UFSCar - UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
DEPARTAMENTO DE ESTATÍSTICA

Projeto 2 - Grupo 1 - Dados 5

Laboratório de Estatística Aplicada

Luis Roberto Ferreira Junior RA: 744864
Luiz Paulo Dal Sasso RA: 631965
Rafael Setti RA: 744870

São Carlos - SP
9 de fevereiro de 2023

Conteúdo

1 Introdução

A mortalidade fetal é um problema de saúde pública muito importante, caracterizado pela morte de um bebê que ainda não nasceu e que está com mais de 20 semanas. Até a vigésima semana de gestação, a morte é considerada um aborto espontâneo. O nascimento de um bebê morto (natimorto) é um problema que atinge com mais frequência países de baixa e média renda, mas que também acontece em países de alta renda. Esse problema, além do número de natimortos não estar caindo nos últimos anos, não é visto como um problema de saúde mundial. Com a mortalidade fetal vem um maior risco de resultados adversos para a saúde materna, bem como o aumento do risco de mortalidade materna. Cardiotocografias (CTGs) medem valores como frequência cardíaca fetal, movimento fetal e contrações uterinas. Dessa forma, CTGs são uma opção simples e acessível para avaliar a saúde fetal, permitindo que os profissionais de saúde tomem medidas para prevenir a mortalidade infantil e materna.

Visto que o risco da mortalidade fetal é um problema tão devastador, o que pode ser feito para diminuir esses números e preservar a saúde materna e fetal? Ao longo desse trabalho, vamos buscar responder a essa questão de como prever os resultados da saúde fetal com base nos dados do CTG. Essas informações podem ser utilizadas por profissionais médicos, especificamente na área de obstetrícia, para minimizar a ocorrência de mortalidade fetal. Embora isso seja indiscutivelmente mais um problema médico do que um problema matemático, as práticas médicas podem se beneficiar muito dessas descobertas, garantindo a melhor saúde possível do paciente.

2 Objetivo

O objetivo desse trabalho é, através dos dados coletados referentes aos resultados obtidos em exames CTGs em gestantes, sermos capazes de ajustar um modelo de classificação que seja suficientemente efetivo para auxiliar na decisão dos profissionais da saúde em decidir o estado de saúde do feto.

3 Materiais e Métodos

3.1 Materiais

A redução da mortalidade infantil reflete-se em vários Objetivos de Desenvolvimento Sustentável das Nações Unidas e é um indicador-chave do progresso humano. A ONU espera que, até 2030, os países acabem com as mortes evitáveis de recém-nascidos e crianças menores de 5 anos, com todos os países visando reduzir a mortalidade de menores de 5 anos para pelo menos 25 por 1.000 nascidos vivos. Para continuarmos avançando em na eficácia dessa identificação de pacientes sob risco, foi desenvolvido os métodos de CTG computadorizados com resposta computadorizada para potencializarmos o poder de diagnóstico. O termo utilizado para mensurar essa variação em curto período de tempo chama-se STV - *Do inglês, Short term variation*, que examina a variabilidade da frequência cardíaca do feto entre as batidas e não podem ser interpretados por análise visual. Esses valores só conseguem ser obtidos a partir de CTG computadorizada.

Os dados que possuímos para desenvolvimento desse estudo são de 2126 fetos que foram avaliados segundo medidas obtidas pelo Cardiotocograma avaliados ao longo de vários estudos observacionais prospectivos de vários centros hospitalares (foram encontrado estudos utlizando valores de STV e coletas de dados de 1983 à 2011) e os valores de STV foram calculados durante o trabalho de parto ativo.

A respeito da nossa base de dados, possuímos 2126 observações e 22 variáveis, em que a variável de interesse desse trabalho é o estado de saúde do feto, que possui 3 classes: 1 - Normal, 2 - Sob Suspeita, 3 - Doente. As variáveis que compõe o banco de estudo, assim como suas descrições, são dadas por:

- Saúde Fetal - Nível de saúde do Feto. 1 - Normal, 2 - Suspect, 3 - Pathological;
- Taxa de batimento basal - Taxa de batimento cardíaco por minuto do feto;
- Acelarações - Número de acelerações do batimento por segundo;
- Movimentos Fetais - Número de movimentos do feto por segundo;
- Contrações uterinas - Número de contrações uterinas por segundo;
- Desacelerações pequenas - Número de pequenas desacelerações dos batimentos por segundo;

- Desacelerações bruscas - Número de desacelerações bruscas dos batimentos por segundo;
- Desacelerações prolongadas - Número de desacelerações prolongadas por segundo;
- Tempo de STV anormal - Percentual de tempo com STV anormal;
- Tempo médio de STV - Tempo médio para STV (variação de curto periodo);
- Tempo de LTV anormal - Percentual de tempo com variação de longo período anormal (LTV);
- Tempo médio de LTV - Tempo médio para LTV (variação de longo período);
- Largura do Histograma - Histograma de frequência cardíaca (Gerado pelo exame);
- Mínimo do Histograma - Mínimo do histograma de FC (Gerado pelo exame);
- Máximo do Histograma - Máximo do histograma de FC (Gerado pelo exame);
- Número de picos do histograma - Número de picos do histograma de FC (Gerado pelo exame);
- Número de zeros - Número de zeros do histograma de FC (Gerado pelo exame);
- Moda do histograma - Moda observada do histograma de FC (Gerado pelo exame);
- Média do histograma - Média observada do histograma de FC (Gerado pelo exame);
- Mediana do histograma - Mediana observada do histograma de FC (Gerado pelo exame);
- Variância do histograma - Variância observada do histograma de FC (Gerado pelo exame);
- Tendência do histograma - Tendência observada do histograma de FC (Gerado pelo exame).

3.2 Métodos

Para iniciar a discussão a respeito da metodologia que aplicaremos, olhemos para distribuição da nossa variável de interesse apresentado na Figura 1 e na Tabela 1.

Visto que classificadores tem dificuldades em lidar com amostras muito desbalanceadas e que a quantidade de indivíduos presentes no grupo dos doentes é bastante inferior,

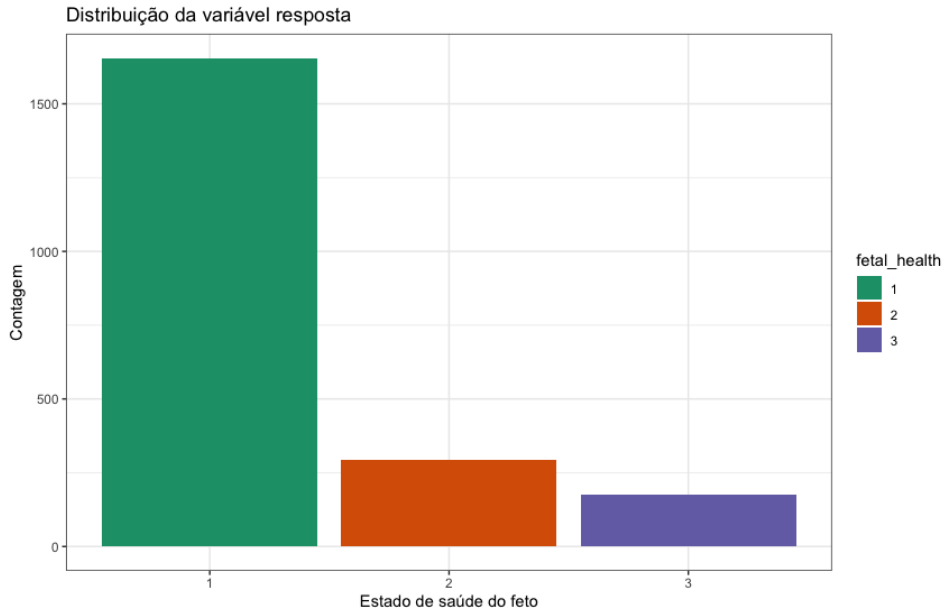

Figura 1: Distribuição do estado de saúde do feto.

