Gabriel Silva de Medeiros

Associação entre as características maternas e do recém-nascido e a macrossomia fetal no Estado da Bahia: uma análise usando aprendizado de máquina

Niterói - RJ, Brasil

18 de dezembro de 2023

Universidade Federal Fluminense

Gabriel Silva de Medeiros

Associação entre as características maternas e do recém-nascido e a macrossomia fetal no Estado da Bahia: uma análise usando aprendizado de máquina.

Trabalho de Conclusão de Curso

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

Orientador: Dr. Jose Rodrigo de Moraes

Coorientadora: Dra. Jessica Pronestino de Lima Moreira

Niterói - RJ, Brasil

18 de dezembro de 2023

Universidade Federal Fluminense

Gabriel Silva de Medeiros

Associação entre as características maternas e do recém-nascido e a macrossomia fetal no Estado da Bahia: uma análise usando aprendizado de máquina.

Monografia de Projeto Final de Graduação sob o título "Associação entre as características maternas e do recém-nascido e a macrossomia fetal no Estado da Bahia: uma análise usando aprendizado de máquina", defendida por Gabriel Silva de Medeiros e aprovada em dezembro de 2023, 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

Profa. Dra. Jessica Pronestino de Lima Moreira
Coorientadora
Faculdade de Farmácia – UFF

Profa. Dra. Patrícia Viana Guimarães Flores
Hospital Federal de Bonsucesso – HFB

Prof. Dr. Bruno Francisco Teixeira Simões
Departamento de Métodos Quantitativos – UNIRIO

Niterói, 18 de dezembro de 2023

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

M488a Medeiros, Gabriel Silva de
Associação entre as características maternas e do recém-nascido e a macrossomia fetal no Estado da Bahia: uma análise usando aprendizado de máquina. / Gabriel Silva de Medeiros. - 2023.
59 f.

Orientador: José Rodrigo de Moraes.
Coorientador: Jessica Pronestino de Lima Moreira.
Trabalho de Conclusão de Curso (graduação)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2023.

1. Modelos Log-Lineares. 2. Razão de Prevalências. 3. Sistema de Informação em Saúde. 4. Macrossomia Fetal. 5. Produção intelectual. I. Moraes, José Rodrigo de, orientador. II. Moreira, Jessica Pronestino de Lima, coorientador. III. Universidade Federal Fluminense. Instituto de Matemática e Estatística. IV. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

A macrossomia fetal é um problema de saúde pública na maioria dos países em desenvolvimento e está associada à ocorrência de complicações que podem aumentar o risco de morbidade e mortalidade da mãe e do bebê. Usando os dados do Sistema de Informações sobre Nascidos Vivos (SINASC), este trabalho teve como objetivo avaliar a associação das características maternas e dos recém-nascidos com o desfecho de macrossomia fetal no Estado da Bahia, durante o ano de 2020. Ajustando o modelo de regressão log-linear de Poisson (com variância robusta), estimou-se a prevalência de macrossomia, a partir de um conjunto de características maternas e do recém-nascido. Considerando ou não o método de redução de viés de Firth no ajuste do modelo, não se observou diferenças substanciais nas estimativas pontuais e intervalares dos parâmetros. Adotando o modelo log-linear de Poisson com o método de Firth, verificou-se que a prevalência de macrossomia foi maior entre bebês nascidos de mães com idade mais avançada (RP = 1,138; p-valor<0,001), não brancas (RP = 1,152; p-valor=0,002), que viviam sem companheiro (RP =1,057; p-valor=0,018), e entre bebês pós-termo (RP= 1,408; p-valor<0,001) e do sexo masculino (RP= 1,675; p-valor<0,001). Observou-se ainda menor prevalência de macrossomia entre bebês de mulheres com nenhuma gestação anterior (RP = 0,613; p-valor<0,001), com ensino superior completo (RP=0,717; p-valor<0,001), nascidos por parto vaginal (RP= 0,487; p-valor<0,001) e com apresentação pélvica ou transversa (RP = 0,663; p-valor<0,001). A partir destes achados, conclui-se sobre a necessidade de desenvolver ações voltadas para a prevenção da macrossomia fetal no Estado da Bahia, a fim de reduzir o risco de complicações materno-infantis. Entre estas ações pode-se citar maiores investimentos em assistência pré-natal priorizando, de modo geral, mulheres grávidas de bebês do sexo masculino, não brancas e com menores níveis socioeconômicos.

Palavras-chave: Modelos Log-Lineares. Razão de Prevalências. Sistema de Informação em Saúde. Macrossomia Fetal

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Quadros

1 Introdução

A macrossomia fetal, de acordo com o Manual Técnico de Gestação de Alto Risco do Ministério da Saúde (2012), é o termo médico utilizado para designar recém-nascidos com peso igual ou superior a 4000 gramas, independentemente da idade gestacional ao nascimento e, durante o pré-natal, deve-se suspeitar desse diagnóstico para os fetos cujo peso estimado seja igual ou maior que o percentil 90. Entretanto, outras definições são utilizadas para designar macrossomia, como por exemplo um peso maior ou igual a 4500 gramas, e, segundo Fiorelli et al. (2007), qualquer recém-nascido com o peso acima do 90º percentil é considerado grande para a sua idade gestacional.

Embora a macrossomia seja relativamente comum, pode trazer consequências significativas para a saúde da mãe e do bebê durante a gestação, parto e pós-parto. Os bebês macrossômicos têm maior risco de morte intrauterina e de complicações perinatais, tais como distocia de ombro, fratura umeral e clavicular, paralisia facial e do plexo braquial, asfixia, aspiração de mecônio, hipoglicemia e hiperbilirrubinemia neonatal, cardiomiopatia hipertrófica e uso da unidade de terapia intensiva (Amorim et al., 2009). As complicações maternas são frequentemente relacionadas à desproporção fetopélvica, e incluem trabalho de parto prolongado, parto cesáreo, hemorragia pós-parto, infecção, lacerações de partes moles de terceiro e quarto grau, eventos tromboembólicos e acidentes anestésicos (Madi et al., 2006; Amorim et al., 2009).

Além das complicações maternas e neonatais, os bebês macrossômicos também podem sofrer efeitos a longo prazo devido ao aumento dos riscos de sequelas neurológicas, obesidade, diabetes e câncer, assim como distúrbios do metabolismo de carboidratos e de lipídios (Evagelidou et al., 2006; Clausen et al., 2005).

A detecção pré-natal da macrossomia é um aspecto importante para o manejo clínico adequado durante a gestação. No entanto, a probabilidade de detectar macrossomia varia amplamente dependendo do método utilizado e do risco gestacional. A detecção pré-natal utilizando métodos ultrassonográficos varia de 15% a 79% e de 40% a 52% com predição clínica em gestações de baixo risco. Já em gestações complicadas pelo diabetes, essa chance diagnóstica atinge cerca de 60%, independentemente do método empregado (Teva et al., 2013).

No entanto, a análise do peso do feto a termo tem uma margem de erro elevada e pode resultar em um aumento indesejável das indicações de parto cesáreo (Amorim et al., 2009).

A macrossomia fetal é influenciada por diversas características, dentre as quais se destacam a multiparidade, não fumar durante a gestação, idade gestacional prolongada, idade materna avançada, sobrepeso materno e paterno, estatura elevada dos pais, etnia, estado civil, história de aborto, IMC pré-gestacional elevado, diabetes gestacional e hipertensão arterial gestacional. Essas características podem contribuir individualmente ou em conjunto para aumentar o risco de macrossomia fetal e suas complicações associadas (Yucra et al., 2022; Oliveira et al., 2008).

A utilização das técnicas de inteligência artificial, extremamente difundida nos últimos tempos, possibilita uma abordagem inovadora para fins preditivos em análise de risco. A aprendizagem de máquina tem se destacado na modelagem preditiva do desfecho de interesse, pois esses algoritmos têm o potencial de aprimorar significativamente a precisão das previsões, ao capturar relações complexas nos dados em questão (Santos et al., 2019). Como ressaltado por Brunialti (2015), a aprendizagem de máquina capacita o computador a aprender com os dados e criar um modelo que melhor descreva ou represente esses dados sob um determinado aspecto, permitindo fazer previsões ou tomar decisões.

Tendo em vista que a macrossomia fetal é um problema de saúde pública na maioria dos países em desenvolvimento e está associado a um aumento no risco de morbidade e mortalidade materna e do bebê (Adugna et al., 2020; Kerche et al., 2005), destaca-se a importância de identificar os fatores associados à ocorrência da macrossomia, por meio do emprego de modelo linear generalizado, implementado através da estratégia de aprendizagem de máquina, a fim de fomentar políticas e ações para a sua prevenção.

2 Objetivos

2.1 Objetivo Geral

O objetivo geral deste trabalho consiste em avaliar a associação entre as características maternas e dos recém-nascidos com o desfecho de macrossomia fetal no Estado da Bahia, durante o ano de 2020.

2.2 Objetivos Específicos

- Analisar as distribuições das características maternas e do recém-nascido;
- Analisar a distribuição do desfecho de macrossomia segundo as características maternas e dos recém-nascidos;
- Identificar quais são as características maternas e do recém-nascido associadas estatisticamente ao desfecho de macrossomia;
- Ajustar um modelo para predição da prevalência de macrossomia fetal, utilizando características da mãe e do recém-nascido;
- Avaliar a capacidade preditiva do modelo para o desfecho de macrossomia fetal.

3 Materiais e Métodos

3.1 Sistema de Informações de Nascidos Vivos

O Sistema de Informações de Nascidos Vivos (SINASC) é um sistema de informações do Ministério da Saúde do Brasil que tem como objetivo registrar e armazenar informações sobre os nascidos vivos no território nacional. Ele foi criado em 1990, com o objetivo de coletar e processar dados demográficos e epidemiológicos sobre o recém-nascido, a mãe, o pré-natal e o parto (Oliveira et al., 2015). Assim, o SINASC propicia condições de se identificar características do recém-nascido e da mãe, além de permitir comparações e análise temporal das informações (Farias et al., 2014).

Além disso, o SINASC é uma fonte de dados essencial para o planejamento estratégico e a gestão eficiente de políticas públicas na área da saúde, permitindo a identificação de áreas de atuação prioritárias, bem como a avaliação da efetividade das políticas implementadas (Pedraza, 2012).

As informações registradas no SINASC são baseadas nos dados contidos na Declaração de Nascido Vivo (DNV), documento oficial e padronizado para todo o país. A DNV é um documento de emissão obrigatória em hospitais e outras instituições de saúde onde ocorrem partos. Além disso, os Cartórios do Registro Civil também têm a obrigação de receber informações sobre os partos que ocorrem no ambiente domiciliar e emitir a DNV correspondente (Pedraza, 2012).

A presente pesquisa foi realizada utilizando dados coletados da base de dados TABWIN/DATASUS no Sistema de Informação de Nascidos Vivos (SISNAC), seguindo os passos: datasus.saude.gov.br → Transferência de arquivos (TABWIN) → Fonte: SINASC - Sistema de Informação de Nascidos Vivos → Modalidade: Dados → Tipo de Arquivo: DN - Declaração de Nascido Vivos 1994 a 2021. Então, foi selecionado o ano e o estado brasileiro de interesse, e em seguida marcada a opção “enviar” para efetuar o download dos arquivos.

3.2 Declaração de Nascido Vivo

A Declaração de Nascidos Vivos (DNV) é um documento emitido “por profissional de saúde responsável pelo acompanhamento da gestação, do parto ou do recém-nascido, inscrito

no Cadastro Nacional de Estabelecimentos de Saúde (CNES) ou no respectivo Conselho Profissional" (art. 3º, § 1º, Lei n.º 12.662, de 5 de junho de 2012). Esse documento é um registro oficial do nascimento da criança e contém informações importantes, como o nome da mãe, do pai, a data e hora do nascimento, o tipo de parto, o peso ao nascer, entre outras informações (Pedraza, 2012).

É importante destacar que a DNV deve ser preenchida de forma completa e correta, pois a falta desse documento pode prejudicar o acesso da criança a direitos como a certidão de nascimento, a identidade, a matrícula em escolas, entre outros (Nunes et al., 2010).

A DNV é um documento padronizado em três vias, numeradas pelo Ministério da Saúde e distribuída gratuitamente às secretarias estaduais de saúde, que as fornecem às instituições de saúde e cartórios de registro civil (Nunes et al., 2010). A primeira via, branca, é enviada aos órgãos regionais das Secretarias Estaduais de Saúde para fins estatísticos. A segunda via, amarela, é destinada à família para apresentação ao Cartório no momento do registro civil do nascimento. Já a terceira via, rosa, permanece arquivada no prontuário da gestante ou do recém-nascido (Nunes et al., 2010; Mishima et al., 1999).

A versão atualmente em uso foi atualizada em 2021 e é composta por 52 variáveis, distribuídas em oito blocos (Brasil, 2022):

Bloco I - Recém-Nascido:

- Nome do Recém-nascido;
- Data e hora do Nascimento;
- Sexo do recém-nascido;
- Raça/Cor do recém-nascido;
- Peso ao nascer (em gramas);
- Índice de Apgar;
- Detecção de alguma anomalia ou defeito congênito.

Bloco II – Local da Ocorrência:

- Local da Ocorrência: Hospital, outro estabelecimento de Saúde, Domicílio, Aldeia Indígena, Outros ou Ignorado;
- Estabelecimento de ocorrência do parto: nome do estabelecimento de saúde, número do Cadastro Nacional de Estabelecimentos de Saúde – CNES;
- Endereço completo de onde ocorreu o parto (logradouro, n°, CEP, município).

Bloco III – Mãe:

• Nome da mãe;

• Cartão SUS;

• Escolaridade;

• Ocupação Habitual;

• Data de nascimento da mãe;

• Idade da mãe;

• Naturalidade da mãe;

• Situação conjugal;

• Raça/Cor da mãe;

• Residência da mãe.

Bloco IV – Responsável Legal:

• Nome do responsável;

• Idade do responsável.

Bloco V - Gestação e Parto:

• Histórico gestacional;

• Data da última menstruação (DUM);

• Número de semanas de gestação, se a DUM for 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;

• Indução (ou não) do trabalho de parto;

• Tipo de parto;

• Cesárea ocorreu ou não antes do trabalho de parto iniciar;

• Nascimento assistido por: médico, enfermagem ou obstetriz, parteira, outro.

Bloco VI – Anomalia Congênita:

• Descrição de todas as anomalias ou defeitos congênitos se observados.

Bloco VII - Identificação do responsável pelo preenchimento:

• Data do preenchimento;

• Nome do responsável pelo preenchimento;

• Função: médico, enfermagem ou obstetriz, parteira, outros;

• Tipo documento;

• Número do documento;

• Órgão emissor.

Bloco VIII – Cartório:

• Nome do cartório, município, registro, data.

