UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

**PEDRO HENRIQUE GONCALVES SILVA**

# APLICAÇÃO DE MODELO REGRESSÃO LOGÍSTICA POLITÔMICA

PRESIDENTE PRUDENTE
2024

**PEDRO HENRIQUE GONCALVES SILVA**

# APLICAÇÃO DE MODELO DE REGRESSÃO LOGÍSTICA POLITÔMICA

Relatório para Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp para aproveitamento na disciplina TCC Ⅱ.
Orientadora: Prof. Dra. Miriam Rodrigues Silvestre.

PRESIDENTE PRUDENTE
2024

S586a

Silva, Pedro Henrique Goncalves
Aplicação de Modelo Regressão Logística Politômica / Pedro Henrique Goncalves Silva. -- Presidente Prudente, 2024
65 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientadora: Miriam Rodrigues Silvestre

1. Regressão Logística. 2. Apgar 5. 3. Dados Desbalanceados. 4. Modelos de Classificação. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

**TERMO DE APROVAÇÃO**

Pedro Henrique Goncalves Silva

APLICAÇÃO DE MODELO REGRESSÃO LOGÍSTICA POLITÔMICA

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora.

**Orientadora**: _____

**Profª. Drª. Miriam Rodrigues Silvestre**
**Departamento de Estatística**

_____

**Prof. Dr. Manoel Ivanildo Silvestre Bezerra**
**Departamento de Estatística**

_____

**Prof. Dr. Fernando Antônio Moala**
**Departamento de Estatística**

Presidente Prudente, 10 de Julho de 2024.

## RESUMO

O desenvolvimento deste estudo teve como objetivo a aplicação dos metodos de regressão logística politômica. Para isso, investigou-se a relação entre os resultados de Apgar 5 minutos (sistema de pontuação para avaliar a saúde de um recém-nascido nos 5 minutos de vida) e um conjunto de variáveis independentes como idade materna, complicações durante a gravidez, histórico médico entre outras. Essas variáveis estão disponíveis no Datasus em SINASC - Sistema de informação de Nascidos Vivos para o ano de 2022. A regressão logística permite identificar quais variáveis independentes têm um impacto significativo na probabilidade de um recém-nascido apresentar um determinado escore de Apgar de 5 minutos. Ao longo do estudo foram realizados tratamentos, análises dos dados e apresentação dos fundamentos teóricos, incluindo os pressupostos necessários para sua aplicação. Essa abordagem proporciona  uma interpretação relevante para auxiliar na identificação de fatores de risco e na implementação de medidas preventivas para melhorar a saúde dos recém-nascidos.

Palavras-chave: Regressão logística politômica; Apgar 5 min; fatores de risco.

**Abstract**

The development of this study aims to apply polytomous logistic regression methods. For this, we will investigate the relationship between the results of Apgar 5 minutes (assessment system to assess the health of a newborn in the 5 minutes of life) and a set of independent variables such as maternal age, complications during pregnancy, medical history among others. These variables are available on Datasus at SINASC - Live Birth Information System for the year 2022. Logistic regression allows identifying which independent variables have a significant impact on the likelihood that a newborn will have a given 5-minute Apgar score. Throughout the study, treatments, data analysis and presentation of theoretical foundations will be carried out, including the necessary orders for its application. This approach provides a relevant interpretation to help identify risk factors and implement preventive measures to improve the health of newborns.

Keywords: Polytomous logistic regression; Apgar 5 min; risk factors.

## LISTA DE FIGURAS

## LISTA DE TABELAS

## SUMÁRIO

# 1 Introdução

O modelo de regressão logística é um modelo amplamente conhecido e utilizado em diversas áreas do conhecimento, tais como: previsão de risco na área tributária – calcular a probabilidade do contribuinte ser inadimplente ou adimplente, risco de crédito para decisão de aprovação de crédito ou negação. Sendo uma técnica mais conhecida para modelos onde a variável dependente é sim ou não, como mostrado nos exemplos acima e sendo pouco estudada para eventos onde a variável dependente é politômica, ou seja, mais de uma opção de escolha como exemplo pode-se citar a escolha da marca de um produto Nike, Puma ou Adidas; para predizer a probabilidade do resultado de uma partida de futebol acabar como derrota, vitória ou empate.

Em Hosmer et al. (2013) é dito que o modelo de regressão logística tem maior influência quando precisamos identificar os fatores sobre uma resposta dicotômica ou politômica, sendo para variáveis independentes categóricas ou não categóricas, na qual é estimado a probabilidade do evento estudado acontecer, através de sua razão de chances.

## 1.1 Objetivos

O objetivo geral deste trabalho é realizar aplicação do modelo de regressão logística politômica em situações em que a variável resposta dependente é politômica, levando em consideração outras variáveis independentes, sejam elas quantitativas ou qualitativas. Em seguida, pretende-se estimar a razão de chances de ocorrência do evento específico em relação à aplicação futura da regressão, realizar o teste dos coeficientes e fornecer a interpretação dos resultados.

O objetivo específico a ser alcançado:

- Aplicação do modelo de regressão logística politômica e interpretação dos resultados para base de dados do SINASC - Sistema de informação de Nascidos Vivos, para a variável resposta Apgar 5 min.

## 1.2 Justificativa

A regressão logística politômica ou multinomial possui vantagens em relação a outras regressões lineares quando se trata de dados categóricos. Nos permite estimar a probabilidade da variável dependente assumir um determinada razão de chances do evento ocorrer com base em seus coeficientes. Os resultados são apresentados em termos de probabilidade no intervalo de 0 a 1 e possibilitam a classificação de indivíduos em categorias.

A regressão logística é usualmente utilizada para variáveis dependentes dicotômicas (0 = “Sim”, 1 = “Não”), e seus métodos são pouco aplicados para quando queremos estudar para variáveis dependentes politômica $(0, 1, 2,\ldots k)$, sendo $k$ o $k$-ésimo termo da variável dependente. Por esse motivo, busca-se aprimorar os conhecimentos dentro da área da regressão logística politômica ou multinomial, a qual possui menos trabalhos realizados em comparação com a regressão logística dicotômica.

Portanto, a regressão logística politômica é alternativa para estudos que exigem a classificação em mais de duas respostas, realizando a interpretação dos resultados da regressão como parte integrante do estudo.

# 2 Referencial Teórico

## 2.1 Regressão Logística

Segundo Mesquita (2014), com o objetivo de realizar classificação de variáveis preditoras ou explicar seus eventos, surgiu o modelo de regressão logística descoberto por volta do século XIX, inicialmente utilizado para a área da saúde, mas com o tempo foi aplicado a diferentes áreas como “Modelagem com regressão logística para análise de concessão de crédito” Beserra (2012) e “Aplicação do Modelo de Regressão Logística num Estudo de Mercado” Cabral(2013).

A partir dos trabalhos de Cox & Snell (1989, apud MESQUITA, 2014) e Hosmer et al. (2013) o modelo de regressão logística teve avanços significativos, sendo mais conhecido pela aplicação *Framingham Heart Study* (1948), com o objetivo de identificar predisposição de problemas cardiovasculares, realizado com a colaboração da Universidade de Boston, conforme Corrar et al (2011).

O modelo de regressão logística é uma técnica amplamente utilizada para classificar objetos em duas ou mais categorias distintas. No entanto, além dessa aplicação comum, também é possível utilizá-la de outras formas. Existem dois tipos principais de regressão logística: binomial ou dicotômica e multinomial ou politômica. A regressão logística binomial é utilizada quando a variável resposta possui apenas duas categorias, enquanto a regressão logística politômica pode lidar com variáveis resposta de natureza ordinal ou nominal. Essas variantes permitem a análise e o ajuste de modelos que envolvem classificações com duas categorias, classificações com ordem e classificações com múltiplas categorias, respectivamente.

Mediante a técnica de regressão logística, é possível identificar as variáveis mais relevantes para prever a ocorrência de um evento de interesse, o que permite estimar de forma direta a probabilidade desse evento acontecer.

## 2.2 Apgar 5 min

O Apgar de 5 minutos, também conhecido como Apgar 5', é um método de avaliação médica usado para avaliar o bem-estar de um recém-nascido logo após o nascimento. Ele foi desenvolvido pela médica anestesiologista Virginia Apgar em 1952 e é amplamente utilizado em todo o mundo como uma ferramenta rápida e eficaz para avaliar o estado de saúde dos recém-nascidos.

O Apgar 5 recebe esse nome porque é realizado aproximadamente 5 minutos após o nascimento do bebê. Durante esse período, o médico ou profissional de saúde avalia cinco critérios fundamentais que fornecem uma visão abrangente do estado geral do recém-nascido. Esses critérios são:

Frequência cardíaca: é verificado o ritmo e a taxa do batimento cardíaco do bebê. Uma frequência cardíaca adequada é um sinal de um sistema circulatório saudável.

Respiração: é observado se o bebê está respirando de forma adequada e regular. A respiração é essencial para garantir a oxigenação adequada dos tecidos do corpo.

Tonicidade muscular: verifica-se o tônus muscular do bebê, ou seja, a sua capacidade de realizar movimentos e se estender ativamente.

Reflexos: são avaliados os reflexos do bebê, como o reflexo de sucção ou o reflexo de retração dos membros em resposta a estímulos.

Coloração da pele: é observada a cor da pele do bebê, que deve ser rosada, indicando uma boa circulação sanguínea e oxigenação adequada.

Cada um desses critérios é avaliado em uma escala de 0 a 2, totalizando um máximo de 10 pontos no escore 'Apgar 5'. Quanto maior o escore, melhor é a condição do recém-nascido.

Segundo Escala (2024), o critério de pontuação dos casos de Apgar de 5 minutos segue a seguinte avaliação.

**Tabela 1 - Pontuação de Apgar**

| Sinal | 0 pontos | 1 ponto | 2 pontos |
|---|---|---|---|
| Cor da pele | Pele azulada ou pálida | Pele rosada, mas mãos ou pés azulados | Pele rosada |
| Frequência cardíaca | Sem batimentos | Menor que 100 batimentos por minuto | Maior que 100 batimentos por minuto |
| Resposta ao estímulo | Não há resposta | Faz caretas quando estimulado | Chora, espirra ou tosse quando estimulado |
| Tônus muscular | Músculos flácidos, sem movimento | Dedos, braços e pernas dobrados, com pouco movimento | Movimenta ativamente |
| Respiração | Ausente | Choro fraco, respiração lenta ou irregular | Choro vigoroso |

Fonte: Escala (2024)

O Apgar 5 é uma ferramenta valiosa para identificar rapidamente qualquer problema potencial de saúde no recém-nascido e tomar as medidas necessárias para garantir seu bem-estar. No entanto, é importante lembrar que o escore Apgar não é uma medida definitiva do futuro desenvolvimento ou saúde do bebê, mas sim uma avaliação inicial que auxilia os profissionais de saúde no cuidado imediato após o nascimento, Lourenço (2022).

## 3 Metodologia

O objetivo principal do estudo será a aplicação do modelo de regressão logística politômica, mais de duas categorias de variável resposta para o modelo, seguindo os métodos estatísticos utilizados na “Regressão logística politômica: revisão teórica e aplicações” Bittencourt (2003).

Dados disponíveis no SINASC - Sistema de informação de Nascidos Vivos. Sobre Apgar 5 min, que são avaliações realizadas em recém nascidos referente a frequência cardíaca, respiração, tonicidade muscular, reflexos e coloração da pele. Cada avaliação tem pontuação de 0 a 2 pontos, depois das 5 avaliações são somadas e tem pontuação de 0 a 10 pontos.

Para a construção do modelo a variável resposta Apgar 5 será divida em três intervalos: 0 a 3, 4 a 7 e 8 a 10. Essa transformação pode ser realizada atribuindo uma nova variável categórica aos dados, onde cada intervalo é representado por uma categoria específica, a divisão dessas categorias é bastante usual na área da saúde. Será realizado o tratamento das variáveis, transformado para variáveis dummy e análise exploratória dos dados no Software *Rstudio*

Realizados toda a preparação para a base de dados, realizamos a regressão logística politômica utilizaremos o Software *Rstudio* para aplicação de técnicas de regressão logística para determinar de maneira mais precisa quais variáveis influenciam na avaliação do bebe recém nascido nos primeiros 5 minutos de vida. Seguindo a metodologia no livro *Applied Logistic Regression de Hosmer et al.* (2013) e estudos de artigos relacionados ao assunto proposto como e “Modelo de decisão sobre o uso de preservativos: Uma regressão logística multinomial” (Batista de Lima, et al., 2016), para aplicação das técnicas estatísticas e acompanhamento para o tema, Fatores associados à morte infantil evitável: uma regressão logística múltipla. (Vidal e Silva, et al., 2018).

# 4 Modelo de Regressão Logística

## 4.1 Regressão Logística Dicotômica Simples

O modelo de Regressão Logística Simples segundo Mesquita (2014) pode ser descrito como técnica que permite explicar a relação entre uma variável dependente dicotômica com “falha” ou “sucesso”, “positivo” ou “negativo”, “sim” ou “não”, “aceitar” ou “rejeitar”, “certo” ou “errado”, “viver” ou “morrer”, “homem” ou “mulher, com outra variáveis independentes tanto qualitativa como quantitativa .

Seja $Y$ a variável resposta que pode assumir dois valores, representados por sucesso $(Y = 1)$ e fracasso $(Y = 0)$.