	Contagem
1	1655
2	295
3	176

Tabela 1: Contagem dos indivíduos por estado de saúde, onde 1 = Saudável, 2 = Sob suspeita, 3 = Doente.

nesse estudo, optaremos pelo agrupamento entre os indivíduos da classe 2 e 3, os posicionando em uma nova classe que classifcaremos de acordo com a qualidade do resultado do CTG que chamaremos de "Não Satisfatórios". Os indivíduos que anteriormente eram "Normais" passarão a se chamar de acordo com o resultado de exame, "Satisfatórios" representado pela classe 0 e "Não Satisfatórios" representado pela classe 1.

Dessa forma, a distribuição do nossa nova variável de interesse se da seguinte forma como indicado na Figura 2:

Como observado, ainda presenciamos um alto grau de desbalanceamento no conjunto de dados, contudo, o classificador adquire mais informações a respeito da classe dos indivíduos com características não satisfatórias em linhas gerais para o exame.

Dessa forma, faremos uma classificação binária da nossa variável de interesse a partir de uma regressão logística de maneira que, sejam p e n números naturais não-nulos tais que $p < n$. Considere um conjunto de dados composto por p covariáveis, que descrevem características de n unidades amostrais, e sejam $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias independentes

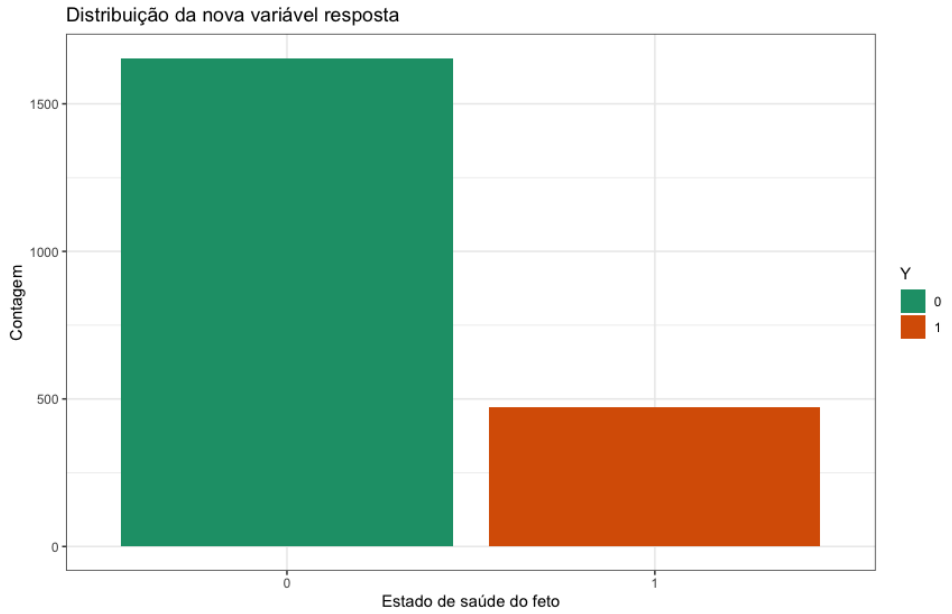

Figura 2: Distribuição do estado de saúde do feto com a nova variável resposta em que 1 = Sob Risco e 0 = Saudaveis.

definidas em um espaço de probabilidade (Ω, F, P) tais que

$$Y_i = \begin{cases} 1, & \text{se a } i\text{-ésima unidade amostral não teve resultado satisfatório no CTG,} \\ 0, & \text{se a } i\text{-ésima unidade amostral teve resultado satisfatório no CTG.} \end{cases}$$

para todo $i = 1, 2, \ldots, n$.

Matricialmente, temos:

$$\boldsymbol{X} := \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{bmatrix} \quad \text{e} \quad \boldsymbol{Y} := \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}.$$

Onde, $\boldsymbol{X}$ é uma matriz de covariáveis com n linhas e p colunas e $\boldsymbol{Y}$ um vetor de tamanho n associado as respostas da variável de interesse.

Dado o conjunto de treinamento, isto é, um subconjunto da amostra observada, desejamos encontrar um classificador que tenha boa performance em discriminar novas unidades amostrais entre as classes de estudo. Assim, modelamos a probabilidade de uma nova unidade amostral pertencer à classe dos resultados "Não satisfatorios", condicionada às observações das covariáveis consideradas no estudo, como uma função dessas covariáveis, ou

seja,

$$P(Y = 1|\boldsymbol{x}) = \pi(\boldsymbol{x}) := \frac{e^{\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}}}{1 + e^{\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}}},$$

em que $\beta_0 \in \mathbb{R}$ representa o intercepto, $\boldsymbol{\beta} := (\beta_1, \beta_2, ..., \beta_p)^T$ é um vetor real de ordem $p \times 1$ que representa os coeficientes associados a cada uma das p covariáveis consideradas no estudo e $\boldsymbol{x}$ é o vetor real de ordem $1 \times p$ que representa o valor de cada uma das p covariáveis observadas na nova unidade amostral em questão.

Neste trabalho, vamos estimar os parâmetros β_0 e $\boldsymbol{\beta}$ do modelo de regressão logística ou, equivalentemente, treinar o classificador logístico, utilizando o método da máxima verossimilhança. Para isso, assumiremos que $Y_1, Y_2, ..., Y_n$ são variáveis aleatórias independentes e $Y_i|\boldsymbol{X_i} = \boldsymbol{x_i} \sim Bernoulli(\pi(\boldsymbol{x}_i))$, em que $\boldsymbol{X_i}$ é o vetor de covariáveis associadas a i-ésima unidade amostral. Sendo assim, podemos escrever a distribuição de $Y_i|\boldsymbol{X_i} = \boldsymbol{x}_i$ da seguinte forma:

$$P(Y_i = y_i|\boldsymbol{x_i}) = \pi(\boldsymbol{x_i})^{y_i}(1 - \pi(\boldsymbol{x_i}))^{1-y_i}\mathbb{I}_{\{0,1\}}(y_i)$$

em que $\mathbb{I}$ denota a função indicadora.

