição amostral assintótica de $\hat{\boldsymbol{\beta}}_{\boldsymbol{MV}}$ é aproximadamente uma normal multivariada, isto é, $\hat{\boldsymbol{\beta}}_{\boldsymbol{MV}} \sim N_k[\boldsymbol{\beta}, \widehat{VAR}(\hat{\boldsymbol{\beta}}_{\boldsymbol{MV}})]$, onde a matriz de variância-covariância dos estimadores dos parâmetros do modelo é dada por:

$$\widehat{VAR}(\hat{\boldsymbol{\beta}}_{\boldsymbol{MV}}) = \hat{\boldsymbol{I}}^{-\mathbf{1}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{MV}})$$

sendo $\hat{\boldsymbol{I}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{MV}}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{u_i}(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{\boldsymbol{MV}}}$, matriz simétrica de dimensão kxk.

3.5.3 Razão de prevalência

A razão de prevalência (RP) é uma medida de associação que visa avaliar o sentido e o grau de associação entre as variáveis explicativas e o desfecho binário em estudos de corte transversal. A partir do modelo de regressão log-linear de Poisson, pode-se obter a prevalência do evento de interesse (Y=1) da i-ésima unidade da amostra, aplicando o exponencial do preditor linear, ou seja:

$$p_i = P(Y_i = 1) = e^{\boldsymbol{x_i'\beta}};\ i = 1, 2, ..., n \tag{3.5}$$

A RP é, portanto, uma medida que indica a prevalência do evento de interesse de um grupo comparativamente a outro grupo, no caso em que a variável explicativa é categórica. Supondo que x_j seja uma variável explicativa binária, matematicamente a razão de prevalência é dada pela divisão (ou razão) entre as probabilidades do evento de interesse (Y=1) obtidas para as categorias $x_j = 1$ e $x_j = 0$ ($x_j = 0$ é a categoria de referência), de forma que:

$$\begin{aligned}
RP_j &= \frac{p_i|x_j = 1}{p_i|x_j = 0} \\
RP_j &= \frac{e^{\beta_0+\beta_1 x_{i1}+...+\beta_j(1)+...+\beta_{k-1}x_{ik-1}}}{e^{\beta_0+\beta_1 x_{i1}+...+\beta_j(0)+...+\beta_{k-1}x_{ik-1}}} \\
RP_j &= \frac{e^{\beta_j(1)}}{e^{\beta_j(0)}} \\
RP_j &= e^{\beta_j}
\end{aligned}$$

Portanto, a estimativa da razão de prevalência, pode ser descrita como:

$$\widehat{RP_j} = e^{\hat{\beta}_j}$$

Quando $\hat{\beta}_j > 0$ temos que $\widehat{RP_j} > 1$, o que significa que a prevalência do evento de interesse na categoria j é $(RP_j - 1)$ x 100% maior que a prevalência do evento de interesse na categoria de referência. Se $\hat{\beta}_j < 0$ e $\widehat{RP_j} < 1$, há indicação de que a prevalência do evento de interesse na categoria j é $(RP_j - 1)$ x 100% menor que a da categoria de referência. E no caso de $\hat{\beta}_j = 0$ e, portanto, $\widehat{RP_j} = 1$, as prevalências do evento de interesse nas duas categorias da variável x_j são iguais.

3.5.4 Inferência estatística dos parâmetros

Utilizando as estimativas pontuais dos parâmetros do modelo, bem como as suas estimativas de variância (ou erros padrão), é possível realizar inferência estatística sobre os parâmetros do modelo, por meio do cálculo de intervalos de confiança e aplicação de testes de hipóteses de significância dos parâmetros. O teste de Wald pode ser empregado para testar hipóteses sobre um único parâmetro ou para múltiplos parâmetros.

3.5.4.1 Teste de Wald de significância individual dos parâmetros do modelo

O teste de Wald de significância individual é utilizado para avaliar a significância de cada parâmetro β_j do modelo, considerando o nível de significância de $100\alpha\%$.

- Hipóteses a serem testadas:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

  onde β_j é o efeito do j-ésimo nível da variável explicativa.

- Estatística de teste

  Sob H_0: $\beta_j = 0$, a estatística de teste é dada por:

$$Z = \frac{\hat{\beta}_j}{\sqrt{\widehat{VAR}(\hat{\beta}_j)}} \overset{a}{\sim} N(0,1).$$

  onde $\hat{\beta}_j$ é o estimador de máxima verossimilhança de β_j e $\sqrt{\widehat{VAR}(\hat{\beta}_j)}$ é o estimador do erro-padrão de $\hat{\beta}_j$. Esta estatística tem distribuição normal padrão para grandes

amostras.

- Região crítica (ou região de rejeição de H_0)

$$RC = \{z \in \mathbb{R} |\ |z| > z_{(1-\alpha/2)}\}$$

onde $z_{(1-\alpha/2)}$ é o valor crítico da distribuição normal padrão no percentil $\left(1 - \frac{\alpha}{2}\right)$.

- Tomada de decisão

Se o valor observado z_{obs} pertencer à região crítica (RC), rejeita-se a hipótese nula H_0 ao nível de significância de $100\alpha\%$, ou seja, existe efeito estatisticamente significante do j-ésimo nível da variável explicativa. Equivalentemente, usando o p-valor do teste de Wald, rejeita-se a hipótese nula H_0 se o p-valor $= 2P(Z > |z_{obs}|) \leq \alpha$. Caso contrário, não há evidências para rejeitar H_0 e conclui-se que o efeito do j-ésimo nível da variável explicativa não é estatisticamente significante ao nível de significância de $100\alpha\%$.

3.5.4.2 Teste de Wald de significância geral dos parâmetros do modelo

O teste de Wald de significância geral é utilizado para avaliar a significância de múltiplos parâmetros do modelo, ao nível de significância de $100\alpha\%$. Desse modo, pode ser usado para testar a significância do efeito de variáveis categóricas com mais de dois níveis (POWERS e XIE, 1999).

Seja $\hat{\boldsymbol{\beta}}_{\boldsymbol{r}}$ um subvetor do vetor de parâmetros estimados $\hat{\boldsymbol{\beta}}$ e $\widehat{\boldsymbol{VAR}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{r}})$ uma submatriz da matriz de variância-covariância de $\hat{\boldsymbol{\beta}}$.

- Hipóteses a serem testadas:

$$\begin{cases} H_0 : \boldsymbol{\beta}_{\boldsymbol{r}} = \mathbf{0} \\ H_1 : \boldsymbol{\beta}_{\boldsymbol{r}} \neq \mathbf{0} \end{cases}$$

- Estatística de teste

Sob $H_0 : \boldsymbol{\beta}_{\boldsymbol{r}} = \mathbf{0}$, a estatística de teste é dada por:

$$W = (\hat{\boldsymbol{\beta}}_{\boldsymbol{r}})' [\widehat{\boldsymbol{VAR}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{r}})]^{-1} (\hat{\boldsymbol{\beta}}_{\boldsymbol{r}}) \sim \chi^2_r$$

onde W é a estatística de Wald que segue uma distribuição Qui-quadrada χ^2 com r graus de liberdade, onde r é a dimensão de $\boldsymbol{\beta}_{\boldsymbol{r}}$.

- Região crítica (ou região de rejeição de H_0)

$$RC = \{w \in \mathbb{R} |\ w > \chi^2_{1-\alpha,r}\}$$

  onde $\chi^2_{1-\alpha,r}$ é o valor crítico da distribuição Qui-quadrada χ^2_r no percentil $(1-\alpha)$.

- Tomada de decisão

Se o valor observado de W pertencer à região crítica (RC), rejeita-se a hipótese nula H_0 ao nível de significância de $100\alpha\%$, ou seja, há evidências de que $\boldsymbol{\beta_r}$ é significativamente diferente de zero. Desse modo, conclui-se que existe efeito estatisticamente significativo da variável explicativa categórica sobre o desfecho do modelo. Caso contrário, não há evidências para rejeitar H_0 ao nível de significância de $100\alpha\%$, ou seja, não há associação significativa entre a variável explicativa categórica e o desfecho do estudo. As mesmas conclusões podem ser obtidas usando o p-valor do teste de Wald de significância geral, dado por: $p - valor = P(W > w_{obs})$. Nesta abordagem, o critério de decisão é o de rejeitar a hipótese nula H_0 se $p - valor \leq 100\alpha\%$, e de não rejeitar a hipótese H_0 se $p - valor > 100\alpha\%$.

3.5.5 Avaliação da qualidade do ajuste e da capacidade preditiva do modelo

3.5.5.1 Pseudo-R^2 de McFadden

Uma forma de avaliar a qualidade global do ajuste do modelo é medir a dimensão do efeito do modelo (MARÔCO, 2007). Em modelos clássicos de regressão linear, a dimensão do efeito das variáveis explicativas sobre a variável resposta Y é avaliada por meio do coeficiente de determinação do modelo (R^2). Este coeficiente indica o quanto da variação total dos valores da variável resposta é explicado pelo modelo ajustado. Em MLGs é comum usar o pseudo-R^2 para essa medição. Os pseudo-R^2 são baseados na comparação do modelo ajustado com o modelo nulo (somente com a constante), e por isso não são medidas da variabilidade explicada pelo modelo (HOSMER e LEMESHOW, 2000).