O valor esperado de $Y$ é dado por:

$$E(Y) = P(Y = 1) = \pi(x). \tag{4.1}$$

A distribuição condicional da variável resposta $Y$ segue uma distribuição binomial e sua média condicional é

$$\pi(x) = E(Y/x). \tag{4.2}$$

Temos que a forma do modelo de Regressão Logística é dado da seguinte forma:

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x_1)}}{1+e^{\beta_0+\beta_1 x_1}}. \tag{4.3}$$

Cabral (2013) usamos a transformação Logito com objetivo de fornecer uma representação linear da relação entre as variáveis independentes e a probabilidade de ocorrência de um evento do modelo de regressão logística. Para isso aplicamos o logaritmo e interpretamos os coeficientes estimados como mudanças nas chances do evento ocorrer e entender como as variáveis independentes afetam essa probabilidade, podendo o valor médio assumir valores dentro do intervalo $(-\infty, +\infty)$.

Temos a seguinte expressão com o logaritmo aplicado:

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1. \quad (4.4)$$

Realizando as operações e substituindo $\pi(x)$, temos a nova forma do modelo logito.

$$g(x) = ln\left(\frac{\frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}}{1-\frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}}\right). \quad (4.5)$$

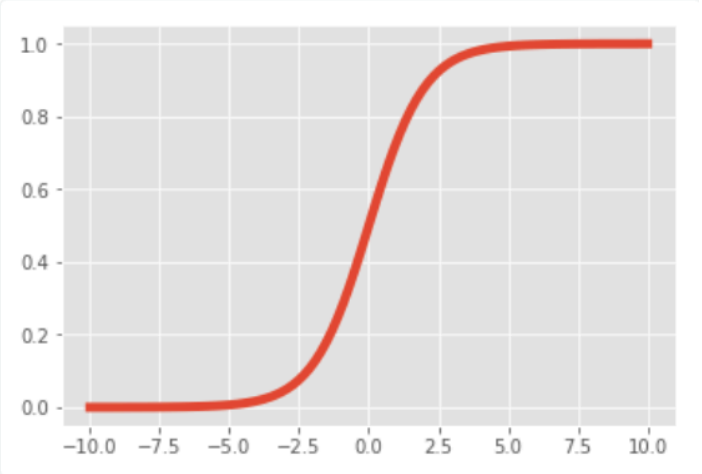
Transformação logit para $\pi(x)$:

$$g(x) = ln\left(e^{\beta_0+\beta_1 x}\right) = \beta_0 + \beta_1. \quad (4.6)$$

Uma das disponibilidades da regressão logística é sua função de comportamento probabilístico no formato de letra “S”, nos valores marcantes no 0 ou 1,que representa a probabilidade de ocorrência do evento de interesse, onde os valores mais perto de 0 indica uma menor probabilidade e os valores mais próximos de 1 indicam uma maior probabilidade de ocorrência do evento de interesse.

Como a função pode ser observada no gráfico da figura 1:

Figura 1 – Gráfico da Função Inversa




Fonte: https://matheusfacure.github.io/2017/02/25/regr-log/

### 4.1.2 Estimação dos Coeficiente Regressão Logística Dicotômica

Na Regressão logística binária a variável $Y$ resposta segue uma distribuição Bernoulli e para a estimação dos coeficientes de $\beta_1$e $\beta_0$ desconhecidos do modelo é realizado o uso da máxima verossimilhança,com o objetivo de encontrar as estimativas mais prováveis dos coeficientes e maximizar a probabilidade de que um evento ocorra.

A variável $Y$ assume uma distribuição de Bernoulli:

$$f(y_i) = \pi(x_i)^{y_i}.[1 - \pi(x_i)]^{1-y_i}, y_i \ \epsilon \{0, 1\}. \tag{4.7}$$

Aplicamos o princípio da máxima verossimilhança supondo que os dados possuem independência dentre suas observações para estimar o valor de $\beta_i$ que maximiza $L(\beta_i)$ . Aplicando o logaritmo, a estimação pode ser obtida da seguinte forma:

$$L(y, \beta_i) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i} . \tag{4.8}$$

Podendo ser representada por:

$$l(y, \beta) = ln|L(\beta)|$$

$$l(\beta) = \sum_{i=1}^{n} \left\{ y_i ln\left[\pi(x_i)\right] + (1 - y_i) ln\left[1 - \pi(x_i)\right] \right\}$$

$$= \sum_{i=1}^{n} \left\{ y_i ln\left[\pi(x_i)\right] + ln\left(1 - \pi_i\right) - y_i ln\left[1 - \pi(x_i)\right] \right\}$$

$$= \sum_{i=1}^{n} \left| y_i ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] + ln\left[1 - \pi(x_i)\right] \right|. \quad (4.9)$$

Substituindo $g(x)$:

$$l(\beta) = \sum_{i=1}^{n} \left[ y_i(\beta_0 + \beta_1 x_1) + ln\left[\frac{1}{1+e^{\beta_0+\beta_1 x_1}}\right] \right]$$

$$= \sum_{i=1}^{n} \left[ y_i(\beta_0 + \beta_1 x_1) + ln(1) - ln\left(1 + e^{\beta_0+\beta_1 x_1}\right) \right]$$

$$= \sum_{i=1}^{n} \left[ y_i(\beta_0 + \beta_1 x_1) - ln\left(1 + e^{\beta_0+\beta_1 x_1}\right) \right]. \quad (4.10)$$

Derivamos $l(\beta)$ em relação a cada parâmetro para encontrar $\left(\beta_0, \beta_1\right)$, para encontrar o valor $\beta$ que maximiza $l(\beta)$.

Derivando a função de $\beta_0$:

$$\frac{\partial l(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} \left| y_i - \frac{1}{1+exp(\beta_0+\beta_1 x_1)} exp(\beta_0 + \beta_1 x_1) \right|$$

$$=$$

$$= \sum_{i=1}^{n} \left[ y_i - \pi(x_i) \right] = 0. \tag{4.11}$$

Derivando a função de $\beta_1$:

$$\frac{\partial l(\beta)}{\partial \beta_1} = \sum_{i=1}^{n} \left[ x_i y_i - \frac{1}{1+exp(\beta_0+\beta_1 x_1)} exp(\beta_0 + \beta_1 x_1) x_1 \right]$$

$$=$$

$$= \sum_{i=1}^{n} x_i \left[ y_i - \pi(x_i) \right] = 0. \tag{4.12}$$

Como as equações não são lineares em $\beta_0$ e $\beta_1$, são necessários métodos iterativos para resolução, estes disponíveis em vários programas computacionais segundo MEYER (1980).

## 4.2 Regressão Logística Dicotômica Múltipla

A regressão logística Múltipla é aplicada quando temos uma variável dependente $Y$ que é binária ou dicotômica e múltiplas variáveis independentes x. Uma das principais vantagens dessa técnica de modelagem é sua capacidade de lidar com várias variáveis independente representadas pelo vetor $x = (x_1, x_2, \ldots, x_p)$ supondo que temos uma coleção de $p$ variáveis, mesmo que elas estejam em diferentes escalas de medida, podemos incluir no modelo diferentes tipos de variáveis, independentemente de suas escalas de medida.

Temos que a forma do logit para a Regressão Logística Múltipla é dada pela equação:

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p. \quad (4.13)$$

Considerando um conjunto de variáveis independentes, denotadas por um vetor $x = \left(x_1, x_2, \ldots, x_p\right)$.

Temos que $E(Y/x) = \pi(x)$ é expressa por :

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}. \quad (4.14)$$

Segundo Hosmer et al. (2013), às variáveis independentes que assumem valores discretos ou nominal, são inapropriadas no modelo, porque seus números colocados para representar as variáveis não são significativos sendo como uma demonstração. Nesta situação, o método de escolha é usar uma coleção de design variables (or dummy variables). exemplo: Codificação da Variável de design para raça, codificadas em três níveis, quando o respondente é “branco”, as duas variáveis de design, $D_1$ e $D_2$, ambos seriam iguais a zero; quando o respondente é “preto”, $D_1$ ser definido igual a 1 enquanto $D_2$ ainda seria igual a 0; quando a raça do entrevistado é “outro”, usamos $D_1$ = 0 e $D_2$ = 1. A Tabela ilustra essa codificação de variáveis de projeto.

**Tabela 2 - Codificação da Variável de design para raça, codificadas em três níveis.**

| Raça | $D_1$ | $D_2$ |
|---|---|---|
| Branca | 0 | 0 |
| Preto | 1 | 0 |
| Outra | 0 | 1 |

Temos que com o uso de variáveis dummy, a equação do Logito é representada da seguinte forma:

$$g(x) = \beta_0 + \beta_1 x_1 + \ldots + \sum_{l=1}^{m-1} \beta_{jl} D_{jl} + \beta_p x_p \tag{4.15}$$

As variáveis de projeto $m - 1$ serão denotadas como $D_{jl}$ e os coeficientes para essas variáveis de projeto serão denotados como $\beta_{jl}$, $l$= 1, 2, ... ,$m - 1$.

## 4.3 Regressão Logística Politômica

Segundo Hosmer et al. (2013), a variável resposta classificada em três ou mais categorias diferentes $Y = y_1, y_2, \ldots, y_k$, possui uma das exigências para aplicação de um modelo de regressão logística politômica, diferente do modelo de regressão logística dicotômica para duas variáveis respostas.

O modelo logístico com variável resposta politômica é uma abordagem em que ajustamos simultaneamente $k - 1$ categorias de variável resposta $Y$. Obtemos o conjunto de parâmetros $\beta_i = \left[\beta_1, \beta_2, \ldots, \beta_p\right]$ correspondentes a categoria da variável Y de referência sendo $p$ o número de covariáveis.

Segundo Bittencourt (2003) para desenvolver o modelo logit, temos $p$ covariáveis e a constante denotado pelo vetor $x = \left(x_0, x_1, x_2, \ldots, x_p\right)$, de dimensão $p + 1$ onde $x_0 = 1$.

Temos que a função logit assumindo o nível $k$, para classes como base $y_k$ , é dada por:

$$g_i(x) = ln\left[\frac{P(Y=y_i|X)}{P(Y=y_k|X)}\right] = \beta_{i0} + \beta_{i1}x_1 + \beta_{i2}x_2 + \ldots + \beta_{ip}x_p \,. \tag{4.16}$$

E sendo $i$ para $i = 1, 2, \ldots, k-1$

Com a probabilidade condicional na forma geral:

$$\pi_j(x) = P(Y = j/x) = \frac{e^{g_j(x)}}{\sum_{k=0}^{i} e^{g_k(x)}}. \tag{4.17}$$

## 4.4 Regressão Logística Politômica Simples

Para realizar o modelo para o caso simples devemos utilizar uma das categorias da variável resposta $Y$ como referência para as outras variáveis respostas. No caso de existir 3 categorias para a variável resposta $Y = (y_1, y_2, y_3)$. Temos o logit para o caso de $Y$ para $k = 3$ variáveis categóricas assumindo $y_1 = 0$, $y_2 = 1$ e $y_3 = 2$, que o modelo de regressão logística politômica terá 2 funções logit segundo Hosmer et al. (2013), para este caso, é necessário assumir uma categoria como base para o logit, neste caso foi assumido o nível $Y = 0$ como a razão entre $Y = 1$ e $Y = 0$ e a razão entre $Y = 2$ e $Y = 0$:

Temos o 1° logit para $Y = 1$ :

$$g_1(x) = ln\left[\frac{P(Y=1|X)}{P(Y=0|X)}\right] = \beta_{10} + \beta_{11}x. \tag{4.18}$$

Temos o 1° logit para $Y = 2$ :

$$g_2(x) = ln\left[\frac{P(Y=2|X)}{P(Y=0|X)}\right] = \beta_{20} + \beta_{21}x. \tag{4.19}$$

## 4.5 Regressão Logística Politômica Múltipla

Neste exemplo temos o logit para o caso Politômica Múltipla de $Y$ para variáveis categóricas assumindo $Y = 0$, $Y = 1$ e $Y = 2$. O modelo de regressão logística politômica terá $K = 2$ funções logit, assumindo uma categoria como base para o logit, neste caso foi assumido o $Y = 0$ como base e $p$ para o n-ésimo termo de variável independente.