Dessa forma, a função ℓ de log-verossimilhança é dada por

$$\ell(\beta_0, \boldsymbol{\beta}|\boldsymbol{X}) := \sum_{i=1}^{n} \left[\log\left(1 - \frac{e^{\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}}}{1 + e^{\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}}}\right) + y_i(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}) \right],$$

e, consequentemente, os estimadores de máxima verossimilhança são definidos da seguinte forma

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) := \arg \min_{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p} -\ell(\beta_0, \boldsymbol{\beta}|\boldsymbol{X}).$$

O classificador logístico obtido através do método de máxima verossimilhança é, portanto, definido da seguinte maneira

$$\hat{\pi}(\boldsymbol{x}^*) := \frac{e^{\hat{\beta}_0 + (\boldsymbol{x}^*)^T\hat{\boldsymbol{\beta}}}}{1 + e^{\hat{\beta}_0 + (\boldsymbol{x}^*)^T\hat{\boldsymbol{\beta}}}}, \tag{1}$$

em que $\boldsymbol{x}^*$ é o vetor de covariáveis observadas em uma nova unidade amostral. Nesse sentido, a classificação $y^* \in \{0, 1\}$ do novo exame é dada a partir da seguinte regra de decisão:

$$y^* = 1 \Leftrightarrow \hat{\pi}(\boldsymbol{x}^*) \geq c,$$

em que c é uma constante pré-fixada.

4 Resultados

4.1 Análise descritiva

Nosso conjunto de dados é composto por 21 covariáveis, sendo uma delas, categórica. A única covariável categórica presente nesse trabalho é a "Tendência do histograma", que assume níveis -1, 0 ou 1. De resto, temos outras 20 covariáveis numéricas:

Variável	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Acelerações	0.000	0.000	0.002	0.003	0.006	0.019
Desacelerações prolongadas	0.000	0.000	0.000	0.0002	0.000	0.005
Desacelerações severas	0.000	0.000	0.000	3.3e-06	0.000	1.0e-03
Desacelerações leves	0.000	0.000	0.000	0.002	0.003	0.015
Movimentos Fetais	0.000	0.000	0.000	0.010	0.003	0.481
Frequência Basal	106	126	133	133,3	140	160
Contrações uterinas	0.000	0.002	0.004	0.0043	0.007	0.015
STV Anormal	12	32	49	47	61	87
STV Medio	0.200	0.700	1.200	1.333	1.700	7.000
% de Tempo STV	0.000	0.000	0.000	9.847	11.00	91.00
LTV Medio	0.00	4.60	7.40	8.20	10.80	50.70
Largura do Histograma	3.00	37.00	67.50	70.45	100.00	180.00
Minimo do Histograma	50.00	67.00	93.00	93.58	120.00	159.00
Maximo do Histograma	122	152	162	164	174	238
Numero picos Histograma	0.000	2.000	3.000	4.068	6.000	18.000
Numero de zeros Histograma	0.000	0.000	0.000	0.3236	0.000	10.000
Moda do histograma	60.0	129.0	139.0	137.50	148.00	187.00
Media do histograma	73.00	125.00	136.00	134.60	145.00	182.00
Mediana do histograma	77.00	129.00	139.00	138.10	148.00	186.00
Variancia do histograma	0.00	2.00	7.00	18.81	24.00	269.00

Primeiramente, é interessante notarmos que uma grande diferença na escala de algumas variáveis entre si. Muitas das variáveis que compõem nosso banco de dados possuem muitos dos valores iguais a zero e variam em escala muito diferente de outras. Isso pode vir a ser um ponto de atenção no caso do desenvolvimento de um modelo preditivo em que covariáveis que possivelmente foram identificadas como significativas componham o modelo final havendo variáveis com escalas muito distintas.

Quanto a distribuição de cada uma das variáveis, é possível observar na Figura 3 que algumas variáveis tem como a grande maioria dos valores iguais a zero, sendo esses valores referentes a todos os tipos de desacelerações cardíacas e acelerações, também a respeito do número de movimentação do feto.

Algumas dessas distribuições são indicadas de maneira ampliada na Figura 4:

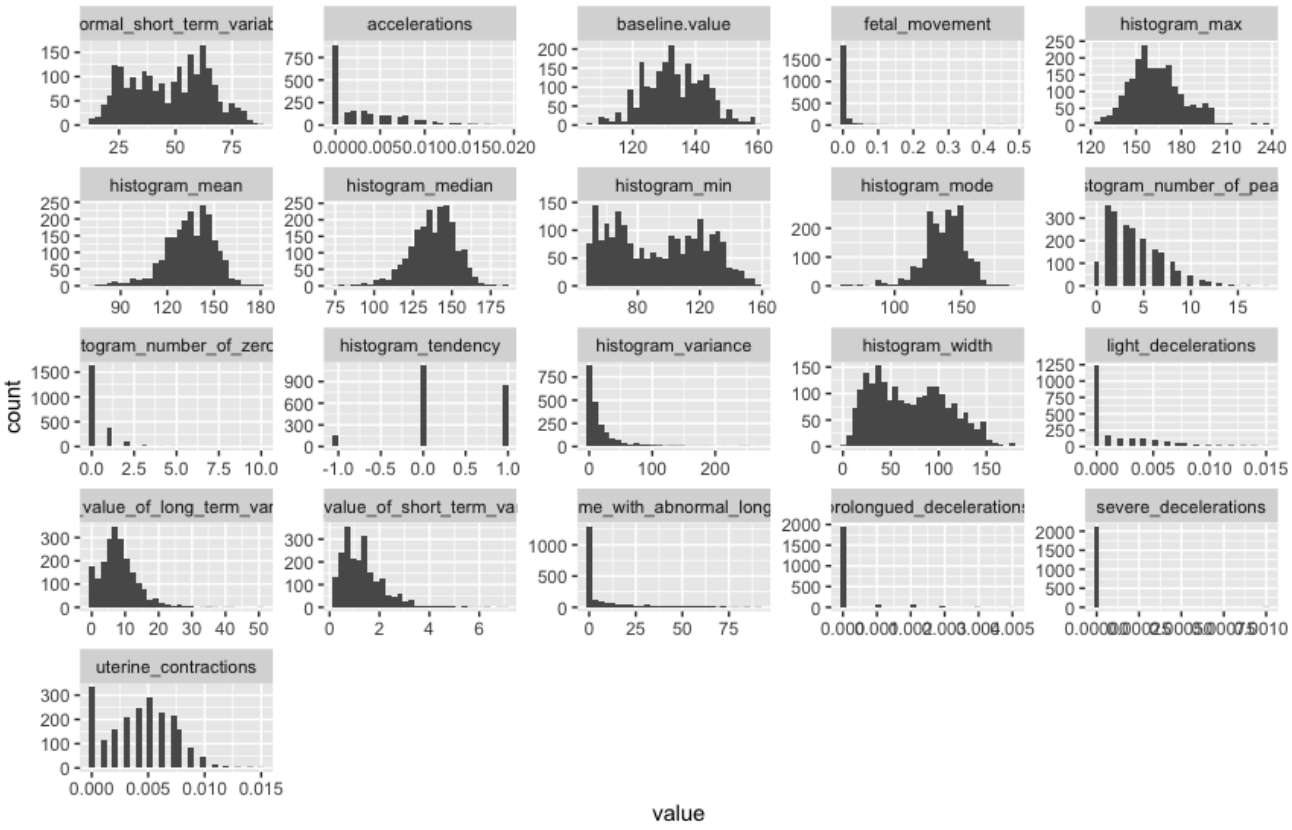

Figura 3: Histograma das variáveis numéricas presentes no estudo.