3.3 População de Estudo

A população de estudo foi composta por todos os recém-nascidos sem anomalia congênita, cujo nascimento foi realizado em ambiente hospitalar no Estado da Bahia, da região Nordeste, em 2020. No entanto, foram excluídos do presente estudo os recém-nascidos com idade gestacional inferior a 37 semanas (bebês pré-termos), com o peso inferior a 500 gramas, que não possuíam informações sobre o peso ao nascer e/ou sobre alguma característica de estudo, bem como os casos de partos gemelares e de recém-nascidos filhos de mães com idade inferior a 15 anos.

3.4 Variáveis de Estudo

O desfecho de macrossomia foi determinado utilizando-se a variável "peso ao nascer (em gramas)" incluída no formulário da DNV, no Bloco I - Recém-Nascido. O desfecho de macrossomia (Y) é uma variável binária, isto é, apresenta duas categorias. Na categoria positiva (Y=1), indica-se que o recém-nascido é macrossômico, ou seja, apresenta um peso ao nascer igual ou superior a 4000 gramas. Já na categoria negativa (Y=0), indica-se que o recém-nascido não é macrossômico, ou seja, apresenta um peso ao nascer inferior a 4000 gramas.

No Quadro 1 estão listadas as variáveis categóricas consideradas no presente estudo, que englobam as características sociodemográficas e de saúde tanto da mãe quanto do recém-nascido.

Quadro 1: Características maternas e do recém-nascido.

Variáveis de estudo	Categorias
Bloco de características sociodemográficas	
Idade da mãe	15 a 29 anos 30 a 34 anos 35 anos ou mais
Raça/Cor da mãe	Branca Não Branca
Sexo do recém-nascido	Masculino Feminino
Escolaridade da mãe	Não tem ensino superior completo Ensino superior completo
Situação conjugal da mãe	Com companheiro Sem companheiro
Bloco de características clínicas	
Número de gestações anteriores, excetuando as gestações com menos de 22 semanas	Nenhuma Uma ou mais
Número de consultas pré-natal	6 ou mais consultas pré-natal Até 5 consultas pré-natal
Tipo de parto	Vaginal Cesáreo
Idade gestacional	A termo (de 37 a 41 semanas) Pós-termo (42 semanas ou mais)
Tipo de apresentação do recém-nascido	Cefálica Pélvica ou Transversa
Macrossomia (desfecho)	Sim Não

3.5 Modelo Linear Generalizado (MLG)

O Modelo Linear Generalizado (MLG) é uma extensão dos modelos de regressão linear que permite lidar com uma ampla variedade de variáveis respostas, incluindo variáveis binárias, categóricas, contagens, proporções, entre outras. Desse modo, no MLG é possível considerar variáveis respostas com outras distribuições de probabilidade, além da distribuição Normal (Dobson e Barnett, 2018).

O Modelo Linear Generalizado pode ser caracterizado por três componentes, isto é, por um conjunto de variáveis aleatórias independentes (componente aleatória), por um preditor linear (componente sistemática) e uma função de ligação (Rencher e Schaalje, 2008; Dobson e Barnett, 2018). Desse modo, no contexto de adoção de um MLG, temos que:

- A componente aleatória é composta por n variáveis aleatórias independentes y_1, y_2, ..., y_n com valor esperado E(y_i) = μ_i, cada uma com função de densidade de probabilidade (ou função de probabilidade) pertencente à família exponencial. Supondo y_i uma variável aleatória discreta, a sua distribuição pertence à família exponencial se a sua função de probabilidade puder ser escrita da seguinte forma:

$$P(y_i; \theta_i) = exp\{a(y_i) \cdot b(\theta_i) + c(\theta_i) + d(y_i)\} \quad (3.1)$$

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

- A componente sistemática é representada pelo preditor linear $\boldsymbol{x}_i^T \boldsymbol{\beta}$ do modelo, definido por um conjunto de parâmetros e variáveis explicativas, isto é:

$$\boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} \ldots + \beta_{k-1} \cdot x_{ik-1} \quad (3.2)$$

$$i = 1, 2, \ldots, n$$

onde $\boldsymbol{\beta}$ é o vetor, de dimensão $k \times 1$, dos parâmetros do modelo e $\boldsymbol{x}_i^T$ é o vetor de variáveis explicativas (covariáveis ou variáveis do tipo dummy para os níveis das variáveis qualitativas), de dimensão $1 \times k$, referente a i-ésima unidade da amostra.

- A função de ligação $g(\mu_i)$, por sua vez, é uma função de ligação monótona e diferenciável que descreve a relação entre a média da variável resposta E(y_i) = μ_i e o preditor linear $\boldsymbol{x}_i^T \boldsymbol{\beta}$ do modelo:

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} \quad (3.3)$$

$$i = 1, 2, \ldots, n$$

3.6 Modelo de regressão log-linear de Poisson (com variância robusta)

O modelo log-linear de Poisson é um MLG usualmente utilizado para analisar dados de contagem (inteiros positivos), onde a variável de resposta é assumida ter distribuição de Poisson (Faraway, 2006).

Embora o modelo de regressão de Poisson seja apropriado para analisar variáveis de contagem (desfecho raro), quando os indivíduos são seguidos ao longo de um período de tempo variável, vem sendo frequentemente adotado como alternativa para modelar respostas binárias, a fim de obter diretamente estimativas de razão de prevalência (RP) em estudos transversais, ou de razão de riscos em estudos longitudinais (Barros e Hirakata, 2003; Coutinho et al., 2008). Ao ajustar o modelo log-linear de Poisson usando dados de estudos transversais, considera-se que o tempo de seguimento é constante. Adicionalmente, uma suposição usada na regressão de Poisson é de que a média e a variância do desfecho são iguais (Dobson e Barnett, 2018).

Quando o modelo de regressão de Poisson é aplicado a desfechos binários (ou binomiais) para estimar razões de prevalência em estudos transversais, os erros-padrão das medidas de associação são superestimados, gerando ainda valores elevados para os p-valores dos testes de significância de Wald e para os intervalos de confiança (Coutinho et al., 2008; Cummings, 2009). Esta superestimação pode ser resolvida através do ajuste do modelo log-linear de Poisson com estimação de variância robusta, isto é, usando o estimador de variância robusta, conhecido, por exemplo, como estimador Huber-White ou estimador sanduíche (Cummings, 2009; Zou, 2004). Com este estimador é possível relaxar a suposição de que os dados seguem uma distribuição de Poisson e, desta forma, obter erros-padrão, p-valores e intervalos de confiança corretos para os parâmetros de interesse (Cummings, 2009).

Além do modelo de regressão de log-linear de Poisson (com variância robusta), cabe mencionar que existem outros MLGs alternativos para a estimação direta de RP, entre as quais pode-se destacar o modelo de regressão log-binomial. Ambos os modelos produzem estimativas corretas de RP e são reportados como melhores alternativas para a análise de desfechos binários em estudos transversais, comparativamente ao modelo logístico binário (Barros e Hirakata, 2003). Bhaskar e Ponnuraja (2021), concluíram que o modelo de Poisson com variância robusta e o modelo log-binomial são os mais adequados para a modelagem de desfechos binários também em estudos clínicos. Apesar do modelo de regressão log-binomial produzir estimativas não viciadas para as medidas de associação, pode ocorrer problema de falta de convergência no processo de estimação dos coeficientes, isto é, o modelo não consegue encontrar uma solução e, portanto, as medidas de associação não podem ser calculadas (Coutinho et al., 2008).

O modelo de regressão log-linear de Poisson com variância robusta, adotado no presente trabalho, foi ajustado utilizando o comando glm do Programa R, mas com o uso dos pacotes "*sandwich*" (Zeileis et al., 2020) e "*lmtest*" (Zeileis e Hothorn, 2002) para a inferência robusta.

3.6.1 Especificação do modelo

O modelo log-linear de Poisson é um MLG, onde a distribuição de probabilidade da variável resposta y_i é a distribuição de Poisson. A função de ligação $g(\mu_i)$ é a função logarítmica (Faraway, 2006). A função de probabilidade da variável aleatória de Poisson é dada por:

$$P(y_i;\mu_i) = \frac{e^{-\mu_i} \cdot {\mu_i}^{y_i}}{y_i!} \; ; \; y_i = 0,1,2,\dots \; ; \; i = 1,2,\dots,n \tag{3.4}$$

A função de probabilidade da variável aleatória de Poisson pode ser expressa na forma da família exponencial.

$$P(y_i;\mu_i) = \exp\{y_i \cdot ln\mu_i - \mu_i - \ln(y_i!)\} \tag{3.5}$$

onde: $a(y_i) = y_i$, $b(\mu_i) = ln\mu_i$, $c(\mu_i) = -\mu_i$, $d(y_i) = -\ln(y_i!)$

Quanto a especificação do modelo, pode-se representar o modelo log-linear de Poisson pela seguinte equação:

$$\ln(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$$
$$i = 1,2,\dots,n$$

Ou alternativamente,

$$\mu_i = exp\{\boldsymbol{x}_i^T\boldsymbol{\beta}\} \tag{3.6}$$

onde:

$ln(\mu_i)$ é o logaritmo neperiano da média da variável resposta y_i referente a i-ésima unidade.

$\mu_i = \mathrm{E}(y_i)$ é a média da variável resposta y_i referente a i-ésima unidade.

$\boldsymbol{x}_i^T$ é o vetor de variáveis explicativas, de dimensão $1 \times k$, referente a i-ésima unidade.

$\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos do modelo de dimensão k x 1.

3.6.2 Método de máxima verossimilhança

O método da máxima verossimilhança (MV) é amplamente utilizado para obter estimativas pontuais e intervalares dos parâmetros em modelos lineares generalizados (Dobson e Barnett, 2018). Considerando que as observações amostrais são independentes e identicamente distribuídas, a estimação por MV busca encontrar os valores dos parâmetros que maximizam a função de verossimilhança da amostra. Em outras palavras, procura-se determinar os valores dos parâmetros que tornam a observação dos dados mais verossímil (Powers e Xie, 1999).

Seja $y_1, y_2, \dots, y_n$ uma amostra aleatória de observações independentes das varáveis aleatórias $Y_1, Y_2, \dots, Y_n$, todas com distribuição de Poisson com parâmetro μ_i. A função de probabilidade de Y_i pode ser expressa por:

$$P(y_i; \mu_i) = \frac{e^{-\mu_i} \cdot {\mu_i}^{y_i}}{y_i!} \; ; \; y_i = 1, 2, \dots \; ; \; i = 1, 2, \dots, n \tag{3.7}$$

A função de verossimilhança da amostra é dada por:

$$L(\boldsymbol{\beta}, \boldsymbol{y}) = \prod_{i=1}^{n} P(y_i; \mu_i) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \cdot {\mu_i}^{y_i}}{y_i!} \tag{3.8}$$

Para encontrar o conjunto de valores dos parâmetros que maximizam a função de verossimilhança, $L(\boldsymbol{\beta}, \boldsymbol{y})$, é mais conveniente maximizar o logaritmo da função verossimilhança, $\ln(L(\boldsymbol{\beta}, \boldsymbol{y}))$. Como o logaritmo é uma transformação estritamente monotônica, os valores que maximizam $L(\boldsymbol{\beta}, \boldsymbol{y})$ também maximizarão o $\ln(L(\boldsymbol{\beta}, \boldsymbol{y}))$, o qual pode ser expresso por:

$$\ln\left(L(\boldsymbol{\beta}, \boldsymbol{y})\right) = l\ln\left(\prod_{i=1}^{n} P(y_i; \mu_i)\right) = l\ln\left(\prod_{i=1}^{n} \frac{e^{-\mu_i} \cdot {\mu_i}^{y_i}}{y_i!}\right)$$
$$= \sum_{i=1}^{n} (-\mu_i + y_i \cdot \ln(\mu_i) - \ln(y_i!))$$

Substituindo μ_i por $e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}$, obtém-se:

$$\ln\left(L(\boldsymbol{\beta},\boldsymbol{y})\right) = \sum_{i=1}^{n}\left(-e^{\boldsymbol{x}_i^T\boldsymbol{\beta}} + y_i \cdot \boldsymbol{x}_i^T\boldsymbol{\beta} - \ln\left(y_i!\right)\right)$$

Para maximizar a log-verossimilhança $\ln\left(L(\boldsymbol{\beta},\boldsymbol{y})\right)$, basta derivá-la em relação a $\boldsymbol{\beta}$, obtendo a função escore $\boldsymbol{U}(\boldsymbol{\beta})$. Ao igualar $\boldsymbol{U}(\boldsymbol{\beta})$ a $\mathbf{0}$, obtém-se um sistema de equações:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial\boldsymbol{\beta}}\ln\left(L(\boldsymbol{\beta},\boldsymbol{y})\right) = \sum_{i=1}^{n}\boldsymbol{x}_i^T\left(y_i - e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right) = \mathbf{0} \qquad (3.9)$$

A solução do sistema de equações é obtida por meio de métodos numéricos iterativos, como o método Escore de Fisher (Dobson e Barnett, 2018). Esta solução é o estimador $\widehat{\boldsymbol{\beta}}_{MV}$ de máxima verossimilhança do vetor $\boldsymbol{\beta}$. Desse modo, $\widehat{\boldsymbol{\beta}}_{MV}$ é um vetor de dimensão k x 1, que contêm os estimadores dos parâmetros do modelo log-linear de Poisson.

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

Para grandes amostras, a distribuição do estimador de máxima verossimilhança $\widehat{\boldsymbol{\beta}}_{MV}$ se aproxima assintoticamente de uma distribuição normal multivariada (Powers e Xie, 1999), com média $E\left(\widehat{\boldsymbol{\beta}}_{MV}\right) = \boldsymbol{\beta}$ e matriz estimada de variância e covariância dos estimadores dos parâmetros, dada por:

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

sendo $\hat{\mathbf{I}}\left(\widehat{\boldsymbol{\beta}}_{MV}\right) = \frac{\partial^2}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^T}\ln\left(L(\boldsymbol{\beta},\boldsymbol{y})\right)$, uma matriz simétrica de dimensão k x k.

3.6.3 Método de máxima verossimilhança penalizada

Como descrito na seção 3.6.2, o método de MV é o método usual para a estimação dos parâmetros de modelos lineares generalizados (Kosmidis et al., 2020); inclusive do modelo log-linear de Poisson adotado neste estudo. Todavia, quando as classes do desfecho são

desbalanceadas (desfecho raro), as estimativas usuais de MV são viciadas (Firth, 1993; Kosmidis e Firth, 2021; Rahman e Sultana, 2017).

