UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

LEONARDO COSTA CASSARO

REGRESSÃO LOGÍSTICA APLICADA A CASOS DE ANOMALIA CONGÊNITA NO ESTADO DE SÃO PAULO EM 2021

Revisado pela Orientadora

Profa. Dra. Miriam R. Silvestre

Assinatura do Orientadora

Data____/____/ 2023

PRESIDENTE PRUDENTE

2023

LEONARDO COSTA CASSARO

REGRESSÃO LOGÍSTICA APLICADA A CASOS DE ANOMALIA CONGÊNITA NO ESTADO DE SÃO PAULO EM 2021

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

PRESIDENTE PRUDENTE

2023

C343r

Cassaro, Leonardo Costa

Regressão logística aplicada a casos de anomalia congênita no estado de São Paulo em 2021 / Leonardo Costa Cassaro. -- Presidente Prudente, 2023

37 p. : 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. Anomalia congênita. 3. Modelo logístico. 4. Analise multivariada. 5. Descritiva. I. Título.

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

Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

LEONARDO COSTA CASSARO

REGRESSÃO LOGÍSTICA APLICADA A CASOS DE ANOMALIA CONGÊNITA NO ESTADO DE SÃO PAULO EM 2021

Relatório 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:

Orientador: _____

Prof. Dra Miriam Rodrigues Silvestre

Departamento de Estatística

_____

Prof. Dr. Sérgio Minoru Oikawa

Departamento de Estatística

_____

Prof. Dr. Edilson Ferreira Flores

Departamento de Estatística

Presidente Prudente, 04 de dezembro de 2023.

RESUMO

O objetivo desse trabalho é estudar os fatores que mais influenciam na má formação congênita, e assim conseguir conscientizar a população sobre os maiores riscos e que devem ser evitados. A escolha desse tema se deve ao fato de ser um assunto muito importante e que gera muito impacto na sociedade.

Foi utilizado principalmente os conceitos práticos e teóricos da regressão logística múltipla para este relatório. Os dados que foram utilizados são do Sinasc (Sistema de informação de nascidosvivos), e para esse trabalho o foco foi nos nascidos vivos no estado de São Paulo no ano de 2021. Os modelos não ficaram bem ajustados. O modelo que teve o menor AIC teve apenas 51% de acurácia, o que indica que o modelo criado não corresponde aos dados reais. Por outro lado, foi possível identificar que as variáveis parto e apgar1 são significantes para o modelo, e que são significantes para a variável dependente idanomal (anomalia congênita sim ou não)

Palavras chave: Anomalia congênita, fatores de risco, regressão logística múltipla.

ABSTRACT

The objective of this work is to study the factors that most influence congenital malformation, and thus be able to raise awareness among the population about the greatest risks that should be avoided. Choosing this theme must be a very important fact that generates a lot of impact on society. The practical and theoretical concepts of multiple logistic regression were mainly used for this work. The data used is from Sinasc (Live Birth Information System), and for this work the focus was on live births in the state of São Paulo in 2021. The models were not well adjusted. The model that had the lowest AIC had only 51% accuracy, which indicates that the model created does not correspond to the real data. On the other hand, it was possible to identify that the variables birth and apgar1 are significant for the model, and that they are significant for the dependent variable idenoma (congenital anomaly yes or no)

Keywords:Congenital anomaly, risk factors, multiple logistic regression.

SUMÁRIO

1. INTRODUÇÃO

Segundo a Organização Panamericana da Saúde (1994) a má formação congênita é definida como toda anomalia funcional ou estrutural no desenvolvimento do feto, decorrente de fatores originados antes do nascimento, sejam esses genéticos, ambientais ou desconhecidos. Ainda que o defeito não seja aparente e de manifestação clínica mais tardia, é considerado má formação congênita. Essas anomalias podem ser estruturais (quando há alteração anatômica), funcionais (se há alteração de funcionamento sem alterações anatômicas), metabólicas e comportamentais. Alguns dos casos de anomalia mais conhecidos são: cardiopatia congênita (quando há uma alteração na estrutura ou função do coração), defeitos de membros (ausentes, ou desenvolvimento alterado), defeitos do tubo neural (como a anencefalia e espinha bífida) e alterações cromossômicas (síndrome de Down).

Um dos maiores problemas gerados pela má formação congênita é o da mortalidade infantil. No Brasil, a anomalia congênita é a segunda maior causa das mortes infantis, ficando atrás apenas da prematuridade. No ano de 1997, os problemas cardiovasculares corresponderam a 39,4% entre os casos de mortalidade infantil gerados pela má formação, seguidos por defeitos do sistema nervoso, que correspondeu a 18,8% dos casos (Victora, Barros, 2001).

Os exames de pré natal possuem um papel fundamental na detecção de anomalias, além de evitar complicações graves e proteger a saúde do bebê. Exames como o ultrassom morfológico conseguem analisar a formação do feto e identificar possíveis problemas no sistema nervoso, membros e coração. Além disso, o pré natal pode desempenhar um papel de auxílio emocional aos pais, já que podem lidar com essa questão antes do nascimento do filho (Kemp, et al., 1998).

É possível prevenir o surgimento de más formações nos fetos através de medidas tanto individuais quanto coletivas. Algumas dessas medidas relacionadas a mãe são: alimentação saudável, controle do peso e prática de exercícios, evitar consumo de álcool e outras drogas que podem influenciar, idade materna elevada, vacinação em dia de doenças infecciosas e tratamento de doenças, tais como

hipertensão, diabetes e AIDS. Para que essas informações cheguem à população, foi criado em 2005 pelo Ministério da Saúde a Agenda de Compromisso com a Saúde Integral da Criança e Redução da Mortalidade Infantil (Ministerio Da Saude, 2005), cujo foco é reduzir o nascimento de bebês com anomalias congênitas ou outras doenças. Para isso, é importante alertar a sociedade, através de propagandas, palestras, aulas, campanhas, principalmente em áreas menos desenvolvidas, onde não há muito conhecimento sobre essas medidas.