Existem diferentes medidas de pseudo-R^2, entre elas pode citar os pseudo-R^2 de Cox-Snell, de Nagelkerke e de McFadden, mas o que apresenta uma melhor interpretabilidade

é o pseudo-R^2 de McFadden, que pode ser representado pela seguinte expressão:

$$R^2_{MF} = 1 - \frac{\ln L(\hat{\boldsymbol{\beta}}|\boldsymbol{y})}{\ln L(\hat{\boldsymbol{\beta}}_{\mathbf{0}}|\boldsymbol{y})}$$

onde $\ln L(\hat{\boldsymbol{\beta}}|\boldsymbol{y})$ é o logaritmo da função de verossimilhança do modelo sob consideração avaliado no ponto $\hat{\boldsymbol{\beta}}$, e $\ln L(\hat{\boldsymbol{\beta}}_{\mathbf{0}}|\boldsymbol{y})$ é o logaritmo da função de verossimilhança do modelo nulo (somente com a constante), isto é, avaliado no ponto $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$.

O R^2_{MF} pode ser interpretado como a proporção de redução do logaritmo da função de verossimilhança do modelo nulo relativamente ao modelo completo, ou seja, a razão do ganho de informação estimada pelo modelo completo, comparativamente ao modelo nulo (MARÔCO, 2007). O pseudo-R^2 de McFadden está compreendido no intervalo [0,1], e quanto maior, melhor o ajuste do modelo sob consideração.

3.5.5.2 Métricas da matriz de confusão

A partir do modelo de regressão log-linear de Poisson ajustado, estima-se a probabilidade de sucesso para cada elemento da amostra. Se a probabilidade estimada $\hat{p}_i$ for maior que o ponto de corte υ ($\hat{p}_i > \upsilon$), o elemento amostral é classificado como "sucesso", e se a probabilidade estimada $\hat{p}_i$ for menor ou igual ao ponto de corte υ ($\hat{p}_i \leq \upsilon$), o elemento é classificado como "fracasso". Qualquer ponto de corte pode ser utilizado para a classificação dos elementos da amostra nos dois grupos ("sucesso" *versus* "fracasso"), entretanto o melhor ponto de corte é aquele que maximiza as medidas de sensibilidade e especificidade.

Para avaliar a capacidade preditiva do modelo, é comum construir uma tabela de contingência, para classificar os n elementos da amostra segundo as categorias observadas e as categorias preditas pelo modelo ajustado .Esta tabela é chamada de matriz de confusão em modelos de classificação em Aprendizado de Máquina (*Machine Learning*), e a partir dela é possível determinar a quantidade de elementos que foram classificados corretamente ou incorretamente pelo modelo. O Quadro 2 ilustra a matriz de confusão, e as possíveis classificações dos elementos da amostra para a avaliação de um modelo com desfecho binário.

Quadro 2: Matriz de confusão para a classificação dos elementos da amostra segundo as categorias observadas e preditas pelo modelo de regressão log-linear de Poisson para um desfecho binário.

Categorias estimadas	**Categorias observadas**		**Total**
	Sucesso Y=1	**Fracasso Y=0**	
Sucesso $\hat{\boldsymbol{Y}} = \mathbf{1}$	f_{11} (Verdadeiros positivos)	f_{12} (Falsos positivos)	$f_{11} + f_{12}$
Fracasso $\hat{\boldsymbol{Y}} = \mathbf{0}$	f_{21} (Falsos negativos)	f_{22} (Verdadeiros negativos)	$f_{21} + f_{22}$
Total	$f_{11} + f_{21}$	$f_{12} + f_{22}$	$n = f_{11} + f_{12} + f_{21} + f_{22}$

Usando a matriz de confusão é possível calcular algumas métricas para a avaliação da capacidade de predição do modelo, conhecidas como sensibilidade, especificidade e acurácia.

A *sensibilidade* (S) é a proporção de verdadeiros positivos entre todos os elementos que satisfazem o evento de interesse (Y=1). A sensibilidade mede a capacidade do modelo de classificar o elemento como possuindo a característica de interesse ($\hat{Y} = 1$), quando de fato a unidade possui a característica (Y=1).

$$S = P(\hat{Y} = 1|Y = 1) = \frac{f_{11}}{f_{11} + f_{21}} \tag{3.6}$$

A *especificidade* (E) é a proporção de verdadeiros negativos entre todos os elementos que não satisfazem o evento de interesse (Y=0). A especificidade mede a capacidade do modelo de classificar o elemento como não tendo a característica de interesse ($\hat{Y} = 0$), quando de fato o elemento não possui a característica de interesse (Y=0).

$$E = P(\hat{Y} = 0|Y = 0) = \frac{f_{22}}{f_{12} + f_{22}} \tag{3.7}$$

De acordo com Marôco (2007), um modelo com boa capacidade preditiva apresenta sensibilidade e especificidade superiores a 80%. Ambas as medidas entre 50% e 80% o modelo tem capacidade preditiva razoável. Abaixo de 50% a capacidade preditiva é considerada ruim.

Por fim, a *acurácia* (A), ou taxa global de classificação corretas, é definida pela razão

entre o número de elementos classificados corretamente pelo modelo (soma do número de verdadeiros positivos e de verdadeiros negativos) e o número total de elementos (tamanho da amostra).

$$A = \frac{f_{11} + f_{22}}{f_{11} + f_{12} + f_{21} + f_{22}} \tag{3.8}$$

Assim como as medidas de sensibilidade e especificidade, a acurácia também pode ser expressa em porcentagem (%).

3.5.5.3 Área sob a curva ROC

Uma outra medida de capacidade preditiva (ou discriminatória) do modelo é a área sob a curva ROC (*Receiver Operating Characteristic)*. A curva ROC é um gráfico cujos eixos Y e X são respectivamente a sensibilidade (3.6) e o complemento da especificidade (3.9) do modelo; Para cada ponto de corte são calculados os valores da sensibilidade (taxa de verdadeiros positivos) e do complemento da especificidade e a união desses pontos forma a curva ROC (MARÔCO, 2007).

O complemento da especificidade é a taxa de falsos positivos, que expressa a capacidade do modelo de classificar o elemento como possuindo a característica de interesse ($\hat{Y} = 1$), quando na realidade este elemento não tem a característica de interesse (Y=0).

$$\bar{E} = 1 - P(\hat{Y} = 0|Y = 0) = P(\hat{Y} = 1|Y = 0) = \frac{f_{12}}{f_{12} + f_{22}} \tag{3.9}$$

A área sob a curva (*AUC, area under curve*) representa a performance global do modelo, permitindo avaliar o seu poder discriminatório. A área varia no intervalo de 0 e 1, e quanto maior a área, maior a capacidade discriminatória do modelo. O poder discriminatório pode ser classificado com base nos valores da área sob a curva ROC, conforme mostrado no Quadro 3:

Quadro 3: Classificação do poder discriminatório do modelo segundo área sob a curva ROC.

Área sob a curva ROC	**Poder discriminatório do modelo**
AUC = 0,5	Sem poder discriminatório
$0,5 \leq \text{AUC} < 0,7$	Discriminação fraca
$0,7 \leq \text{AUC} < 0,8$	Discriminação aceitável
$0,8 \leq \text{AUC} < 0,9$	Discriminação boa
$\text{AUC} \geq 0,9$	Discriminação ótima

Fonte: Adaptado de MARÔCO et al. (2007)

3.6 Aprendizado de Máquina Supervisionado

Aprendizado de Máquina (*Machine Learning*) é uma área da Inteligência Artificial (IA) cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado, tal como a construção de sistemas capazes de adquirir conhecimentos de forma automática (MONARD e BARANAUSKAS, 2003). O aprendizado de máquina supervisionado é uma técnica na qual o algoritmo aprende uma função a partir dos dados da base de treinamento (AYODELE, 2010). Na abordagem de aprendizado de máquina supervisionado, se extrai significado dos dados com base no treinamento de um modelo em dados rotulados, tendo, portanto, como objetivo a construção de um modelo para predizer uma variável resposta a partir de um conjunto de variáveis explicativas ou preditoras (BRUCE e BRUCE, 2019).

Neste contexto, existe uma variável resposta (output), que no presente trabalho é uma variável qualitativa binária, cujos valores queremos predizer a partir de um conjunto de variáveis explicativas (inputs) que representam as características da mãe e do bebê. A partir dos dados de treinamento, onde são conhecidos tanto os valores da variável de saída quanto os valores da variável de entrada para as unidades da amostra, o objetivo é construir um modelo de predição que seja capaz de predizer o valor da variável resposta (output) para um novo conjunto de unidades não observadas. Como apontado por Hastie et al.(2009), um bom modelo de predição é aquele que prediz com acurácia um determinado desfecho. Então, deve-se escolher o modelo que tem o menor erro de predição (ou a maior acurácia) entre um conjunto de modelos candidatos.

3.6.1 Validação Cruzada (*Cross Validation*)

A Validação Cruzada é uma ferramenta utilizada para obter uma estimativa mais realista do erro de predição do modelo, cuja utilização vem crescendo devido ao aumento

da potência e velocidade computacional (EFRON E TIBSHIRANI, 1994).

Visando avaliar o desempenho de diferentes modelos a fim de escolher o melhor, bem como para o modelo selecionado avaliar a sua capacidade preditiva para novos dados, Hastie et al.(2009) recomendam dividir aleatoriamente a amostra total em três partes, mantendo os seguintes percentuais de dados em cada parte: 50% para a amostra treino, 25% para a amostra de validação e os 25% restantes dos dados para a amostra teste. Como mencionados por estes autores a amostra treino é usada para ajustar o modelo de predição; a amostra de validação para avaliar o desempenho do modelo de predição, por meio da estimação do erro de predição ou da acurácia da predição; e, por fim, a amostra teste é utilizada para avaliar a capacidade de generalização do modelo final escolhido, sendo a amostra teste composta por dados novos. Shalev-Shwartz e Ben-David (2014), também abordam a divisão da base de dados em treino-validação-teste, e mencionam que uma estimativa adequada do erro de predição pode ser obtida usando alguns dos dados de treinamento sem sobreposição para compor a amostra de validação.