Segundo Firth (1993), o vício das estimativas de MV pode ser reduzido por meio da modificação da função escore $\boldsymbol{U}(\boldsymbol{\beta})$ definida na equação (3.9). Para tanto, para a estimação dos parâmetros do modelo adiciona-se um termo de penalização na função de verossimilhança, conhecido como a priori de Jeffreys (Kosmidis e Firth, 2021). Neste caso, os parâmetros do modelo são estimados por máxima verossimilhança penalizada (MVP), usando o método de redução de viés de Firth, cuja função de verossimilhança penalizada é dada por:

$$L^*(\boldsymbol{\beta}, \boldsymbol{y}) = L(\boldsymbol{\beta}, \boldsymbol{y}) \cdot |\mathbf{I}(\boldsymbol{\beta})|^{\frac{1}{2}},$$

onde $L(\boldsymbol{\beta}, \boldsymbol{y})$ é a função de verossimilhança definida na equação (3.8) e $|\mathbf{I}(\boldsymbol{\beta})|^{\frac{1}{2}}$ é o termo de penalização da função de verossimilhança conhecido como a priori de Jeffreys, e onde $|\mathbf{I}(\boldsymbol{\beta})| = \det[\mathbf{I}(\boldsymbol{\beta})]$ é o determinante da matriz de informação de Fisher.

O logaritmo da função de verossimilhança penalizada, por sua vez, é determinado do seguinte modo (Firth, 1993; Joshi et al., 2022):

$$\ln L^*(\boldsymbol{\beta}, \boldsymbol{y}) = \ln L(\boldsymbol{\beta}, \boldsymbol{y}) + \frac{1}{2} \cdot \ln|\mathbf{I}(\boldsymbol{\beta})| \tag{3.10}$$

A partir da equação (3.10), obtém-se a função de escore modificada (Firth,1993), denotada por $\boldsymbol{U}^*(\boldsymbol{\beta})$, como mostrado a seguir:

$$\boldsymbol{U}^*(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \ln L^*(\boldsymbol{\beta}, \boldsymbol{y}) = \boldsymbol{U}(\boldsymbol{\beta}) + \frac{1}{2} \cdot traço \left[\mathbf{I}^{-1}(\boldsymbol{\beta}) \cdot \left(\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{I}(\boldsymbol{\beta}) \right) \right]$$

Fazendo $\boldsymbol{U}^*(\boldsymbol{\beta}) = \mathbf{0}$, obtêm-se um sistema de equações, cuja solução é obtida pelo método iterativo de Escore de Fisher. Partindo de estimativas iniciais $\widehat{\boldsymbol{\beta}}^{(\mathbf{0})}$ para os parâmetros do modelo, as estimativas são atualizadas em cada iteração até que as diferenças entre as estimativas de uma iteração e a seguinte sejam desprezíveis. Na r-ésima iteração, as estimativas de MVP são calculadas do seguinte modo:

$$\widehat{\boldsymbol{\beta}}^{(r+1)} = \widehat{\boldsymbol{\beta}}^{(r)} + \left[\mathbf{I}\left(\widehat{\boldsymbol{\beta}}^{(r)}\right)\right]^{-\mathbf{1}} \cdot \boldsymbol{U}^*\left(\widehat{\boldsymbol{\beta}}^{(r)}\right)$$

No presente trabalho, todas as análises estatísticas foram desenvolvidas usando o programa R (R Core Team, 2022). O modelo de regressão log-linear de Poisson (com variância robusta), usando o método de redução de viés de Firth, foi ajustado com o comando glm especificando o método "brglmFit" e o tipo "MPL_Jeffreys", após a instalação do pacote “brglm2” (Kosmidis, 2023). O pacote brglm2, por ser mais geral que o pacote brglm (Kosmidis, 2021), possibilita a estimação e inferência para diferentes tipos de modelos lineares generalizados empregando diferentes métodos de redução de viés, inclusive o método de Firth, que utiliza a priori de Jeffreys.

3.6.4 Razão de prevalência

A razão de prevalência (RP) é uma medida de associação utilizada em estudos de corte transversal para avaliar a relação entre variáveis explicativas e um desfecho binário. Ela possibilita determinar se a variável de exposição está associada a maior ou menor prevalência do desfecho (Coutinho et.al, 2008).

No contexto do modelo de regressão log-linear de Poisson, pode-se obter a prevalência (p_i) do evento de interesse (Y = 1) para a i-ésima unidade da amostra, aplicando o exponencial do preditor linear.

$$p_i = e^{x_i^T \boldsymbol{\beta}} ; \ i = 1, 2, \dots, n$$

A razão de prevalência (RP) indica a prevalência do evento de interesse em um grupo, em comparação com outro grupo de referência, quando a variável explicativa é categórica. Considerando x_{ij} como uma variável explicativa binária, a razão de prevalência pode ser matematicamente calculada pela divisão entre a prevalência do evento de interesse entre os expostos (x_{ij} = 1) e a referida prevalência entre os não expostos (x_{ij} = 0, sendo x_{ij} = 0 a categoria de referência). Portanto, pode-se expressar a razão de prevalência da seguinte maneira:

$$RP_j = \frac{p_i(x_{ij} = 1)}{p_i(x_{ij} = 0)}$$

$$RP_j = \frac{e^{\beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_j(1) + \cdots + \beta_{k-1} \cdot x_{i,k-1}}}{e^{\beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_j(0) + \cdots + \beta_{k-1} \cdot x_{i,k-1}}}$$

$$RP_j = \frac{e^{\beta_j(1)}}{e^{\beta_j(0)}} = e^{\beta_j(1)-\beta_j(0)} = e^{\beta_j}$$

Do exposto, a razão de prevalência estimada é dada por:

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

Se $\hat{\beta}_j > 0$, isso implica que $\widehat{RP_j} > 1$, indicando que a prevalência do evento de interesse na categoria j é $(RP_j - 1) \cdot 100\%$ maior do que a prevalência do evento de interesse na categoria de referência. Por outro lado, se $\hat{\beta}_j < 0$, então $\widehat{RP_j} < 1$, isso sugere que a prevalência do evento de interesse na categoria j é $\left|(RP_j - 1)\right| \cdot 100\%$ menor do que na categoria de referência. E no caso de $\hat{\beta}_j = 0$ e, consequentemente, $\widehat{RP_j} = 1$, as prevalências do evento de interesse nas duas categorias da variável x_j são iguais.

3.6.5 Testes de significância dos parâmetros

O teste de significância dos parâmetros é um procedimento de inferência estatística utilizado para avaliar se os coeficientes em um modelo de regressão são estatisticamente diferentes de zero. Esses coeficientes representam os efeitos das variáveis explicativas sobre o desfecho em estudo.

3.6.6 Teste de Wald significância individual

O teste de Wald de significância individual é utilizado para avaliar a significância dos parâmetros do modelo, considerando um determinado nível de significância $100 \cdot \alpha\%$. Esse teste estatístico permite determinar a significância de cada parâmetro do modelo, isto é, avalia se o efeito de cada nível da variável explicativa sobre o desfecho em estudo é estatisticamente diferente de zero ou não.

- Hipóteses:

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

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

- Estatística teste:

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

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

onde, $\hat{\beta}_j$ é o estimador de MV do parâmetro β_j e $\widehat{VAR}(\hat{\beta}_j)$ é o estimador da variância de $\hat{\beta}_j$.

- Região Crítica:

$$RC = \left\{ z \in \mathbb{R} \mid |z| > z_{1-\frac{\alpha}{2}} \right\}$$

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

- Tomada de decisão:

A um nível de significância de $100 \cdot \alpha\%$, rejeita-se a hipótese nula (H_0) quando o valor observado (z_{obs}) da estatística de teste Z pertence a região crítica (RC); nesse caso, conclui-se que o j-ésimo nível da variável explicativa é estatisticamente diferente de zero. Por outro lado, se o valor observado (z_{obs}) não estiver na RC, não há evidências para rejeitar H_0 ao nível de significância de $100 \cdot \alpha\%$, o que indica que o j-ésimo nível da variável explicativa não é estatisticamente diferente de zero.

O p-valor do teste de Wald pode ser utilizado também como critério para tomar decisões estatísticas. No caso do teste de Wald de significância individual, rejeita-se a hipótese nula (H_0) se o p-valor, calculado como $2 \cdot P(Z > |z_{obs}|)$, for menor ou igual ao nível de significância α. Por outro lado, caso o p-valor > α, não há evidências suficientes para rejeitar H_0 ao nível de significância de $100 \cdot \alpha\%$.

3.6.7 Teste de Wald de significância geral

O teste de significância geral de Wald é uma abordagem mais ampla que permite testar se múltiplos parâmetros são estatisticamente iguais a zero, considerando o nível de significância de $100 \cdot \alpha\%$. Esse teste é especialmente útil para avaliar se os efeitos de variáveis categóricas com mais de dois níveis são estatisticamente significativos ou não (Powers e Xie, 1999).

Seja $\boldsymbol{\beta}_r$ um subvetor de dimensão $r \times 1$, do vetor de parâmetros $\boldsymbol{\beta}$. Para testar se $\boldsymbol{\beta}_r$ é significativamente diferente de zero, pode-se utilizar as seguintes etapas:

- Hipóteses:

$$\begin{cases} H_0 : \boldsymbol{\beta}_r = 0 \\ H_1 : \boldsymbol{\beta}_r \neq 0 \end{cases}$$

- Estatística teste:

Sob a hipótese nula ($H_0 : \boldsymbol{\beta}_r = 0$), a estatística de teste é dada por:

$$W = \widehat{\boldsymbol{\beta}}_{\boldsymbol{r}}^{\boldsymbol{T}} \boldsymbol{V}_{\boldsymbol{d}}^{-\mathbf{1}} \widehat{\boldsymbol{\beta}}_{\boldsymbol{r}} \sim \chi_{\boldsymbol{r}}^{\mathbf{2}}$$

onde, W é a estatística de Wald que segue uma distribuição Qui-quadrada χ^2 com r graus de liberdades, sendo que r é a dimensão de $\boldsymbol{\beta}_r$, e $\boldsymbol{V}_{\boldsymbol{d}} = \widehat{\boldsymbol{VAR}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{r}})$ uma submatriz da matriz de variância-covariância estimada de $\widehat{\boldsymbol{\beta}}$.

- Região Crítica:

$$RC = \left\{ w \in \mathbb{R} \mid w > \chi_{\mathbf{1}-\alpha;r}^{\mathbf{2}} \right\}$$

onde, $\chi_{\mathbf{1}-\alpha;r}^{\mathbf{2}}$ é o valor crítico da distribuição Qui-quadrado $\chi_{\boldsymbol{r}}^{\mathbf{2}}$ no percentil $(1-\alpha)$.

- Tomada de decisão:

Ao nível de significância de $100 \cdot \alpha\%$, rejeita-se a hipótese nula (H_0) quando o valor observado (w_{obs}) pertence a região crítica (RC); nesse caso conclui-se que $\boldsymbol{\beta}_r$ é estatisticamente diferente de zero. Por outro lado, se o valor observado (w_{obs}) não estiver na RC, não há evidências para rejeitar H_0 ao nível de significância de $100 \cdot \alpha\%$, o que indica que $\boldsymbol{\beta}_r$ não é estatisticamente diferente de zero.

No caso do teste de Wald de significância geral, rejeita-se a hipótese nula (H_0) se o p-valor, calculado como $2 \cdot P(W > |w_{obs}|)$, for menor ou igual ao nível de significância α. Por outro lado, caso o p-valor > α, não há evidências estatísticas suficientes para rejeitar H_0 ao nível de significância de $100 \cdot \alpha\%$.

Os p-valores do teste de Wald são fornecidos nas saídas de softwares estatísticos.

3.6.8 Medidas de qualidade do ajuste de modelos

3.6.8.1 Matriz de confusão (Confusion Matrix)

A matriz de confusão é uma ferramenta utilizada para avaliar a capacidade preditiva do modelo. Tem a função de organizar os resultados preditivos do modelo em uma tabela que contabiliza as classificações corretas e incorretas em relação aos rótulos reais dos dados (Quadro 2).

A partir do modelo log-linear de Poisson ajustado, pode-se calcular as probabilidades estimadas de sucesso para cada um dos n elementos da amostra. Definindo um ponto de corte δ, tal que $0 \leq \delta \leq 1$, o elemento da amostra pode ser classificado como "sucesso" ($\hat{Y}_i = 1$) se a probabilidade estimada $\hat{p}_i$ for maior que o ponto de corte δ ($\hat{p}_i > \delta$). Caso contrário, se a probabilidade estimada for menor ou igual ao ponto de corte ($\hat{p}_i \leq \delta$), então o elemento amostral é classificado como "fracasso" ($\hat{Y}_i = 0$).

Quadro 2: Classificação dos elementos segundo as categorias observadas e preditas da variável resposta Y.

Categoria Observadas	Categorias Preditas		Total
	Sucesso ($\hat{Y}_i = 1$)	Fracasso ($\hat{Y}_i = 0$)	
Sucesso ($Y_i = 1$)	VP (Verdadeiro Positivo)	FN (Falso Negativo)	VP + FN
Fracasso ($Y_i = 0$)	FP (Falso Positivo)	VN (Verdadeiro Negativo)	FP + VN
Total	VP + FP	FN + VN	n = VP + FP + FN + VN

A partir da matriz de confusão, é possível avaliar a qualidade do ajuste do modelo de classificação, permitindo a análise de métricas como acurácia, sensibilidade, especificidade, entre outras.

3.6.8.2 Acurácia

A acurácia é uma métrica que mede a taxa de acertos do modelo, representando a proporção das classificações feitas corretamente pelo modelo em relação ao total de elementos (tamanho da amostra).

$$A = P(\hat{Y} = 0, Y = 0) + P(\hat{Y} = 1, Y = 1) = \frac{VP + VN}{\text{VP} + \text{FP} + \text{FN} + \text{VN}}$$

3.6.8.3 Sensibilidade

A sensibilidade (ou taxa de verdadeiros positivos), é uma métrica que avalia a capacidade do modelo em prever corretamente os elementos que possuem a característica de interesse. Ela mede a proporção de elementos corretamente classificados como possuindo a característica, em relação ao total de elementos que realmente possuem a característica de interesse ($Y = 1$).

$$S = P(\hat{Y} = 1 | Y = 1) = \frac{VP}{VP + FN}$$

3.6.8.4 Especificidade

A especificidade (ou taxa de verdadeiros negativos) é uma métrica que avalia a capacidade do modelo em prever corretamente os elementos que não possuem a característica de interesse. Ela mede a proporção de elementos corretamente classificados como não possuindo a característica, em relação ao total de elementos que de fato não possuem a característica de interesse ($Y = 0$).

$$E = P(\hat{Y} = 0 | Y = 0) = \frac{VN}{FP + VN}$$

3.6.8.5 Curva ROC

A Curva ROC (*Receiver Operating Characteristic*) é uma representação gráfica utilizada para avaliar e comparar o desempenho de modelos, mostrando a relação entre a taxa de verdadeiros positivos (sensibilidade) e a taxa de falsos positivos (1 - especificidade) para diferentes pontos de corte (Polo et al., 2020).