A escala de APGAR é uma medida utilizada para avaliar de maneira imediata o recém-nascido. Esse índice é realizado em dois momentos, no primeiro e quinto minuto pós nascimento. Para poder classificar a saúde do bebê, é necessário checar 5 componentes, sendo eles:

- cor da pele (aparência);
- frequência cardíaca (pulso);
- resposta ao estímulo (gesticulação);
- tônus muscular (atividade);
- respiração

Cada uma das variáveis possui pontuação de 0 a 2, sendo 0 a identificação de algum problema maior momentâneo, e 2 significando normalidade. Para determinar o estado do bebê é realizado uma soma dessas 5 pontuações. Se o índice estiver entre 7 e 10 é considerado baixo risco, entre 4 e 6 é definido como risco moderado, e de 0 a 3 é determinado como alto risco. É importante realizar em 2 etapas essa escala (primeiro e quinto minuto pós nascimento) pois muitas vezes o próprio recém-nascido consegue apresentar evolução em sua saúde nesse curto período de tempo. É necessário dizer que apresentar um baixo índice de APGAR não significa que os bebês vão ter problemas no futuro, relacionados a saúde, comportamento ou inteligência, e sim que irão precisar de um cuidado médico maior naquele determinado momento, garantindo que possam se desenvolver de maneira saudável.

Nesse estudo pretende-se identificar as variáveis que mais influenciam no surgimento de anomalias congênitas em nascidos vivos no estado de São Paulo no ano de 2021. Para realizar as análises desse projeto, serão utilizadas técnicas de regressão logística para 2 classes (com e sem má formação congênita) e multivariada

que possam nos dizer quais são os principais fatores de risco para o surgimento desses casos.

A melhor maneira de diminuir o surgimento de anomalias congênitas é sabendo justamente o que as causa e assim criar soluções para enfrentar essa ocorrência. Com o auxílio de técnicas estatísticas, será possível criar um modelo que identifique os principais fatores de risco para o surgimento dessas malformações?

O principal objetivo nesse projeto é utilizar a regressão logística para estudar e analisar as variáveis que mais influenciam nos casos de má formação genética no estado de São Paulo em 2021.

Para a realização desse estudo traçamos três objetivos específicos que irão nos orientar durante a realização do projeto. São eles:

1. Revisar e aprofundar o conhecimento em regressão logística;
2. Estudar como se comportam as variáveis do banco de dados encontrado;
3. Estudar a aplicação de modelos de regressão logística para 2 classes (com e sem má formação congênita).

A anomalia congênita é um tema que vem ganhando cada vez mais importância no Brasil, passando de quinto para segundo lugar em casos de óbito em menores de um ano entre os anos 1980 e 2000 (Horovitz, 2005).

Estudos mostram que há uma relação entre aumento de tristeza materna com filhos com malformações (Kemp, et al, 1998), visto que alguém nessa situação pode demandar mais atenção dos pais.

Para aprimorar ainda mais esse tema, será essencial a visão de um especialista no assunto, para que possa agregar ainda mais conhecimento sobre a má formação congênita, e como ela pode impactar na vida do bebê e da família. Para o relatório final a ideia é obter essa informação.

Por se tratar de um tema muito relevante, é importante realizar um estudo através da coleta e análise dos dados em que possamos identificar quais são as variáveis que mais se associam ao problema, e assim conseguir conscientizar as pessoas ao redor sobre as melhores maneiras de evitar esse tipo de situação.

Portanto, por ser um tema grave e que vem ganhando cada vez mais importância, justifica-se um projeto de Trabalho de Conclusão de Curso.

2. REFERENCIAL TEÓRICO

A regressão logística é uma técnica estatística em que seu objetivo é a criação de um modelo através de uma variável categórica (normalmente binária) em função de uma ou mais variáveis independentes. A regressão logística é a mais importante para dados com resposta categórica (Agresti, 2002).

Os exemplos mais comuns de aplicação dessa técnica são: análise de crédito (bom ou mau pagador), área da saúde (possui ou não possui a doença), e outros temas que utilizem a variável dependente como: sim ou não, sucesso ou fracasso e 0 ou 1. Nesse projeto será trabalhado a variável dicotômica (com ou sem anomalia congênita)

A regressão logística múltipla é utilizada quando a variável categórica está em função de duas ou mais variáveis independentes. Serão mostrados abaixo dois exemplos da aplicação dessa técnica estatística.

No projeto realizado por Souza (2013), seu objetivo era utilizar a regressão logística para determinar os fatores de risco na ocorrência de anomalia congênita em recém nascidos em uma maternidade de João Pessoa nos anos de 2011 e 2012. Foram utilizadas as seguintes variáveis independentes para estudo: escolaridade da mãe, uso de corticoide no pré Natal, APGAR (1º e 5º minuto), idade da mãe e tipo de parto. Em sua conclusão foi definido que todas essas variáveis influenciaram no surgimento de bebês com malformação genética.