Temos o 1° logit para $Y = 1$ :

$$g_1(x) = ln\left[\frac{P(Y=1|X)}{P(Y=0|X)}\right] = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \ldots + \beta_{1p}x_p. \quad (4.20)$$

Temos o 1° logit para $Y = 2$ :

$$g_2(x) = ln\left[\frac{P(Y=2|X)}{P(Y=0|X)}\right] = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \ldots + \beta_{2p}x_p. \quad (4.21)$$

## 4.6 Estimação dos coeficientes Regressão Logística.

### 4.6.1 Estimação dos coeficientes Regressão Logística Dicotômica Múltipla.

Estimação dos parâmetros da regressão logística Dicotômica, pelo método da máxima verossimilhança para um vetor $\beta^t = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ . Supondo que os dados possuem independência dentre suas variáveis.

$$L(\beta) = ln[L(\beta)] = \sum_{i=1}^{n} \{y\, ln[\pi(x)] + (1 - y) ln[1 - \pi(x)]\}$$

$$= \ldots = \sum_{i=1}^{n} \left[\left(y\,\beta_0 + y\beta_1 x_1 + y\beta_2 x_2 + \ldots + y\beta_p x_p\right) - ln(1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p})\right] \quad (4.22)$$

Após a realização da derivação da função log de verossimilhança, é obtido $p$ + 1 equações referentes aos $p$ + 1 coeficientes. Temos então:

$$\sum_{i=1}^{n}\left\{\left(y_i - \widehat{\pi}_i\left(x_i\right)\right)\right\} = 0 \quad \text{e} \tag{4.23}$$

$$\sum_{i=1}^{n}\left\{x_{ij}\left(y_i - \widehat{\pi}\left(x_i\right)\right)\right\} = 0\text{. para j = 1,2,....,p.} \tag{4.24}$$

Realizando o cálculo da estimação dos parâmetros através de softwares estatísticos, temos a estimação dos parâmetros, Logito e $\pi(x)$, temos então:

$$\widehat{\pi}(x) = \frac{e^{\beta_{20}+\beta_{21}x_1+\beta_{22}x_2+\ldots.+\beta_{2p}x_p}}{1+e^{\beta_{20}+\beta_{21}x_1+\beta_{22}x_2+\ldots.+\beta_{2p}x_p}}. \tag{4.25}$$

$$\widehat{g}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_p x_p. \tag{4.26}$$

### 4.6.2 Estimação dos coeficientes Regressão Logística Politômica.

Como na Regressão Logística Dicotômica a Politômica, também é utilizado o método máxima verossimilhança para estimação dos coeficientes do modelo. Segundo Hosmer et al. (2013) Para construir a função de verossimilhança, criamos três variáveis binárias codificadas como 0 ou 1 para indicar a participação no grupo de uma observação. Notamos que essas variáveis são introduzidas apenas para esclarecer a função de verossimilhança e não são usados na análise de regressão logística multinomial. As variáveis são codificadas da seguinte forma:$Y = 0$ então $Y_0 = 1, Y_1 = 1$, e $Y_2 = 0$ e $Y = 0$ então $Y_0 = 0$, $Y_1 = 0$, e $Y_2 = 0$; e se $Y = 2$ então $Y_0 = 0$, $Y_1 = 0$, e $Y_2 = 1$. Usando a notação segue que a função de verossimilhança condicional para uma amostra de n independentes observações é:

$$l(\beta) = \prod_{i=1}^{n}\left[\pi_0(x_i)^{y0i}\pi_1(x_i)^{y1i}\pi_2(x_i)^{y2i}\right]. \tag{4.27}$$

A função do log:

$$L(\beta) = \sum_{i=1}^{n} y_{1i}g_1(x_i) + y_{2i}g_2(x_i) - ln(1 + e^{g1(x_i)} + e^{g2(x_i)}). \tag{4.28}$$

A forma geral da equação:

$$\frac{\partial L(\beta)}{\partial \beta_{jk}} = \sum_{i=1}^{n} x_{ki}(y_{ji} - \pi_{ji}). \text{ de } j = 1, 2 \text{ e } k = 1, 2,..., p \text{ , com } x_{0i} = 1 . \tag{4.29}$$

Segunda derivada parciais.

$$\frac{\partial^2 L(\beta)}{\partial \beta_{jk} \partial \beta_{jk`}} = \sum_{i=1}^{n} x_{k`i} x_{ki} \pi_{ji}(1 - \pi_{ji}). \tag{4.30}$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_{jk} \partial \beta_{j`k`}} = \sum_{i=1}^{n} x_{k`i} x_{ki} \pi_{ji} \pi_{j`i} . \text{de} j \; e \; j` = 1, 2 \text{ e } k \; e \; k` = 1, 2, \ldots, p. \tag{4.31}$$

### 4.6.3 Teste Da Razão de Verossimilhança

No modelo de regressão Multinomial, o ajuste do modelo também está baseado na comparação dos valores observados da variável resposta com os respectivos valores preditos pelo modelo ajustado, e esta comparação é feita para cada uma das funções Logit, de maneira análoga ao da Regressão logística Binária Hosmer et al. (2013).

Segundo Bittencourt (2003) depois de obter o modelo, é necessário realizar o teste da máxima verossimilhança, onde a hipótese de que há pelo menos um dos parâmetros $\beta_{ij}$ diferente de zero.

Hipótese:

$$H_0 : \beta_{ij} = \ldots = \beta_{kp} = 0$$

$$H_a : \beta_{ij} \neq \ldots \neq \beta_{kp} \neq 0$$

Para isso encontramos a estatística D chamada de Deviance, uma medida que avalia a qualidade do ajuste do modelo.

Expressa da seguinte forma:

$$D = -2ln\left[\frac{Função\ de\ máxima\ verossimilhança\ do\ modelo\ corrente}{Função\ de\ máxima\ verossimilhança\ do\ modelo\ saturado}\right]. \quad (4.32)$$

Para testar a significância do coeficiente de uma variável no modelo comparando os valores observados da variável resposta com os preditos, para testar a hipótese nula do efeito das covariáveis para o efeito do modelo.

O modelo é considerado saturado quando possui todas as variáveis disponíveis, enquanto um modelo ajustado é aquele que contém variáveis significantes para o estudo. Quanto menor o deviance indica que o modelo se ajusta de forma mais precisa aos dados sendo considerado um bom ajuste ao modelo.

Estatística do teste:

$$G = -2ln\left[\frac{modelo\ corrente}{modelo\ saturado}\right] \sim sob\ H_0\ \chi^2(p-1). \quad (4.33)$$

Ao rejeitarmos a hipótese nula, assumimos estatisticamente que existe ao menos um coeficiente diferente de zero. Sendo significativos para o nosso modelo

### 4.6.4 Teste de Wald

O teste de Wald é um teste de hipótese usado na estatística para avaliar se um ou mais parâmetros de um modelo são significativamente diferentes de um valor específico de maneira individual . Utilizado em modelos de regressão para testar a significância individual dos coeficientes de regressão.

A distribuição de $Wald$ possui distribuição Normal, realizando a verificação para cada covariável sendo a hipótese nula o coeficiente particular de $\beta_{ij}$ é igual a zero.

Hipótese:

$$H_0: \beta_{ij} = 0 \quad vs \qquad H_a: \beta_{ij} \neq 0$$

Sendo $i = 1,\ldots,k$ para a quantidade de variáveis $logit$ e $j = 1,\ldots,p$ para quantidade de variáveis independentes.

Estatística do teste definida como a estimativa de máxima verossimilhança para o coeficiente e seu respectivo erro-padrão (EP)

Estatística do teste:

$$W_j = \left( \frac{\hat{\beta}_{ij}}{EP\left(\hat{\beta}_{ij}\right)} \right)^2 \qquad \sim \quad \chi^2_1 \, gl. \tag{4.34}$$

### 4.6.5 Análise de Desempenho

A Matriz de Confusão é uma tabela que traz frequências referentes a cada classe preditiva e observada do modelo, e possibilita o desenvolvimento de algumas métricas de desempenho.

Segundo Guanis et al. (2022) para dados binários, onde existem apenas duas possibilidades de classificação (verdadeiro positivo e verdadeiro negativo), a matriz de confusão é bastante simples, como mostrado na tabela 3.

**Tabela 3 - Matriz de Confusão 2x2**

| | Predição | |
|---|---|---|
| Referência | Verdadeiro Positivo (VP) | Falso Negativo (FN) |
| | Falso Positivo (FP) | Verdadeiro Negativo (VN) |

Fonte: Própria (2024)

Esta matriz de confusão 2x2 é a clássica tabela de contingência utilizada em diagnósticos de exames clínicos, onde os resultados podem ser VP (verdadeiro positivo), VN (verdadeiro negativo), FP (falso positivo) e FN (falso negativo).

Para o caso quando temos três possibilidades de classificação utilizamos a matriz de confusão realizada na Tabela 4.

**Tabela 4 - Matriz de confusão 3x3**

| | | Previsão | | |
|---|---|---|---|---|
| | | Positivo | Negativo | Neutro |
| Referência | Positivo | Verdadeiro Positivo (VP) | Falso Negativo (FN) | Falso Neutro (FNeu) |
| | Negativo | Falso Positivo (FP) | Verdadeiro Negativo(VN) | Falso Neutro (FNeu) |
| | Neutro | Falso Positivo (FP) | Falso Negativo (FN) | Verdadeiro Neutro (VNeu) |

Fonte: Próprio (2024)

Segundo o Medium (2022), com a matriz de confusão visualizamos de forma tabular a quantidade de erros e acertos do modelo. Podendo assim tirar as informações necessárias para realização das métricas de Acurácia, Recall e Precisão.

**Acurácia** indica a precisão do modelo em todas as suas previsões, sendo expressa pela razão entre o total de previsões corretas e o total de todas as previsões.

$$ACC = \frac{VP+VN+VNeu}{FP+FN+FNeu+VP+VN+VNeu} \times 100. \quad (4.35)$$

**Recall**, também conhecido como sensibilidade, representa a proporção de eventos corretamente identificados pelo modelo, medindo sua habilidade em prever verdadeiros positivos. É calculado pela razão entre verdadeiros positivos e a soma de verdadeiros positivos e falsos negativos.

*Caso positivo*:

$$Precision \ = \frac{VP}{VP+FN+FNeu} \times 100. \tag{4.36}$$

*Caso Negativo*:

$$Precision \ = \frac{VN}{VN+FP+FNeu} \ \times 100 \tag{4.37}$$

*Caso Neutro*:

$$Precision \ = \frac{VNeu}{VNeu+FP+FNeu} \ \times 100. \tag{4.38}$$

**Precisão** é a medida da proporção de eventos corretamente classificados de uma categoria em relação ao total de classificações feitas para essa mesma categoria. É representada pela razão entre verdadeiros positivos e a soma de verdadeiros positivos e falsos positivos.

*Caso positivo:*

$$Precision \ = \frac{VP}{VP+FP} \ \times 100. \tag{4.39}$$

*Caso Negativo:*

$$Precision \ = \frac{VN}{VN+FN} \ \times 100. \tag{4.40}$$

*Caso Neutro*:

$$Precision \ = \frac{VNeu}{VNeu+FNeu} \ \times 100. \tag{4.41}$$

Temos que o cálculo de métricas vai nos mostrar como está o desempenho do nosso modelo para multiclasses.

### 4.6. 6 Curva Receiver Operating Characteristic (ROC)

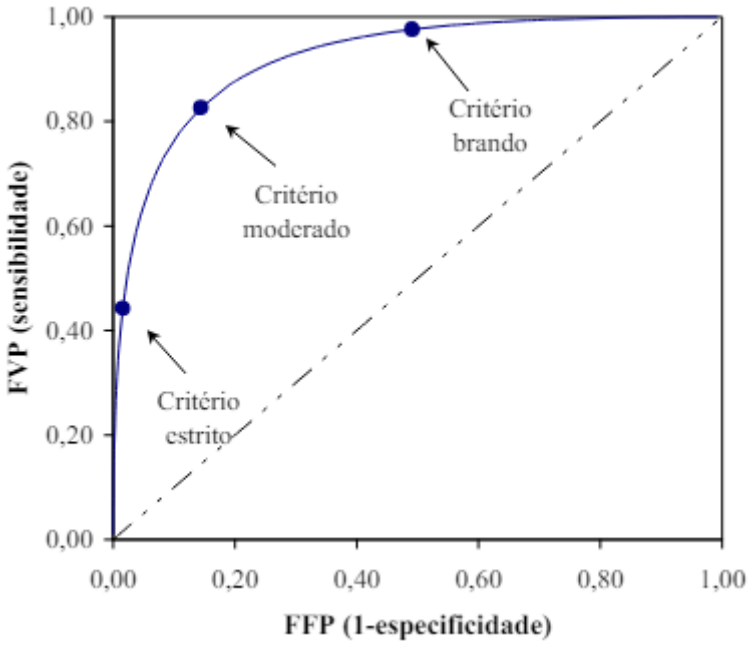
Depois de realizado os teste dos parâmetros dos modelos, vamos construir e analisar a curva ROC e a área sob a curva (AUC), que através dessas métricas para a comparação dos modelos encontrados é possível identificar qual possui maior desempenho.

A curva ROC mostra a relação entre a taxa de verdadeiros positivos (sensibilidade) e a taxa de falsos positivos (1 - especificidade) em diferentes pontos de corte para a classificação.

Segundo Braga(2000) A sensibilidade é a taxa de verdadeiros positivos, ou seja, mede a capacidade do modelo de prever corretamente a ocorrência de um evento. A especificidade é a taxa de verdadeiros negativos, avaliando a capacidade do modelo de prever corretamente a não ocorrência de um evento.