A construção da Curva ROC envolve variar o limiar de classificação e calcular a sensibilidade e a especificidade correspondentes para cada ponto de corte, onde cada ponto na curva representa um equilíbrio entre a sensibilidade e a especificidade. A curva ROC pode ser utilizada para auxiliar na identificação do melhor ponto de corte (Polo et al., 2020).

A partir da Curva ROC, outra forma de avaliar o desempenho de um modelo é utilizando a métrica AUC (*Area Under Curve*), que é a área sob a curva ROC. A AUC é uma medida que 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 AUC, melhor o desempenho do modelo em distinguir entre as classes do desfecho. Segundo Dinov (2018), um valor de AUC maior ou igual a 0,7 já é indicativo de que o modelo tem um desempenho aceitável/razoável.

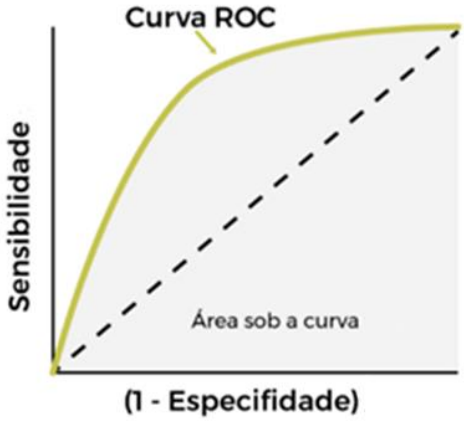


Figura 1: Esboço da curva ROC

3.7 Aprendizado de Máquinas

Aprendizado de Máquina (*Machine Learning*) é uma área da Inteligência Artificial (*IA*) cujo objetivo principal é desenvolver técnicas computacionais para o aprendizado, permitindo a construção de sistemas capazes de adquirir conhecimento de forma automática (Monard e Baranaukas, 2003). No Aprendizado de Máquina Supervisionado, um algoritmo de aprendizado é alimentado com um conjunto de dados de treinamento, no qual os objetos estão associados a rótulos conhecidos, onde cada caso é descrito por vetores de valores de atributos e um rótulo de classe (Von Luxburg e Schölkopf, 2008). O objetivo do algoritmo de aprendizado é determinar corretamente a classe de novos exemplos ainda não rotulados. Esse problema é chamado de classificação, para rótulos de classe discretos/categóricos, ou regressão, quando se trata de valores contínuos (Monard e Baranaukas, 2003; Lee e Monard, 2003). Já no aprendizado de máquinas não-supervisionado os dados de treinamento consistem apenas em vetores de valores de atributos, não havendo rótulo de classe. Nesse caso, a ênfase está em descobrir padrões ou estruturas nos dados sem ter uma resposta específica pré-definida (Von Luxburg e Schölkopf, 2008; Monard e Baranaukas, 2003).

No presente estudo, como o objetivo é prever os valores de uma variável resposta (output) binária, que representa uma característica qualitativa observada, a partir de um conjunto de variáveis explicativas (inputs), utilizou-se o modelo supervisionado. Através dos dados de treinamento, nos quais têm-se informações tanto sobre os valores da variável de saída quanto sobre os valores das variáveis de entrada para as unidades amostrais, busca-se construir um modelo de previsão capaz de predizer o valor da variável resposta (output) para um novo conjunto de unidades não observadas.

3.8 Validação-Cruzada (*Cross-Validation*)

A validação cruzada é uma abordagem estatística para validar métodos de previsão, modelos de classificação e técnicas de agrupamento. Ela avalia a confiabilidade e estabilidade dos resultados das análises estatísticas correspondentes (por exemplo, previsões, classificações, prognósticos) com base em conjuntos de dados independentes. Para prever tendências, associações, agrupamentos e classificações, um modelo geralmente é treinado em um conjunto de dados (dados de treinamento) e posteriormente testado com base em um novo conjunto de dados (dados de teste ou validação) (Dinov, 2018).

Para selecionar o melhor modelo e avaliar sua capacidade preditiva em novos dados, é comum realizar uma divisão aleatória do conjunto de dados em duas ou três partes. De acordo com Dinov (2018) e Hastie et al. (2009), em situações com dados abundantes, a abordagem recomendada é dividir o conjunto de dados em três partes: treinamento, validação e teste.

O conjunto de treinamento é usado para ajustar os modelos, o conjunto de validação é utilizado para estimar o erro de previsão e selecionar o melhor modelo, enquanto o conjunto de teste é utilizado para avaliar o erro de generalização do modelo final escolhido. Essa divisão pode seguir uma proporção de 50% para treinamento, 25% para validação e 25% para teste, mas outras proporções podem ser escolhidas a critério do pesquisador, que deve levar em conta complexidade dos modelos ajustados (Hastie et al., 2009). Por outro lado, em algumas situações, a abordagem de duas partes também pode ser adotada, com um conjunto para treinamento e outro para teste. Nessa abordagem, o conjunto de treinamento é usado para ajustar o modelo, e o conjunto de teste é utilizado para a seleção do melhor modelo com base nos erros estimados de previsão, onde o modelo com os menores erros é escolhido. Essa divisão pode seguir, por exemplo, a proporção de 80% para treinamento e 20% para teste (Figura 2). Para fazer a divisão aleatória da amostra em treino e teste utilizou-se a função createDataPartition do pacote Caret do programa R (Kuhn, 2008).

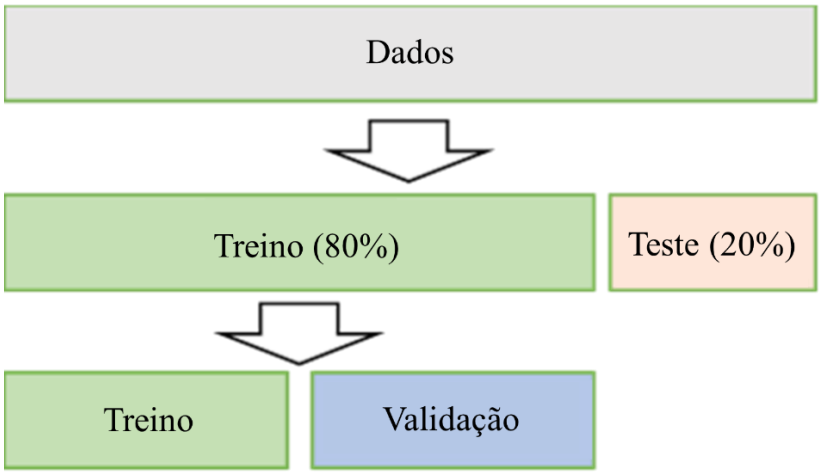


Figura 2: Representação da separação do conjunto de dados

Fonte: Adaptado do autor FERREIRA (2018)

3.8.1 Método de reamostragem k-fold

O método de k-fold é uma técnica de validação cruzada que envolve a divisão da base de dados em k partes iguais, gerando assim novas amostras (Figura 3). Esse processo é repetido k vezes, em cada iteração, uma das partes é utilizada como amostra de treino, enquanto as outras k-1 partes são utilizadas como amostras de validação (Hastie et al., 2009). Para avaliar o erro de predição, é calculada a média de todos os erros nas diferentes repetições das amostras de validação.

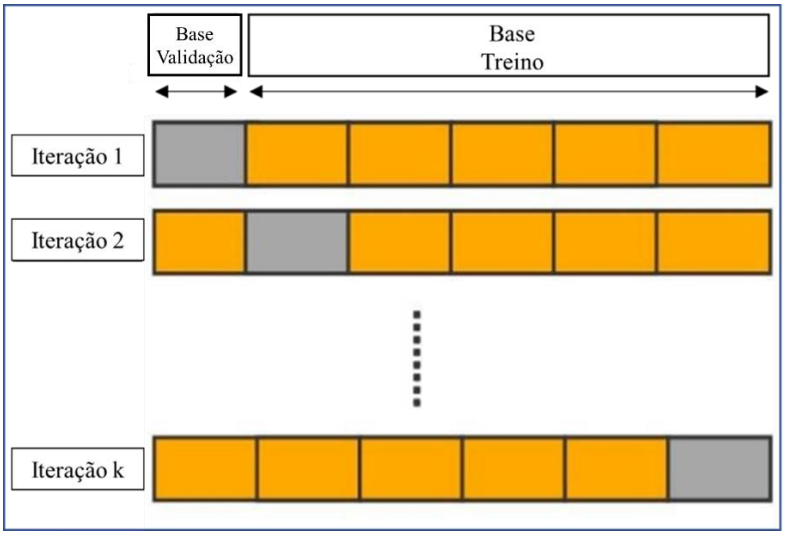


Figura 3:Representação do método de k-fold

Fonte: Adaptado do autor Alaoui (2018)

Quanto maior o valor escolhido para k, menor será o viés do modelo, porém a variância será maior. Isso resulta em uma estimativa mais precisa entre os valores observados e os previstos, mas com uma maior variabilidade. Por outro lado, ao usar um valor pequeno de k, o viés será maior e a variância será menor. Isso leva a estimativas menos precisas para a variável resposta, porém com uma menor variabilidade (Kohavi, 1995).

Cabe salientar que no presente estudo, para prever o desfecho de interesse, usando abordagem de aprendizado de máquina supervisionado (modelo log-linear de Poisson), a base de dados foi inicialmente dividida, de modo aleatório, em duas partes distintas. Uma parte foi dedicada ao ajuste de modelos (*amostra treino*) e a outra parte foi destinada para avaliar a performance do modelo (*amostra teste*). No conjunto de treinamento, representando 80% da amostra completa, foram aplicadas rotinas de validação cruzada k-fold, considerando k = 10, uma escolha baseada no estudo de Kohavi (2001). Isso levou à divisão deste conjunto de treinamento original em 10 partes (folds) mutuamente exclusivas de tamanhos iguais, em que 9/10 dessas partes foram empregadas para ajustar o modelo (treinamento), e os 1/10 restantes foram destinados à validação do modelo, o que resultou na criação da curva ROC. De acordo com Santos (2019), a área sob a Curva ROC (AUC) é uma métrica útil para a comparação das performances de modelos com diferentes preditores. Portanto, o modelo com maior valor médio de AUC é selecionado como tendo melhor capacidade preditiva.

A princípio, foram ajustados três modelos log-linear de Poisson, com diferentes blocos de preditores, objetivando determinar qual tipo de variável (bloco ou dimensão) proporcionaria um ajuste mais apropriado aos dados observados. O primeiro modelo é o modelo com o bloco sociodemográfico e clínico, isto é, que incorpora todas as variáveis de estudo. O segundo modelo, por sua vez, inclui o bloco de características clínicas, como idade gestacional, tipo de parto, número de consultas pré-natais, número de gestações anteriores e o tipo de apresentação do recém-nascido. O terceiro modelo é o que inclui apenas o bloco de variáveis sociodemográficas, tais como idade, escolaridade, situação conjugal e raça da mãe e o sexo do recém-nascido.

Após a seleção do melhor modelo (isto é, do melhor bloco de variáveis), através do processo de validação cruzada 10-fold, o modelo log-linear de Poisson (com variância robusta), foi ajustado utilizando a amostra de treinamento original. O modelo com as variáveis selecionadas foi então aplicado na amostra teste para avaliar a sua capacidade preditiva.

Por fim, o modelo log-linear de Poisson (com variância robusta), com o mesmo bloco selecionado de variáveis explicativas, foi ajustado na amostra de treinamento original usando o método de redução de viés de Firth, a fim de levar em consideração a baixa prevalência do desfecho (desbalanceamento dos dados).

4 Análise dos Resultados

A amostra total referente a população de estudo no banco do SINASC foi de 157.498 recém-nascidos. Após a exclusão de dados faltantes (missings), a amostra ficou composta por 141.515 recém-nascidos, dos quais 132.699 (93,8%) não apresentam macrossomia, enquanto apenas 8.816 (6,2%) apresentam essa condição clínica. Como no Estado da Bahia, a prevalência de macrossomia é baixa (desfecho raro), os dados são considerados desbalanceados.

A Figura 4 apresenta a distribuição percentual de recém-nascidos na amostra completa segundo a presença ou não de macrossomia.

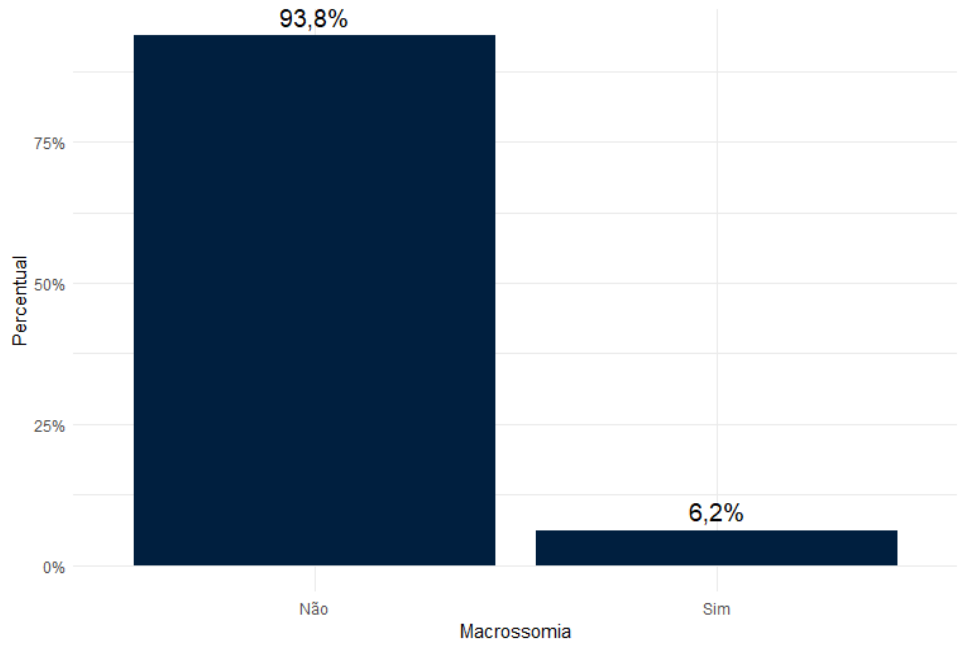


Figura 4: Distribuição percentual de recém-nascidos na amostra completa segundo a presença ou não de macrossomia.

A Tabela 1 apresenta a distribuição percentual dos recém-nascidos com ou sem macrossomia segundo as variáveis de estudo.

Ao analisar a Tabela 1, pode-se afirmar que população de estudo há predominância de recém-nascidos a termo, nascidos por parto vaginal, do sexo masculino e com apresentação cefálica. Além disso, a maioria dos recém-nascidos é proveniente de mães com as idades de 15 a 29 anos, de raça não branca, com educação até o ensino superior incompleto, vivendo sem companheiros, com uma ou mais gestações anteriores e que fizeram seis ou mais consultas pré-natais.