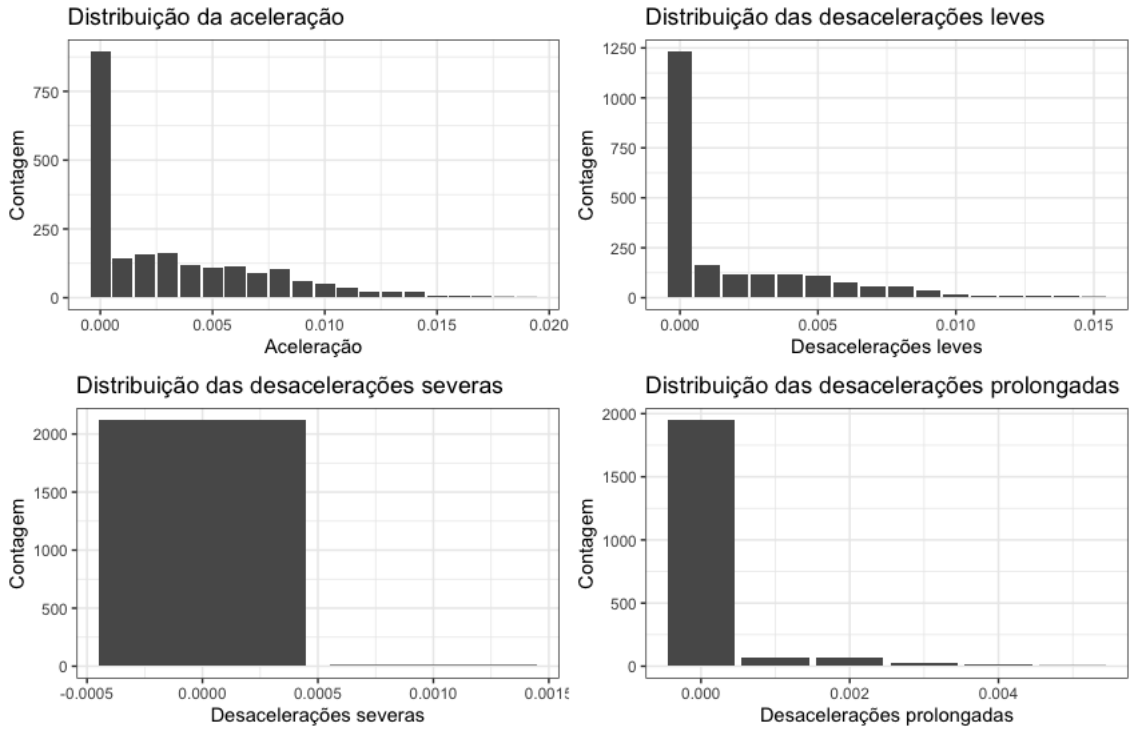

Figura 4: Histogramas das acelerações e desacelerações.

Também é interessante olharmos exclusivamente para algumas outras medidas que são referentes as variações de Frequências cardíacas a longo e curto prazo, comparando os dois grupos de interesse (“Resultados satisfatórios” e “Resultados não satisfatórios”) conforme indicado na Figura 5.

A respeito da Figura 5, “Variações anormais de curto período”, “frequência cardíaca basal” e “percentual de variações anormais mais prolongadas” aparentam distinguir bastante em suas distribuições quando comparadas interclasses. Novamente é interessante olharmos para variável acelerações que possui uma distribuição de valores bastante distinta quando comaparada entre o grupo dos resultados “Satisfatórios” e “Não Satisfatórios”.

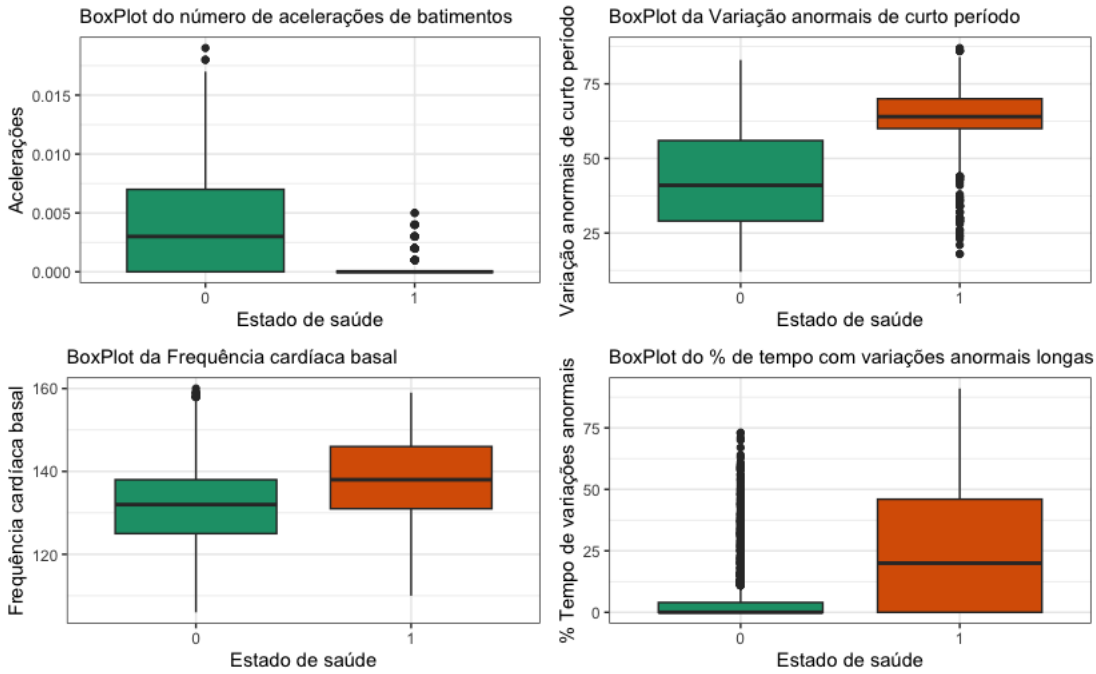

Figura 5: BoxPlot de variáveis numéricas referentes a frequência cardíaca.

Algumas medidas de resumo a respeito de taxas cardíacas mais comuns de serem mensuradas (Número de acelerações e Frequência cardíaca basal) podem ser visualizadas na Tabela 4 e na Tabela 3.

Estatística	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Resultados Satisfatórios	0.000	0.000	0.003	0.0039	0.0070	0.0190
Resultados não Satisfatórios	0.000	0.000	0.000	0.0003	0.0000	0.0050

Tabela 2: Estatísticas resumo da variável número de acelerações dos batimentos.