Já no projeto realizado por Martins e Velásquez-Melendez (2004), o objetivo era associar fatores de risco a mortalidade neonatal na cidade de Montes Claros (MG) no período de 1997 a 1999. A variável dependente utilizada foi 1 (ocorrência do óbito) e 0 (não ocorrência do óbito). As variáveis independentes utilizadas para esse estudo foram em relação ao recém-nascido (sexo, peso, APGAR 1º e 5º minuto, idade gestacional), em relação ao parto (tipo de gravidez e parto, número de consultas pré natal e local do nascimento), e por fim em relação à mãe (escolaridade, idade, filhos e abortos). Em sua conclusão foi destacado que sexo e tipo de parto não tiveram relação com a mortalidade neonatal, enquanto a variável pré termo (nascidos em

menos de 37 semanas de gestação) foi a que mostrou a maior associação a variável dependente.

Além disso, outro fator importante é entender como está a distribuição das suas variáveis do banco de dados, e se é necessário agrupar ou não para melhor compreensão.

No projeto realizado por Vanassi et al. (2022), foram considerados apenas 2 classes de APGAR (0 a 7 e 8 a 10). Já no projeto realizado por Vidal e Silva et al. (2018), a variável APGAR possuía 3 classes (0 a 3, 4 a 7 e 8 a 10), e a variável peso ao nascer estava dividido nos seguintes grupos: menor que 1000 g, de 1000 g a 1499 g, de 1500 g a 2499 g, de 2500 g a 3999 g, 4000 g ou mais. E por último, no artigo feito por Leal et al. (2020), o peso foi classificado como baixo para menor que 2500 gramas ao nascer. Essa discussão é importante pois no banco de dados dos nascidos vivos no estado de São Paulo em 2021, as variáveis APGAR1, APGAR5, PESO e IDADEMAE podem ser agrupadas, futuramente será analisado se possuem melhores resultados com grupos ou separadas.

3. METODOLOGIA

Para a realização desse projeto, foram utilizados os dados dos Nascidos Vivos no Estado de São Paulo em 2021, a variável dependente, que classifica se o indivíduo possui ou não anomalia congênita foi chamada de IDANOMAL (1: sim 2:não), com 6699 casos de anomalia congênita e 515810 sem a malformação. Algumas variáveis independentes que foram estudadas, relacionadas ao bebê, como: PESO, APGAR 1, APGAR 5 e SEXO, relacionadas a mãe, por exemplo: ESCMAE (escolaridade), IDADEMAE (idade) e também em relação ao parto, tais como: GESTACAO (número de semanas), PARTO (tipo de parto), CONSULTAS (quantidade de consultas pré natal). A metodologia foi dividida em quatro etapas:

Na primeira etapa o foco foi em revisar os conhecimentos em regressão logística.

Na segunda etapa foi realizada uma análise descritiva do banco de dados (nascidos vivos no estado de São Paulo no ano de 2021) para a compreensão das variáveis independentes.

A terceira etapa o foco foi na criação do modelo. Para isso, será necessário executar técnicas de análise multivariada, utilizando ferramentas como a validação cruzada. Será utilizado o software R para a construção desse projeto.

A quarta etapa é a fase em que foram definidos os melhores modelos criados através de medidas de classificação, como por exemplo a acurácia, precisão, recall, F1-Score.

4. REGRESSÃO LOGÍSTICA MÚLTIPLA

Para adquirir um pouco mais de conhecimento sobre essa técnica estatística, será necessário mostrar um pouco mais de sua estrutura.

Supondo que Ω seja um experimento aleatório com m repetições de natureza binária, ou seja, a variável dependente possui apenas 2 resultados: 0 (fracasso) e 1 (sucesso). Como as repetições possuem as mesmas condições, as probabilidades podem ser dadas por: $P(1) = \Phi$ e $P(0) = 1 - \Phi$ para as m repetições. Se Y for a variável aleatória de interesse que representa o número de vezes nas m repetições em que ocorre sucesso, Y pode assumir os seguintes valores: 0, 1, 2, ..., m. E a cadaum desses valores será associado uma probabilidade P(Y=y), com y assumindo os seguintes valores: 0,1, ..., m. (Barreto, 2011).

Portanto, com essas definições utilizadas para Ω, nota-se que possui propriedades da distribuição Binomial, com parâmetros m e Φ para a variável aleatória Y. Logo, pode-se determinar a sua função de probabilidade, o valor esperado e a sua variância.

$$f(y; = P(Y_i = y) = (\tbinom{m}{y})\Phi^y(1-\Phi)^{m-y} \quad [1]$$

A esperança e variância possuem a seguinte expressão:

$$E(Y_i) = m\Phi \quad [2]$$

$$Var(Y_i) = m\Phi(1-\Phi) \quad [3]$$

Em casos onde ocorra apenas uma repetição, ou seja, m=1, então a distribuição passa a ser uma Bernoulli, e a variável Y_i assumirá apenas o valor 0 ou 1.

Nos modelos lineares generalizados em casos de variáveis aleatórias contínuas, não há restrição para a resposta estimada pelo modelo. Já no caso do experimento Ω, em que a variável aleatória Y_i pertence ao intervalo [0,1] a resposta esperada será uma probabilidade, ou seja, P(Y_i = 1). Para a correção dessa limitação é importante que haja uma transformação sobre Y_i, deixando assim seu domínio contínuo na reta real. A transformação que é normalmente utilizada em variáveis dicotômicas é a função exponencial, transformando assim o valor esperado davariável Y em uma função de ligação logística (Barreto, 2011).

$$E(Y_i) = \frac{\exp(n_i)}{1+\exp(n_i)} \qquad [4]$$

Nota-se que a transformação foi concluída, já que η_i pode assumir qualquer valor real, Φ_1 sendo uma probabilidade segue restrito ao intervalo [0,1]. Então, o valor estimado pela modelo será η_i e possui *p-1* variáveis explicativas, e *p* parâmetros, resultando na seguinte função:

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1}; i = 1,2,\ldots,n \qquad [5]$$

sendo que:

η_i é a resposta média para a i-esima observação;

X_i é o valor da variável preditora para a i-ésima observação;

β_k (k=0,1,...,p-1) é o coeficiente da regressão logística

Substituindo a equação [5] em [4], tem-se que o modelo de regressão logística múltipla é da seguinte forma:

$$E(Y_i) = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{ip-1})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{ip-1})} \quad [6]$$

Assume-se que o Y_i no modelo escrito em [6] são variáveis aleatórias Bernoulli independentes, com parâmetro E(Y_i) = Φ_i

A diferença desse modelo para um modelo de regressão simples é que na regressão simples possuiria apenas uma variável explicativa e somente dois parâmetros a serem estimados, β_0 e β_1.

O próximo passo é explicar mais sobre o odds de sucesso. Quando a probabilidade de sucesso for igual a 0,5 o odds vale 1. Já se a probabilidade for inferior a 0,5 o o odds será inferior a 1. E quando a probabilidade for maior que 0,5 o odds será maior que 1.

Será desenvolvido abaixo $odds_1$.

$$\hat{o}dds_1 = \frac{\Phi}{1 + \hat{\Phi}} \quad [7]$$

Substituindo $\hat{\Phi}$ temos a seguinte equação:

$$\hat{o}dds_1 = \frac{\frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)}}{1 - \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)}} \quad [8]$$

Portanto, $\hat{o}dds_1$ é dado pela seguinte equação:

$$\hat{o}dds_1 = \exp(b_0 + b_1 X) \quad [10]$$

É possível também desenvolver $odds_2$

$$\hat{o}dds_2 = \frac{\frac{\exp(b_0 + b_1(X+1))}{1 + \exp(b_0 + b_1(X+1))}}{1 - \frac{\exp(b_0 + b_1 X + b_1)}{1 + \exp(b_0 + b_1(X+1))}} \quad [11]$$

Seguindo, tem-se:

$$\hat{o}dds_2 = \frac{\exp(b_0 + b_1 X + b_1)}{1 + \exp(b_0 + b_1 X + b_1) - \exp(b_0 + b_0 X + b_{1)})} \quad [12]$$

Seguindo, tem-se a seguinte expressão:

$$\hat{o}dds_2 = \exp(b_0 + b_1 X)\exp(b_1) \quad [13]$$

Então, tem-se que:

$$\hat{o}dds_2 = \hat{o}dds_1 \exp(b_1) \quad [14]$$

Portanto, pode ser concluído que a razão das chances, que determina a taxa de variação dos odds equivale a:

$$\exp(b_1) = \frac{\hat{o}dds_2}{\hat{o}dds_1} \quad [15]$$

Um método para estimar os parâmetros quando a função de distribuição de probabilidade é conhecida, é o da máxima verossimilhança. Seu foco é encontrar estimados que maximizem os parâmetros populacionais.

A verossimilhança é dado pela função de probabilidade conjunta com n observações amostrais, ao redor dos parâmetros de um modelo *β*, sendo denotado por *L(β)*. Assim, o método escolhe o maior valor possível para a função *L(β)*.

Já que a função de probabilidade de um modelo de regressão logístico é conhecido, é possível utilizar o método de máxima verossimilhança. Se houver apenas uma repetição para cada variável aleatória Y_i então será utilizado a distribuição de Bernoulli. Portanto, tem-se que a função de probabilidade de Y_i é dado por:

$$P(Y_i = k) = \Phi_i^k (1 - \Phi)^{1-k}; \;\; k = 0,1 \quad [16]$$

Portanto, assumindo que as variáveis aleatórias Y_i são independentes, a função de probabilidade conjunta pode ser representada por:

$$g(Y_{\mathrm{i}}, \ldots, Y_{\mathrm{n}}; \Phi_i, \ldots, \Phi_n) = \prod_{i=1}^{n} \Phi_i^{ki} (1 - \Phi_i)^{1-ki}; k = 0{,}1 \; e \; i = 1{,}2, \ldots, n \qquad [17]$$

Aplicando ln na função, tem-se que:

$$\ln[g(Y_i, \ldots, Y_n; \Phi_i, \ldots, \Phi_n)] = \ln \prod_{i=1}^{n} \Phi_i^{ki} (1 - \Phi_i)^{1-ki} \qquad [18]$$

Utilizando as propriedades do logaritmo natural a fórmula pode ser escrita da seguinte maneira

$$\ln[g(Y_i, \ldots, Y_n; {}^{\Phi}{}_i, \ldots, {}^{\Phi}n)] = \sum_{i=1}^{n} Y_i ln \frac{{}^{\Phi}i}{1 - {}^{\Phi}i} + \sum_{i=1}^{n} \ln (1 - {}^{\Phi}i) \qquad [19]$$

Portanto, podemos reescrever a equação da seguinte forma:

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \ln [1 + \exp(\beta_0 + \beta_1 X_i)] \qquad [20]$$

Essa função é conhecida como log-verossimilhança para casos de regressão logística simples. As estimativas de máxima verossimilhança dos parâmetros são dados pelos valores de β_0 e β_1 que maximizam essa função (Barreto, 2011). Assim, com a obtenção de MV para β_0 e β_1 em uma regressão logística simples, é possível dizer qual será o valor ajustado para o i-ésimo caso:

$$\hat{\Phi}i = \frac{\exp\,(b_0 + b_1 X)}{1 + \exp\,(b_0 + b_1 X)} \qquad [21]$$

E essa equação obviamente pode ser estendida para a regressão logística múltipla.