Segundo Braga(2000) O ponto de corte, que otimiza a sensibilidade e a especificidade, é o ponto da curva mais próximo do canto superior esquerdo do gráfico. A Figura 2 mostra como esse gráfico se comporta em diferentes situações

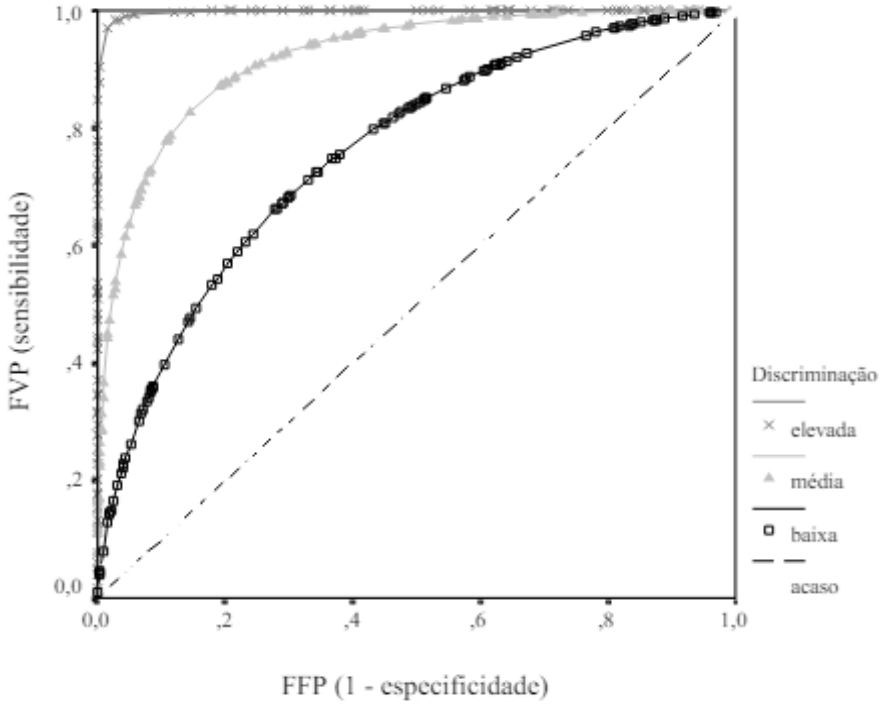
Figura 2 - Comportamento Curva Roc




Fonte: Braga (2000)

Cada ponto na curva ROC representa uma classificação diferente. E para um modelo com score alto terá uma curva ROC que passa pelo canto superior esquerdo do gráfico, indicando alta sensibilidade e baixa taxa de falsos positivos.

Figura 3 - Níveis Curva Roc




Fonte: Braga (2000)

A AUC é a área sob a curva ROC e fornece uma medida numérica da qualidade geral do modelo de classificação. Varia entre 0 e 1, sendo que um valor de 1 indica um modelo perfeito e um valor de 0,5 indica um modelo que não tem capacidade discriminatória melhor do que o acaso. A Tabela 5 mostra as faixas de classificação do ROC segundo Hosmer et al. (2013).

**Tabela 5 - Classificação curva ROC**

| Área sob a curva | Classificação |
|---|---|
| ROC =0.5 | Discriminação pobre. |
| 0.5< ROC <0.7 | Discriminação pobre, não muito melhor. |
| 0.7≤ ROC < 0.8 | Discriminação aceitável. |
| 0.8 ≤ ROC < 0.9 | Excelente discriminação. |
| ROC ≥ 0.9 | Discriminação excepcional. |

## 5 Dados Desbalanceados

Segundo Medium (2022) dados desbalanceados ocorrem quando há uma discrepância significativa no número de observações entre as diferentes classes. Isso pode ser observado em diversas áreas, como em estudos sobre doenças raras, análise de inadimplência de clientes e detecção de fraudes. No entanto, é importante ressaltar que o desbalanceamento pode prejudicar a eficácia dos modelos, pois as classes minoritárias, que representam a minoria dos dados, podem não ser adequadamente reconhecidas, como visualizado no gráfico desbalanceamento dos dados. Figura 4 ilustra essa metodologia.

Figura 4 - Gráfico de dispersão Desbalanceamento

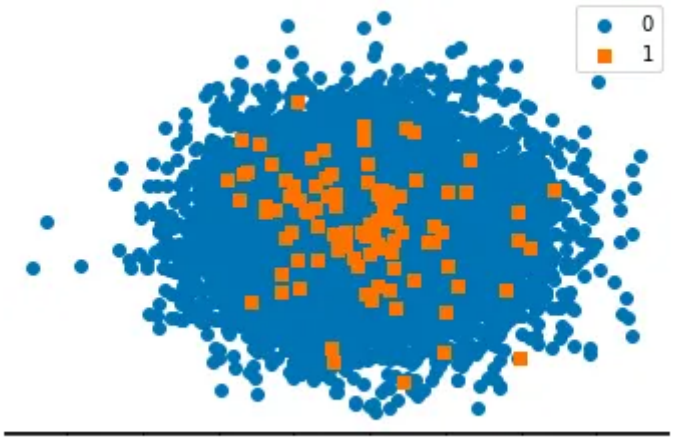



Gráfico de dispersão Desbalanceamento

**Fonte:** Medium (2022)

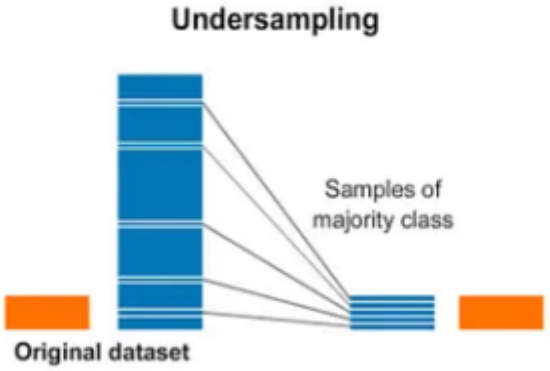
Para contornar essa questão, é necessário aplicar técnicas de balanceamento, sendo elas métodos de undersampling (subamostragem) e oversampling (sobreamostragem). Essas técnicas ajustam a distribuição dos dados para equilibrar as classes, garantindo que o modelo seja capaz de reconhecer adequadamente todas as classes presentes no conjunto de dados. Caso contrário,

a modelagem pode favorecer as classes majoritárias pela falta de informação das classes minoritárias, resultando em classificadores que não são capazes de acertar as previsões, podendo apresentar alta acurácia geral, mas falha em prever corretamente as classes minoritárias.

### 5.1 Método Undersampling

Segundo Medium (2022) Undersampling, também conhecido como método de subamostragem, reduz a quantidade de dados na classe majoritária para equilibrar as classes até que a proporção entre elas seja de 1:1. Isso pode ser feito de duas maneiras: removendo elementos aleatoriamente mas que resulta em uma perda maior de dados e a segunda maneira combinando uma ou mais observações da classe majoritária em uma única observação pelo método NearMiss e K-NN (algoritmo dos k-vizinhos mais próximos). Figura 5 ilustra essa metodologia.

Figura 5 - Dados balanceados pelo undersampling

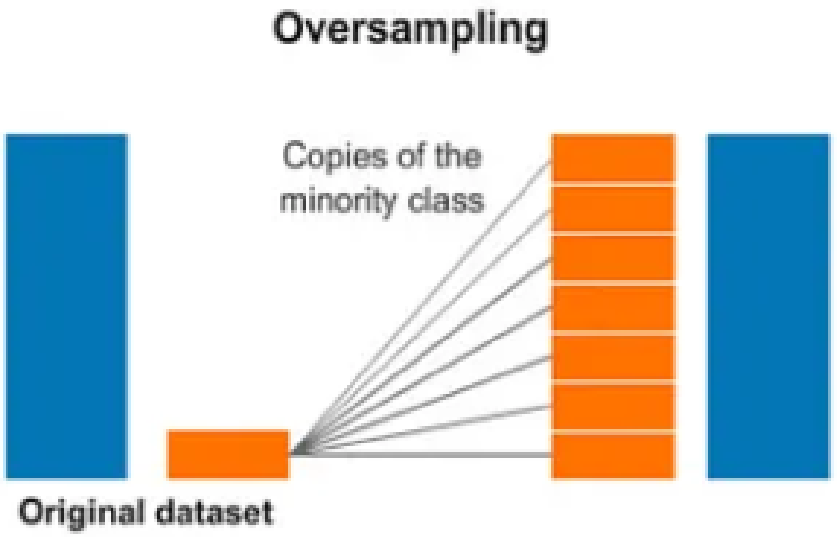



**Fonte:** Medium (2022)

## 5.2 Método Oversampling

Segundo Medium (2022) Oversampling envolve a criação de novas observações sintéticas na classe minoritária para equalizar a proporção das categorias. Uma abordagem inicial para realizar o Oversampling é duplicar dados já existentes na classe minoritária. Por exemplo, se nossos dados consistem em três configurações (x, y, z), o Oversampling poderia ser realizado selecionando algumas entradas e simplesmente aumentando o número de vezes que essas entradas são duplicadas. Essa abordagem pode fornecer resultados satisfatórios quando a classe minoritária apresenta pouca variação quantitativa em seus parâmetros. No entanto, caso haja uma ampla gama de variação, o modelo pode se tornar eficiente na identificação de casos específicos da classe minoritária em vez de reconhecer a categoria como um todo. Figura 6 ilustra essa metodologia.

Figura 6 - Dados balanceados pelo Oversampling.




**Fonte:** Medium (2022)

# 6 Análise dos Dados

## 6.1 Descrição dos Dados

Para o presente estudo, será utilizada a base de dados que está disponível no Datasus em SINASC - Sistema de informação de Nascidos Vivos, para o ano de 2022 para o estado de São Paulo. A Tabela 6 apresenta a descrição de algumas variáveis presentes no banco de dados, contabilizando o total de 13 variáveis independentes e a variável resposta APGAR 5 que será utilizada para realização do modelo.

**Tabela 6 - Descrição das variáveis**

| Variável | Descrição |
|---|---|
| IDADEMAE | Idade da mãe em anos. |
| QTDFILVIVO | Número de filhos vivos |
| QTDFILMORT | Número de filhos mortos. |
| GRAVIDEZ | Tipo de gravidez, conforme a tabela: 9: Ignorado 1: Única 2: Dupla 3: Tripla e mais |
| PARTO | Tipo de parto, conforme a tabela: 9: Ignorado 1: Vaginal 2: Cesáreo |
| CONSULTAS | Número de consultas de pré-natal: 1: Nenhuma 2: de 1 a 3 3: de 4 a 6 4: 7 e mais 9: Ignorado |
| SEXO | Sexo, conforme a tabela: 0: Ignorado 1: Masculino 2: Feminino |
| APGAR 1 | Apgar no primeiro minuto 00 a 10 |
| APGAR 5 | Apgar no quinto minuto 00 a 10 |
| PESO | Peso ao nascer, em gramas |
| IDANOMAL | Anomalia congênita: 9-Ignorado 1=Sim 2=Não |
| TPAPRESENT | Tipo de apresentação do RN. Valores: 1– Cefálico; 2– Pélvica ou podálica; 3– |

| | |
|---|---|
| | Transversa; 9– Ignorado. |

Fonte: Própria (2024)

Analisando a tabela 6, temos que algumas variáveis serão transformadas para dummy como, número de consultas de pré-natal, peso e a variável resposta Apgar de 5 minutos, que varia numa escala de 0 a 10 pontos. Para utilizar a variável resposta Apgar 5 min no modelo de regressão logística, é necessário realizar a classificação da variável. Para este caso vamos realizar a categorização para 3 resposta em Alto, Médio e Baixo. O nível Alto, possuiu um alto risco de complicação do bebe após o parto, seguindo o mesmo conceito para os casos de Médio e Baixo. O ideal para o nascimento é que o bebe nasça com baixo risco para não ocorrência de alguma possível complicação após o nascimento.

Para a variável peso, sendo uma característica muito importante dos bebês ao nascer, optou-se por realizar a categorização de acordo com o Método abordado Canguru (2016).

**Tabela 7 - Classificação do Peso Recém nascidos.**

| Classificação | Peso em Gramas (g) |
|---|---|
| EXTREMO | 0 a 999 |
| MUITO BAIXO | 1000 a 1499 |
| BAIXO | 1500 a 2499 |
| NORMAL | 2500 a 3999 |
| GRANDE | 4000 ou superior |

Tabela 8, mostra os intervalos e classificações em 3 níveis baixo, médio e alto risco. Os valores da avaliação do nível de adaptação do bebe recém nascido seguem respectivamente a pontuação de 8 a 10 uma boa vitalidade, 4 a 6 asfixia moderada, 0 a 3 asfixia grave.

Na qual os valores dos intervalos dos riscos estão classificados logo abaixo.

**Tabela 8 - Classificação Apgar 5 minutos**

| Alto Risco | Médio Risco | Baixo Risco |
|---|---|---|
| 0 a 3 | 4 a 7 | 8 a 10 |

Fonte: Própria (2024)

A Construção do modelo de regressão logística politômica será realizada considerando a classe Alto risco como faixa de referência para as outras variáveis.

**Tabela 9 - Classificação dos níveis de risco**

| Alto Risco | Médio Risco | Baixo Risco |
|---|---|---|
| $y = 0$ | $y = 1$ | $y = 2$ |

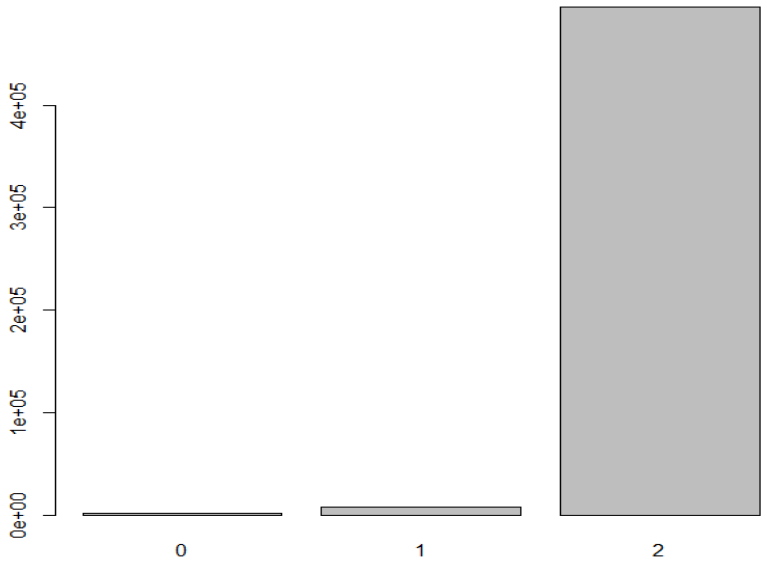
Fonte: Própria (2024)

## 6.2 Tratamento dos Dados

Tratando dados ausentes, optou-se por uma abordagem de exclusão de 6.252 das observações que contenham pelo menos uma variável com dados ausentes. Essa decisão foi tomada devido à insignificância dos valores ausentes em relação ao tamanho expressivo de 525.239 da base de dados do Datasus utilizada no estudo. Antes da exclusão, foi realizada uma análise detalhada da quantidade de dados ausentes em cada variável para garantir que a exclusão não introduzisse viés nos resultados. Para a análise foi usado 505.084 dados, devido a retirada de casos ausentes e ignorados da base de dados.