Essas diferenças entre as distribuições associados as classes da variável resposta ilustrados na Figura 5 nos dá indícios de possíveis variáveis relevantes para composição do nosso modelo final e, ao que parece, a maioria das variáveis que dizem respeito aos batimentos

Estatística	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Resultados Satisfatórios	106	125	132	132	138	160
Resultados não Satisfatórios	110	131	138	137,9	146	159

Tabela 3: Estatísticas resumo da frequência cardíaca basal.

cardíacos e tais variações se mostram possivelmente relevantes para constiruição do nosso modelo final.

Uma outra medida interessante de olharmos que é a única variável que não diz respeito à frequência cardíaca e suas variações, é o número de contrações uterinas por segundo. Assim, olhemos para o comportamento das distribuição comparando as classes de estudo conforme apresentado na Figura 6:

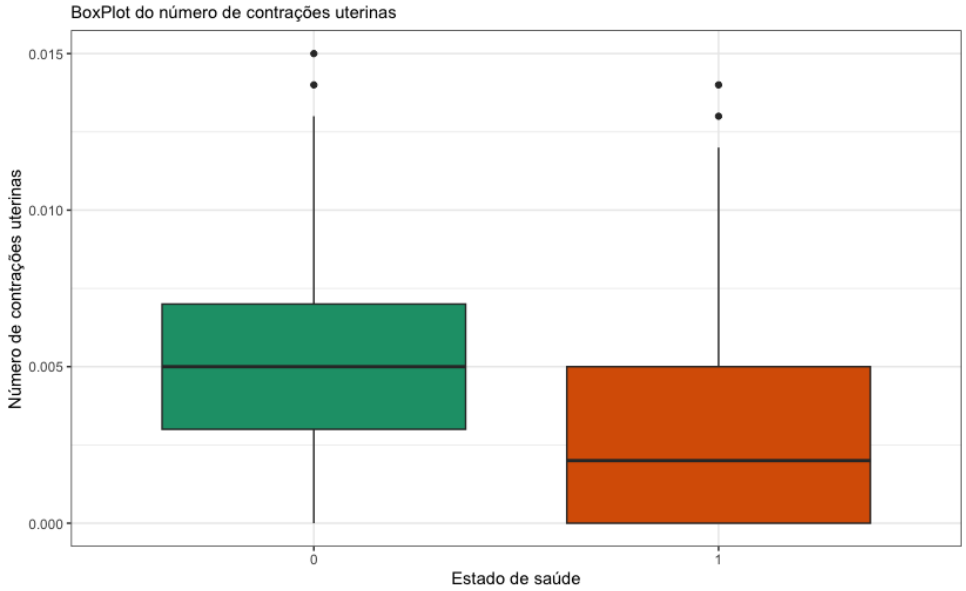

Figura 6: BoxPlot do número de contrações uterinas.

Tal como a maioria das outras distribuições apresentadas, para número de contrações uterinas por segundo, também foi possível notar que, normalmente tais contrações ocorrem com mais frequência para quando o paciente tem “Satisfatórios” no exame, nos dando indícios que essa variável fisiológica possa ser relevante para compor nosso modelo final.

Estatística	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Resultados Satisfatórios	0.000	0.003	0.005	0.0048	0.007	0.015
Resultados não Satisfatórios	0.000	0.000	0.002	0.0030	0.005	0.014

Tabela 4: Estatísticas resumo da variável número de contrações uterinas.

Tendo em vista a presença de várias medidas relativas a variações cardíacas do feto, é de noss interesse interpretar a respeito da correlação entre as variáveis quantitativas presentes em nosso problema de estudo. A matriz de correlação está indicada na Figura 7.

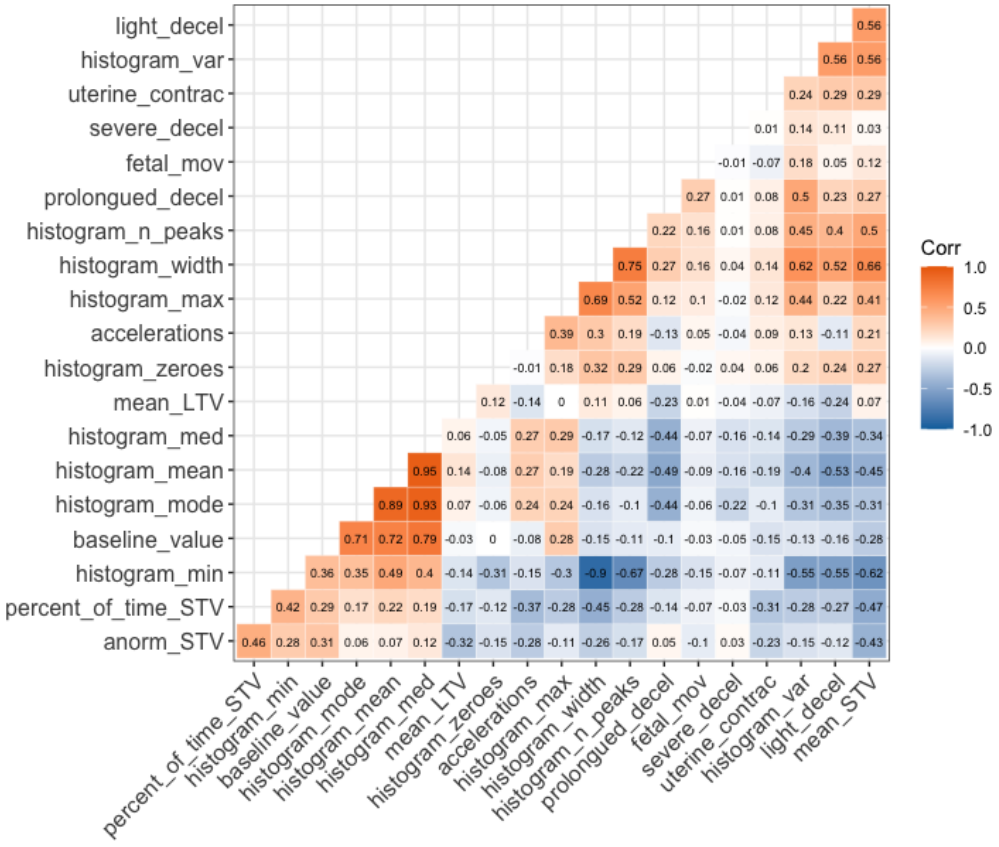

Figura 7: Matriz de correlação das variáveis quantitativas.

Como visto na Figura 7, as medidas de tendência central que dizem respeito a distribuição da frequência cardíaca são altamente correlacionadas entre si.

Dessa forma, passa a surgir indícios que não possamos incluir todas essas variáveis no modelo final, já que a inclusão de todas elas pode nos levar a enfrentar problemas de multicolinearidade.

De maneira geral, é possível observar que há muitas variáveis que se correlacionam entre si de alguma forma, tanto positivamente, quanto negativamente. Dessa forma, temos mais alguns indícios que precisamos prosseguir com alguns métodos de seleção de variáveis afim de lidarmos com eventuais problemas de multicolinearidade.