Ao comparar o percentual de macrossomia entre os grupos definidos por cada variável de estudo, é evidente que a macrossomia é mais predominante entre os recém-nascidos que são pós-termos (8,9%), do sexo masculino (7,8%), com apresentação cefálica (6,3%) e que nasceram de parto vaginal (8,4%). A ocorrência de macrossomia também foi mais frequente em mães com idade superior a 35 anos (7,5%), de raça não branca (6,3%), com educação até o ensino superior incompleto (6,4%), com uma ou mais gestações anteriores (7,2%) e entre as mães que realizaram seis ou mais consultas pré-natais (6,2%).

Tabela 1: Distribuição dos recém-nascidos na amostra completa por presença ou não de macrossomia, segundo as características da mãe e do recém-nascido. Bahia, 2020.

Características da mãe e do recém-nascido	Percentual de recém-nascidos (n = 141.515)	Macrossomia	
		Não (n = 132.699)	Sim (n = 8.816)
Idade da mãe			
15 a 29 anos	58,3%	94,6%	5,4%
30 a 34 anos	21,2%	92,8%	7,1%
35 anos ou mais	20,6%	92,5%	7,5%
Raça/Cor da mãe			
Branca	7,8%	94,7%	5,3%
Não Branca	92,2%	93,7%	6,3%
Idade gestacional			
A termo	96,4%	93,9%	6,1%
Pós-termo	3,6%	91,2%	8,9%
Escolaridade materna			
Até ensino superior incompleto	88,6%	93,6%	6,4%
Ensino superior completo	11,4%	94,8%	5,2%
Situação conjugal da mãe			
Com companheiro	44,2%	93,7%	6,3%
Sem companheiro	55,8%	93,8%	6,2%
Número de gestações anteriores			
Uma ou mais	62,4 %	92,7%	7,2%
Nenhuma	37,6%	95,6%	4,4%
Número de consultas pré-natal			
6 ou mais consultas	99,1%	93,8%	6,2%
Até 5 consultas	0,9%	94,2%	5,8%

Tabela 1 (continuação)

Características da mãe e do recém-nascido	Percentual de recém-nascidos (n = 141.515)	Macrossomia	
		Não (n = 132.699)	Sim (n = 8.816)
Tipo de Parto			
Vaginal	52,7%	91,6%	8,4%
Cesáreo	47,3%	95,7%	4,3%
Sexo do recém-nascido			
Masculino	51,0%	92,2%	7,8%
Feminino	49,0%	95,4%	4,6%
Tipo de apresentação do recém-nascido			
Cefálico	97,2%	93,7%	6,3%
Pélvico ou Transversa	2,8%	94,6%	5,4%

Primeiramente, a amostra completa foi dividida aleatoriamente em dois conjuntos de dados, respeitando a proporção de 80% para treinamento (n=113.215) e 20% para teste (n=28.300).

A Figura 5 mostra a distribuição percentual de recém-nascidos, segundo a presença ou não de macrossomia, após a divisão em amostra treino e amostra teste.

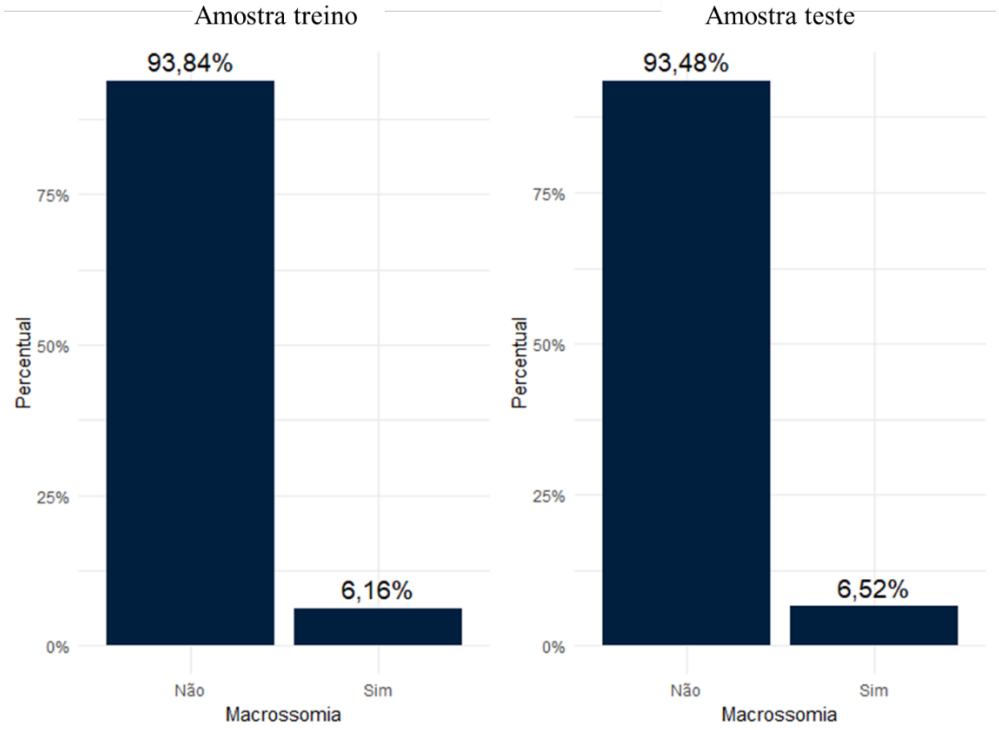


Figura 5: Distribuição percentual de recém-nascidos na amostra treino e na amostra teste, segundo a presença ou não de macrossomia.

Para o conjunto de treinamento foi aplicado, o método de validação cruzada 10-fold, ou seja, a amostra de treinamento foi dividida em k=10 partes (folds) mutuamente exclusivas de tamanhos iguais. Nove dessas partes (9/10) foram utilizadas nos ajustes dos modelos log-linear de Poisson (treino), considerando três diferentes blocos de variáveis ("clínico+sociodemográfico", "clínico" e "sociodemográfico"). A parte restante (1/10), por sua vez, foi utilizada para a validação dos modelos, utilizando como métrica a área sob curva ROC (AUC). A Tabela 2 fornece os valores de AUC, obtidos para os modelos ajustados considerando os três diferentes blocos de variáveis mencionados, usando o método de validação cruzada k-fold.

Tabela 2: Valores de AUC dos ajustes dos modelos log-linear de Poisson, com três diferentes blocos de variáveis, para predição do desfecho de macrossomia, usando o método de reamostragem 10-fold.

Iteração (10-folds)	**Bloco 1: Clínico + Sociodemográfico**	**Bloco 2: Clínico**	**Bloco 3: Sociodemográfico**
	AUC	**AUC**	**AUC**
1	0,654	0,609	0,612
2	0,652	0,619	0,605
3	0,656	0,629	0,590
4	0,688	0,650	0,611
5	0,665	0,641	0,590
6	0,646	0,600	0,608
7	0,647	0,610	0,602
8	0,656	0,621	0,597
9	0,673	0,647	0,589
10	0,649	0,602	0,604

A partir dos diferentes valores de AUC (Tabela 2) obtidos com base na amostra de validação, definida através do método de validação cruzada 10-fold, foram calculadas a média, o desvio-padrão e o coeficiente de variação (Tabela 3).

Desse modo, conforme evidenciado na Tabela 3, os três modelos geraram valores de AUC com baixo grau de variabilidade, com coeficientes de variação inferiores a 3%. No entanto, o modelo que inclui conjuntamente as características clínicas e sociodemográficas se destaca ao apresentar um AUC médio ligeiramente superior, em comparação aos modelos com apenas o bloco de variáveis clínicas ou apenas o bloco sociodemográfico. Isso sugere que, ao utilizar validação cruzada 10-fold, o modelo considerando o bloco "clínico + sociodemográfico" teve melhor desempenho relativo na amostra de validação, tornando a sua escolha preferencial.

Tabela 3: Média, desvio-padrão e coeficiente de variação dos valores de AUC calculados na amostra de validação, para os modelos log-linear de Poisson ajustados com os três diferentes blocos de variáveis.

Modelos log-linear de Poisson	**Amostra de validação**		
	Média de AUC	Desvio-padrão de AUC	CV de AUC
Bloco 1: Clínico+Demográfico	0,658	0,013	2,0%
Bloco 2: Clínico	0,623	0,018	2,9%
Bloco 3: Sociodemográfico	0,601	0,009	1,5%

Após a seleção do melhor modelo (bloco 1), a partir da validação cruzada 10-fold, o modelo log-linear de Poisson (com variância robusta), foi ajustado utilizando a totalidade da amostra de treinamento. Na Tabela 4, são fornecidos os resultados do ajuste do modelo completo com todas as variáveis clínicas e sociodemográficas, bem como os resultados do modelo com os preditores selecionados. Entre estes resultados estão as razões de prevalência (RP), seus intervalos de 95% de confiança e os p-valores do teste de Wald.

Tabela 4: Resultados do ajuste do modelo log-linear de Poisson (com variância robusta) para a predição do desfecho de macrossomia.

Características da mãe e do recém-nascido	**Modelo completo**			**Modelo com as variáveis selecionadas**		
	RP	**IC 95%**	**p-valor (Wald)**	**RP**	**IC 95%**	**p-valor (Wald)**
Idade da mãe			<0,001			<0,001
15 a 29 anos	1	-	-	1	-	-
30 a 34 anos	1,124	[1,061;1,192]	<0,001	1,124	[1,061;1,192]	<0,001
35 anos ou mais	1,145	[1,078;1,216]	<0,001	1,145	[1,078;1,216]	<0,001
Raça/Cor da mãe						
Branca	1	-		1	-	-
Não Branca	1,216	[1,105;1,337]	<0,001	1,216	[1,105;1,337]	<0,001
Idade gestacional						
A termo	1	-	-	1	-	-
Pós-termo	1,451	[1,311;1,607]	<0,001	1,452	[1,311;1,608]	<0,001
Escolaridade da mãe						
Até ensino superior incompleto	1	-	-	1	-	-
Ensino superior completo	0,695	[0,639;0,757]	<0,001	0,696	[0,639;0,757]	<0,001

Tabela 4 (continuação)

Características da mãe e do recém-nascido	Modelo completo			Modelo com as variáveis selecionadas		
	RP	IC 95%	p-valor (Wald)	RP	IC 95%	p-valor (Wald)
Situação Conjugal						
Com companheiro	1	-	-	1	-	-
Sem companheiro	1,070	[1,021;1,121]	0,005	1,069	[1,020;1,121]	0,005
Nº de gestações anteriores						
Uma ou mais	1	-	-	1	-	-
Nenhuma	0,614	[0,582;0,648]	<0,001	0,614	[0,582;0,648]	<0,001
Nº de consultas pré-natais						
Até 5 consultas	1		-			
6 ou mais consultas	0,932		0,586			
Tipo de Parto						
Cesáreo	1	-	-	1	-	-
Vaginal	0,494	[0,471;0,519]	<0,001	0,494	[0,471;0,519]	<0,001
Sexo do recém-nascido						
Feminino	1	-		1	-	-
Masculino	1, 713	[1,633;1,796]	<0,001	1,712	[1,633;1,796]	<0,001
Tipo de apresentação						
Cefálico	1	-	-	1	-	-
Pélvico ou Transversa	0,680	[0,586,790]	<0,001	0,680	[0,586;0,790]	<0,001

No modelo completo, apenas o número de consultas de acompanhamento pré-natal não apresentou uma associação estatisticamente significativa com o desfecho de macrossomia (p-valor = 0,586). Após a sua remoção, todas as variáveis restantes mantiveram uma associação estatisticamente significativa com o desfecho, a um nível de significância de 5% (Tabela 4).

Em relação às características maternas a prevalência de macrossomia em bebês nascidos de mães com idade de 30 a 34 anos foi 12,4% maior em comparação a nascidos de mães com 15 a 29 anos de idade (RP = 1,124; p-valor<0,001). E para bebês filhos de mães na

faixa etária de 35 anos ou mais, observou-se uma prevalência de macrossomia 14,5% superior a bebês com mães na faixa etária de 15 a 29 anos (RP = 1,145; p-valor<0,001).

Quanto à raça/cor, a prevalência de macrossomia em recém-nascidos de mulheres não brancas foi 21,6% maior do que em bebês de mães brancas (RP = 1,216; p-valor<0,001). Com relação à situação conjugal, observou-se uma prevalência de macrossomia 6,9% maior para filhos de mulheres que vivem sem companheiro (RP =1,069; p-valor=0,005).

No que se refere à paridade, filhos de mulheres com nenhuma gestação anterior apresentaram uma prevalência de macrossomia 38,6% menor do que os nascidos de mulheres que já tiveram uma ou mais gestações anteriores (RP = 0,614; p-valor<0,001).

Já no tocante à escolaridade materna, quando se consideram recém-nascidos de mães com ensino superior completo, a prevalência de macrossomia foi 30,4% menor em comparação com recém-nascidos de mães que não completaram o ensino superior (RP=0,696; p-valor<0,001).

Com relação às características do recém-nascido, observou-se que bebês nascidos pós-termo apresentaram prevalência de macrossomia 45,2% maior do que os bebês nascidos a termo (RP= 1,452; p-valor<0,001). Além disso, a prevalência de macrossomia entre bebês do sexo masculino foi 71,2% maior do que entre bebês do sexo feminino (RP= 1,712; p-valor<0,001).

No que tange ao tipo de parto, os bebês que nasceram de parto vaginal apresentaram uma prevalência de macrossomia 50,6% menor do que os nascidos de parto cesáreo (RP= 0,494; p-valor<0,001). Quanto ao tipo de apresentação do bebê, observou-se que quando a apresentação é pélvica (sentado) ou transversa (atravessado) a prevalência de macrossomia foi 32,0% menor do que a dos bebês com apresentação cefálica (RP = 0,680; p-valor<0,001).

O modelo log-linear de Poisson (com variância robusta), selecionado e ajustado na amostra de treinamento, foi avaliado usando os dados da amostra teste (Tabela 5).

Tabela 5: Métricas de avaliação do modelo log-linear de Poisson (com variância robusta) usando os dados da amostra teste.

Métricas* de avaliação do modelo com as variáveis selecionadas	**Amostra Teste**
Acurácia	53,6%
Sensibilidade	68,4%
Especificidade	52,6%
AUC	0,648

*Utilizando o ponto de corte ótimo ($\delta = 0{,}0575$).