A validação cruzada é um método comumente utilizado na regressão, seu foco é mostrar como o modelo se comporta, utilizando dados nunca visto antes, e assim podendo simular uma situação real. Para esse trabalho, foi utilizado a técnica k-fold.

Inicialmente os dados são separados em treino e teste na proporção desejada, e depois há a definição de qual será o valor de K (quantidade de grupos que serão analisados).

O próximo passo é dividir os K grupos e fazer uma iteração entre cada grupo. Após a separação do conjunto para teste, o restante é utilizado para treinamento. Em sequência, o modelo com os dados de treinamento é treinado, testa-se com os dados do teste e assim é possível observar as métricas, e assim sucessivamente, até que todos os grupos possam ser observados, e assim é possível determinar qual dos grupos apresenta as melhores métricas.

Um passo importante para a determinação do melhor modelo, se desenvolve

através do conceito da matriz de confusão. Ela determina a quantidade de falsos positivos (FP) , falsos negativos (FN), verdadeiros positivos (VP) e verdadeiros negativos (VN) na comparação do modelo com os dados reais, e é assim que é possível obter informações valiosas.

A acurácia é utilizada para definir o quanto o seu modelo está acertando em comparação aos dados reais, e a sua fórmula é dada por:

$$Acurácia = \frac{VP + VN}{VP + VN + VN + FP}$$

Outro medidor importante que é obtivo através da matriz de confusão é a precisão, cujo foco é determinar a porcentagem de acertos de classe positivo

$$Precisão = \frac{VP}{VP + FP}$$

Já o recall tem o foco em determinar a taxa de acerto entre os dados de classe positivo como esperado.

$$Recall = \frac{VP}{VP + FN}$$

O F1-Score é utilizado para criar uma média harmônica entre a precisão e recall, e também é um medidor muito importante na hora de comparar os modelos criados

$$F1Score = \frac{2 * Precisão * Recall}{Precisão + Recall}$$

Agora que foi mostrado um pouco mais sobre a regressão logística, simples e múltipla, o próximo tópico terá como foco mostrar os resultados já obtidos pelo banco de dados, mostrando a frequência absoluta e relativa de cada variável dependente e independente. Além disso, será apresentado uma tabela já relacionando IDANOMAL (anomalia sim ou não) e as outras variáveis.

5. RESULTADOS

5.1 Apresentação dos Dados

Serão apresentados algumas variáveis do banco de dados, e que podem ser utilizadas para a criação do modelo. Ao todo tem-se 525239 indivíduos nascidos vivos no ano de 2021 no estado de São Paulo, porém cada variável terá uma quantidade de valores com dados faltantes, porém é um número baixo quando comparado ao tamanho do banco de dados.

As tabelas serão mostradas por frequência absoluta e frequência relativa de cada grupo dentro da variável, seja ela dependente ou independente

- Variável Dependente

  IDANOMAL:

  1: Anomalia congênita (SIM)

  2: Anomalia congênita (NÃO)

Tabela 1: Frequência de anomalia

Anomalia	Frequência absoluta	Frequência relativa
Sim	6699	1,28%
Não	515810	98,72%

Fonte: Própria (2023)

- Variáveis Independentes

  SEXO:

  1: Masculino

  2: Feminino

Tabela 2: Frequência de gênero

Gênero	Frequência absoluta	Frequência relativa
Masculino	268029	51,04%
Feminino	257138	48,96%

Fonte: Própria (2023)

PESO:

Na variável peso do banco de dados dos nascidos vivos no estado de São Paulo no ano de 2021 não há agrupamento para essa variável, portanto será dividido aqui somente para uma melhor visualização, pois ainda precisa ser discutido e analisado para ver qual forma irá obter os melhores resultados (agrupado ou não)

1: 0 a 999 gramas

2: 1000 a 1499 gramas

3: 1500 a 2499 gramas

4: 2500 a 4000 gramas

5: 4000 gramas ou mais

Tabela 3: Frequência de peso

Peso	Frequência absoluta	Frequência relativa
0 a 999 gramas	3644	0,69%
1000 a 1499 gramas	4563	0,87%
1500 a 2499 gramas	41944	7,99%
2500 a 3999 gramas	453306	86,31%
4000 gramas ou mais	21769	4,14%

Fonte: Própria (2023)

APGAR 1:

Apgar1 e apgar5 possuem a mesma discussão da variável peso, aqui será classificado apenas em 2 grupos, baixo ou adequado, mas será necessário avaliar futuramente se o agrupamento é a melhor escolha.

0 a 7: baixo

8 a 10: adequado

Tabela 4: Frequência de APGAR 1

Classificação	Frequência absoluta	Frequência relativa
Baixo	58792	11,24%
Adequado	464486	88,76%

Fonte: Própria (2023)

APGAR 5:

0 a 7: baixo

8 a 10: adequado

Tabela 5: Frequência de APGAR 5

Classificação	Frequência absoluta	Frequência relativa
Baixo	9725	1,86%
Adequado	513624	98,14%

Fonte: Própria (2023)

ESCMAE (escolaridade da mãe, em anos)

1: Nenhuma

2: 1 a 3 anos

3: 4 a 7 anos

4: 8 a 11 anos

5: 12 e mais

Tabela 6: Frequência de ESCMAE

Escolaridade mãe	Frequência absoluta	Frequência relativa
Nenhuma	437	0,08%
1 a 3 anos	2064	0,39%
4 a 7 anos	30336	5,79%
8 a 11 anos	347314	66,24%
12 e mais	144160	27,5%

Fonte: Própria (2023)

IDADEMAE (idade da mãe, em anos):

No banco de dados utilizado, essa variável não possui agrupamentos, portanto, foi gerado alguns grupos apenas para a melhor visualização da distribuiçãodos dados, porém, serão necessários estudos para identificar se de fato é melhor optar pelo agrupamento.