4.2 Regressão Logística

Conforme foi introduzido em nossa metodologia, faremos uma classificação binária usando da regressão logística.

Primeiramente, vamos prosseguir com um modelo utilizando todas as covariáveis presentes no banco de dados com excessão de incluir todas as variáveis que dizem respeito ao histograma gerado pelo exame, pois estas são, em geral, correlacionadas entre si.

Dessa forma, demos prosseguimento em um modelo inicial que inclui somente algumas das variáveis referentes ao comportamento do Histograma gerado pelo exame. Através da análise descritiva, optamos por retirar do primeiro modelo as variáveis "Tendência do Histograma", "Media do Histograma", "Mediana do Histograma", "Maximo do Histograma", "Numero de zeros do Histograma" e "Largura do Histograma".

Posteriormente, após um ajuste de modelo incluindo todas as variáveis que selecionamos a partir da análise descritiva, os níveis de significância das 3 variáveis, sendo elas "Desaceleracoes severas", "Desaceleracoes leves" e "Movimentos fetais", não se mostraram signficativas para compor o modelo a um nível de significância de $\alpha = 0,05$.

Assim, removendo do modelo final as covariáveis citadas anteriormente, foi desenvolvido um primeiro modelo geral a partir da Regressão Logística contendo 12 covariáveis, sendo todas elas numéricas. Os coeficientes e suas significativas correspondentes podem ser visualizadas na Tabela 5. Assim, é possível observar que todas as variáveis se mostraram signficativas para compor um modelo final.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-15.3206	1.8558	-8.26	0.0001
accelerations	-599.6186	96.0819	-6.24	0.0001
prolongued_decelerations	1982.0540	305.2418	6.49	0.0001
baseline.value	0.0885	0.0225	3.93	0.0001
uterine_contractions	-234.6758	41.9873	-5.59	0.0001
abnormal_short_term_variability	0.0839	0.0101	8.32	0.0001
mean_value_of_short_term_variability	-0.5414	0.2284	-2.37	0.0177
Perc_T_with_abnormal_long_term_variability	0.0265	0.0065	4.09	0.0001
mean_value_of_long_term_variability	0.0694	0.0299	2.32	0.0204
histogram_min	0.0231	0.0061	3.78	0.0002
histogram_number_of_peaks	0.1845	0.0542	3.40	0.0007
histogram_mode	-0.0371	0.0162	-2.28	0.0224
histogram_variance	0.0419	0.0082	5.08	0.0001

Tabela 5: Coeficientes do modelo geral.

Na sequência, como multicolinearidade era uma preocupação em nossa pesquisa, vamos observar como se dão os valores de VIF (Fator de inflação de variância), conforme indicado na Tabela 6.

	VIF
accelerations	1.20
prolongued_decelerations	1.78
baseline.value	3.89
uterine_contractions	1.23
abnormal_short_term_variability	1.47
mean_value_of_short_term_variability	2.99
Perc_T_with_abnormal_long_term_variability	1.48
mean_value_of_long_term_variability	1.55
histogram_min	3.22
histogram_number_of_peaks	2.19
histogram_mode	4.95
histogram_variance	2.65

Tabela 6: VIFs do modelo geral.

Conforme indicado através dos valores observados do VIF, as covariáveis que se mostram mais correlacionadas entre si são “Moda do Histograma” e “Frequência cardíaca Basal”, entretanto, ainda sim permaneceram em um nível satisfatório para prosseguirmos com a interpretação dos parâmetros.

4.2.1 Interpretação dos Parâmetros

Os coeficientes da regressão logística são interpretados por meio da odds ratio, que nada mais é que a exponencial dos parâmetros estimados, denotado por $\Psi(\beta_i)$. Interpreta-se esta medida como a razão entre as chances da variável resposta ocorrer (evento positivo) e não ocorrer (evento negativo). Ou seja, para cada um dos parâmetros estimados do nosso modelo, temos a mudança que ocorre na probabilidade da i-ésima unidade amostral não ter resultado satisfatório no CTG a cada unidade de aumento na variável correspondente a esse parâmetro. Os valores aproximados para as odds ratio dos parâmetro estimados do modelo geral são apresentados na Tabela 7.

Também é importante ressaltar que essa interpretação depende de fatores como a escala das covariáveis e a presença de multicolinearidade. Portanto, é importante avaliar cuidadosamente o contexto de cada problema antes de realizar esse tipo de interpretação dos coeficientes.

	Odds Ratio
(Intercept)	0.00
accelerations	0.00
prolongued_decelerations	Inf
baseline.value	1.09
uterine_contractions	0.00
abnormal_short_term_variability	1.09
mean_value_of_short_term_variability	0.58
Perc_T_with_abnormal_long_term_variability	1.03
mean_value_of_long_term_variability	1.07
histogram_min	1.02
histogram_number_of_peaks	1.2
histogram_mode	0.96
histogram_variance	1.04

Tabela 7: Odds Ratio para os parâmetros do modelo.

Analisando a Tabela 7, podemos destacar alguns pontos: há três coeficientes (um deles é o intercepto) com valor 0 e um como infinito. Isso acontece devido à escala das variáveis em questão e a grande quantidade de zeros no banco de dados. Por exemplo, no banco de dados, a variável "Desacelerações prolongadas" possui valores entre 0 e 0.005 em que quase 92% desses valores é 0, o que gerou um parâmetro estimado muito grande comparado aos outros. Assim, quando calculamos a exponencial da estimativa, o valor tende a infinito. Ou seja, caso aumentássemos uma unidade nessa variável o aumento na probabilidade da unidade amostral não ter resultado satisfatório no CTG "aumentaria infinitamente", porém isso não acontece na prática porque os valores dessa variável são muito próximos de zero. De forma análoga para as odds ratio com valor 0.

Para o restante das variáveis, podemos concluir, por exemplo, que a probabilidade da unidade amostral não ter resultado satisfatório no CTG aumenta aproximadamente 9% a cada unidade de aumento na variável "Taxa de batimento basal", haja vista que $\hat{\Psi} \approx 1.09$; e que a probabilidade da unidade amostral não ter resultado satisfatório no CTG diminui aproximadamente 42% a cada unidade de aumento na variável referente à média observada no histograma de FC, haja vista que $\hat{\Psi} \approx 0.52$.

4.2.2 Performance do modelo e Análise de Diagnóstico

Para avaliarmos a performance do modelo, faremos uso da Matriz de Confusão para mensurar os resultados obtidos pela previsão. A Matriz de Confusão está apresentada na Tabela 8.

	0 (Predito)	1 (Predito)
0 (Observado)	439	58
1 (Observado)	11	130

Tabela 8: Matriz de Confusão do modelo geral.

Baseado nessa matriz de confusão, uma medida famosa de performance do modelo é a curva ROC. A Curva ROC observada para o primeiro modelo proposto é indicada na Figura 8.