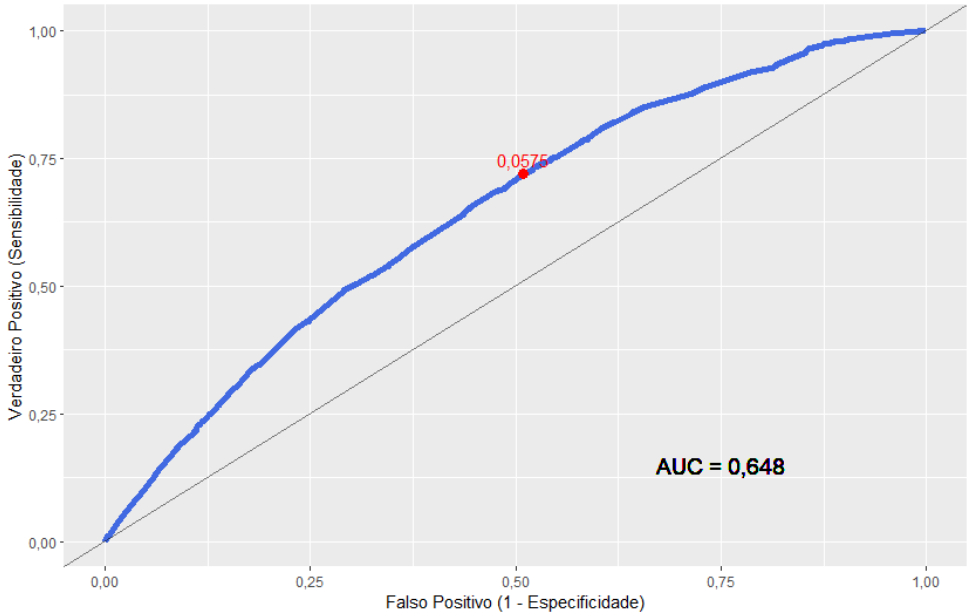


Figura 6: Curva ROC, AUC e o ponto de corte ótimo para o modelo log-linear de Poisson ajustado por MV.

O desempenho do modelo final revelou medidas de sensibilidade e especificidade razoáveis, na faixa de 50% a 80%. Apesar disto, mesmo utilizando o ponto de corte ótimo para a probabilidade preditiva, visando reduzir o impacto do desbalanceamento dos dados no desempenho do modelo, a acurácia do modelo foi de 53,6%, indicando que um pouco mais da metade dos bebês foi corretamente classificado pelo modelo selecionado. Além disso, a área sob a curva ROC (AUC=0,648), não atingiu o mínimo desejado de 0,70 para considerar o modelo como tendo um desempenho preditivo satisfatório.

Na tentativa de mitigar o viés das estimativas de máxima verossimilhança (MV) dos coeficientes em função do desfecho de macrossomia ser raro, optou-se por treinar o mesmo modelo log-linear de Poisson (variância robusta), utilizando o método de redução de viés de Firth. Este modelo foi então ajustado por máxima verossimilhança penalizada (MVP), com o método de Firth, e os principais resultados são apresentados na Tabela 6. Cabe destacar que a estimativas de MVP dos coeficientes do modelo não apresentaram diferenças substânciais das estimativas usuais de MV. Pode-se notar ainda que os p-valores do teste de Wald também não apresentaram diferenças substanciais.

Tabela 6: Resultados do ajuste do modelo log-linear de Poisson (com variância robusta), utilizando o método de redução de viés de Firth, para a predição do desfecho de macrossomia.

Características da mãe e do recém-nascido	Modelo completo			Modelo com as variáveis selecionadas		
	RP	**IC 95%**	**p-valor (Wald)**	**RP**	**IC 95%**	**p-valor (Wald)**
Idade da mãe			<0,001			<0,001
15 a 29 anos	1	-	-	1	-	-
30 a 34 anos	1,115	[1,053;1,181]	<0,001	1,116	[1,054;1,183]	<0,001
35 anos ou mais	1,137	[1,072;1,207]	<0,001	1,138	[1,073;1,208]	<0,001
Raça/Cor da mãe						
Branca	1	-		1	-	-
Não Branca	1,152	[1,052;1,261]	0,002	1,152	[1,152;1,261]	0,002
Idade gestacional						
A termo	1	-	-	1	-	-
Pós-termo	1,406	[1,270;1,556]	<0,001	1,408	[1,272;1,559]	<0,001
Escolaridade da mãe						
Ensino superior incompleto	1	-	-	1	-	-
Ensino superior completo	0,715	[0,658;0,777]	<0,001	0,717	[0,660;0,779]	<0,001
Situação conjugal						
Com companheiro	1	-	-	1	-	-
Sem companheiro	1,059	[1,011;1,111]	0,016	1,057	[1,009;1,107]	0,018
Nº de gestações anteriores						
Uma ou mais	1	-	-	1	-	-
Nenhuma	0,616	[0,584;0,649]	<0,001	0,613	[0,581;0,646]	<0,001
Nº de consultas pré-natais						
Até 5 consultas	1	-	-			
6 consultas ou mais	0,882	[0,687;1,132]	0,324			

Tabela 6 (continuação)

Características da mãe e do recém-nascido	Modelo completo			Modelo com as variáveis selecionadas		
	RP	IC 95%	p-valor (Wald)	RP	IC 95%	p-valor (Wald)
Tipo de parto						
Cesáreo	1	-	-	1	-	-
Vaginal	0,487	[0,464;0,511]	<0,001	0,487	[0,464;0,511]	<0,001
Sexo do recém-nascido						
Feminino	1	-	-	1	-	-
Masculino	1,675	[1,599;1,755]	<0,001	1,675	[1,599;1,755]	<0,001
Tipo de apresentação						
Cefálico	1	-	-	1	-	-
Pélvico ou Transversa	0,663	[0,572;0,769]	<0,001	0,663	[0,572;0,769]	<0,001

No modelo completo, utilizando o método de redução de viés de Firth, assim como no modelo completo ajustado por MV, apenas o número de consultas de acompanhamento pré-natal não apresentou uma associação estatisticamente significativa com o desfecho de macrossomia (p-valor=0,324). Após a sua remoção, todas as variáveis restantes mantiveram uma associação estatisticamente significativa com a prevalência de macrossomia, a um nível de significância de 5% (Tabela 6).

A prevalência de macrossomia em bebês nascidos de mães com idade de 30 a 34 anos foi 11,6% maior em comparação aos nascidos de mães com 15 a 29 anos (RP = 1,116; p-valor<0,001). Para os bebês filhos de mães na faixa etária de 35 anos ou mais, observa-se uma prevalência de macrossomia 13,8% superior, comparativamente aos nascidos de mães com idade de 15 a 29 anos (RP = 1,138; p-valor<0,001).

A prevalência de macrossomia nos recém-nascidos de mulheres não brancas foi 15,2% maior do que nos recém-nascidos de mulheres brancas (RP = 1,152; p-valor=0,002). Quanto à situação conjugal, observou-se que a prevalência de macrossomia foi 5,7% maior para recém-nascidos de mães que vivem sem companheiro (RP =1,057; p-valor=0,018).

Os bebês filhos de mulheres com nenhuma gestação anterior apresentaram prevalência de macrossomia 38,7% menor em comparação aos filhos de mulheres que tiveram uma ou mais gestações anteriores (RP = 0,613; p-valor<0,001).

Nos recém-nascidos de mães com ensino superior completo, a prevalência de macrossomia foi 28,3% menor em comparação aos recém-nascidos de mães que não completaram o ensino superior (RP=0,717; p-valor<0,001).

Com relação às características dos recém-nascidos, observou-se que entre os bebês pós-termo a prevalência de macrossomia foi 40,8% maior do que os bebês a termo (RP= 1,408; p-valor<0,001). Além disso, a prevalência de macrossomia entre bebês do sexo masculino foi 67,5% maior do que entre bebês do sexo feminino (RP= 1,675; p-valor<0,001).

No que tange ao tipo de parto, os bebês que nasceram de parto vaginal apresentaram prevalência de macrossomia 51,3% menor do que aqueles nascidos de parto cesáreo (RP= 0,487; p-valor<0,001). Quanto a apresentação do bebê é pélvica (sentado) ou transversa (atravessado) a prevalência de macrossomia foi 33,7% menor comparativamente a apresentação cefálica (RP = 0,663; p-valor<0,001).

O modelo log-linear de Poisson (com variância robusta), utilizando o método de Firth, selecionado e ajustado na amostra de treinamento, foi avaliado usando os dados da amostra teste por meio das métricas disponibilizadas na Tabela 7.

Tabela 7: Métricas de avaliação do modelo log-linear de Poisson (com variância robusta), utilizando o método de Firth, a partir da amostra teste.

Métricas* de avaliação do modelo com as variáveis selecionadas	**Amostra Teste**
Acurácia	49,1%
Sensibilidade	75,4%
Especificidade	47,4%
AUC	0,662

*Utilizando o ponto de corte ótimo ($\delta = 0{,}0513$).

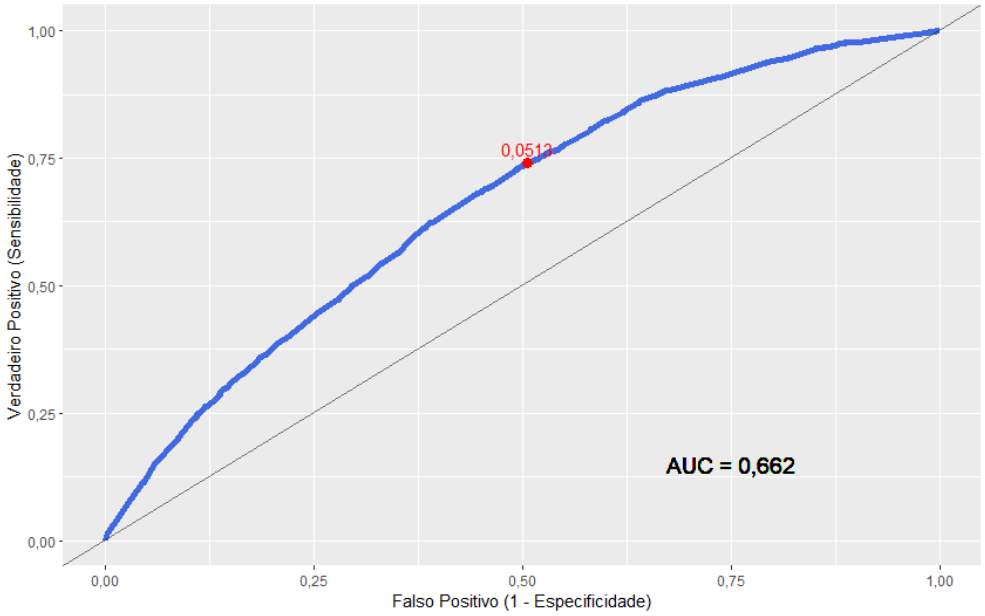


Figura 7: Curva ROC, AUC e ponto de corte ótimo para o modelo log-linear de Poisson ajustado por MVP, utilizando método de Firth.

O modelo log-linear de Poisson utilizando o método de Firth apresentou uma sensibilidade de 75,4%, indicando que 75,4% dos recém-nascidos com macrossomia foram classificados corretamente por este modelo como sendo macrossômicos. Esta sensibilidade foi superior ao do modelo ajustado por MV, mas os valores de especificidade e acurácia foram ligeiramente menores. Além disso, observou-se apenas um pequeno aumento na AUC ao adotar o método de Firth, mas não atingiu o limiar desejado de 0,7 necessário para classificar este modelo como tendo um desempenho satisfatório.

5 Discussão e Conclusão

No presente trabalho buscou-se avaliar a associação entre a macrossomia fetal e um conjunto de características das mães e dos recém-nascidos no Estado da Bahia, localizada na região Nordeste do Brasil, no ano de 2020.

A macrossomia pode ser definida como um recém-nascido com peso ao nascer acima de 4500 gramas, e neste caso a prevalência varia de 1,3 a 1,5% em países desenvolvidos, enquanto a definição de macrossomia baseada num peso maior ou igual a 4000 gramas apresenta segundo Campbel (2014) uma prevalência em torno de 7%. Há críticas quanto ao uso do limite de 4500 g, pois o risco de distocia de ombro já é alto para bebês com pesos ao nascer de 4000 e 4250 g mesmo para partos não complicados por diabetes materna (Campbel, 2014; Nesbitt et al., 1998). No presente estudo, o desfecho de macrossomia foi construído usando o peso ao nascer igual ou maior que 4000 gramas, independentemente da idade gestacional, como adotado em outros estudos (Adugna et al., 2020; Madi et al., 2006; Ruiz-Canchucaja e Cano-Cardenas, 2022). Baseada nesta definição, a prevalência de macrossomia na amostra estudada foi de 6,2%. No estudo conduzido por Amorim et al. (2009) também em formato de corte transversal, envolvendo 551 mulheres internadas no Instituto de Saúde Elpídio de Almeida (ISEA), em Campina Grande (Paraíba-Brasil), em 2007, foi observada uma prevalência de macrossomia de 5,4%. Carvalhaes et al. (2013) também realizaram um estudo em formato de corte transversal, em um município do interior paulista, com 212 mulheres atendidas em unidades básicas de saúde em dias de atendimento pré-natal no período de maio de 2009 a dezembro de 2010, e obtiveram uma prevalência de macrossomia de 5,1%.

Usando o modelo log-linear de Poisson com ou sem o método de Firth, observou-se uma maior prevalência de macrossomia entre os recém-nascidos pós-termo, do sexo masculino, filhos de mães com idade avançada (35 anos ou mais anos de idade), de raça/cor não branca, sem ensino superior completo, vivendo sem companheiro e que haviam experimentado uma ou mais gestações anteriores. Também se observou uma maior prevalência de macrossomia entre bebês que nasceram por parto cesáreo e que estavam na apresentação cefálica.

A gravidez em idade materna avançada (35 anos ou mais) é um fator de risco para desfechos maternos e perinatais adversos (Glick, 2021). Segundo Blomberg et al.(2014) para o grupo de mulheres com idade avançada, há maior risco de lacerações perineais, pré-eclâmpsia, descolamento prematuro da placenta, hemorragia pós-parto e resultados neonatais desfavoráveis. A associação entre a maior faixa etária da mãe e a presença de macrossomia fetal também foi observada por Oliveira et al. (2008), Madi et al. (2006), Fiorelli e Zugaib

(2007), e Ruiz-Canchucaj e Cano-Cardenas (2022). Uma explicação apontada para esta associação é que a gestação em idade avançada acarreta uma alteração no metabolismo da mãe, que por sua vez, geraria um maior crescimento do feto. Embora a idade avançada seja considerada um fator de risco associado à macrossomia fetal, vale ressaltar que há estudos, que não encontraram associação significativa entre idade materna e macrossomia (Amorim et al., 2009; Kerche et al., 2005).

Chung et al.(2022), usando análise de regressão logística multivariada, verificaram que filhos tanto de mães quanto de pais com mais de 12 anos de estudo apresentaram menores chances de macrossomia. Na presente análise foi considerada apenas a escolaridade materna, indicando no mesmo sentido que o nível educacional mais elevado (ensino superior completo) está associado a uma menor ocorrência (prevalência) de macrossomia.