Segundo Efron e Tibshirani (1994), a razão para divisão da amostra deve-se ao fato de que quando se utiliza uma mesma amostra tanto para ajustar o modelo quanto para avaliá-lo usando as estimativas dos parâmetros do modelo obtidas na própria amostra, o erro de predição do modelo (ou a acurácia do modelo) é otimista, isto é, tende a subestimar o verdadeiro erro de predição.

Para fins de ilustração, suponha que um modelo de regressão linear simples da forma $y_i = f(x_i) + \epsilon_i$, foi ajustado utilizando a amostra treino:$(x_1, y_1), (x_2, y_2)...., (x_n, y_n)$, obtendo os valores estimados $\hat{y}_1 = \hat{f}(x_1), \hat{y}_2 = \hat{f}(x_2), ..., \hat{y}_n = \hat{f}(x_n)$. Como ressaltado por James et al. (2013), em aprendizado estatístico, não se tem interesse em avaliar se os valores estimados na amostra treino são aproximadamente iguais aos valores observados. O que interessa é que o modelo ajustado, $\hat{f}$, na amostra treino quando aplicado aos dados da amostra de validação, $(x_1^*, y_1^*), (x_2^*, y_2^*)...., (x_k^*, y_k^*)$ produza valores estimados próximos dos valores observados, isto é, $\hat{y}_1^* = \hat{f}(x_1^*) \approx y_1^*, \hat{y}_2^* = \hat{f}(x_2^*) \approx y_2^*, ..., \hat{y}_k^* = \hat{f}(x_k^*) \approx y_k^*$. A separação da base de dados é representada na Figura 1:

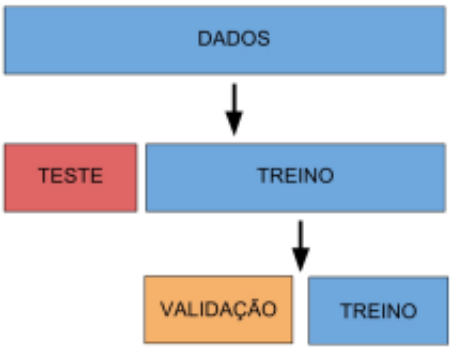


Figura 1: Representação da separação da base de dados em treino, validação e teste. Fonte: Adaptado de FERREIRA (2018).

3.6.2 Método de reamostragem - Bootstrap

O Bootstrap é um método de reamostragem utilizado para gerar novas amostras de forma aleatória da base de dados, com reposição, e pode ser usado para avaliar a precisão da estimativa de um parâmetro qualquer, sendo que em modelos de aprendizado de máquina supervisionado, este parâmetro pode ser uma medida de acurácia do desempenho do modelo ou o erro de predição (HASTIE et al., 2009; RASCHKA, 2018). Segundo Raschka (2018), a acurácia da predição, já definida na seção 3.5.5.2, como a razão entre o número total de predições (classificações) corretas e o tamanho da amostra (n), é o complementar do erro de predição. Usando uma notação matemática mais formal, a acurácia de predição é dada por:

$$A = 1 - Erro = 1 - \frac{\sum_{i=1}^{n} L(\hat{y}_i, y_i)}{n}$$

onde:

$$L(\hat{y}_i, y_i) = \begin{cases} 1, \text{se } \hat{y}_i \neq y_i \\ 0, \text{se } \hat{y}_i = y_i \end{cases}$$

sendo $\hat{y}_i$ e y_i os valores estimado e observado (real) da variável resposta para a i-ésima unidade da amostra, respectivamente.

Considerando a acurácia como medida de desempenho do modelo como abordado em RASCHKA (2018), o método Bootstrap consiste em:

1. Gerar B amostras aleatórias com reposição e de mesmo tamanho da base de dados de treinamento de tamanho n;

2. Ajustar o modelo para cada amostra bootstrap (amostra treino) e calcular a sua

acurácia usando os dados da amostra de validação, gerando assim um conjunto de B estimativas de acurácia, denotadas por: $A_1, A_2, \ldots, A_B$;

3. Calcular a estimativa da acurácia do modelo como a média das B medidas de acurácia estimadas usando cada uma das B amostras bootstrap, da seguinte forma:

$$\overline{A} = \frac{\sum_{i=1}^{B} A_i}{B}$$

onde A_i é a estimativa da acurácia obtida na b-ésima amostra bootstrap (amostra de validação).

As estimativas da variância e do coeficiente de variação (CV) da acurácia média podem ser facilmente obtidas usando as seguintes expressões:

$$\widehat{VAR}(\overline{A}) = \frac{\sum_{i=1}^{B}(A_i - \overline{A})^2}{B-1}$$

e

$$\widehat{CV}(\overline{A}) = \frac{\sqrt{\widehat{VAR}(\overline{A})}}{\overline{A}}\text{x}100$$

Segundo Efron e Tibshirani (1994), a utilização de 50 a 200 amostras bootstrap é suficiente para obter boas estimativas para os erros padrão de estimadores.

A realização do método bootstrap é resumida na Figura 2.

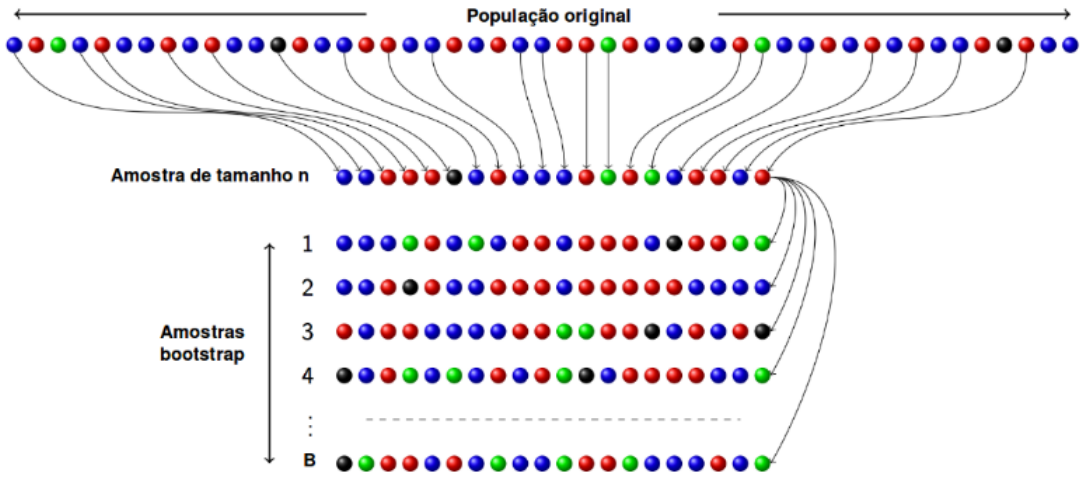


Figura 2: Geração de B amostras bootstrap de uma base de dados de tamanho n.
Fonte: Adaptado de FERREIRA (2018).

4 Análise dos Resultados

Inicialmente, para realizar a análise dos dados, foram excluídos do banco de dados os recém-nascidos que possuíam alguma informação faltante (*missings*), que corresponderam a 7,8% dos recém-nascidos que compõem a população de estudo do presente trabalho. Após esta exclusão, a base de dados, considerada a amostra total deste estudo, ficou com 180.159 recém-nascidos. Dentre esses bebês, 17.110 eram prematuros, ou seja, a prevalência de prematuridade foi de 9,5%. A distribuição percentual dos recém-nascidos para cada categoria das variáveis explicativas está disponível na Tabela 1.

Com relação às características da mãe, pode-se identificar um maior percentual de mães solteiras (62,4%), entre 20 a 34 anos (68,8%), não brancas (66,0%), nulíparas (sem gestações anteriores)(38,9%), que possuiam ensino médio ou superior incompleto (59,0%), que realizaram 6 ou mais consultas de acompanhamento pré-natal (84,2%) e que tiveram parto cesáreo (57,1%). Quanto às características dos recém-nascidos, pode-se destacar o maior percentual de bebês com índice de Apgar no 5° minuto normal (99,1%), sem baixo peso ao nascer (92,3%) e com apresentação cefálica (97,1%).

Ainda na Tabela 1, ao analisar a distribuição do desfecho de prematuridade segundo as características maternas e do bebê, pode-se observar percentuais de bebês prematuros ligeiramente maiores entre as mães na faixa etária mais baixa (19 anos ou menos; 10,4%) e na mais alta (35 anos ou mais; 11,8%), que possuiam até ensino fundamental (10,1%), viúvas ou separadas/divorciadas (12,2%), que já tiveram duas ou mais gestações anteriores (multíparas; 10,2%), que realizaram no máximo 5 consultas prá-natal (18,2%) e que tiveram parto cesáreo (10,5%). Além disso, foram observados maiores percentuais de prematuridade entre os bebês com índice de Apgar no 5° minuto classificados como muito baixo (43,1%) e baixo (40,7%), com baixo peso ao nascer (59,6%) e com apresentação pélvica ou podálica (20,1%). Verifica-se também, na análise descritiva, que o percentual de prematuridade foi muito próximo entre bebês do sexo masculino e feminino, assim como entre mães de cor branca e não branca (Tabela 1).

Tabela 1: Distribuição dos recém-nascidos na amostra completa por presença ou não de prematuridade, segundo as características da mãe e do recém-nascido. Estado do Rio de Janeiro,2019.