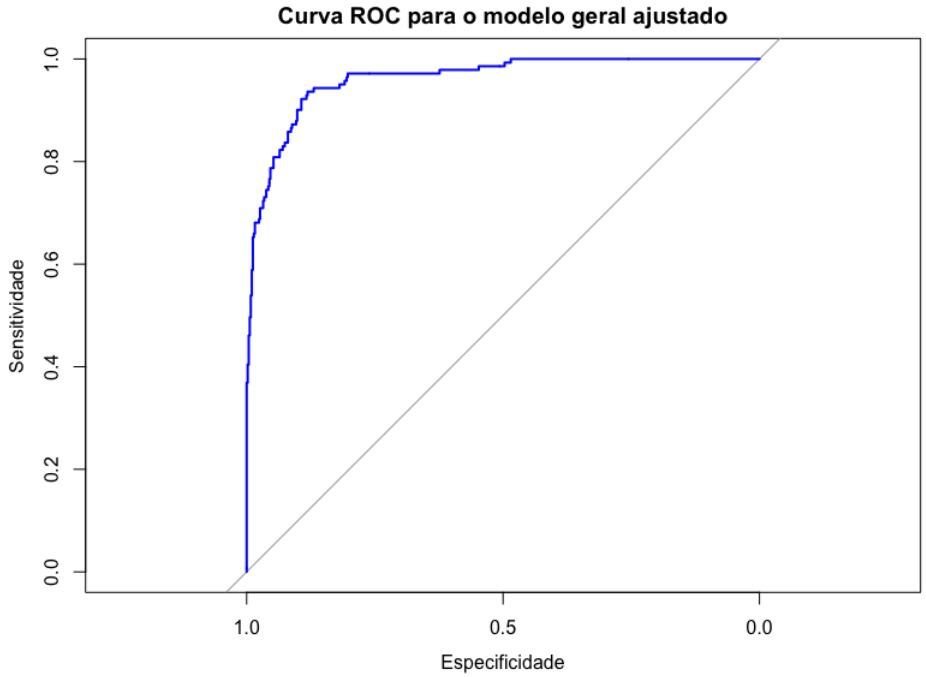

Figura 8: Curva ROC para o modelo geral.

Baseado na curva ROC, podemos calcular o valor do AUC (Area Under de Curve) que é uma medida da área abaixo da curva gerada para os diferentes níveis de cortes propostos. Para o problema em questão, obtivemos um valor de $AUC = 0,9614$. O que, por convenção, em problemas de aprendizado de máquina, costuma indicar uma boa performance do modelo ajustado.

Em seguida, faremos uma análise de diagnóstico a fim de entender se o modelo performou de acordo com as suposições que um modelo de regressão logística deve seguir e, caso esteja de acordo, já teremos um modelo suficiente.

Conforme apresentado na Figura 9, é possível ver que, embora o modelo esteja performando bem de acordo com a curva ROC e com o AUC, não cumpriu com as suposições que um modelo de regressão logística deve seguir como indicado em cada um dos gráficos da Figura 9.

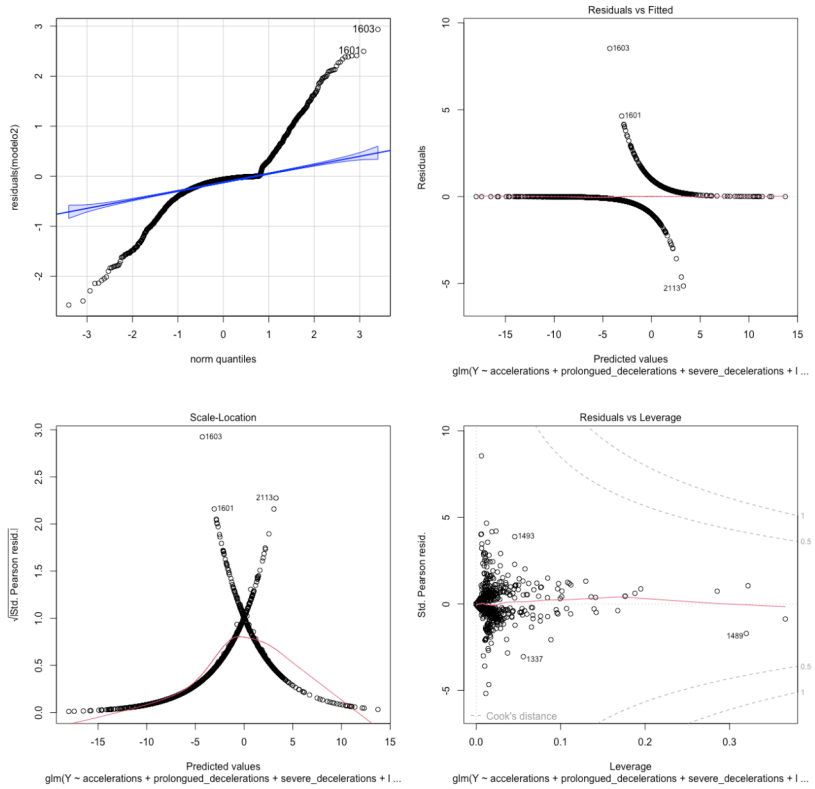

Figura 9: Análise de diagnóstico para o modelo geral.

Contudo, como nosso objetivo é em estabelecer um modelo que performe bem no sentido de previsão, podemos salientar que, em geral, mesmo havendo problemas com a análise de diagnóstico, também é interessante que observemos a respeito do poder preditivo do modelo que, segundo as medidas tomadas a partir da curva ROC, podemos ver que coletamos resultados bastante satisfatórios.

4.3 Árvore de decisão

Uma alternativa para o modelo de Regressão Logística que não foi capaz de se ajustar bem aos nossos dados quanto ao seu diagnóstico, é trabalhar com a mesma abordagem preditiva, mas ao invés de conduzirmos uma Regressão Logística, prosseguir com a metodologia baseado em uma Árvore de Decisão.

Para ilustrar o funcionamento da nossa árvore de decisão temos o diagrama ilustrado na Figura 10.

No topo da árvore vista na Figura 10, podemos ver a probabilidade total do exame não ser satisfatório. Ou seja, mostra a proporção de exames não satisfatórios, 22%. Esse nó verifica se a ”Variação média de curto prazo” é maior ou igual a $0,55$, caso seja, descemos