Foi observado que grande maioria das observações de Apgar aos 5 minutos correspondia a Baixo Risco (8 a 10), com poucos casos de Médio e Alto Risco.

Figura 9 - Gráfico para Apgar 5 minutos




Fonte: Própria (2024)

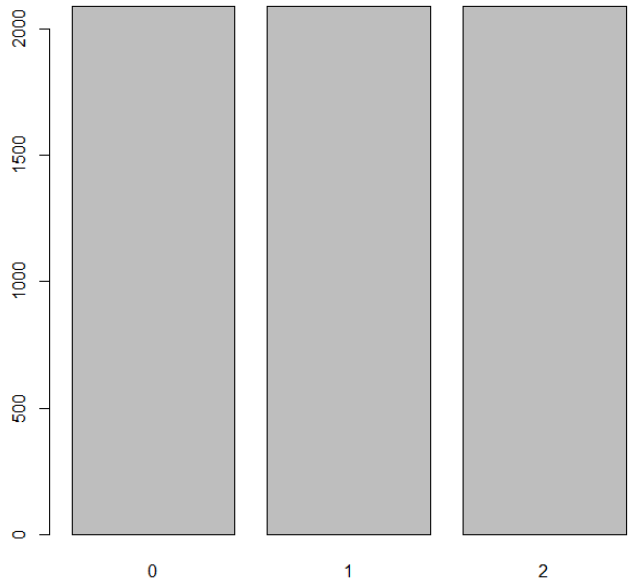
**Tabela 10 - Total para caso de Apgar 5 Minutos**

| Alto Risco | Médio Risco | Baixo Risco | Total |
|---|---|---|---|
| 2.089 | 7.310 | 495.685 | 505.084 |

Fonte: Própria (2024)

Para o desenvolvimento dos modelos, devido ao desbalanceamento, optou-se por aplicar o método de undersampling devido os dados de Alto Risco terem a menor quantidade. Neste método, as amostras da classes majoritárias (Baixo e Médio risco de Apgar) foram reduzidas para equilibrar com a classe minoritária (Alto risco de Apgar), garantindo que todas as classes tivessem a mesma quantidade de amostras.

Figura 7 - Gráfico com balanceamento Undersampling




Fonte: Própria (2024)

**Tabela 11 - Balanceamento para Regressão Logística Politômica.**

| Alto Risco | Médio Risco | Baixo Risco | Total |
|---|---|---|---|
| 2.089 | 2.089 | 2.089 | 6267 |

Abaixo a Tabela 12 nos mostra a quantidade de dados balanceados para o modelo de regressão logística binomial, para o modelo binomial os casos de Alto risco estão classificados para 0 a 5 e Baixo risco para 6 a 10. Para o modelo de regressão logístico binomial as classificações de Alto risco foram 3.089 e para Baixo risco foram 501.995. Realizamos o método de undersamplig para os dados do modelo de regressão logística binomial.

**Tabela 12 - Balanceamento para Regressão Logística Binomial.**

| Alto Risco | Baixo Risco | Total |
|---|---|---|
| 3.089 | 3.089 | 6178 |

Fonte: Própria (2024)

Foram criadas tabelas comparativas das variáveis explicativas antes e depois da aplicação do método de balanceamento (undersampling). A tabela 10 inicial reflete o desbalanceamento natural dos dados, enquanto a tabela 11 após o balanceamento apresenta valores ajustados para o conjunto de dados equilibrado, o que é bom para o desenvolvimento preciso do modelo de regressão logística.

**Tabela 13 - Todas Variáveis separadas por Apgar 5 min**

| **Variáveis** | **1** , N = 2.089 [1] | **2** , N = 7.310 [1] | **3** , N = 495.685 [1] |
|---|---|---|---|
| **TPAPRESENTE** | | | |
| 1. Cefálico; | 1.732 (83%) | 6.443 (88%) | 475.771 (96%) |
| 2.Pélvica ou podálica; | 338 (16%) | 806 (11%) | 19.109 (3,9%) |
| 3.Transversa; | 19 (0,9%) | 61 (0,8%) | 805 (0,2%) |
| **GRAVIDEZ** | | | |
| 1. Única | 1.918 (92%) | 6.778 (93%) | 482.900 (97%) |
| 2.Dupla | 162 (7,8%) | 516 (7,1%) | 12.508 (2,5%) |
| 3.Tripla+ | 9 (0,4%) | 16 (0,2%) | 277 (<0,1%) |
| **PARTO** | | | |
| 1.VAGINAL | 1.198 (57%) | 3.077 (42%) | 199.628 (40%) |
| 2.CESÁREO | 891 (43%) | 4.233 (58%) | 296.057 (60%) |
| **PESO** | | | |
| 1. BAIXO | 313 (15%) | 1.420 (19%) | 41.309 (8,3%) |
| 2.EXTREMO | 756 (36%) | 901 (12%) | 2.046 (0,4%) |
| 3.GRANDE | 35 (1,7%) | 244 (3,3%) | 18.113 (3,7%) |
| 4.MUITO_BAIXO | 151 (7,2%) | 580 (7,9%) | 3.784 (0,8%) |
| 5.NORMAL | 834 (40%) | 4.165 (57%) | 430.433 (87%) |
| **SEXO** | | | |

| | | | |
|---|---|---|---|
| 1.HOMEM | 1.176 (56%) | 4.086 (56%) | 253.133 (51%) |
| 2.MULHER | 913 (44%) | 3.224 (44%) | 242.552 (49%) |
| **CONSULTAS** | | | |
| 1. NENHUMA | 120 (5,7%) | 184 (2,5%) | 3.645 (0,7%) |
| 2. 1 a 3 | 376 (18%) | 595 (8,1%) | 16.441 (3,3%) |
| 3. 4 a 6 | 648 (31%) | 1.795 (25%) | 69.461 (14%) |
| 4. 7 + | 945 (45%) | 4.736 (65%) | 406.138 (82%) |
| **IDANOMAL( ANOMALIA )** | | | |
| 1. SIM | 261 (12%) | 461 (6,3%) | 6.143 (1,2%) |
| 2. NÃO | 1.828 (88%) | 6.849 (94%) | 489.542 (99%) |

Fonte: Própria (2024)

**Tabela 14 - Volumetria das variáveis**

| Característica | Mínimo | 1º Quartil (Q1) | Mediana | Média | 3º Quartil (Q3) | Máximo |
|---|---|---|---|---|---|---|
| APGAR1 | 0 | 8 | 9 | 8.39 | 9 | 10 |
| IDADEMAE | 11 | 23 | 28 | 28.6 | 34 | 65 |

Fonte: Própria (2024)

A tabela 15 mostra as tabelas com os dados balanceados.

**Tabela 15 - Para os Balanceados usando Undersampling**

| **Variáveis** | **1** , N = 2.089 [1] | **2** , N = 2.089 [1] | **3** , N = 2.089 [1] |
|---|---|---|---|
| **TPAPRESENTE** | | | |
| 1.Cefálico; | 1.732 (83%) | 1.839 (88%) | 2.005 (96%) |
| 2.Pélvica ou podálica; | 338 (16%) | 235 (11%) | 80 (3,8%) |
| 3.Transversa; | 19 (0,9%) | 15 (0,7%) | 4 (0,2%) |
| **PARTO** | | | |
| 1. VAGINAL | 1.198 (57%) | 898 (43%) | 873 (42%) |
| 2. CESÁREO | 891 (43%) | 1.191 (57%) | 1.216 (58%) |
| **PESO** | | | |
| 1.BAIXO | 313 (15%) | 394 (19%) | 197 (9,4%) |
| 2.EXTREMO | 756 (36%) | 273 (13%) | 3 (0,1%) |
| 3.GRANDE | 35 (1,7%) | 66 (3,2%) | 78 (3,7%) |
| 4.MUITO_BAIXO | 151 (7,2%) | 168 (8,0%) | 20 (1,0%) |
| 5.NORMAL | 834 (40%) | 1.188 (57%) | 1.791 (86%) |
| **SEXO** | | | |
| 1.HOMEM | 1.176 (56%) | 1.202 (58%) | 1.075 (51%) |
| 2.MULHER | | 887 (42%) | 1.014 (49%) |
| **CONSULTAS** | | | |
| 1. NENHUMA | 120 (5,7%) | 59 (2,8%) | 10 (0,5%) |
| 2. 1 a 3 | 376 (18%) | 173 (8,3%) | 82 (3,9%) |
| 3. 4 a 6 | 648 (31%) | 506 (24%) | 304 (15%) |
| 4. 7 + | 945 (45%) | 1.351 (65%) | 1.693 (81%) |

| | | | |
|---|---|---|---|
| **IDANOMAL** | | | |
| 1.SIM | 261 (12%) | 117 (5,6%) | 30 (1,4%) |
| 2. NÃO | 1.828 (88%) | 1.972 (94%) | 2.059 (99%) |

Fonte: Própria (2024)

**Tabela 16 - Volumetria das variáveis balanceadas**

| Característica | Mínimo | 1º Quartil (Q1) | Mediana | Média | 3º Quartil (Q3) | Máximo |
|---|---|---|---|---|---|---|
| APGAR1 | 0 | 2 | 5 | 4.906 | 8 | 10 |
| IDADEMAE | 12 | 23 | 28 | 28.36 | 34 | 51 |

Fonte: Própria (2024)

Podemos notar que a realização do método Undersampling, trouxe uma proporção parecida em alguns casos, em relação a base desbalanceadas.

### 6.3 Variáveis Dummy

Para as variáveis que apresentam mais de uma categoria foi criado a variável dummy. Mesmo que não utilizamos todas as variáveis dummy no modelo, sendo necessário escolher uma categoria como referência. No caso de uma variável contínua, o primeiro passo para criar variáveis dummy é a categorização, como foi realizado para as variáveis escolhidas.

Com base nos insights obtidos durante o estudo da tabela, foram definidas as categorias das variáveis, conforme apresentado na tabela a seguir para criar as variáveis dummy.

**Tabela 17 - Variáveis transformadas para dummy**

| | | |
|---|---|---|
| Tipo de apresentação | TPAPRESENT | TPAPRESENT_1,<br>TPAPRESENT_2,<br>TPAPRESENT_3. |
| Peso ao nascer, em gramas | PESO | PESO_BAIXO,<br>PESO_EXTREMO,<br>PESO_GRANDE,<br>PESO_MUITO_BAIXO,<br>PESO_NORMAL. |
| Tipo de gravidez, | GRAVIDEZ | GRAVIDEZ_1,<br>GRAVIDEZ_2,<br>GRAVIDEZ_3. |
| Número de consultas de pré-natal | CONSULTAS | CONSULTAS_1,<br>CONSULTAS_2 ,<br>CONSULTAS_3 ,<br>CONSULTAS_4. |
| Sexo | SEXO | SEXO_ FEMININO,<br>SEXO_MASCULINO. |
| Parto vaginal ou cesáreo | PARTO | PARTO_VAGINAL<br>PARTO_CESAREO |
| Anomalia congênita | IDANOMAL | IDANOMAL_PRESENTE,<br>IDANOMAL_AUSENTE |

Fonte: Própria (2024)

Após a categorização das variáveis, foram criadas $c - 1$ variáveis binárias para cada variável, onde $c$ é o número de classes da divisão. Por exemplo, para a

variável "Gravidez", duas novas variáveis foram introduzidas. A primeira classe de cada variável categorizada é usada como referência ao interpretar o modelo.

# 7 Aplicação do Modelo

## 7.1 Qualidade do Ajuste

Antes de iniciar as manipulações, os dados foram divididos em duas partes: uma parte para treinar o modelo, ou seja, para estimar os parâmetros, e outra para testar as predições. A base de treino corresponde a 70% do total, resultando em 4.386 linhas, enquanto a base de teste compreende os 30% restantes, totalizando 1.881 linhas.

## 7.2 Modelo de Regressão Logística

Inicialmente, o modelo foi construído incorporando todas as variáveis dummy geradas. Posteriormente, empregou-se o método de seleção de variáveis, especificamente o stepwise, o qual resultou na exclusão de algumas variáveis dummy consideradas não significativas no teste Z-Wald. Além disso, utilizou-se o método VIF (Variance Inflation Factor) para avaliar a multicolinearidade entre as variáveis. O VIF ajuda a identificar se alguma variável independente está altamente correlacionada com outras, o que pode distorcer os resultados do modelo. Variáveis com VIF alto foram examinadas e, se necessário, removidas para melhorar a robustez do modelo.

Com o intuito de evitar alguma confusão, o modelo está apresentado em forma de tabela para simplificar a compreensão de cada variável individualmente. Podendo mostrar uma visualização melhor.