Assim como em Madi et al. (2006), foi observada uma associação significativa entre a maior paridade (número de partos anteriores) e a macrossomia. Adugna et al. (2020), ao realizarem um estudo transversal de base institucional, usando uma amostra de 491 mães e seus recém-nascidos, também constataram que o sexo masculino foi um dos fatores significativamente associados com uma maior ocorrência de macrossomia. O sexo masculino e a maior paridade também tiveram associação significativa com este desfecho no estudo de Cunha et al., (2017), baseando-se nos dados do peso ao nascer de 6121 crianças menores de cinco anos investigadas na Pesquisa Demográfica e de Saúde da Família (ENDES) de 2013. No entanto, ainda há carência de explicação clínica para essa associação, tanto para bebês do sexo masculino, quanto para filhos de mulheres multíparas. De acordo com Oliveira et al. (2008), apesar dessas duas características serem não modificáveis, sugere-se que mulheres a partir da segunda gestação ou grávidas de bebês do sexo masculino devam atentar para outros fatores controláveis que podem influenciar na ocorrência de macrossomia (Oliveira et al., 2008).

No presente estudo, identificou-se associação estatisticamente significativa das variáveis raça/cor e situação conjugal com o desfecho de macrossomia. Esta constatação difere dos resultados obtidos por Oliveira et al. (2008) e Kac e Velásquez-Meléndez (2005), cujas pesquisas foram conduzidas em uma unidade de rede básica de saúde no município do Rio de Janeiro, no período de junho de 2005 a abril de 2007, e no próprio município do Rio de Janeiro, no período de maio de 1999 a abril de 2001, respectivamente. Nestes estudos anteriores, não foi observada associação significativa entre a raça/cor e a situação conjugal da mãe e o desfecho macrossomia.

Outra característica investigada neste estudo e que também se revelou associada de forma significativa à macrossomia foi parto cesáreo (Fiorelli e Zugaib, 2007; Cunha et al.,

2017). Entretanto, Fiorelli e Zugaib (2007) demostraram que as frequências de intercorrências materno-fetais não diferiram entre os partos cesáreo e vaginal em fetos macrossômicos. Do exposto, estes autores ressaltam que a gestação não complicada com indícios de macrossomia não é uma indicação de primeira cesárea ou indução do trabalho de parto.

A maior prevalência de macrossomia associada ao pós-datismo também foi observada em outros estudos como o de Ávila et al. (2013) e Fiorelli e Zugaib (2007).

Para fetos com apresentação transversa ou pélvica em estágio final da gestação, há um maior risco no momento do parto (Brasil, 2013), e estes tipos de apresentação se mostraram associados à uma maior prevalência de prematuridade (idade gestacional < 37 semanas) no estudo de Souza (2021); e entre bebês prematuros há maior prevalência de baixo peso ao nascer (peso < 2500g) como constatado por Santos (2021), o que pode explicar o sentido da associação encontrada entre a posição fetal transversa ou pélvica e a menor prevalência de macrossomia (peso ≥ 4000 g).

Com relação a potencialidade deste estudo, destaca-se o uso do modelo log-linear de Poisson com variância robusta. Este modelo é comumente empregado em estudos de natureza transversal, quando a variável resposta é binária. Nesse contexto, é atribuído um valor unitário ao tempo de seguimento de cada observação, para a obtenção de razão de prevalência de macrossomia, ao invés de razão de chance de macrossomia (*odds ratio* – OR), onde esta última medida é comumentemente calculada em modelos de regressão logística (Coutinho et al., 2008). Como ressaltado por Barros e Hirakata (2003), utilizar o modelo log-linear de Poisson sem o ajuste de variância robusta para análise de dados transversais resulta em consideráveis erros nas estimativas de intervalo. Por outro lado, ao empregá-lo com a correção de variância robusta, o modelo log-linear de Poisson fornece estimativas pontuais e intervalares corretas, e vale destacar que a vantagem de se adotar a razão de prevalência como medida de associação reside na sua maior interpretabilidade e no maior interesse nesta medida por pesquisadores da área da saúde que trabalham com dados de estudos transversais. Coutinho et al. (2008) demonstraram que, se interpretados como estimativas de Razão de Prevalência (RP), os *Odds Ratios* (OR) tenderiam a superestimar as associações para desfechos com prevalência baixa, intermediária e alta em 13%, quase 100% e quatro vezes mais, respectivamente. Além disso, embora o modelo log-linear de Poisson (com variância robusta) seja considerado adequado para estimar a razão de prevalência (Coutinho et al., 2008; Barros e Hirakata, 2003), foi ajustado ainda o modelo log-linear de Poisson (com variância robusta) por máxima verossimilhança penalizada (MVP), com o uso do método de Firth, já que para desfechos raros, as estimativas usuais de máxima verossimilhança dos parâmetros dos modelos são viciadas (Gosho et al., 2023; King e Zeng, 2001). Apesar disso, vale destacar que não foram observadas

diferenças substanciais entre as estimativas pontuais dos modelos ajustados por MV e por MVP.

Outra potencialidade deste estudo foi a incorporação de métodos de aprendizado de máquinas. Conforme ressaltado por Moreira et al. (2020), a aplicação de técnicas de aprendizado de máquinas em artigos voltados à área da saúde tem crescido notavelmente (Zacharaki et al., 2009; David e Arun, 2020; Burns et al., 2011), trazendo consigo benefícios na otimização dos processos e na potencial redução de custos de tratamentos, posto que é possível desenvolver um conjunto de ferramentas de alerta e gerenciamento de riscos, que vão desde a predição de possíveis surtos de epidemias à geração de resultados de forma ágil. A base de dados, num primeiro momento, foi dividida aleatoriamente em duas partes (treino e teste), e aplicado o método de validação cruzada k-fold na amostra de treino, sendo o uso de k=10 (divisão aleatória dos dados em dez subconjuntos aleatórios chamados *folds*) uma prática comum, como ressaltado por Dinov (2018), visando avaliar o melhor bloco de variáveis para a predição do desfecho de macrossomia. Como mencionado por Efron e Tibshirani (1994), esta divisão é importante, pois ao utilizar a mesma amostra para ajustar o modelo e avaliar seu desempenho na predição de resultados, há uma tendência de otimismo no erro de predição do modelo, ou seja, uma propensão a subestimar o verdadeiro erro de predição.

Este estudo apresenta algumas limitações, uma vez que sua análise é restrita aos dados contidos no banco de dados do Sistema de Informações sobre Nascidos Vivos (SINASC). A confiabilidade está intrinsecamente ligada à precisão e abrangência do preenchimento do formulário de Declaração de Nascidos Vivos (DNV). Com o SINASC, por ter caráter transversal, não é possível estabelecer uma relação causal entre a macrossomia e as características maternas e do recém-nascido. Outra limitação se refere a não inclusão de outras informações importantes para predição da macrossomia, por não fazerem parte do escopo da fonte utilizada como diabetes materno, diabetes gestacional, ganho de peso materno durante a gravidez, obesidade pré-gestacional, índice de massa corporal (IMC) materno prévio à gravidez, tabagismo, histórico de macrossomia, aborto prévio, hipertensão arterial, nível glicêmico e intensidade da atividade física. A inclusão de algumas destas variáveis, entre elas diabetes materno, poderia ter melhorado a capacidade preditiva do modelo, fazendo com o valor da AUC alcançasse o patamar mínimo desejável de 0,7.

A partir destes achados, conclui-se sobre a necessidade de desenvolver ações voltadas para a prevenção da macrossomia fetal no Estado da Bahia, a fim reduzir o risco de complicações materno-infantis. Entre estas ações pode-se citar maiores investimentos em assistência pré-natal, recomendações nutricionais e de práticas de atividade física priorizando,

de modo geral, mulheres grávidas de bebês do sexo masculino, não brancas e com menores níveis socioeconômicos.

Referências

ADUGNA, Dagnew Getnet; ENYEW, Engidaw Fentahun; JEMBERIE, Molla Taye. Prevalence and associated factors of macrosomia among newborns delivered in University of Gondar Comprehensive Specialized Hospital, Gondar, Ethiopia: an institution-based cross-sectional study. *Pediatric health, medicine and therapeutics*, p. 495-503, 2020.

ALAOUI, Safae Sossi; LABSIV, Y.; AKSASSE, Brahim. *Classification algorithms in data mining*. Int. J. Tomogr. Simul, v. 31, n. 4, p. 34-44, 2018.

AMORIM, Melania Maria Ramos de; LEITE, Debora Farias Batista; GADELHA, Tarcísia Gonçalves Nóbrega; MUNIZ, Anna Gabriella Viana; MELO, Adriana Suely de Oliveira; ROCHA, Aline da Mota. Fatores de risco para macrossomia em recém-nascidos de uma maternidade-escola no nordeste do Brasil. *Revista Brasileira de Ginecologia e Obstetrícia*, v. 31, p. 241-248, 2009.

ASEVEDO, J.M.; MARTINEZ, L. Factores de riesgo asociados a macrosomia fetal en el Hospital JB Iturraspe de la ciudad de Santa Fe. *Fac Ciencias Médicas UNL Área*, 2017.

ÁVILA, R.R.; HERRERA, P.M.; SALAZAR, C.C.I.; CAMACHO, R.R.I. Factores de riesgo del recién nacido macrosómico. *Pediatría de México*, v. 15, n. 1, p. 6-11, 2013.

BARROS, Aluísio J.D.; HIRAKATA, Vânia N. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC medical research methodology*, v. 3, n. 1, p. 1-13, 2003.

BHASKAR, A.; PONNURAJA, C. Comparison of Regression Models for Binary Outcome Variables in Clinical Trials. *Current Science*, v. 119, n. 12, p. 2010-2013, 2021. https://doi.org/10.18520/cs/v119/i12/2010-2013

BLOMBERG, M.; BIRCH, Tyrberg R.; KJØLHEDE, P. Impact of maternal age on obstetric and neonatal outcome with emphasis on primiparous adolescents and older women: a Swedish Medical Birth Register Study. *BMJ Open*, v. 4, n. 11. p. e005840, 2014. Doi: 10.1136/bmjopen-2014-005840

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise Epidemiológica e Vigilância de Doenças Não Transmissíveis. *Declaração de Nascido Vivo Manual de instruções para preenchimento / Ministério da Saúde*, Secretaria de Vigilância em

Saúde. Departamento de Análise Epidemiológica e Vigilância de Doenças Não Transmissíveis. – 4.ed. – Brasília: Ministério da Saúde, 2022.

BRASIL. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Ações Programáticas Estratégicas. Gestação de alto risco: manual técnico / Ministério da Saúde, Secretaria de Atenção à Saúde, Departamento de Ações Programáticas Estratégicas. – 5. ed. – Brasília : Editora do Ministério da Saúde, 2012. 302 p.

BRASIL. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Atenção Básica. Atenção ao pré-natal de baixo risco [recurso eletrônico] – 1. ed. rev. – Brasília: Editora do Ministério da Saúde, 2013.

BRUNIALTI, Lucas F.; FREIRE, Valdinei; PERES, Sarajane M.; LIMA, Clodoaldo A.M. Aprendizado de máquina em sistemas de recomendação baseados em conteúdo textual: uma revisão sistemática. *Anais do XI Simpósio Brasileiro de Sistemas de Informação*, p. 203-210, 2015.

BURNS, M.; BEGALE, M.; DUFFECY, J.; GERGLE, D.; KARR, C.; GIANGRANDE, E.; MOHR, D. Harnessing Context Sensing to Develop a Mobile Intervention for Depression. *Journal of medical Internet research*, v. 13, n. 3, p. e1838, 2011. URL: https://www.jmir.org/2011/3/e55 doi: 10.2196/jmir.1838

CAMPBELL, Stuart. Fetal macrosomia: a problem in need of a policy. *Ultrasound in Obstetrics & Gynecology*, v. 43, n. 1, p. 3-10, 2014. doi: 10.1002/uog.13268. PMID: 24395685

CARVALHAES, Maria Antonieta de Barros Leite; GOMES, Caroline de Barros; MALTA, Maria Barreto; PAPINI, Silvia Justina; PARADA, Cristina Maria Garcia de Lima. Sobrepeso pré-gestacional associa-se a ganho ponderal excessivo na gestação. Revista Brasileira de Ginecologia e Obstetrícia, v. 35, p. 523-529, 2013.

CHUNG, Y.H.; HWANG, I.S.; JUNG, G.; KO, H.S. Advanced parental age is an independent risk factor for term low birth weight and macrosomia. *Medicine*, v. 101, n. 26, 2022. doi: 10.1097/MD.0000000000029846. PMID: 35777059; PMCID: PMC9239628.

CLAUSEN, T.; BURSKI, T.K.; ØYES, N.; GODANG, K.; BOLLESRSLEV, J. Maternal anthropometric and metabolic factors in the first half of pregnancy and risk of neonatal macrosomia in term pregnancies. A prospective study, *Society of the European Journal of Endocrinology*, v. 153, n. 6, p. 887-894, 2005.

COUTINHO, Leticia; SCAZUFCA, Marcia; MENEZES, Paulo R. Métodos para estimar razão de prevalência em estudos de corte transversal. *Revista de Saúde Pública*, v. 42, 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.

CUNHA, A.J.L.A. da; TORO, M.S.; GUTIÉRREZ, C.; ALARCÓN-VILLAVERDE, J. Prevalence and associated factors of macrossomia in Peru, 2013. Rev *Peru Med Exp Salud Publica*. v. 34 n. 1 p. 36-42, 2017.

DAVID, D.S.; ARUN, L. Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. *Artech J. Eff. Res. Eng. Technol*, v. 1, p. 57-63, 2020.

DINOV, Ivo D. Data science and predictive analytics. *Cham, Switzerland*, 2018.

DOBSON, A.J.; BARNETT, A.G. *An Introduction to Generalized Linear Models*. 4. ed. USA: Chapman and Hall/CRC, 2018. ISBN 9781138741515.

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

EVAGELIDOU, E.N.; KIORTSIS, D.N.; BAIRAKTARI, E.T.; GIAPROS, V.I; CHOLEVAS, V.K.; TZALLAS, C.S.; ANDRONIKOU, S.K. Lipid Profile, Glucose Homeostasis, Blood Pressure, and Obesity-Anthropometric Markers in Macrosomic Offspring of Nondiabetic Mothers, *Diabetes Care*, v.29, n. 6, p. 1197-1201, 2006.

FARAWAY, J.J. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. [S.l.]:* CRC press, 2016.

FARIAS, Maria do Carmo Andrade Duarte de; OLIVEIRA, Karla Maria Duarte Silva; DINIZ, Alcidez da Silva; MAIA, Paula Christianne Gomes Souto; ABRANTES, Kennia Sibelly Marques de; ABREU, Luiz Carlos de. Entre a captação e a divulgação de dados: a importância da dnv e do seu adequado preenchimento. *Journal of Human Growth and Development*, v. 24, n. 2, p. 150 – 156, 2014.

FERREIRA, E.V. *Método 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.

FIORELLI, Lilian Renata; ZUGAIB, Marcelo. Resultado perinatal na macrossomia fetal. *Revista de Medicina*, v. 86, n. 3, p. 144-147, 2007.