1: 11 a 18 anos

2: 19 a 26 anos

3: 27 a 34 anos

4: 35 a 42 anos

5: 43 a 50 anos

6: 51 ou mais

Tabela 7: Frequência de IDADEMAE

Idade da mãe	Frequência absoluta	Frequência relativa
11 a 18 anos	32608	6,21%
19 a 26 anos	186967	35,60%
27 a 34 anos	197760	37,65%
35 a 42 anos	102571	19,53%
43 a 50 anos	5261	1%
51 ou mais	67	0,01%

Fonte: Própria (2023)

Nota-se que a maior parte das mães possui idade entre 27 a 34 anos, e somente 0,01% possui 51 ou mais anos de idade

GESTACAO (quantidade de semanas de gestação do bebê):

1: Menos de 22 semanas

2: 22 a 27 semanas

3: 28 a 31 semanas

4: 32 a 36 semanas

5: 37 a 41 semanas

6: 42 semanas ou mais

Tabela 8: Frequência de gestação

Semanas gestação	Frequência absoluta	Frequência relativa
Menos de 22 semanas	262	0,05%
22 a 27 semanas	2884	0,55%
28 a 31 semanas	5527	1,05%
32 a 36 semanas	51951	9,90%
37 a 41 semanas	457440	87,20%
42 semanas ou mais	6510	1,24%

Fonte: Própria (2023)

PARTO (tipo de parto):

1: Vaginal

2: Cesárea

Tabela 9: Frequência de tipo de parto

Tipo de Parto	Frequência absoluta	Frequência relativa
Vaginal	217242	41,37%
Cesárea	307920	58,63%

Fonte: Própria (2023)

CONSULTAS (quantidade de consultas pré natal)

1: Nenhuma

2: 1 a 3 consultas

3: 4 a 6 consultas

4: 7 ou mais

Tabela 10: Frequência de consultas

Quant consultas	Frequência absoluta	Frequência relativa
Nenhuma	4451	0,85%
1 a 3	17795	3,40%
4 a 6	72724	13,88%
7 ou mais	429011	81,88%

Fonte: Própria (2023)

As variáveis possuem valores totais diferentes, e isso se deve ao fato de que alguns dados estão sem informação ou sem preenchimento, porém, é um número pequeno comparando a quantidade total do banco de dados dos nascidos vivos no estado de São Paulo no ano de 2021.

O próximo passo será mostrar um pouco mais da relação da variável dependente com as variáveis independentes apresentadas. Para isso, será mostrado abaixo uma tabela com a distribuição de frequência dos fatores de risco utilizando como referência a variável IDANOMAL, que possui as classes sim e não para anomalia congênita, assim será possível compreender um pouco mais sobre a relação da variável dependente com as variáveis independentes, o que irá facilitar para as análises futuras desse projeto, já que para uma análise completa, será necessário obter outras informações, como precisão, acurácia, F1 score, recall entre outras técnicas que possibilitem identificar o melhor modelo, e assim identificar os principais fatores de risco, que é o objetivo principal desse projeto.

Tabela 11: Tabela de distribuição IDANOMAL e Fator de Risco

Fatores de Risco		IDANOMAL					
		1: SIM		2: NÃO		TOTAL	
		Freq	%	Freq	%	Freq	%
Gênero	Masculino	3725	56,14	262941	50,98	266666	51,04
	Feminino	2910	43,86	252869	49,02	255779	48,96
	TOTAL	6635	100	515810	100	522445	100
Escolaridade mãe	Nenhuma	6	0,09	427	0,08	433	0,08
	1 a 3 anos	33	0,49	2023	0,39	2056	0,39
	4 a 7 anos	433	6,47	29734	5,78	30167	5,79
	8 a 11 anos	4341	64,92	341084	66,24	345425	66,22
	12 e mais	1874	28,03	141685	27,51	143559	27,52
	TOTAL	6687	100	514953	100	521640	100
Semanas Gestação	Menos de 22 semanas	12	0,18	242	0,05	254	0,05
	22 a 27 semanas	123	1,84	2736	0,53	2859	0,55
	28 a 31 semanas	245	3,66	5259	1,02	5504	1,05
	32 a 36 semanas	1330	19,88	50414	9,78	51744	9,91
	37 a 41 semanas	4932	73,72	450148	87,37	455080	87,3
	42 semanas ou mais	48	0,72	6401	1,24	6449	1,23
	TOTAL	6690	100	515200	100	521890	100

Tipo de Parto	Vaginal	2288	34,17	214178	41,53	216466	41,43
	Cesárea	4409	65,83	301571	58,47	305980	58,57
	TOTAL	6697	100	515749	100	522446	100
Consultas	Nenhuma	78	1,17	4323	0,84	4401	0,84
	1 a 3	301	4,5	17398	3,38	17699	3,39
	4 a 6	1093	16,37	71288	13,85	72381	13,88
	7 ou mais	5206	77,96	421584	81,93	426790	81,8
	TOTAL	6678	100	514593	100	521721	100

Fonte: Própria (2023)

É possível identificar alguns fatos curiosos, como a variável semanas de gestação, em quebebês nascidos com menos de 36 semanas possuem uma porcentagem maior de casos de anomalia congênita do que os casos sem a má formação. Porém, essas informações isoladas ainda não conseguem definir nada, por isso é tão importante terminar o estudo, para assim conseguir interpretar os resultados da maneira correta.