Características da mãe e do recém nascido	**Percentual de recém-nascidos n = 180159**	**Prematuridade (%)**	
		Sim (n=17110)	**Não (n=163049)**
Idade da mãe			
19 anos ou menos	13,9	10,4	89,6
20 a 34 anos	68,8	8,7	91,3
35 anos ou mais	17,3	11,8	88,2
Raça/Cor da mãe			
Branca	34,0	9,6	90,4
Não Branca	66,0	9,5	90,6
Escolaridade da mãe			
Até ensino fundamental	25,3	10,1	89,9
Ensino médio ou superior incompleto	59,0	9,2	90,8
Ensino superior completo	15,6	9,6	90,4
Situação conjugal da mãe			
Solteira	62,4	9,4	90,6
Casada/União estável	35,9	9,5	90,5
Viúva/Separada /Divorciada	1,7	12,2	87,8
Paridade			
Nulípara	38,9	9,8	90,2
Primípara	30,9	8,5	91,5
Multípara	30,1	10,2	89,8
Número de consultas pré-natal			
6 ou mais consultas pré-natal	84,2	7,9	92,1
Até 5 consultas pré-natal	15,8	18,2	81,8
Tipo de parto			
Cesáreo	57,1	10,5	89,5
Vaginal	42,9	8,2	91,8
Sexo do recém-nascido			
Masculino	51,0	9,9	90,2
Feminino	49,0	9,1	90,9
Índice Apgar no quinto minuto			
Muito baixo	0,2	43,1	56,9
Baixo	0,7	40,7	59,3
Normal	99,1	9,2	90,8
Baixo peso ao nascer			
Sim	7,7	59,6	40,4
Não	92,3	5,3	94,7
Tipo de apresentação do bebê			
Cefálico	97,1	9,2	90,8
Pélvica ou podálica/Transversa	2,9	20,1	72,7

Para predizer o desfecho de prematuridade, a partir das características maternas e do bebê, usando a abordagem de aprendizado de máquina supervisionado para a aplicação do modelo log-linear de Poisson (com variância robusta), o primeiro passo foi separar a base de dados em três partes. Uma parte foi empregada para treinar o modelo (amostra treino), outra parte para avaliar o desempenho do modelo (amostra de validação), e, por fim,uma terceira parte foi reservada para avaliação da capacidade de generalização do modelo final escolhido (amostra teste). A Tabela 2 mostra as variáveis explicativas selecionadas nos modelos ajustados considerando cada uma das amostras treino geradas usando o método de re-amostragem bootstrap, apresentado na seção Material e Métodos.

Como pode ser visualizado na Tabela 2, foram geradas 50 amostras treino, e portanto 50 modelos de regressão log-linear de Poisson foram ajustados incluindo todas as onze variáveis explicativas descritas no Quadro 1. Aplicando o teste de Wald de significância geral, com nível de significância de $\alpha = 5\%$, pode-se observar que do total das amostras treino geradas, em 86% (43) delas, apenas a variável "escolaridade" não apresentou associação significativa com o desfecho de prematuridade, sendo excluída destes modelos. Na amostra treino 17, além da "escolaridade", também foi excluída a variável "paridade"; e em seis amostras treino (6,18,30,40,46,49) foi excluída a "cor da mãe", além da "escolaridade". Por fim, para cada modelo selecionado na amostra treino calculou-se o ponto de corte ótimo.

A Tabela 2 mostra, ainda, três medidas de qualidade do ajuste (acurácia, sensibilidade e especificidade) calculadas para cada amostra de validação gerada pelo método de reamostragem bootstrap, a fim de avaliar a capacidade dos modelos selecionados na predição da variável resposta

Tabela 2: Processo de treinamento do modelo log-linear de Poisson (com variância robusta) para predição do desfecho de prematuridade, usando o método de reamostragem bootstrap.

Amostra	**Processo de treinamento do modelo (75% da amostra total, n=135.120)**				
	Amostra treino (50%, n=90.081)		**Amostra de validação (25%,n=45.039)**		
	Váriaveis removidas	**Ponto de corte ótimo**	**A**	**S**	**E**
1	ESCOLARIDADE	0,088	0,907	0,506	0,949
2	ESCOLARIDADE	0,087	0,906	0,518	0,946
3	ESCOLARIDADE	0,082	0,904	0,516	0,945
4	ESCOLARIDADE	0,084	0,896	0,521	0,936
5	ESCOLARIDADE	0,086	0,903	0,495	0,946

Tabela 2 (continuação)

Amostra	Processo de treinamento do modelo (75% da amostra total, n=135.120)				
	Amostra treino (50%, n=90.080)		Amostra de validação (25%,n=45.040)		
	Váriaveis removidas	Ponto de corte ótimo	A	S	E
6	ESCOLARIDADE; RAÇA/COR MÃE	0,086	0,906	0,503	0,949
7	ESCOLARIDADE	0,085	0,901	0,521	0,941
8	ESCOLARIDADE	0,089	0,908	0,504	0,949
9	ESCOLARIDADE	0,086	0,905	0,498	0,948
10	ESCOLARIDADE	0,083	0,894	0,509	0,934
11	ESCOLARIDADE	0,094	0,912	0,485	0,957
12	ESCOLARIDADE	0,086	0,902	0,500	0,943
13	ESCOLARIDADE	0,087	0,905	0,501	0,948
14	ESCOLARIDADE	0,090	0,910	0,522	0,950
15	ESCOLARIDADE	0,087	0,910	0,501	0,953
16	ESCOLARIDADE	0,086	0,906	0,501	0,948
17	ESCOLARIDADE; PARIDADE	0,096	0,914	0,493	0,957
18	ESCOLARIDADE; RAÇA/COR MÃE	0,086	0,908	0,495	0,952
19	ESCOLARIDADE	0,086	0,905	0,503	0,948
20	ESCOLARIDADE	0,097	0,914	0,486	0,960
21	ESCOLARIDADE	0,086	0,904	0,506	0,945
22	ESCOLARIDADE	0,083	0,901	0,508	0,942
23	ESCOLARIDADE	0,084	0,901	0,505	0,943
24	ESCOLARIDADE	0,085	0,905	0,516	0,945
25	ESCOLARIDADE	0,085	0,905	0,501	0,948
26	ESCOLARIDADE	0,084	0,901	0,506	0,942
27	ESCOLARIDADE	0,085	0,905	0,509	0,946
28	ESCOLARIDADE	0,082	0,900	0,509	0,942
29	ESCOLARIDADE	0,091	0,912	0,487	0,957
30	ESCOLARIDADE; RAÇA/COR MÃE	0,086	0,909	0,495	0,952
31	ESCOLARIDADE	0,088	0,909	0,505	0,952
32	ESCOLARIDADE	0,086	0,903	0,502	0,945
33	ESCOLARIDADE	0,084	0,904	0,510	0,946
34	ESCOLARIDADE	0,087	0,899	0,524	0,938
35	ESCOLARIDADE	0,086	0,903	0,511	0,943
36	ESCOLARIDADE	0,084	0,904	0,515	0,944
37	ESCOLARIDADE	0,086	0,903	0,508	0,944
38	ESCOLARIDADE	0,084	0,902	0,512	0,944
39	ESCOLARIDADE	0,085	0,903	0,504	0,945
40	ESCOLARIDADE; RAÇA/COR MÃE	0,086	0,906	0,508	0,948
41	ESCOLARIDADE	0,093	0,913	0,494	0,958

Tabela 2 (continuação)

Amostra	Processo de treinamento do modelo (75% da amostra total, n=135.120)				
	Amostra treino (50%, n=90.080)		Amostra de validação (25%,n=45.040)		
	Váriaveis removidas	Ponto de corte ótimo	A	S	E
42	ESCOLARIDADE	0,086	0,906	0,508	0,948
43	ESCOLARIDADE	0,085	0,904	0,507	0,945
44	ESCOLARIDADE	0,085	0,904	0,507	0,945
45	ESCOLARIDADE	0,084	0,899	0,513	0,939
46	ESCOLARIDADE; RAÇA/COR MÃE	0,087	0,908	0,496	0,952
47	ESCOLARIDADE	0,083	0,901	0,507	0,942
48	ESCOLARIDADE	0,083	0,900	0,510	0,942
49	ESCOLARIDADE; RAÇA/COR MÃE	0,086	0,908	0,501	0,951
50	ESCOLARIDADE	0,087	0,903	0,520	0,945

Com relação a avaliação do desempenho dos 50 modelos selecionados na amostra treino e aplicados nas amostras de validação, observa-se que todos eles apresentaram medidas altas de acurácia e de especificidade, superiores a 89%, embora as medidas de sensibilidade tenham sido de aproximadamente 50%.

Na Tabela 3, são fornecidas as médias, os desvios-padrão e os coeficientes de variação (CV) das três medidas de qualidade de ajuste, estimadas na amostra de validação.

Tabela 3: Média, Desvio-padrão e coeficiente de variação para as distribuições das 50 medidas de sensibilidade, especificidade e acurácia estimadas nas amostras de validação.

Medidas de qualidade	Amostra de validação		
	Média	Desvio-padrão	CV
Sensibilidade (S)	0,506	0,009	1,78%
Especificidade (E)	0,947	0,005	0,53%
Acurácia (A)	0,905	0,004	0,44%

As três medidas de qualidade de ajuste apresentaram um grau de variabilidade muito baixo, sendo a acurácia a medida com menor coeficiente de variação (CV), indicando uma pequena variação relativa das acurácias dos modelos em torno da acurácia média. Como nem todas as amostras treino geraram modelos com o mesmo subconjunto de variáveis explicativas, fez-se necessário escolher qual subconjunto de variáveis será utilizado para o ajuste do modelo final, cuja capacidade de generalização será avaliado com novos dados (amostra teste).