### 7.2.1 Modelo de Regressão Logística Binária Simples

Para o modelo de regressão logística binária simples foram usados a variável resposta APGAR 5 min, e como variaveis explicativas Sexo do recem nascido.

**Tabela 18 - Modelo de Regressão Logístico Simples**

| **Coeficiente** | **Estimativa** | **std.error** | **statistic** | **p.value** | **2.5% ; 97.5%** |
|---|---|---|---|---|---|
| (Intercept) | -0.103 | 0.041 | -2.49 | 0.0127 | -0.103, -0.022 |
| SEXO_MASC | 0.226 | 0.061 | 3.69 | 0.000221 | 0.226, 0.346 |

Fonte: Própria (2024)

Observa-se que todos os p-valores são estatisticamente significativos ao nível de significância de 5%. Além disso, ao analisar os intervalos de confiança, nota-se que nenhum deles contém o valor 0. No entanto, para afirmar que o modelo é adequado e tirar conclusões, é necessário avaliar seu desempenho por meio da matriz de confusão.

**Tabela 19 - Matriz de Confusão Modelo Binomial Simples**

| | **ALTO** | **BAIXO** |
|---|---|---|
| **ALTO** | 102 | 17 |
| **BAIXO** | 825 | 910 |

Fonte: Própria (2024)

**Tabela 20 - Métricas de desempenho  modelo binomial simples**

| Métrica | Resultado |
|---|---|
| **Acurácia** | 0.54 |
| **Precisão** | 0.85 |
| **Recall** | 0.11 |

Fonte: Própria (2024)

Ao observar a matriz de confusão, nota-se que o modelo de regressão logística binomial não previu adequadamente os casos de Apgar 5 minutos, acertando apenas 54% das previsões. A precisão do modelo é 0,85, indicando uma boa proporção de previsões corretas entre as realizadas. No entanto, o Recall é de apenas 0,11, revelando que o modelo identificou corretamente apenas 11% dos casos reais de Apgar 5. Esses resultados sugerem que, apesar da alta precisão, o modelo possui um desempenho insatisfatório em termos de cobertura dos casos reais, o que o torna um modelo preditivo inadequado para esta aplicação específica.

### 7.2.2 Modelo de Regressão Logística  Binomial Múltiplo

Para construir o modelo de regressão logística binomial múltipla, utilizaremos a base de treino contendo todas as variáveis dummy. Em seguida, realizaremos os testes e as análises de desempenhos,  para selecionar as variáveis mais relevantes para o modelo, este método também foi seguido para a realização do modelo de regressão logística politômica ou multinomial.

**21 - Modelo de Regressão Logístico Binomial Múltiplo**

| **Coeficiente** | **Estimativa** | **std.error** | **statistic** | **p.value** | **2.5% ; 97.5%** |
|---|---|---|---|---|---|
| (Intercept) | -6.61192 | 0.41600 | -15.894 | < 2e-16 | -7.445, -5.815 |
| APGAR1 | 0.81247 | 0.0253 | 32.102 | < 2e-16 | 0.764, 0.863 |
| PESO_EXTREMO | -0.9360 | -0.93601 | -2.819 | 0.004822 | -1.619, -0.312 |
| IDANOMAL_AUSENTE | 1.03568 | 0.31977 | 3.239 | 0.001200 | 0.418 , 1.670 |

Fonte: Própria (2024)

Todas as variáveis analisadas apresentaram significância estatística ao nível de 5%, conforme indicado pelos p-valores e pelos intervalos de confiança, que não incluem o valor 0. Esses resultados reforçam a relevância estatística das variáveis no modelo. Diante disso, prosseguiremos com a análise de desempenho do modelo, utilizando a matriz de confusão para avaliar sua eficácia preditiva..

**Tabela 22 - Matriz de confusão modelo binomial múltiplo**

| | REFERÊNCIA | |
|---|---|---|
| **PREDIÇÃO** | **ALTO** | **BAIXO** |
| **ALTO** | 807 | 35 |
| **BAIXO** | 120 | 892 |

Fonte: Própria (2024)

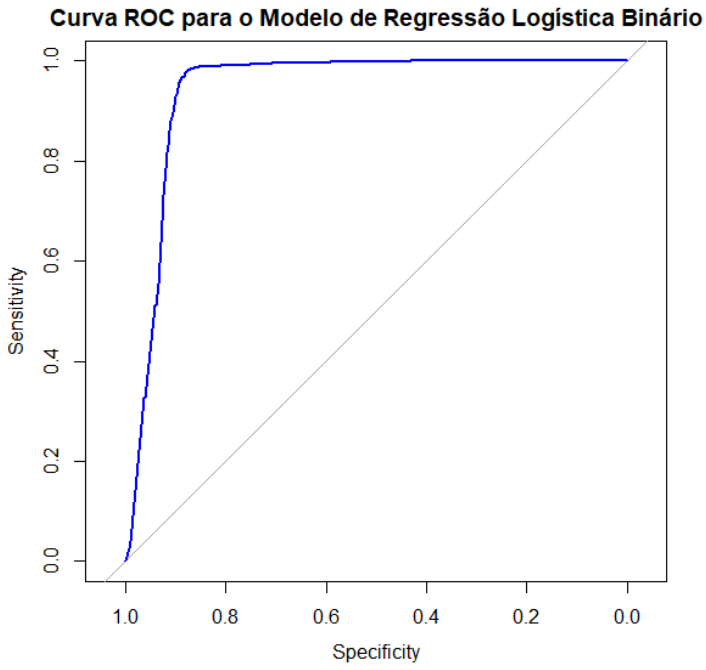
**Tabela 23 - Métricas de desempenho das predições do modelo**

| Métrica | Resultado |
|---|---|
| **Acurácia** | 0.92 |
| **Precisão** | 0.96 |
| **Recall** | 0.95 |

Fonte: Própria (2024)

A acurácia do modelo alcança 92%, refletindo uma alta taxa geral de acertos. A precisão é de 96,22%, indicando que 96% dos eventos de interesse foram corretamente identificados. O recall é de 95%, destacando que o modelo capturou 95% dos eventos relevantes. Para avaliar a eficácia do modelo na diferenciação entre a classificação dos casos de Apgar 5 minutos, com base nas taxas de verdadeiros positivos e falsos positivos, foi gerada a curva ROC e calculada a AUC, utilizando um ponto de corte de 0,5.

Figura 8 - Curva ROC Regressão logística binomial múltipla




Fonte: Própria (2024)

Realizando a curva ROC, temos que a curva ficou acima do ponto do corte e mostrando pela área da curva  um desempenho significante. Sugerindo que o modelo acerta AUC de 94% das predições feitas com a base de dados usada para estimar seus parâmetros. Podemos concluir que o modelo performou muito bem para realização da discriminação. Vamos realizar Odds Ratio para entender o quanto cada variável se comporta para o aumento ou diminuição  dos casos de Apgar.

#### 7.2.2.1 Odds Ratio:

**Tabela 24 - Odds  Ratio do Modelo Binomial Simples**

| Característica | 0R[1] | IC 95% [1] | valor p |
|---|---|---|---|
| APGAR1 | 2,22 | 2,07; 2,39 | <0,001 |
| PESO_EXTREMO | 0,27 | 0,10; 0,65 | 0,006 |
| IDANOMAL_AUSENTE (Ausência de  Anomalia) | 2,76 | 1,18; 6,63 | 0,021 |

OR = Razão de Chances, IC = Intervalo de Confiança

Fonte: Própria (2024)

Realizando a interpretação das Odds Ratio para o modelo temos que um aumento na pontuação Apgar de 1 minuto está associado a um aumento de 122% nas chances de estar no grupo de baixo risco de Apgar de 5 minutos em comparação ao grupo de alto risco. Este resultado é altamente significativo, indicando que quanto maior a pontuação Apgar de 1 minuto, maior a probabilidade de o bebê estar no grupo de baixo risco.

Bebês com peso extremo têm 73% menos chances de estar no grupo de baixo risco de Apgar de 5 minutos em comparação ao grupo de alto risco. Este

resultado é estatisticamente significativo, sugerindo que bebês com peso extremo estão mais propensos a estar no grupo de alto risco de Apgar de 5 minutos.

A ausência de anomalias congênitas está associada a um aumento de 176% nas chances de estar no grupo de baixo risco de Apgar de 5 minutos em comparação ao grupo de alto risco. Este resultado é estatisticamente significativo, indicando que a presença de anomalias congênitas aumentam a probabilidade de o bebê estar no grupo de alto risco.

Os resultados mostram que a pontuação Apgar de 1 minuto é um forte preditor de estar no grupo de baixo risco de Apgar de 5 minutos e peso extremo têm uma chance significativamente menor de estar no grupo de baixo risco, sugerindo um maior risco associado ao peso extremo

### 7.3. Modelo de Regressão Logística Multinomial

Para realização dos modelos multinomial as variáveis preditoras foram classificadas em alto, médio e baixo risco, como mostrado na Tabela com os dados Balanceados pelo método Undersampling. A variável de referência para as demais no modelo foi definida como os casos com Apgar de 5 minutos e alto risco de complicação. Como foi realizado no modelo de regressão binária, também seguiremos realizando o mesmo método, construindo o modelo para todas variáveis dummy geradas.

Empregou-se o método de seleção de variáveis, especificamente o stepwise, o qual resultou na exclusão de algumas variáveis dummy consideradas não significativas no teste Z-Wald. E em seguida, utilizou-se o método VIF (Variance Inflation Factor) para avaliar a multicolinearidade entre as variáveis. O VIF ajuda a identificar se alguma variável independente está altamente correlacionada com outras, o que pode distorcer os resultados do modelo. Variáveis com VIF alto foram examinadas e, se necessário, removidas para melhorar a robustez do modelo.

### 7.3.1 Modelo de Regressão Logística Multinomial Simples

Para o modelo de regressão logística multinomial simples foram usados a variável reposta APGAR 5 min, e como variaveis explicativas Sexo do recem nascido.

**Tabela 25 - Modelo de Regressão logístico Multinomial Simples**

| y | term | estimate | std.error | statistic | p.value | 2.5%, 97.5% |
|---|---|---|---|---|---|---|
| 1 | Médio Risco | | | | | |
| 1 | (Intercept) | 0.033 | 0.049 | 0.668 | 0.504 | -0.064 0.130 |
| 1 | SEXO_Mas | -0.075 | 0.074 | -1.01 | 0.315 | -0.221 0.071 |
| 2 | Baixo Risco | | | | | |
| 2 | (Intercept) | -0.087 | 0.051 | -1.71 | 0.088 | -0.187 0.013 |
| **2** | **SEXO_Mas** | **0.184** | **0.074** | **2.48** | **0.013** | **0.039 0.329** |

Fonte: Própria (2024)

Temos que somente a varivel sexo para o modelo com predição para os casos de Baixo risco de Apgar, foi estatisticamente significativa para p-valores menor que 0.05, para as outras variáveis o p-valor não foi significativo.Vamos realizar a análise de desempenho do modelo.

**Tabela 26 - Matriz de confusão do Modelo de Multinomial Simples**

| | **Alto** | **Médio** | **Baixo** |
|---|---|---|---|
| **Alto** | 372 | 348 | 340 |
| **Médio** | 0 | 0 | 0 |
| **Baixo** | 255 | 279 | 287 |

Fonte: Própria (2024)

Dado que não identificamos nenhuma variável significativa para o modelo de predição de Médio Risco de Apgar 5, não foi possível realizar nenhuma previsão para esse cenário específico. No entanto, para os demais casos, conseguimos efetuar as previsões com sucesso, devido a variável ser considerada significativa.

**Tabela 27 - Métricas de desempenho das predições Multinomial simples**

| **Métrica** | **Alto** | **Médio** | **Baixo** |
|---|---|---|---|
| **Acurácia** | 0.33 | 0 | 0.33 |
| **Precisão** | 0.35 | 0 | 0.34 |
| **Recall** | 0.59 | 0 | 0.45 |

Fonte: Própria (2024)

O modelo possui métricas ruins, devido não ser eficiente para os casos de médio risco, o modelo está errando mais da metade dos casos.

### 7.3.2 Modelo de Regressão Logística Politômica Múltiplo

O modelo de regressão Politômica foi realizado considerando todas as variáveis disponíveis. Temos abaixo o modelo completo com todas as variáveis.

Na elaboração de um modelo de regressão Politômica , é importante ressaltar que algumas variáveis podem não ser significativas para a predição de um fator, porém, podem desempenhar um papel significativo no outro modelo preditivo.