FIRTH, D. Bias Reduction of Maximum Likelihood Estimates. *Biometrika*, v. 80, n. 1, p. 27-38, 1993. doi: 10.1093/biomet/80.1.27

GLICK, Itamar; KADISH, Ela; ROTTENSTREICH, Misgav. Management of pregnancy in women of advanced maternal age: Improving outcomes for mother and baby. *International journal of women's health*, p. 751-759, 2021. doi: 10.2147/IJWH.S283216

GOSHO, Masahiko; ISHII, Ryota; NOMA, Hisashi; MARUO, Kazushi. A comparison of bias-adjusted generalized estimating equations for sparse binary data in small-sample longitudinal studies. *Statistics in Medicine*, 2023.

HASTIE, Trevor; FRIEDMAN, Jerome; TIBSHIRANI, Robert. *The elements of statistical learning: data mining, inference, and prediction*. New York: springer, 2009.

JOSHI, Ashwini; GEROLDINGER, Angelika; JIRICKA, Lena; SENCHAUDHURI, Pralay; CORCORAN, Christopher; HEINZE, Georg. Solutions to problems of nonexistence of parameter estimates and sparse data bias in Poisson regression. *Statistical Methods in Medical Research*, v. 31, n. 2, p. 253-266, 2022. doi: 10.1177/09622802211065405.

KAC, Gilberto; VELÁSQUEZ-MELÉNDEZ, Gustavo. Ganho de peso gestacional e macrossomia em uma coorte de mães e filhos. *Jornal de pediatria*, v. 81, p. 47-53, 2005. doi: 10.2223/JPED.1282

KERCHE, Luciane Teresa Rodrigues Lima; ABBADE, Joelsio Francisco; COSTA, Roberto Antonio Araújo; RUDGE, Marilza Vieira Cunha; CALDERON, Iracema de Mattos Paranhos. Fatores de risco para macrossomia fetal em gestações complicadas por diabete ou por hiperglicemia diária. *Revista Brasileira de Ginecologia e Obstetrícia*, v. 27, p. 580-587, 2005. doi: 10.1590/S0100-72032005001000003

KING, Gary; ZENG, Langche. Logistic regression in rare events data. *Political analysis*, v. 9, n. 2, p. 137-163, 2001.

KOHAVI, Ron. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. In: Ijcai. 1995. p. 1137-1145.

KOSMIDIS, I. brglm: Bias Reduction in Binary-Response Generalized Linear Model. *R package version 0.7.2*, 2021. <https://cran.r-project.org/package=brglm>

KOSMIDIS, I. brglm2: Bias Reduction in Generalized Linear Models_. *R package version 0.9.2*, 2023. <https://CRAN.R-project.org/package=brglm2>

KOSMIDIS, Ioannis; FIRTH, David. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*, v. 108, n. 1, p. 71-82, 2021.

KOSMIDIS, Ioannis; KENNE PAGUI, Euloge Clovis; SARTORI, Nicola. Mean and median bias reduction in generalized linear models. *Statistics and Computing*, v. 30, n. 1, p. 43-59, 2020.

KUHN, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, v. 28, n. 5, p. 1–26, 2008. https://doi.org/10.18637/jss.v028.i05

LEE, Huei Diana; MONARD, Maria Carolina. Seleção de atributos para algoritmos de aprendizado de máquina supervisionado utilizando como filtro a dimensão fractal. *Revista de La Sociedad Chilena de Ciencia de La Computación*, v. 4, n. 1, p. 1-8, 2003.

MADI, José Mauro; ROMBALDI, Renato Luís; FILHO, Petrônio Fagundes de Oliveira; ARAÚJO, Breno Fauth; ZATI, Helen; MADI, Sônia regina Cabral. Fatores maternos e perinatais relacionados à macrossomia fetal. *Revista Brasileira de Ginecologia e Obstetrícia*, v. 28, p. 232-237, 2006. doi: 10.1590/S0100-72032006000400005

MISHIMA, Flávia Cristiane; SCOCHI, Carmen Gracinda Silvan; FERRO, Maria Alice Rossato; LIMA, Regina Aparecida Gracia de; COSTA, Isabel Aparecia Ribeiro. Declaração de nascido vivo: análise do seu preenchimento no Município de Ribeirão Preto, São Paulo, Brasil. *Cadernos de Saúde Pública*, v. 15, p. 387-395, 1999.

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

MOREIRA, Paulo Sergio da Conceição; BYANCA, Neumann Salerno; DENISE, Fukumi Tsunoda. *Internet das coisas e aprendizado de máquina na área da saúde: uma análise bibliométrica da produção científica de 2009 a 2019.* 2020.

NESBITT, Thomas S.; GILBERT, William M.; HERRCHEN, Beate. Shoulder dystocia and associated risk factors with macrosomic infants born in California. *American journal of obstetrics and gynecology*, v. 179, n. 2, p. 476-480, 1998.

NUNES, L.M.N.; PEREIRA, A.C.; QUELUZ, D.P. Fissuras orais e sua notificação no sistema de informação: análise da Declaração de Nascido Vivo (DNV) em Campos dos Goytacazes, Rio de Janeiro, 1999-2004. *Ciência & Saúde Coletivo*, v. 15, n. 2, p. 345–352, 2010. doi: 10.1590/S1413-81232010000200009

OLIVEIRA, Lívia Costa de; PACHECO, Alice Helena de Resende Nóra; RODRIGUES, Patricia Lima; SCHLÜSSEL, Michael Maia; SPYRIDES, Maria Helena Constantino; KAC, Gilberto. Fatores determinantes da incidência de macrossomia em um estudo com mães e filhos atendidos em uma Unidade Básica de Saúde no município do Rio de Janeiro. *Revista Brasileira de Ginecologia e Obstetrícia*, v. 30, p. 486-493, 2008.

OLIVEIRA, Max Moura de; ANDRADE, Silvânia Suely Caribé de Araújo; DIMECH, George Santiago; OLIVEIRA, João Carlos Guedes de; MALTA, Deborah Carvalho; NETO, Dácio de Lyra Rabello; MOURA, Lenilto de. Avaliação do sistema de informações sobre nascidos vivos. Brasil, 2006 a 2010. *Epidemiologia e Serviços de Saúde*, v. 24, p. 629-640, 2015. doi: 10.5123/S1679-49742015000400005

PEDRAZA, Dixis Figueroa. Qualidade do Sistema de Informações sobre Nascidos Vivos (Sinasc): análise crítica da literatura. *Ciência & Saúde Coletiva*, v. 17, p. 2729-2737, 2012.

POLO, Tatiana Cristina Figueira; MIOT, Hélio Amante. Aplicações da curva ROC em estudos clínicos e experimentais. *Jornal Vascular Brasileiro*, v. 19, 2020.

POWERS, Daniel; XIE, Yu. *Statistical methods for categorical data analysis*. Emerald Group Publishing, 2008.

R Core Team (2022). R: A language and environment for statistical computing. R foundation for Statistical computing, Vienna, Austria. URL https://www.R-project.org/.

RAHMAN, M. Shafiqur; SULTANA, Mahbuba. Performance of Firth-and logF-type penalized methods in risk prediction for small or sparse binary data. *BMC medical research methodology*, v. 17, n. 1, p. 1-15, 2017.

RENCHER, Alvin C.; SCHAALJE, G. Bruce. *Linear models in statistics*. John Wiley & Sons, 2008.

RIBEIRO, Soraia Pereira; COSTA, Ricardo Barros; DIAS, Clara Paz. Macrossomia neonatal: fatores de risco e complicações pós-parto. *NASCER E CRESCER-BIRTH AND GROWTH MEDICAL JOURNAL*, v. 26, n. 1, p. 21-30, 2017.

RUIZ CANCHUCAJA, Angela; CANO CARDENAS, Luis A. Maternal factors associated with fetal macrosomia according to the national survey of demographics and family health 2020. *Revista de la Facultad de Medicina Humana*, v. 22, n. 3, p. 489-496, 2022.

SANTOS, Bernardo José Moura Fonseca dos. *Modelagem do baixo peso ao nascer para bebês nascidos no Estado do Rio de Janeiro, em 2019*. Orientador: José Rodrigo de Moraes, 2021. 51f. TCC (Graduação) – Curso Estatística, Instituto de Matemática e Estatística, Universidade Federal Fluminense, Niterói, 2021. Disponível em: https://estatistica.uff.br/tcc-2020/

SANTOS, Hellen Geremias dos; NASCIMENTO, Carla Ferreira do; IZBICKI, Rafael; DUARTE, Yeda Aparecida de Oliveira; CHIAVEGATTO FILHO, Alexandre Dias Porto. Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. *Cadernos de Saúde Pública*, v. 35, p. e00050818, 2019. doi: 10.1590/0102-311X00050818

SOUZA, Julia 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*. Orientador: José Rodrigo de Moraes, Coorientadora: Patrícia Viana Guimarães Flores, 2021. 55f. TCC (Graduação) – Curso

Estatística, Instituto de Matemática e Estatística, Universidade Federal Fluminense, Niterói, 2021. Disponível em: https://estatistica.uff.br/tcc-2021/

TEVA G., María Jesús; REDONDO A., Rosario; RODRÍGUEZ G., Isabel; MARTÍNEZ C., Sara; ABULHAJ M., Mariam. Análisis de la tasa de detección de fetos macrosómicos mediante ecografía. *Revista chilena de obstetricia y ginecología*, v. 78, n. 1, p. 14-18, 2013.

VON LUXBURG, Ulrike; SCHÖLKOPF, Bernhard. Statistical learning theory: Models, concepts, and results. In: *Handbook of the History of Logic*. North-Holland, 2011. p. 651-706.

YUCRA, René Mamani; TUDELA, Luzbeth Lipa; HUANCA-AROHUANCA, Jesús Wiliam. Factores de riesgo materno perinatal asociados a macrosomía en recién nacidos en los hospitales EsSalud Juliaca-Puno. *Revista Científica de Salud UNITEPC*, v. 9, n. 1, p. 25-37, 2022.

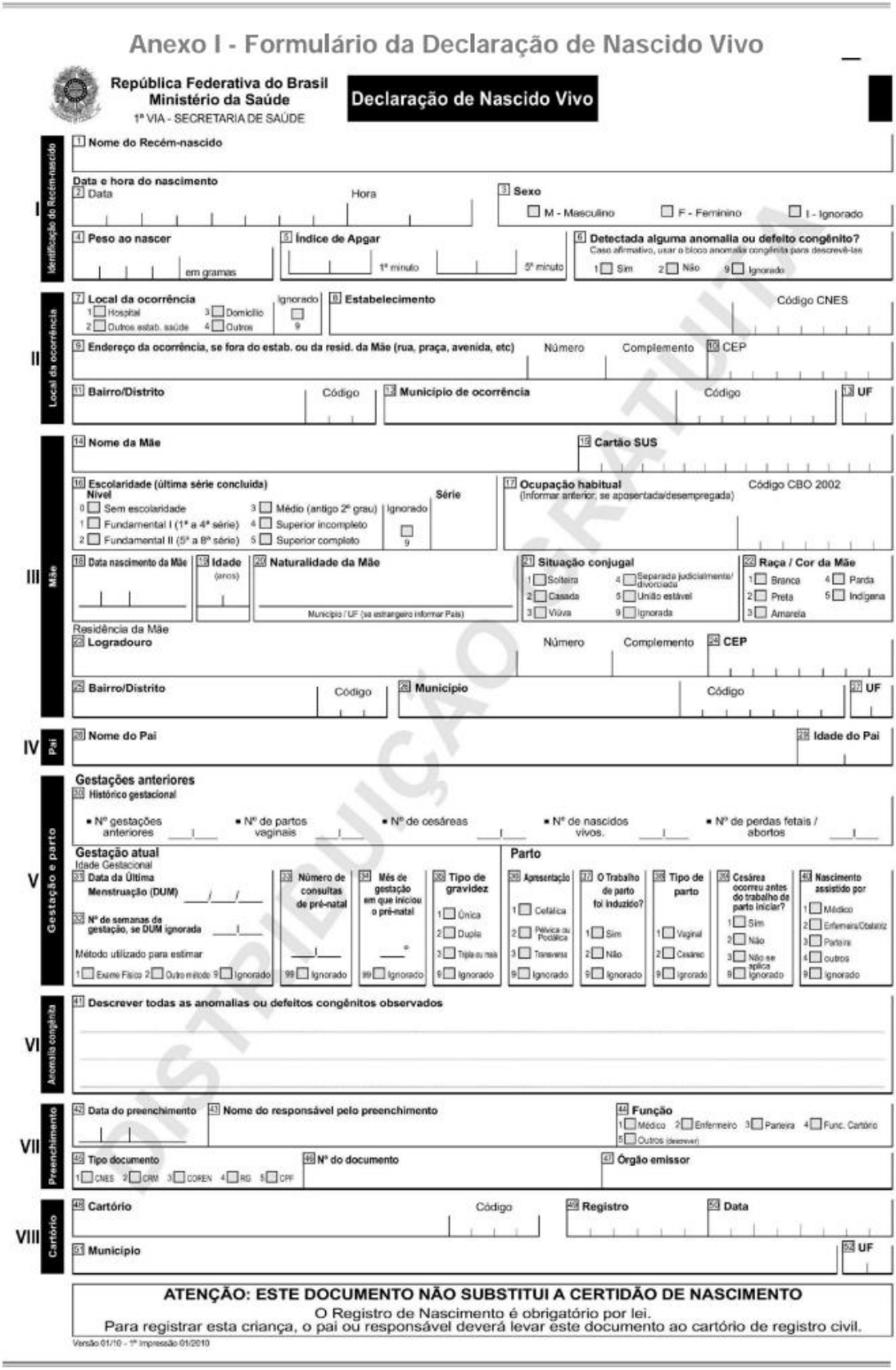
ZACHARAKI, Evangelia I.; WANG, Sumei; CHAWLA, Sanjeev; YOO, Dong Soo; Wolf, Ronald; MELHEM, Elias R.; DAVATZIKOS, Christos. Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, v. 62, n. 6, p. 1609-1618, 2009. doi: 10.1002/mrm.22147

ZEILEIS, A.; HOTHORN, T. Diagnostic Checking in Regression Relationships. *R News*, v. 2, n. 3, p. 7–10, 2002. https://CRAN.R-project.org/doc/Rnews/.

ZEILEIS, A.; KÖLL, S.; GRAHAM. N. Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R. *Journal of Statistical Software*, v. 95, n. 1, p. 1–36, 2020. doi:10.18637/jss.v095.i01.

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 I - Formulário 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 ou defeito congênito? (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

29 Nome do Pai | 28 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

35 Tipo de gravidez: 1 ☐ Única 2 ☐ Dupla 3 ☐ Tripla ou mais 9 ☐ Ignorado

Parto

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 ou defeitos congênitos observados

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 - 1ª Impressão 01/2010