Como todos os modelos geraram medidas de acurácia altas e próximas, a princípio qualquer um dos subconjuntos de variáveis explicativas poderia ser escolhido. Tendo em vista que a acurácia média é uma boa estimativa da acurácia do modelo, escolheu-se o subconjunto de variáveis do modelo ajustado na amostra treino 24, por ter gerado uma acurácia na amostra de validação mais próxima da acurácia média. Então, o modelo final é composto por todas as variáveis de estudo, com exceção da variável escolaridade materna.

A Tabela 4 mostra os resultados do ajuste do modelo log-linear de Poisson usando os 75% dos dados da amostra total (amostra treino + amostra de validação), com o subconjunto selecionado de variáveis, isto é, o subconjunto que não contém apenas a variável "escolaridade materna".

Tabela 4: Resultados do ajuste do modelo log-linear de Poisson para predição do desfecho de prematuridade, com o subconjunto selecionado de variáveis, considerando os 75% de dados da amostra total (n=135.120).

	Resultados do ajuste do modelo (n=135120)			
Características da mãe e do recém nascido	**Estimativa pontual**	**Razão de Prevalência (RP)**	**Erro-padrão robusto**	**p-valor (Wald)**
Idade materna				**<0,001**
19 anos ou menos	0	1	-	-
20 a 34 anos	-0,052	0,949	0,023	0,022
35 anos ou mais	0,127	1,136	0,028	<0,001
Raça da mãe				**<0,001**
Branca	0	1	-	-
Não branca	-0.060	0,942	0,016	<0,001
Situação conjugal da mãe				**<0,001**
Solteira	0	1	-	-
Casada/União estável	0,100	1,105	0,017	<0,001
Viúva/Separada/Divorciada	0,179	1,196	0,050	<0,001
Paridade				**<0,001**
Núlipara	0	1	-	-
Primípara	-0,057	0,944	0,019	0,002
Multípara	0,034	1,035	0,019	0,067
N° de consultas pré-natal				**<0,001**

Tabela 4 (continuação)

Características da mãe e do recém nascido	Resultados do ajuste do modelo (n=135120)			
	Estimativa pontual	**Razão de Prevalência (RP)**	**Erro-padrão robusto**	**p-valor (Wald)**
Até 5 consultas	0	1	-	-
6 ou mais consultas	-0,443	0,642	0,016	<0,001
Tipo de parto				**<0,001**
Vaginal	0	1	-	-
Cesárea	0,188	1,207	0,016	<0,001
Sexo do bebê				**<0,001**
Masculino	0	1	-	-
Feminino	-0,153	0,858	0,015	<0,001
Índice Apgar no 5° minuto				**0,020**
Muito baixo	0	1	-	-
Baixo	0,033	1,033	0,061	0,589
Normal	-0,033	0,722	0,056	<0,001
Baixo peso ao nascer				**<0,001**
Sim	0	1	-	-
Não	-2,291	0,101	0,015	<0,001
Tipo de apresentação do bebê				**<0,001**
Cefálico	0	1	-	-
Pélvico ou Podálica/Transversa	0,232	1,260	0,026	<0,001

Na Tabela 4, pode-se observar que todas as variáveis explicativas apresentam associação estatisticamente significante com o desfecho de prematuridade considerando o nível de significância de 5%.

Com relação às características das mães, observa-se que na faixa etária materna de 20 a 34 anos houve uma prevalência de prematuridade 5,1% menor do que na faixa etária materna de 19 anos ou menos (RP=0,949; $p\text{-}valor = 0{,}022$). Já na faixa etária materna de 35 anos ou mais verificou-se uma prevalência de prematuridade 13,6% maior do que na faixa etária materna de referência (RP=1,136; $p\text{-}valor < 0,001$).

Quanto à raça/cor da mãe, entre as mães não brancas a prevalência de prematuridade foi 5,8% menor do que no grupo das mães brancas (RP=0,942; $p\text{-}valor < 0,001$).

Com relação à situação conjugal, observa-se que mães viúvas, separadas ou divorciadas tiveram prevalência de prematuridade 19,6% maior do que mães solteiras (RP=1,196; *p-valor* $< 0,001$). Além disso, foi observado que as mães casadas ou com união estável apresentaram prevalência de prematuridade 10,5% maior, comparativamente às mães solteiras (RP=1,105; *p-valor* $< 0,001$).

No que tange à paridade, a prevalência de prematuridade no grupo das mães que tiveram uma gestação anterior (primíparas) foi 5,6% menor que no de mães que não tiveram nenhuma gestação prévia (RP=0,944; *p-valor* $= 0,002$).Não observou-se diferenças na prevalência de prematuridade na comparação entre mães que tiveram mais de uma gestação anterior (multíparas) e mães que não tiveram nenhuma gestação anterior (nulíparas) (RP=1,035; *p-valor* $= 0,067$).

Observou-se que mães que realizaram 6 ou mais consultas de acompanhamento pré-natal tiveram prevalência de prematuridade 35,8% menor do que mães que realizaram no máximo 5 consultas (RP=0,642; *p-valor* $< 0,001$). Mães que realizaram parto cesáreo apresentaram prevalência de prematuridade 20,7% maior que mães que realizaram parto vaginal (RP=1,207; *p-valor* $< 0,001$).

No que se refere às características dos recém-nascidos, observa-se que a prevalência de prematuridade entre bebês do sexo feminino foi 14,2% menor do que bebês do sexo masculino (RP=0,858; *p-valor* $< 0,001$).

Considerando o índice Apgar no 5° minuto, verifica-se que os bebês que receberam a avaliação "normal"tiveram prevalência de prematuridade 27,8% menor do que os bebês com índice classificado como "muito baixo"(RP=0,722; *p-valor* $< 0,001$). Verificou-se também, que bebês que não possuíam baixo peso ao nascer tiveram prevalência de prematuridade 89,9% menor do que bebês que nasceram com baixo peso (RP=0,101; *p-valor* $< 0,001$).

Quanto ao tipo de apresentação do bebê, observa-se que a prevalência de prematuridade foi 26,0% maior entre recém-nascidos com apresentação pélvica ou podálica/transversa quando comparados àqueles com apresentação cefálica (RP=1,260; *p-valor* $< 0,001$).

Com relação ao pseudo-R^2 de Mcfadden, verifica-se que o ganho de informação do modelo final selecionado foi de 18,4% comparativamente ao modelo nulo.

O modelo final selecionado foi aplicado na base teste (25% dos dados restantes da amostra total), para avaliar como o modelo se comporta com dados nunca antes utilizados, a fim de conhecer a sua capacidade de generalização. As medidas de qualidade para o

modelo final selecionado estão na Tabela 5 e na Figura 3, que apresenta a curva ROC.

Tabela 5: Medidas de avaliação da capacidade do modelo log-linear de Poisson selecionado para a predição do desfecho de prematuridade, quando aplicado na amostra teste (n=45.039).

Medida de avaliação da capacidade preditiva do modelo (amostra teste)	**Valor**
Sensibilidade (S)	51,0%
Especificidade (E)	94,8%
Acurácia (A)	90,6%
Área sob a curva ROC (AUC)	0,766

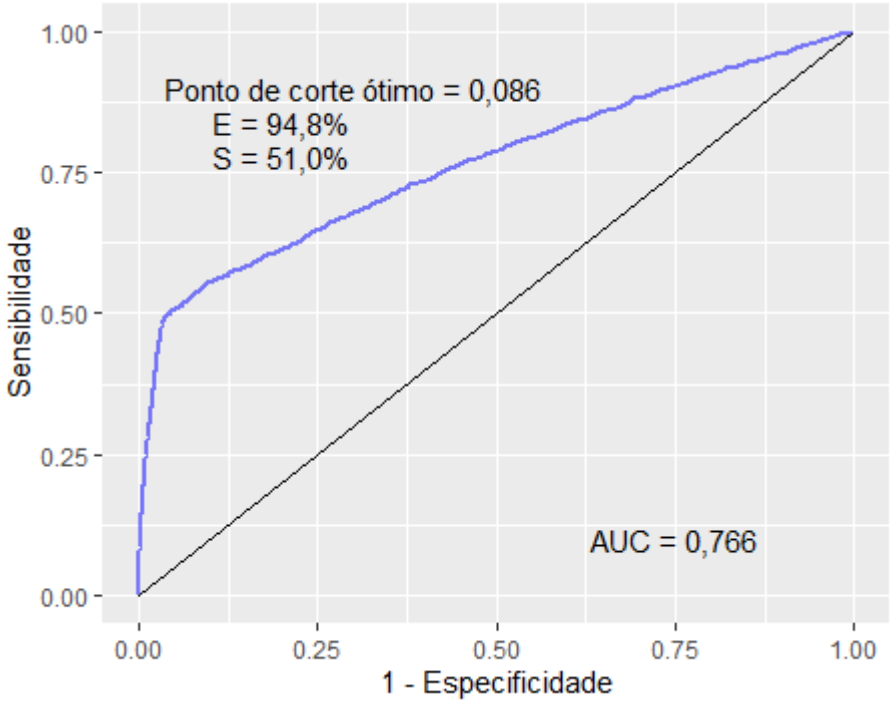


Figura 3: Curva ROC e da capacidade preditiva do modelo final selecionado quando aplicado na amostra teste.

O modelo final selecionado apresentou uma acurácia alta, indicando que um total de 90,6% do total de bebês foram classificados corretamente. Percebe-se que 94,8% são classificados corretamente como não prematuros pelo modelo selecionado e que 51,0% dos recém-nascidos são classificados corretamente como prematuros pelo modelo selecionado. A área sob a curva ROC foi de 0,766, o que indica que o modelo selecionado possui uma capacidade para a predição do desfecho de prematuridade considerada aceitável.