5.2 Aplicação do modelo

O primeiro passo para a aplicação do modelo é verificar se há um certo desbalanceamento de classes. No conjunto de dados em questão, é possível notar que será necessário uma alteração. É comum isso acontecer em estudos sobre fraudes ou doenças raras, visto que são eventos com uma baixa chance de ocorrer.

O problema em um conjunto de dados em que há um desbalanceamento grande entre as classes, é que a acurácia pode sair prejudicada, e consequentemente a análise e interpretação dos resultados podem ficar ruins.

Existem hoje diversas técnicas que podem ser utilizadas para resolver esse problema. O método oversampling aumenta a quantidade de registros com a frequência mais baixa, seja através de dados duplicados ou valores próximos. Essa aplicação não pôde ser utilizada em este projeto. Os computadores que foram utilizados para gerar os modelos não conseguiram rodar tal técnica.

Uma solução seria trabalhar com menos variáveis, porém a ideia do trabalho é tentar encontrar algum fator de risco ‘diferente’, e por isso é importante trabalhar com o máximo possível de conjunto de dados.

Portanto, iremos utilizar outra técnica, em que o processo é o inverso da oversampling. O método undersampling consiste em reduzir os dados da classe majoritária no conjunto de classe minoritária. Essa técnica pode apresentar problemas, visto que há uma relevante perda de informação. Porém, pela falta de recurso, é a que melhor se encaixou para a conclusão do trabalho.

5.3 Modelo Completo

O primeiro modelo que será apresentado, utilizará todas as variáveis independentes apresentadas até então (apgar1, apgar5, gestação, parto, idade da mãe, escolaridade da mãe, quantidade de consultas pré natal e tipo de parto).

$$glm\ (formula = IDANOMAL \sim PARTO + CONSULTAS + SEXO + APGAR1 + APGAR5 + GESTACAO + IDADEMAE + ESCMAE$$

Tabela 12: ANOVA Modelo Completo

V.A INDEPENDENTE	DF	Chisq	Pr(>Chisq)
PARTO	1	5.2333	0.02216
CONSULTAS	3	0.6827	0.87727
SEXO	1	0.0418	0.83809
ESCMAE	4	3.4460	0.48613
IDADEMAE	37	34.0195	0.60594
APGAR1	10	18.6162	0.04542
APGAR5	10	9.1122	0.52149
GESTACAO	5	3.0796	0.68771

Fonte: Própria (2023)

Ao realizar a ANOVA, chega-se a conclusão que somente as variáveis PARTO e APGAR1 são significantes para o modelo.

Algumas das métricas para esse modelo são:

Tabela 13: Métricas Modelo Completo

MÉTRICAS	Porcentagem
ACURÁCIA	52,76%
PRECISÃO	51,78%
RECALL	52,82%
F1-SCORE	52,31%

Fonte: Própria (2023)

Portanto, é possível perceber que o modelo não está bem ajustado, e que o modelo não corresponde aos dados reais. Costuma-se esperar uma acurácia de no mínimo 70% para que um modelo seja bem avaliado. Alguns fatores podem ter influenciado nesses valores, como a técnica utilizada para o desbalanceamento, mas o principal fator seria a própria variável dependente IDANOMAL, em que até hoje existem muitas dúvidas sobre os principais fatores que podem influenciar em um recém nascido com ou sem a anomalia.

5.4 Modelo 2

O segundo modelo se baseia no resultado gerado pelo primeiro, que determinou que somente as variáveis PARTO e APGAR1 são significantes para o modelo.

$$glm(formula = IDANOMAL \sim APGAR1 + PARTO$$

Tabela 14: ANOVA Modelo 2

V.A INDEPENDENTE	DF	Chisq	Pr(>Chisq)
PARTO	1	4.9607	0.02593
APGAR1	10	17.9478	0.04685

Fonte: Própria (2023)

X1 = PARTO

X2 = APGAR1

$$y = -0{,}89 + 0{,}02X1 + 0{,}875X2$$

Foi realizado o ANOVA para determinar se as duas variáveis irão continuar significantes para o modelo em questão. Podemos concluir que as 2 variáveis seguem como significantes para o modelo.

Tabela 15: Métricas Modelo 2

MÉTRICAS	Porcentagem
ACURÁCIA	51,29%
PRECISÃO	57,83%
RECALL	51,14%
F1-SCORE	54,28%

Fonte: Própria (2023)

A conclusão é muito parecida com o primeiro modelo, infelizmente o modelo não está representando de maneira positiva os dados reais.

5.5 Modelo stepwise

Por último, o modelo stepwise foi criado, para verificar de fato se apenas a variável parto será significante para o trabalho realizado.

O modelo que apresentou o menor AIC, teve a seguinte estimativa:

Tabela 16: ANOVA Modelo stepwise

V.A INDEPENDENTE	DF	Chisq	Pr(>Chisq)
PARTO	1	4.9607	0.04784

Fonte: Própria (2023)

5.6 Melhor modelo

Para a comparação do melhor modelo, foi utilizado os valores de AIC:

Tabela 17: Comparação AIC

Modelo	AIC
Modelo1	18620
Modelo2	18570
Stepwise	18572

Fonte: Própria (2023)

Portanto, como o modelo 2 apresentou o menor AIC, podemos afirmar que possui o melhor modelo entre os testados.