**Tabela 28 - Modelo de Regressão Logístico Multinomial Múltiplo**

| yl | term | estimate | std.error | statistic | p.value | 2.5%, 97.5% |
|---|---|---|---|---|---|---|
| 1 | Médio Risco | | | | | |
| 1 | (Intercept) | -1.790 | 0.321 | -5.57 | 2.52e-08 | -2.419 -1.160 |
| 1 | IDADEMAE | 0.00583 | 0.00654 | 0.892 | 3.73e-01 | -0.007 0.019 |
| **1** | **QTDFILVIVO** | **-0.0787** | **0.0356** | **-2.21** | **2.72e-02** | **-0.149 -0.009** |
| 1 | QTDFILMORT | 0.0408 | 0.0601 | 0.679 | 4.97e-01 | -0.077 0.159 |
| **1** | **APGAR1** | **0.305** | **0.0194** | **15.7** | **8.07e-56** | **0.267 0.343** |
| 1 | TPAPRESENT_2 | 0.0287 | 0.128 | 0.224 | 8.23e-01 | -0.222 0.279 |
| 1 | TPAPRESENT_3 | -0.152 | 0.452 | -0.335 | 7.37e-01 | -1.037 0.734 |
| 1 | SEXO_MAS | -0.0781 | 0.0833 | -0.937 | 3.49e-01 | -0.241 0.085 |
| 1 | GRAVIDEZ_2 | 0.268 | 0.161 | 1.66 | 9.67e-02 | -0.048 0.583 |
| 1 | GRAVIDEZ_3 | 0.0835 | 0.771 | 0.108 | 9.14e-01 | -1.427 1.594 |
| **1** | **PESO_EXTREMO** | **-0.989** | **0.145** | **-6.83** | **8.24e-12** | **-1.273 -0.705** |

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | PESO_GRANDE | 0.174 | 0.279 | 0.624 | 5.33e-01 | -0.372 0.720 |
| 1 | PESO_MUITO_BAIXO | 0.0118 | 0.172 | 0.0688 | 9.45e-01 | -0.325 0.349 |
| 1 | PESO_NORMAL | -0.153 | 0.120 | -1.28 | 2.00e-01 | -0.388 0.081 |
| **1** | **PARTO_CESAREO** | **0.332** | **0.0874** | **3.79** | **1.50e-04** | **0.160 0.503** |
| 1 | CONSULTAS_2 | 0.00545 | 0.254 | 0.0215 | 9.83e-01 | -0.492 0.503 |
| 1 | CONSULTAS_3 | 0.170 | 0.241 | 0.706 | 4.80e-01 | -0.302 0.642 |
| 1 | CONSULTAS_4 | 0.277 | 0.240 | 1.15 | 2.49e-01 | -0.194 0.748 |
| **1** | **IDANOMAL_AUSENTE** | **0.752** | **0.146** | **5.15** | **2.54e-07** | **0.466 1.037** |
| 2 | Baixo risco | | | | | |
| 2 | (Intercept) | -9.040 | 0.767 | -11.8 | 4.66e-32 | -10.538, -7.533 |
| 2 | IDADEMAE | -0.00313 | 0.0107 | -0.293 | 7.70e-01 | -0.024 0.018 |
| 2 | QTDFILVIVO | -0.0389 | 0.0608 | -0.639 | 5.23e-01 | -0.158 0.080 |
| 2 | QTDFILMORT | 0.148 | 0.0994 | 1.49 | 1.37e-01 | -0.047 0.343 |
| **2** | **APGAR1** | **1.240** | **0.0394** | **31.4** | **9.47e-216** | **1.158 1.313** |
| 2 | TPAPRESENT_2 | -0.0735 | 0.281 | -0.261 | 7.94e-01 | -0.625 0.478 |
| 2 | TPAPRESENT_3 | 0.504 | 1.210 | 0.418 | 6.76e-01 | -1.859 2.867 |
| 2 | SEXO_MAS | -0.0233 | 0.131 | -0.177 | 8.59e-01 | -0.281 0.234 |
| 2 | GRAVIDEZ_2 | 0.00812 | 0.338 | 0.0240 | 9.81e-01 | -0.654 0.671 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2 | GRAVIDEZ_3 | 1.290 | 2.130 | 0.605 | 5.45e-01 | 2.879 5.452 |
| **2** | **PESO_EXTREMO** | **-3.080** | **0.812** | **-3.80** | **1.47e-04** | **-4.677 -1.492** |
| 2 | PESO_GRANDE | 0.716 | 0.394 | 1.82 | 6.95e-02 | -0.057 1.488 |
| 2 | PESO_MUITO_BAIXO | -0.697 | 0.440 | -1.58 | 1.13e-01 | -1.561 0.166 |
| 2 | PESO_NORMAL | 0.178 | 0.195 | 0.914 | 3.61e-01 | -0.203 0.559 |
| 2 | PARTO_CESAREO | 0.161 | 0.140 | 1.15 | 2.49e-01 | -0.113 0.434 |
| 2 | CONSULTAS_2 | -0.343 | 0.576 | -0.596 | 5.51e-01 | -1.472 0.785 |
| 2 | CONSULTAS_3 | 0.0152 | 0.527 | 0.0288 | 9.77e-01 | -1.018 1.048 |
| 2 | CONSULTAS_4 | 0.419 | 0.517 | 0.811 | 4.18e-01 | -0.594 1.431 |
| **2** | **IDANOMAL_AUSENTE** | **1.670** | **0.439** | **3.81** | **1.40e-04** | **0.812 2.533** |

Fonte: Própria (2024)

Temos que, para o modelo completo, as variáveis Anomalias Congênitas (IDANOMAL), Peso Extremo, e APGAR 1 foram estatisticamente significativas para ambos os casos da variável preditora, com p-valores < 0,05.A variável QTDFILVIVO foi significativa para os casos de Médio Risco e não foi significativa para os casos de Baixo Risco.

**Tabela 29 - Modelo Multinomial pelo Stepwise**

| y | term | estimate | std.error | statistic | p.value | I.C.2.5%,97.5% |
|---|---|---|---|---|---|---|
| 1 | Médio Risco | | | | | |
| 1 | (Intercept) | -1.55 | 0.226 | -4.57 | 7.51e-19 | -2.199, -1.207 |
| **1** | **QTDFILVIVO** | **-0.178** | **0.0507** | **-1.72** | **3.35e-02** | **-0.132, -0.075** |
| **1** | **APGAR1** | **0.235** | **0.0305** | **10.5** | **1.28e-25** | **0.178, 0.294** |
| **1** | **PESO_EXTREMO** | **-0.875** | **0.192** | **-7.98** | **5.21e-14** | **-1.216, -0.513** |
| **1** | **PARTO_CESAREO** | **0.655** | **0.132** | **2.22** | **1.39e-05** | **0.402, 0.934** |
| **1** | **CONSULTAS_4** | **0.284** | **0.198** | **2.08** | **5.02e-02** | **-0.0008, 0.575** |
| **1** | **IDANOMAL_AUSENTE** | **0.935** | **0.145** | **1.24e-07** | **1.24e-07** | **0.482, 1.349** |
| 2 | Baixo Risco | | | | | |
| 2 | (Intercept) | -8.31 | 0.554 | -13.3 | 1.72e-70 | -10.343, -8.286 |
| 2 | QTDFILVIVO | -0.150 | 0.0830 | -2.29 | 4.39e-01 | -0.148, 0.064 |
| **2** | **APGAR1** | **1.26** | **0.0592** | **21.2** | **9.35e-22** | **1.085, 1.319** |
| **2** | **PESO_EXTREMO** | **-2.49** | **0.797** | **-2.28** | **2.54e-04** | **-4.427 , -0.438** |
| 2 | PARTO_CESAREO | 0.819 | 0.207 | 0.210 | 2.24e-01 | 0.419, 1.225 |
| **2** | **CONSULTAS_4** | **0.505** | **0.403** | **0.819** | **1.20e-03** | **0.056, 0.910** |
| 2 | IDANOMAL_AUSENTE | 0.770 | 0.435 | 3.88 | 1.05e-04 | -0.338, 1.842 |

Fonte: Própria (2024)

Observa-se que nem todos os p-valores são estatisticamente significativos ao nível de significância de 5% para ambas as predições, tanto para casos de Médio

Risco quanto para Baixo Risco. Algumas variáveis foram significativas em um modelo, mas não no outro. Ao analisar os intervalos de confiança, nota-se que para as variáveis que não foram significativas, o valor 0 faz parte do intervalo de confiança.

Devido à significância presente em pelo menos um dos modelos, optamos por manter essas variáveis, já que sua inclusão não prejudica o desempenho do outro modelo. Para confirmar a adequação do modelo aos dados, é necessário avaliar seu desempenho. Vamos proceder com essa avaliação por meio da matriz de confusão e suas métricas.

**Tabela 30 - Matriz de confusão Modelo Multinomial Múltiplo**

| Reference \ Prediction | ALTO | MÉDIO | BAIXO |
|---|---|---|---|
| **ALTO** | 445 | 146 | 3 |
| **MÉDIO** | 99 | 438 | 36 |
| **BAIXO** | 83 | 43 | 588 |

Fonte: Própria (2024)

**Tabela 31 - Métricas de desempenho Modelo Multinomial Múltiplo**

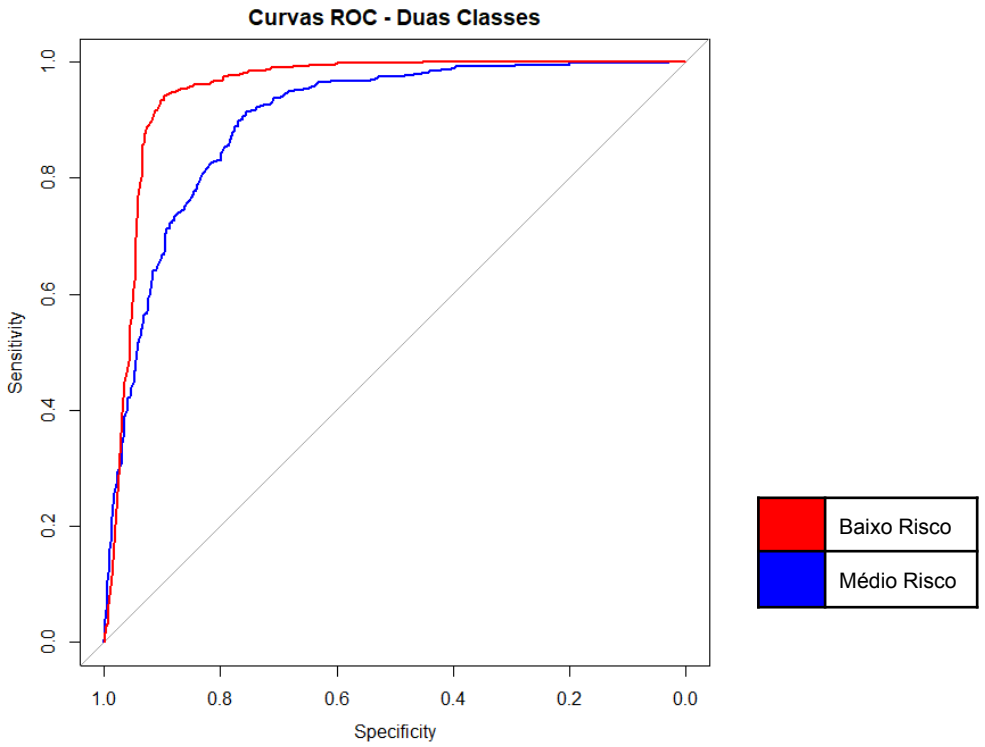
| Métrica\Predição | Alto | Médio | Baixo |
|---|---|---|---|
| **Acurácia** | 0.782 | 0.782 | 0.782 |
| **Precisão** | 0.749 | 0.764 | 0.823 |
| **Recall** | 0.881 | 0.698 | 0.937 |

Fonte: Própria (2024)

Durante a análise da matriz de confusão, foram extraídas métricas fundamentais: acurácia, precisão e recall. A acurácia alcançou 78,2%, indicando que aproximadamente 78% das previsões foram corretas no geral. A precisão foi registrada como 0,749 para casos de Alto Risco, 0,764 para Médio Risco e 0,823 para Baixo Risco, sugerindo que o modelo acertou mais de 74% das vezes ao prever a ocorrência de um evento. Quanto ao recall, os valores foram 0,881 para Alto Risco, 0,698 para Médio Risco e 0,937 para Baixo Risco, mostrando que cerca de 69% das situações relevantes foram capturadas pelo modelo. Esses insights são essenciais para uma avaliação precisa da eficácia do modelo desenvolvido.

Para avaliar a eficácia do modelo na distinção entre as classificações geradas, foram construídas a curva ROC e calculada a AUC, utilizando um ponto de corte de 0,5, com base nas taxas de verdadeiros positivos e falsos positivos.

Figura 9 - Curva ROC do Modelo de Regressão Logística Multinomial




Fonte: Própria (2024)

Nota-se que a AUC foi de 0,94 para o modelo referente a Baixo Risco de Apgar aos 5 minutos e de 0,899 para o modelo referente a Médio Risco de Apgar aos 5 minutos. Ambos os modelos estão significativamente acima do ponto de corte. Esses valores são bastante satisfatórios para modelos de classificação, indicando que o modelo possui uma excelente discriminação. Podemos afirmar que o modelo acerta 94% dos casos de Baixo Risco e 89% dos casos de Médio Risco. Observa-se que o modelo está discriminando melhor os casos de Apgar aos 5 minutos de Baixo Risco de desenvolver problemas.

Para concluir a interpretação, calculou-se as Odds Ratio estimadas, dispostas:

**Tabela 32 - Odds Ratio modelo Multinomial Múltiplo**

| Característica | OR [1] | IC 95% [1] | valor p |
|---|---|---|---|
| **1 Médio Risco** | | | |
| QTDFILVIVO | 0,84 | 0,75, 0,93 | <0,001 |
| APGAR1 | 1,26 | 1,20, 1,34 | <0,001 |
| PESO_EXTREMO | 0,42 | 0,30, 0,59 | <0,001 |
| PARTO_CESAREO | 1,93 | 1,50, 2,48 | <0,001 |
| CONSULTAS_4 | 1,33 | 1,00, 1,77 | 0,051 |
| IDANOMAL_AUSENTE | 2,55 | 1,64, 3,94 | <0,001 |
| **2 Baixo Risco** | | | |
| QTDFILVIVO | 0,86 | 0,73, 1,01 | 0,068 |
| APGAR1 | 3.33 | 2,97, 3,73 | <0,001 |
| PESO_EXTREMO | 0,08 | 0,01, 0,65 | 0,018 |
| PARTO_CESAREO | 2.26 | 1,52, 3,36 | <0,001 |
| CONSULTAS_4 | 1,67 | 1,06, 2,65 | 0,028 |
| IDANOMAL_AUSENTE | 2.16 | 0,71, 6,55 | 0,2 |

[1] OR = Razão de Chances, CI = Intervalo de Confiança

Fonte : Própria(2024)

Realizando a interpretação da Odds Ratio, primeiramente para os casos de Apgar de 5 minutos para Médio Risco de complicações ao nascer. Quantidade de filho vivo, para cada filho vivo adicional está associado a uma redução de 16% nas chances de estar no grupo de médio risco em comparação ao grupo de alto risco, com essa redução sendo estatisticamente significativa.