5 Discussão e Conclusão

No presente trabalho, foi observado que no estado do Rio de Janeiro, no ano de 2019, houve maior prevalência de prematuridade nas mães com 35 anos ou mais de idade, brancas, viúvas ou separadas/divorciadas, que realizaram no máximo 5 consultas de acompanhamento pré-natal e que tiveram parto cesáreo. As variáveis maternas aqui associadas à maior prevalência de prematuridade já foram objetos de outros estudos acerca do tema.

A maior prevalência de prematuridade nos filhos de mães com idade mais avançada, tal como observado por Oliveira et al. (2016) e Souza et al. (2019), é compatível com a probabilidade aumentada de complicações clínicas durante a gestação neste grupo, assim como a maior possibilidade de outras morbidades pré-existentes, condições estas que podem levar ao trabalho de parto prematuro ou à necessidade da interrupção prematura da gestação.

Guimarães et al. (2017), ao analisar os dados do SINASC 2008 a 2011 relativos a uma região de Minas Gerais, também encontraram associação entre prevalência de prematuridade e realização de menos de seis consultas pré-natal e ressaltaram que há possibilidade da presença de viés nesta associação, uma vez que, com a interrupção precoce da gestação, não haveria tempo hábil para a realização de um número mais elevado de consultas. Porém, cabe lembrar que se o acompanhamento pré-natal for iniciado no primeiro trimestre da gestação, com retornos mensais até a 28ª semana de gestação conforme preconizado pelo Ministério da Saúde (BEZERRA et al., 2006; BRASIL, 2019), a gestante alcançaria as cinco consultas nessa idade gestacional. Tal estimativa reforça a possibilidade de haver, de fato, uma associação entre acompanhamento pré-natal inadequado ou insuficiente e a ocorrência de prematuridade.

Na amostra estudada, as mães que se declararam como viúvas, separadas ou divorciadas foram aquelas que apresentaram maior prevalência de prematuridade, seguidas pelo grupo de mães casadas ou com união estável, quando comparadas ao grupo das mães que se declararam solteiras. Souza et al. (2019) encontraram resultados semelhantes, porém

outros estudos descreveram maior prevalência de prematuridade entre mães sem companheiros, sugerindo que o envolvimento da figura paterna poderia ter um impacto positivo na prevenção de desfechos indesejados na gestação (BEZERRA et al., 2006; ALIO et al., 2011).

A maior prevalência de prematuridade entre os nascidos vivos por parto cesáreo, encontrada no presente estudo, assim como em outros estudos (GUIMARÃES et al., 2017; SOUZA et al., 2019) pode estar relacionada a duas situações. A primeira situação seria a presença de intercorrências durante a gestação, como corioamnionite, doença hipertensiva da gestação ou sofrimento fetal, que podem levar à necessidade da interrupção precoce da gestação por meio do parto cesáreo, para maior segurança para a gestante e o concepto. A outra situação é a frequência elevada de nascimentos prematuros tardios associados aos partos cesáreos indicados de forma eletiva, antes de iniciado o trabalho de parto espontâneo (LEAL et al., 2016b). As informações disponíveis no SINASC não permitem analisar de forma mais aprofundada a natureza da associação entre o parto cesáreo e a maior prevalência de prematuridade.

Outra variável frequentemente associada a maior prevalência de prematuridade é a baixa escolaridade materna (SILVEIRA et al., 2010; OLIVEIRA et al., 2016; SOUZA et al., 2019). Entretanto, assim como nos estudos de Assunção et al. (2012) e Guimarães et al. (2017), no presente estudo o baixo nível de instrução materna não se mostrou uma variável associada de modo significativo a maior prevalência de prematuridade.

Com relação às características do recém-nascido associadas à maior prevalência de prematuridade, no presente trabalho, podem ser citadas: sexo masculino, índice de Apgar no 5° minuto com escore muito baixo ou baixo, baixo peso ao nascer e apresentação pélvica ou podálica.

Quanto ao sexo do bebê, a maior prevalência de prematuridade entre os meninos também foi observada em outros estudos, como os de Almeida et al. (2012), Basso et al. (2012) e Jesus et al. (2019). A hipótese de que existam variáveis biológicas que prejudicam o desenvolvimento de meninos pode ser corroborada pelos achados de Wainstock et al.(2015) de que os fetos masculinos apresentam pior resposta adaptativa a situações de estresse intrauterino pré-natal.

De encontro com os resultados do presente trabalho, o índice de Apgar com escore muito baixo também é associado com a prevalência de prematuridade no estudo de Oliveira et al.(2016). Baixos índices de Apgar nos recém-nascidos prematuros devem-se à imaturidade fisiológica, com diminuição de irritabilidade reflexa e incapacidade para res-

ponder de forma autônoma às funções cardiovasculares e respiratórias (ILIODROMITI et al., 2014).

A associação encontrada de baixo peso ao nascer e maior prevalência de prematuridade é coerente com a relação direta existente entre o aumento do peso do feto e o aumento da idade gestacional, que ocorre de forma mais acelerada nas últimas semanas da gestação. Assim sendo, o BPN é associado com a prevalência de prematuridade por conta da interrupção do gestação e a impossibilidade do feto de ganho nutricional. O baixo peso ao nascer é um determinante importante da desnutrição e influencia o crescimento e desenvolvimento da criança (RAMOS e CUMAN, 2009).

Com relação a potencialidade do presente estudo, pode-se destacar o uso do modelo de regressão log-linear de Poisson, que produz medidas de razão de prevalência de prematuridade, ao invés de medidas de razão de chance de prematuridade (odds ratio - OR) calculadas em modelos de regressão logística. Como apontado por Barros e Hirakata (2003) a razão de prevalência (RP) é uma medida natural e de maior interesse em estudos de natureza transversal, e concluiram que o modelo log-linear de Poisson com variância robusta é uma das melhores alternativas para a modelagem estatística de desfechos binários em estudos desta natureza quando comparado ao modelo logístico. Coutinho et al. (2008), mostraram que os OR, se interpretados como estimativas de RP, superestimariam as associações para os desfechos com prevalência baixa, intermediária e alta em 13%, quase 100% e quatro vezes mais, respectivamente.

Ainda como potencialidade, este estudo utilizou o método de aprendizado de máquina supervisionado, dividindo a base de dados em treino-validação-teste como recomendado por Shalev-Shwartz e Ben-David (2014) e Hastie et al. (2009), a fim de obter um modelo mais adequado para predição do desfecho de prematuridade. Como destacado por Efron e Tibshirani (1994), quando se usa a mesma amostra para ajustar o modelo e avaliar o seu desempenho na predição do desfecho, o erro de predição do modelo tende a ser otimista, isto é, tende a subestimar o verdadeiro erro de predição. O uso do método de aprendizado de máquina tem mostrado resultados promissores na área dos estudos epidemiológicos em saúde (BELUZO et al., 2020; WIEMKEN e KELLEY, 2020) e o presente estudo encontra-se entre os pioneiros no tema, no Brasil.

Com relação as limitações, este estudo se baseou exclusivamente em dados disponíveis no banco de dados do SINASC 2019, ficando dependente do preenchimento correto e completo do formulário de Declaração de Nascido Vivo (DN), além de não contar com mais características de saúde e estilo de vida da mãe, importantes para uma análise mais

aprofundada da questão dos fatores associados à maior prevalência de prematuridade. Ressalta-se que, por tratar-se de um estudo descritivo, do tipo transversal, não é possível estabelecer e comprovar associações de causalidade entre as variáveis analisadas

A prematuridade, apesar de cada vez mais discutida, ainda se faz muito presente no cenário atual. Os resultados do estudo evidenciam a associação da prematuridade com desfechos neonatais indesejáveis, como baixo peso e piores escores de Apgar no 5º minuto de vida. Conclui-se que é necessário maior atenção, por parte dos gestores e profissionais de saúde, às gestantes maiores de 35 anos, brancas, separadas, que tiveram menos consultas de acompanhamento pré-natal que o recomendado. Orientar as gestantes sobre a importância da qualidade do pré-natal por meio de campanhas públicas e implantação de políticas que visem diminuir o número de partos via cesariana, são recomendações a fim de reduzir o número de nascimentos prematuros e, consequentemente, o número de mortes neonatais.

Recomenda-se, também, a realização de mais estudos, com foco nas características associadas com a prematuridade, segundo classificação do prematuro - prematuros extremos, prematuros muito pré-termo e prematuros moderados ou tardios, a fim de compreender melhor a relação das características com o problema insistente da prematuridade.

Referências

AGUIAR, A. S. C. d.; CARDOSO, M. V. L. M. L.; LUCIO, I. M. L. Teste do reflexo vermelho: forma de prevenção à cegueira na infância. *Revista Brasileira de Enfermagem*, SciELO Brasil, v. 60, n. 5, p. 541–545, 2007.

ALEXANDER, G. R.; ALLEN, M. C. Conceptualization, measurement, and use of gestational age. i. clinical and public health practice. *Journal of Perinatology*, v. 16, n. 1, p. 53–59, 1996.

ALIO, A. P.; MBAH, A. K.; GRUNSTEN, R. A.; SALIHU, H. M. Teenage pregnancy and the influence of paternal involvement on fetal outcomes. Pediatr Adolesc Gynecol., v. 24, n. 6, p. 404-409, 2011.

ALMEIDA, C. G. M.; RODRIGUES, O. M. P. R.; SALGADO, M. H. Diferenças no desenvolvimento de meninos e meninas em condições de risco. *Bol. psicol*, São Paulo , v. 62, n. 136, p. 1-14, jun. 2012.