Por fim, foi realizado dentro do modelo 2 a razão de chances, para verificar quais grupos possuem maior chance de estar relacionados a presença da anomalia congênita.

Tabela 18: Comparação AIC

Razão de Chance	Porcentagem
APGAR1: 1 ponto	199,3%
APGAR1: 2 pontos	143,43%
APGAR1: 3 pontos	48,89%
PARTO2: Cesárea	8,14%

Fonte: Própria (2023)

Portanto, quando a pontuação de APGAR no primeiro minuto for 1, a chance do indíviduo ter a anomalia congênita aumenta em cerca de 199,3%. Já quando APGAR1 for igual a 2, a chance aumenta em 143,43%. E por fim, quando o parto é realizado através de uma cesárea, há um aumento de 8,14% de chance do indivíduo ter a anomalia congênita.

Portanto, é possível identificar alguns padrões dentro do modelo 2, com as 2 variáveis que se mostraram significantes.

6. CONSIDERAÇÕES FINAIS

Esse tema vem ganhando cada vez mais importância à medida que a ciência evolui e busca conhecer mais sobre o ser humano. O foco do trabalho foi encontrar osprincipais fatores de riscos para a formação da anomalia congênita. O banco de dadosutilizado foi do Sinasc (Sistema de Informação de Nascidos Vivos), e foi feito com dados do estado de São Paulo no ano de 2021.

Os resultados não foram muito bons. O modelo que apresentou a melhor acurácia teve somente 51% de acurácia, o que é considerado baixo. Algumas literaturas citam que para um modelo ser considerado adequado, é necessário que a acurácia tenha no mínimo 70%. O principal motivo se deve ao fato de que a anomalia congênita ainda é um tema de muitas incertezas, e que até hoje não se sabe exatamente a origem de todos os casos.

Embora os resultados não sejam promissores, foi possível observar que para esse banco de dados, as variáveis PARTO e APGAR1 influenciam nos resultados. Ou seja, o objetivo do trabalho foi alcançado, visto que a ideia era identificar os principais fatores de risco para a presença de anomalia congênita.

Foi de extrema importância realizar esse trabalho de conclusão, principalmente para conhecer novas técnicas computacionais, e relembrar todo o aprendizado na graduação. Além disso, foi extremamente importante conhecer mais sobre esse tema da saúde, e que pode afetar a sociedade como um todo. Ao longo do ano, pude aprender muito mais sobre as anomalias, maneiras de evitá-las principalmente através da educação.

Para projetos futuros, esse banco de dados é bem completo, visto que há ao todo 61 variáveis, incluindo algumas com datas e locais. E dentro do tema da anomalia congênita, outros bons trabalhos seriam utilizando algumas outras variáveis, e assim podendo verificar quais podem ser significantes para a má formação.

REFERÊNCIAS

AGRESTI, A. **Categorycal Data Analysis.** 2nd ed. New Jersey: John Wiley e Sons, 2002.

BARRETO, A. S. Modelos de Regressão: **Teoria e Aplicação com o Programa Estatístico R**, Edição do Autor, 1 Edição, p. 109-114, Brasília 2011.

HOROVITZ, D. D. G.; LLERENA, J. C.; MATTOS, R. A. Birth defects and health strategies in Brazil: an overview. **Cadernos de Saúde Pública**, v. 21, n. 4, p. 1055-1064, 2006.

KEMP J.; DAVENPORT, M.; PERNET A. Antenatal diagnosed surgical anomalies: the psychological effect of parental antenatal counseling. **Journal of Pediatric Surgery**, v. 33, n. 9, p. 1376-1379, 1998.

LEAL, et al. (2020). **Prenatal care in the Brazilian public health services**. Revista De SaúdePública, 54, 08.
Disponível em: https://doi.org/10.11606/s1518-8787.2020054001458. Acesso em 17 agosto 2023

MANUAL da Classificação Estatística Internacional de Doenças, Lesões e Causas de Óbito; 9a revisão. São Paulo, **Centro Brasileiro de Classificação de Doenças**, 1980.

MARTINS, E. F; VELÁSQUEZ-MELÉNDEZ, **Determinantes da mortalidade neonatal a partir de uma coorte de nascidos vivos, Montes Claros, Minas Gerais, 1997-1999**. **Rev Bras Saúde Matern Infant**, v. 4, n. 4, p. 405-412, 2004.

Organização Panamericana de Saúde. **Saúde materno infantil:** atenção primária nas Américas. Organização Panamericana de Saúde: Washington, DC; 1994.

SOUZA, L. D. A. **Uso de Regressão Logística para Identificar os Fatores de Risco associados à Ocorrência de Anomalias Congênitas em Recém-nascidos.** João Pessoa, 2013. Monografia (Graduação em Estatística). UFPB.
Disponível em: https://repositorio.ufpb.br/jspui/handle/123456789/823. Acesso em: 21 abril 2023.

VANASSI, B. M et al. Congenital anomalies in Santa Catarina: case distribution and trends in 2010–2018. Revista Paulista de Pediatria [online]. 2022, v. 40. Disponível em: https://doi.org/10.1590/1984-0462/2022/40/2020331 . Acesso em: 17 agosto 2023

Vidal e Silva et al (2018). **Factors associated with preventable infant death: a multiple logistic regression**. Revista De Saúde Pública, 52, 32.
Disponível em: https://doi.org/10.11606/S1518-8787.2018052000252. Acesso em: 17 agosto 2023

VICTORA C. G., BARROS F. C.; Infant mortality due to perinatal causes in Brazil: trends, regional patterns and possible interventions. **Sao Paulo Medical Jornal**, p 33-42, 2001.