Para casos de Apgar 1 minuto,cada aumento na pontuação Apgar de 1 minuto está associado a um aumento de 26% nas chances de estar no grupo de médio risco em comparação ao grupo de alto risco, com essa diferença sendo estatisticamente significativa.

Bebês com peso extremo têm 58% menos chances de estar no grupo de médio risco em comparação ao grupo de alto risco, com essa redução sendo estatisticamente significativa.

O parto tipo Cesáreo está associado a um aumento de 93% nas chances de estar no grupo de médio risco em comparação ao grupo de alto risco, com essa diferença sendo estatisticamente significativa.

Ter quatro ou mais consultas pré-natais está associado a um aumento de 33% nas chances de estar no grupo de médio risco em comparação ao grupo de alto risco. Este resultado é marginalmente significativo.

A ausência de anomalias congênitas está associada a um aumento de 155% nas chances de estar no grupo de médio risco em comparação ao grupo de alto risco, com essa diferença sendo estatisticamente significativa.

Realizando a interpretação do Baixo Risco de Apgar de 5 minutos sobre Alto Risco de Apgar de 5 minutos. Cada filho vivo adicional está associado a uma redução de 14% nas chances de estar no grupo de baixo risco em comparação ao grupo de alto risco. Este resultado não é estatisticamente significativo.

Cada aumento na pontuação Apgar de 1 minuto está associado a um aumento de 233% nas chances de estar no grupo de baixo risco em comparação ao grupo de alto risco, com essa diferença sendo estatisticamente significativa.

Bebês com peso extremo têm 92% menos chances de estar no grupo de baixo risco em comparação ao grupo de alto risco, com essa redução sendo estatisticamente significativa.

O parto tipo Cesáreo está associado a um aumento de 126% nas chances de estar no grupo de baixo risco em comparação ao grupo de alto risco, com essa diferença sendo estatisticamente significativa.

Ter quatro ou mais consultas pré-natais está associado a um aumento de 67% nas chances de estar no grupo de baixo risco em comparação ao grupo de alto risco, com essa diferença sendo estatisticamente significativa.

A ausência de anomalias congênitas está associada a um aumento de 116% nas chances de estar no grupo de baixo risco em comparação ao grupo de alto risco. Este resultado não é estatisticamente significativo.

### 7.4 Método de comparação dos modelos utilizando BIC e AIC

Para avaliar e comparar o desempenho dos modelos de regressão logística, os critérios de informação AIC (Akaike Information Criterion) e BIC (Bayesian Information Criterion) são ferramentas essenciais. Esses critérios equilibram a qualidade do ajuste e a complexidade do modelo, ou seja a quantidade de parâmetros , aplicando penalidades pela inclusão de mais parâmetros.

**Tabela 33 - Comparação dos Modelos utilizando BIC e AIC**

| Modelos | AIC | BIC |
|---|---|---|
| Modelos Binário | | |
| Mod.Binário Simples | 5984.67 | 5997.41 |
| Mod.Binário Múltiplo | 2068.50 | 2113.11 |
| Modelos Multinomial | | |
| Mod.Multinomial Simples | 4137.37 | 4159.53 |
| Mod.Multinomial Múltiplo | 2425.98 | 2525.69 |

Fonte: Própria (2024)

Com base nos resultados, podemos concluir que o modelo multinomial Múltiplo é mais eficiente em termos de ajuste aos dados, sendo assim o modelo preferido para o método Multinomial. Para o modelo binário o melhor modelo é o modelo de regressão logístico binário. Utilizar ambos os critérios garante uma escolha robusta, promovendo modelos que são precisos. Se fossemos comparar todos os modelos temos que o modelo de regressão logística binomial múltiplo foi melhor, mas como o modelo de regressão logística politômica está prevendo para mais variáveis preditoras não é correta a comparação.

# 8 CONSIDERAÇÕES FINAIS

Neste trabalho, alcançou-se o objetivo de estudar a aplicação da técnica de Regressão Logística, com foco no modelo multinomial, para detectar casos de Apgar aos 5 minutos com alto risco de complicações. Adicionalmente, adquiriu-se conhecimento sobre como lidar, tratar e trabalhar com dados desbalanceados, resultando em um modelo eficiente capaz de prever casos de Apgar aos 5 minutos.

Ao longo do estudo, foram realizadas quatro aplicações utilizando diferentes abordagens de Regressão Logística: Binomial Simples, Binomial Múltipla, Multinomial Simples e Multinomial Múltipla, todos os modelos com dados balanceados por meio da técnica de undersampling.

Os resultados mostraram que os modelos de regressão logística binomial múltipla e regressão logística multinomial múltipla apresentaram a melhor acurácia, respectivamente, de 94% e 78%. Para o modelo de regressão logística binomial, essa acurácia indica uma discriminação excelente, enquanto para o modelo de regressão logística multinomial, a discriminação é aceitável. É importante considerar que o modelo de regressão logística multinomial prevê para mais de uma variável preditora, fornecendo informações mais detalhadas sobre cada variável.

Entre as variáveis analisadas, os modelos apresentaram diferenças significativas para as variáveis QTDFILVIVO, PESO_EXTREMO, PARTO_CESAREO e CONSULTAS_4 mostraram-se significativas para a discriminação dos casos de Apgar aos 5 minutos no modelo de regressão logística multinomial, diferentemente do modelo de regressão logística binomial, onde não foram significativas.

Os modelos de regressão logística simples não tiveram um desempenho significativo em termos de acurácia, não sendo precisos para uso informativo. No modelo de regressão logística multinomial simples, a variável sexo foi significativa apenas para os casos de baixo nível de Apgar.

Concluímos que o modelo de regressão logística multinomial pode ser utilizado para a previsão de casos de Apgar aos 5 minutos, visto que passou em todos os testes. Esse modelo forneceu Odds Ratios interessantes para cada variável

significativa, podendo ser utilizado na tomada de decisões para obter recém-nascidos sem risco de complicações de Apgar.

# Referências


BATISTA DE LIMA, I. M.; CRUZ DE OLIVEIRA, A. E.; MOURA DE ANDRADE, J.; CAMPOS COÊLHO, H. F.; SILVA LIMA, K. Modelo de decisão sobre o uso de preservativos: Uma regressão logística multinomial. Tempus – **Actas de Saúde Coletiva**, v. 10, n. 2, p. Pág. 67-80, 7 jul. 2016

BESERRA, RS.; BARBOSA, NFM.; PEIXOTO, APB.; MORAIS XAVIER, Érika F. .; JALE, JS.; XAVIER JÚNIOR, SFA Modelagem com regressão logística para análise de concessão de crédito. **Investigação, Sociedade e Desenvolvimento** , *[S. l.]* , v. 11, n. 7, pág. e15211729761, 2022. DOI: 10.33448/rsd-v11i7.29761. Disponível em: https://rsdjournal.org/index.php/rsd/article/view/29761. Acesso em: 16 jan. 2023.

BITTENCOURT, Hélio Radke. Regressão logística politômica: revisão teórica e aplicações. **Acta Scientiae**, v. 5, n. 1, p. 77-86, 2003.

BRAGA, A. C. Curva ROC: Aspectos funcionais e aplica¸c˜oes, 2000. Tese (Doutorado em Estatística) - Universidade do Minho, Braga, 2000.

CABRAL Cleidy Isolete Silva. **Aplicação do Modelo de Regressão Logística num Estudo de Mercado.**2013.Trabalho de projeto de mestrado em Matemática Aplicada à Economia e à Gestão, apresentada à Universidade de Lisboa, através da Faculdade de Ciências, http://hdl.handle.net/10451/10671, Acesso em 20 mar 2023.

CORRAR, LUIZ J. **Análise Multivariada:** para os cursos de administração, ciências contábeis e economia / FIPECAPI – Fundação Instituto de pesquisas Contábeis, Atuariais e Financeiras; Luiz J. Corrar, Edílson Paulo, José Maria Dias Filho (coordenadores).1 ed. – São Paulo: Atlas, 2011.

COX, D.R.; SNELL, E.J. **Analysis of Binary Data.** London: Chapman & Hall, 2ª Edição, 1989.

DANONE BABY**, Teste de Apgar:** o que avalia e como é realizado. Disponível no site: https://www.danonenutricia.com.br/infantil/primeiros-meses/saude/teste-de-apgar 26 Jun 2017.

FÁVERO, Luiz Paulo Lopes et al. **Análise de dados:** modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier. . Acesso em: 16 jan. 2023. , 2009

HOSMER, D.; LEMESHOW, S.; STURDIVANT, X. **Applied Logistic Regression.**3th ed. Canada: John Wiley Sons, 2013.

LOURENÇO, A. L C Magalhães "**Proporção e fatores associados a Apgar menor que 7 no 5º minuto de vida: de 1999 a 2019, o que mudou**?" Faculdade de Ciências Médicas, Universidade do Estado do Rio de Janeiro., Acesso disponivel em: https://www.scielo.br/j/csc/a/SRKqK3M38f5TV7YhFSgggMB/ 26 jul 20222

MESQUITA, P. S. B. **Um modelo de Regressão Logística para Avaliação de Programas de Pós-Graduação no Brasil**. Master Thesis. Universidade Estadual do Norte Fluminense, Campos dos . . . , 2014

SILVEIRA, mbg da.; BARBOSA,nfm.; PEIXOTO,apb.; XAVIER, Érika fm.; XAVIER JÚNIOR, sfa Aplicação da regressão logística na análise do fator de risco associado à hipertensão arterial. **Investigação, Sociedade e Desenvolvimento** , *[S. l.]* , v. 10, n. 16, pág. e20101622964, 2021. DOI: 10.33448/rsd-v10i16.22964. Disponível em: https://rsdjournal.org/index.php/rsd/article/view/22964. Acesso em: 17 jan. 2023.

SILVA, L. S. R.; CAVALCANTE, A. N.; CARNEIRO, J. K. R.; OLIVEIRA, M. A. S. Índice de Apgar correlacionado a fatores maternos, obstétricos e neonatais a partir de dados coletados no Centro de Saúde da Família do bairro Dom Expedito Lopes situado no município de Sobral/CE. **Revista Científica da Faculdade de Medicina de Campos**, *[S. l.]*, v. 15, n. 1, p. 25–30, 2020. DOI: 10.29184/1980-7813.rcfmc.232.vol.15.n1.2020. Disponível em: https://revista.fmc.br/ojs/index.php/RCFMC/article/view/232. Acesso em: 24 jun. 2023

VIDAL E SILVA, Sandra Maria Cunha e cols. **Fatores associados à morte infantil evitável:** uma regressão logística múltipla. Disponivel:https://doi.org/10.11606/S1518-8787.2018052000252 Revista de saúde pública , v. 52, p. 32 de 2018.

COMO LIDAR COM DADOS DESBALANCEADOS EM PROBLEMAS DE CLASSIFICAÇÃO. Disponível em: https://medium.com/data-hackers/como-lidarcom-dados-desbalanceados-em-problemas-de-classifica%C3%A7%C3%A3o17c4d4357ef9. Acesso em: 06 jan. 2022.

GUANIS B. VILELA Júnior, BRÁULIO N. Lima, ADRIANO A. Pereira, MARCELO F. Rodrigues,José Ricardo L. OLIVEIRA , Luís F. SILIO, Anderson S. CARVALHO, Heros R.FERREIRA, Ricardo P. PASSOS. Métricas utilizadas para avaliar a eficiência de classificadores em algoritmos inteligentes. Revista **CPAQV,** V.14, n.2. (2022) DOI: 10.36692/v14n2-01R

Brasil. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Ações Programáticas Estratégicas. Guia de orientações para o Método Canguru na Atenção Básica: cuidado compartilhado / Ministério da Saúde, Secretaria de Atenção à Saúde, Departamento de Ações Programáticas Estratégicas. – Brasília : **Ministério da Saúde**, 2016. 56 p. Disponível em: https://bvsms.saude.gov.br/bvs/publicacoes/guia_orientacoes_metodo_canguru.pdf. Acesso: 20 mar 2024.

Matriz de Confusão para análise de predições em modelos de Aprendizado de Máquina voltado à Análise de Sentimentos em Processamento de Linguagem Natural Disponivel em : https://medium.com/@fcorreiagon/matriz-de-confus%C3%A3o-para-an%C3%A1lise-de-predi%C3%A7%C3%B5es-em-modelos-de-aprendizado-de-m%C3%A1quina-voltado-%C3%A0-an%C3%A1lise-b46b8d491e5b

Escala de Apgar: o que é, como é feito, tabela e resultado, Disponível: https://www.tuasaude.com/escala-de-apgar/ , Acesso em : 10 jun 24.