ASSUNÇÃO, P. L.; NOVAES, H. M.; ALENCAR, G. P. Fatores associados ao nascimento pré-termo em Campina Grande, Paraíba, Brasil: um estudo caso-controle. *Cad. Saúde Pública*, v. 28, n. 6, p. 1078-1090, 2012.

AYODELE, T. O. Machine learning overview. *New Advances in Machine Learning*, v. 2, 2010.

BARROS, A.J.D; HIRAKATA, V.N. Alternatives for logistic regression in cross sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC medical research methodology*, 3(1): 13, 2003.

BASSO, C. G.; NEVES, E. T.; SILVEIRA, a. Associação entre realização de pré-natal e morbidade neonatal. *Texto Contexto Enferm*, v. 21, n. 2, p. 269-276, 2012.

BECK, S. et al. The worldwide incidence of preterm birth: a systematic review of maternal mortality and morbidity. *Bulletin of the World Health Organization*, SciELO Public Health, v. 88, p. 31–38, 2010.

BELUZO, C. E. et al. Towards neonatal mortality risk classification: a data-driven approach using neonatal, maternal, and social factors. *Informatics in Medicine Unlocked*, v. 20, 2020, 100398.

BEZERRA, L. C. et al. Prevalência e fatores associados à prematuridade entre gestantes submetidas à inibição de trabalho de parto prematuro.*Rev. Bras. Saude Mater. Infant.*, v. 6, n. 2, p. 223-229, 2006.

BITTAR, R. E.; ZUGAIB, M. Indicadores de risco para o parto prematuro. *Revista Brasileira de Ginecologia e Obstetrícia*, SciELO Brasil, v. 31, n. 4, p. 203–209, 2009.

BLENCOWE, H. et al. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The lancet, Elsevier*, v. 379, n. 9832, p. 2162–2172, 2012.

BLENCOWE, H. et al. Preterm birth–associated neurodevelopmental impairment estimates at regional and global levels for 2010. *Pediatric research*, Nature Publishing Group, v. 74, n. 1, p. 17–34, 2013.

BRANDI, L. D. A. et al. Fatores de risco materno-fetais para o nascimento pré-termo em hospital de referência de Minas Gerais. *Rev. Médica de Minas Gerai*, v. 30, n. 4, p. 41–47, 2020.

BRASIL. Ministério da Saúde. Fundação Nacional de Saúde. *Manual de procedimentos do sistema de informações sobre nascidos vivos.* - Brasília: Ministério da Saúde, 2001.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação de Saúde. *Manual de Instruções para o preenchimento da Declaração de Nascido Vivo/Ministério da Saúde*, Secretaria de Vigilância em Saúde, Departamento de Análise de Situação de Saúde.–Brasília: Ministério da Saúde, 2011.

BRASIL. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Ações Programáticas Estratégicas. *Atenção à saúde do recém-nascido: guia para os profissionais de saúde / Ministério da Saúde*, Secretaria de Atenção à Saúde, Departamento de Ações Programáticas Estratégicas.–2.ed.– Brasília: Ministério da Saúde, 2012.

BRASIL. Ministério da Saúde. Sociedade Beneficente Israelita Brasileira Albert Einstein. *NOTA TÉCNICA PARA ORGANIZAÇÃO DA REDE DE ATENÇÃO À SAÚDE COM FOCO NA ATENÇÃO PRIMÁRIA À SAÚDE E NA ATENÇÃO AMBULATORIAL ESPECIALIZADA – SAÚDE DA MULHER NA GESTAÇÃO, PARTO E PUERPÉRIO.* / Sociedade Beneficente Israelita Brasileira Albert Einstein. São Paulo: Hospital Israelita Albert Einstein: Ministério da Saúde, 2019.

BRUCE, P.; BRUCE, A. Estatística Prática para Cientistas de Dados. Rio de Janeiro: Alta Books, 2019.

BULHÕES, T. R. B. de et al. Prevalência de recém nascidos pré-termo de mães adolescentes. *Id on Line Revista de Psicologia*, v. 12, n. 39, p. 84–96, 2018.

CASCAES, A. M. et al. Prematuridade e fatores associados no estado de Santa Catarina, Brasil, no ano de 2005: análise dos dados do sistema de informações sobre nascidos vivos. *Cadernos de Saúde Pública*, SciELO Public Health, v. 24, p. 1024–1032, 2008.

CASTRO, M. P.; RUGOLO, L. M. S. S.; MARGOTTO, P. R. Sobrevida e morbidade em prematuros com menos de 32 semanas de gestação na região central do Brasil. *Revista Brasileira de Ginecologia e Obstetrícia*, SciELO Brasil, v. 34, n. 5, p. 235–242, 2012.

COUTINHO, L.; SCAZUFCA, M.; MENEZES, P. R. Métodos para estimar razão de prevalência em estudos de corte transversal. *Revista de Saúde Pública*, SciELO Brasil, v. 42, n. 6, p. 992–998, 2008.

CUMMINGS, P. Methods for estimating adjusted risk ratios. *The Stata Journal*, SAGE Publications Sage CA: Los Angeles, CA, v. 9, n. 2, p. 175–196, 2009.

EFRON, B.; TIBSHIRANI, R. J. *An introduction to the bootstrap*. CRC press, 1994.

FARAWAY, J. J. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models.* [S.l.]: CRC press, 2006.

FERRAZ, T. R.; NEVES, E. T. Fatores de risco para baixo peso ao nascer em maternidades públicas: um estudo transversal. *Revista Gaúcha de Enfermagem*, SciELO Brasil, v. 32, n. 1, p. 86–92, 2011.

FERREIRA, E. V. *Métodos de Reamostragem*. Material de apoio à disciplina de Machine Learning para Cientista de Dados, lecionada na LEG/UFPR, 2018. Disponível em: <http://cursos.leg.ufpr.br/ML4all/slides/Reamostragem.pdf>

FRANÇA, E. B. et al. Principais causas da mortalidade na infância no brasil, em 1990 e 2015: estimativas do estudo de carga global de doença. *Revista brasileira de epidemiologia*, SciELO Public Health, v. 20, p. 46–60, 2017.

GUIMARÃES, E. A. A. et al. Prevalência e fatores associados à prematuridade em Divinópolis, Minas Gerais, 2008-2011: análise do sistema de informações sobre nascidos vivos. *Epidemiologia e Serviços de Saúde*, SciELO Public Health, v. 26, p. 91–98, 2017.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, v. 2, Springer, 2009.

HOSMER, D. W.; LEMESHOW, S. *Applied logistic regression*. [S.l.]: John Wiley & Sons, New York, 2000.

ILIODROMITI S, MACKAY DF, SMITH GC, PELL JP, NELSON SM. Apgar score and the risk of cause-specific infant mortality: a population-based cohort study. *Lancet*. 2014;

JAMES, G. et al. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.

JESUS, R. L. R. Et al. Caracterização dos recém-nascidos pré-termo nascidos no estado do Piauí entre 2011 a 2015. *Arch Health Invest*, v. 8, n. 4, p. 217-223, 2019.

KUHN, M. caret: Classification and Regression Training. R package version 6.0-88. https://CRAN.R-project.org/package=caret, 2021.

LEAL, M. C. et al. Prevalence and risk factors related to preterm birth in brazil.*Reproductive health*, Springer, v. 13, n. 3, p. 163–174, 2016a.

LEAL, M. C. et al. Provider-initiated late preterm births in Brazil: differences between public and private health services. *PLOS One*, v. 11, n. 5, e0155511, 2016b.

MARÔCO, J. *Análise estatística: com utilização do SPSS* (3ª edição). [S.l.]: Edições Sílabo, 2007.

MARTINS, M. G. et al.Associação de gravidez na adolescência e prematuridade. *Rev. Bras. Ginecol. Obstet.*, v. 33, n. 11, p. 354-360, 2011.

MELLO, R. R.; MEIO, M. D. B. B. Follow-up de recém-nascidos de risco. *Editora FIOCRUZ*, p. 179–184, 2003.

MELLO JORGE, M. H. P. d. et al . Avaliação do sistema de informação sobre nascidos vivos e o uso de seus dados em epidemiologia e estatísticas de saúde. *Rev. Saúde Pública*, São Paulo, v. 27, supl. p. 1-46, 1993.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, v. 1, n. 1, p. 32, 2003.

MORAES, C. L.; REICHENHEIM, M. E. Validade do exame clínico do recém-nascido para a estimação da idade gestacional: uma comparação do escore new ballard com a data da última menstruação e ultra-sonografia.*Cadernos de Saúde Pública*, SciELO Public Health, v. 16, p. 83–94, 2000.

MWAMAKAMBA, L. W.; ZUCCHI, P. Estimativa de custo de permanência hospitalar para recém-nascidos prematuros de mães adolescentes em um hospital público brasileiro. *Einstein (São Paulo)*, SciELO Brasil, v. 12, n. 2, p. 223–229, 2014.

NOMURA, R. M. Y. A. et al. Avaliacão da maturidade fetal em gestações de alto risco: análise dos resultados de acordo com a idade gestacional. *Revista da Associação Médica Brasileira*,SciELO Brasil, v. 47, n. 4, p. 346–351, 2001.

OLIVEIRA, L. L. d. et al. Fatores maternos e neonatais relacionados à prematuridade. *Revista da Escola de Enfermagem da USP*, SciELO Brasil, v. 50, n. 3, p. 382–389, 2016.

POWERS, D. A.; XIE, Y. *Statistical Methods for Categorical Data Analysis*. Academic Press, 1999.

RADES, É.; BITTAR, R. E.; ZUGAIB, M. Determinantes diretos do parto prematuro eletivo e os resultados neonatais. *Revista Brasileira de Ginecologia e Obstetrícia*, SciELO Brasil, v. 26, n. 8, p. 655–662, 2004.

RAMOS, H. d. C.; CUMAN, R. K. N. Fatores de risco para prematuridade: pesquisa documental.*Esc Anna Nery Rev Enferm*, SciELO Brasil, v. 13, n. 2, p. 297–304, 2009.

RASCHKA, S. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv*:1811.12808, 2018

RECHIA, I. C. et al. Efeitos da prematuridade na aquisição da linguagem e na maturação auditiva: revisão sistemática. *CoDAS*. SciELO BRASIL, v. 28, n. 6, p. 843–854, 2016.

RENCHER, A. C. *Linear models in statistics*. [S.l.]: John Wiley Sons, 2008.

SILVEIRA, M. F. et al. Determinants of preterm birth: Pelotas, Rio Grande do Sul State, Brazil, 2004 birth cohort. *Cad. Saúde Pública*, Rio de Janeiro, v. 26, n. 1, p. 185-194, 2010.

SOUSA, D. S. et al. Morbidade em recém-nascidos prematuros de extremo baixo peso em unidade de terapia intensiva neonatal. *Revista Brasileira de Saúde Materno Infantil*, SciELO Brasil, v. 17, n. 1, p. 139–147, 2017.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. *Understanding Machine Learning: FromTheory to Algorithms*. USA: Cambridge University Press, 2014.

SOUZA, D. M. L. de et al. Prevalência de prematuridade e fatores associados no estado do Rio Grande do Sul.*Brazilian Journal of Health Review*, v. 2, n. 5, p. 4052–4070, 2019.

TEIXEIRA, G. A. et al. Perfil de mães e o desfecho do nascimento prematuro ou a termo.*Cogitare Enfermagem*, v. 23, n. 1, 2018.

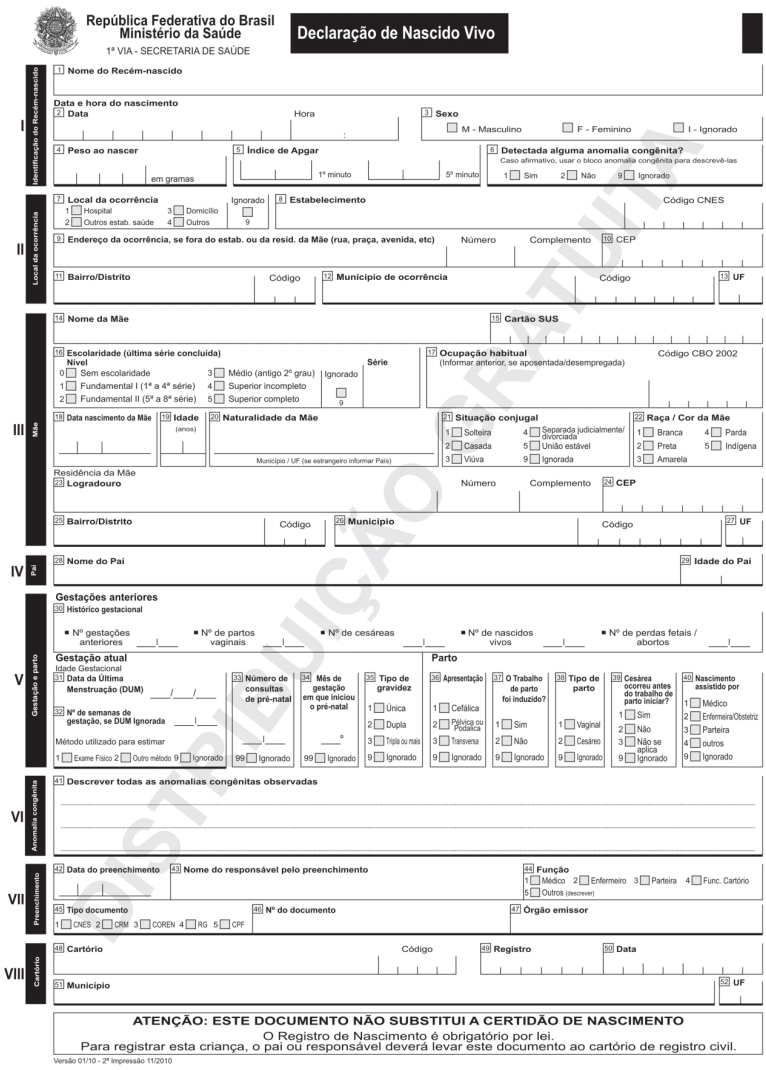
WAINSTOCK, T; SHOHAM-VARDI, I.; GLASSER, S.; ANTEBY, E.; LERNER-GEVA, L. Fetal sex modifies effects of prenatal stress exposure and adverse birth outcomes. *Stress*, v. 18, n. 1, p. 49-56, 2015.

WIEMKEN, T. L.; KELLEY, R. R. Machine Learning in Epidemiology and Health Outcomes Research. *Annual Review of Public Health*, v. 41, p. 21-36, 2020.

ZOU, G. A modified poisson regression approach to prospective studies with binary data. American journal of epidemiology, *Oxford University Press*, v. 159, n. 7, p. 702–706, 2004.

ANEXO 1 – Declaração de Nascido Vivo

ANEXO A - Modelo da Declaração de Nascido Vivo



República Federativa do Brasil
Ministério da Saúde
1ª VIA - SECRETARIA DE SAÚDE

Declaração de Nascido Vivo

I – Identificação do Recém-nascido

1 Nome do Recém-nascido

Data e hora do nascimento
2 Data — Hora :
3 Sexo: M - Masculino; F - Feminino; I - Ignorado
4 Peso ao nascer (em gramas)
5 Índice de Apgar: 1º minuto; 5º minuto
6 Detectada alguma anomalia congênita? (Caso afirmativo, usar o bloco anomalia congênita para descrevê-las): 1 Sim; 2 Não; 9 Ignorado

II – Local da ocorrência

7 Local da ocorrência: 1 Hospital; 2 Outros estab. saúde; 3 Domicílio; 4 Outros; Ignorado 9
8 Estabelecimento — Código CNES
9 Endereço da ocorrência, se fora do estab. ou da resid. da Mãe (rua, praça, avenida, etc) — Número — Complemento — 10 CEP
11 Bairro/Distrito — Código
12 Município de ocorrência — Código
13 UF

III – Mãe

14 Nome da Mãe
15 Cartão SUS
16 Escolaridade (última série concluída) — Nível: 0 Sem escolaridade; 1 Fundamental I (1ª a 4ª série); 2 Fundamental II (5ª a 8ª série); 3 Médio (antigo 2º grau); 4 Superior incompleto; 5 Superior completo; Ignorado 9; Série
17 Ocupação habitual (Informar anterior, se aposentada/desempregada) — Código CBO 2002
18 Data nascimento da Mãe
19 Idade (anos)
20 Naturalidade da Mãe — Município / UF (se estrangeiro informar País)
21 Situação conjugal: 1 Solteira; 2 Casada; 3 Viúva; 4 Separada judicialmente/ divorciada; 5 União estável; 9 Ignorada
22 Raça / Cor da Mãe: 1 Branca; 2 Preta; 3 Amarela; 4 Parda; 5 Indígena

Residência da Mãe
23 Logradouro — Número — Complemento — 24 CEP
25 Bairro/Distrito — Código
26 Município — Código
27 UF

IV – Pai

28 Nome do Pai
29 Idade do Pai

V – Gestação e parto

Gestações anteriores
30 Histórico gestacional
■ Nº gestações anteriores ___ | ___ ■ Nº de partos vaginais ___ | ___ ■ Nº de cesáreas ___ | ___ ■ Nº de nascidos vivos ___ | ___ ■ Nº de perdas fetais / abortos ___ | ___

Gestação atual
Idade Gestacional
31 Data da Última Menstruação (DUM) ___/___/___
32 Nº de semanas de gestação, se DUM Ignorada ___ | ___
Método utilizado para estimar: 1 Exame Físico; 2 Outro método; 9 Ignorado
33 Número de consultas de pré-natal: 99 Ignorado
34 Mês de gestação em que iniciou o pré-natal: 99 Ignorado

Parto
35 Tipo de gravidez: 1 Única; 2 Dupla; 3 Tripla ou mais; 9 Ignorado
36 Apresentação: 1 Cefálica; 2 Pélvica ou Podálica; 3 Transversa; 9 Ignorado
37 O Trabalho de parto foi induzido?: 1 Sim; 2 Não; 9 Ignorado
38 Tipo de parto: 1 Vaginal; 2 Cesáreo; 9 Ignorado
39 Cesárea ocorreu antes do trabalho de parto iniciar?: 1 Sim; 2 Não; 3 Não se aplica; 9 Ignorado
40 Nascimento assistido por: 1 Médico; 2 Enfermeira/Obstetriz; 3 Parteira; 4 outros; 9 Ignorado

VI – Anomalia congênita

41 Descrever todas as anomalias congênitas observadas

VII – Preenchimento

42 Data do preenchimento
43 Nome do responsável pelo preenchimento
44 Função: 1 Médico; 2 Enfermeiro; 3 Parteira; 4 Func. Cartório; 5 Outros (descrever)
45 Tipo documento: 1 CNES; 2 CRM; 3 COREN; 4 RG; 5 CPF
46 Nº do documento
47 Órgão emissor

VIII – Cartório

48 Cartório — Código
49 Registro
50 Data
51 Município
52 UF

ATENÇÃO: ESTE DOCUMENTO NÃO SUBSTITUI A CERTIDÃO DE NASCIMENTO
O Registro de Nascimento é obrigatório por lei.
Para registrar esta criança, o pai ou responsável deverá levar este documento ao cartório de registro civil.

Versão 01/10 - 2ª Impressão 11/2010