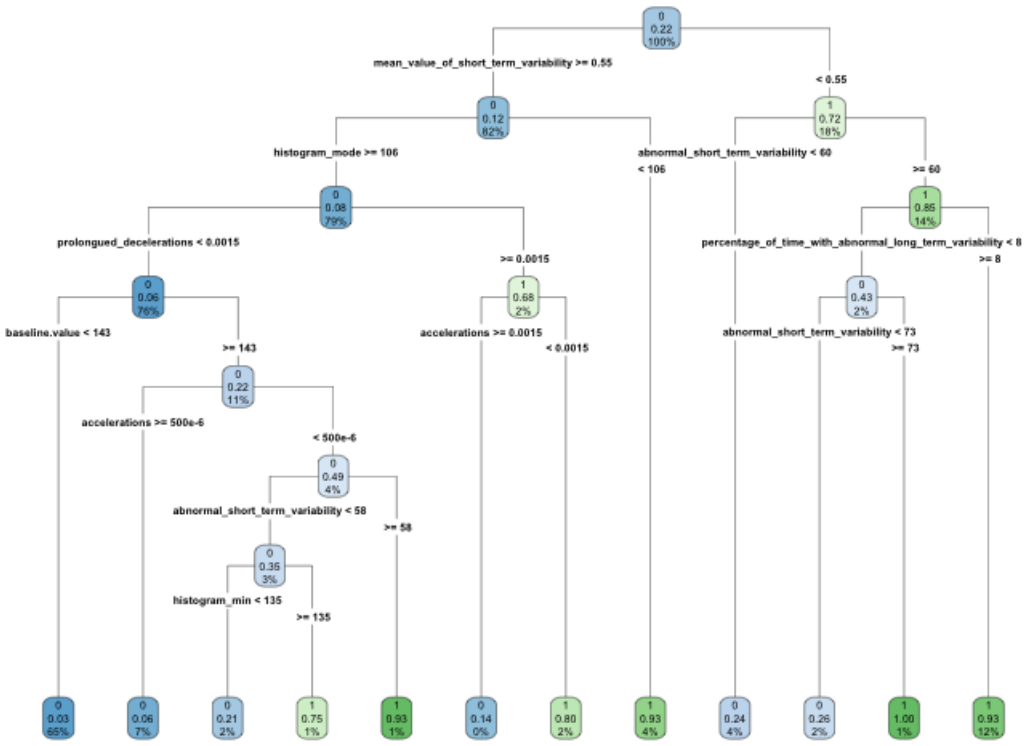

Figura 10: Resumo da árvore de decisão ajustada.

para o nó filho à esquerda. Verifica-se que 82% possuem ”Variação média de curto prazo” maior ou igual à $0,55$, com uma probabilidade do exame não ser satisfatório igual à 0,12.

No segundo nó, verifica-se se a ”Moda do Histograma” é maior ou igual à 106. Caso seja, então descemos novamente para o nó filho à esquerda, onde temos uma proporção de exames não satisfatórios de 8%.

Adiante, verifica-se se a ”Desacelerações prolongadas” é menor que 0,0015. Caso seja, descemos para o nó filho à esquerda, ontem temos uma proporção de não satisfatórios de 6%.

Após isso, verifica-se se a ”Frequência cardíaca basal” é menor que 143. Caso seja, chegamos à primeira folha da árvore, onde temos 65% de toda a massa utilizada no desenvolvimento da árvore, com uma proporção de exames não satisfatórios de apenas 3%.

A interpretação para os demais nós e folhas segue a mesma lógica utilizada na interpretação acima, que foi feita olhando sempre para o nó filho à esquerda.

Podemos notar algumas folhas onde a proporção de exames não satisfatórios é extremamente alta, algumas com proporções acima de 90%. Temos uma folha que possui uma proporção de 100% de não satisfatórios. Essa folha dá-se pelas seguintes condições: ”Varia-

ção média de curto prazo" menor que 0,55; "Variações anormais de curto prazo" maior ou igual à 60; "Percentual de tempo com variações anormais" menor que 8; e por fim, "Variações anormais de curto prazo" maior ou igual à 73. Apesar de conseguir distinguir um público com uma proporção extremamente alta de resultados não satisfatórios, podemos notar que essa folha representa apenas 1% dos exames.

A partir do modelo de árvore de decisão, obtivemos a seguinte matriz de confusão:

	0 (Predito)	1 (Predito)
0 (Observado)	483	27
1 (Observado)	14	114

Tabela 9: Matriz de Confusão do modelo baseado em árvore de decisão.

Fruto dessa matriz de confusão, foi observado um $AUC = 0,9188$. Um pouco menor que quando comparado ao modelo de Regressão Logística ajustado anteriormente.

	Variable	Importância
1	mean_value_of_short_term_variability	85.76
2	abnormal_short_term_variability	54.91
3	histogram_mode	45.93
4	Perc_T_with_abnormal_long_term_variability	39.93
5	histogram_variance	30.95
6	histogram_min	26.94
7	prolongued_decelerations	20.55
8	baseline.value	9.63
9	accelerations	8.91
10	mean_value_of_long_term_variability	7.27
11	histogram_number_of_peaks	5.07
12	severe_decelerations	3.47
13	uterine_contractions	2.28
14	fetal_movement	1.48
15	light_decelerations	0.82

Tabela 10: Importância de cada covariável no modelo de Arvore de decisão.

Na Tabela 10, temos a importância das variáveis que compõe nossa Árvore de decisão. A interpretação do valor de importância é baseada no princípio de que uma variável relevante é aquela que tem um grande impacto na projeção dos dados. A importância é medida com base na quantidade de informação que cada variável contribui para a projeção. Os resultados obtidos foram observados a partir da função "vi" (*Variable Importance*) do pacote "vip" no R e, em resumo, a função "vi" retorna a importância das variáveis em modelos de aprendizado de máquina baseados em projeção, e o valor de importância de uma variável é proporcional à quantidade de informação que ela contribui para a projeção.

Dessa forma, foi possível observar que a variável ”Média da variação de curto período” foi a mais importante para classificação dos indivíduos de acordo com nossa árvore de decisão. A respeito das medidas de desaceleração, apenas ”Desacelerações prolongadas” contribuiu com um aumento significativo de informação para o nosso modelo.

Após a observação de ambos os resultados coletados pelos nossos modelos de classificação ajustados, é importante salientar que existe um tipo de erro que é mais grave que o outro.

Basicamente, temos 2 tipos de erros:

- Erro 1: Classificar o resultado do exame como não satisfatório, dado que o resultado foi satisfatório;
- Erro 2: Classificar o resultado do exame como satisfatório, dado que o resultado na verdade não foi satisfatório

Dado o contexto do problema de estudo, é interessante estarmos atentos ao quanto os nossos modelos estão errando em cada um desses erros elencados anteriormente.

O ”Erro 1” possui um custo operacional de alocar mais recursos médicos a pacientes que são potencialmente saudáveis. Já o ”Erro 2”, pode acarretar no descuido de pacientes que estão na zona de risco e como consequência, pode levar ao óbito das crianças. Dessa forma, para concluir a respeito de nosso modelo final, priorizaremos modelos que estejam com a menor taxa de ”Erro 2”, por mais que isso possa significar aumentar a incidência do ”Erro 1”. O ”Erro 2” no ajuste da árvore foi ligeiramente superior comparado ao modelo logístico.

5 Conclusão

Tendo em vista o contexto do problema em que esse trabalho foi construido, para a escolha do modelo final, é necessário que pensemos no objetivo inicial que foi proposto.

O objetivo proposto nesse trabalho envolve o desenvolvimento de um modelo de classificação que seja suficientemente efetivo para auxiliar na decisão dos profissionais da saúde e, pensando no que tange os objetivos médicos, é preciso avaliarmos a performance dos dois modelos que propusemos em relação aos tipos de erros observados no processo de classificação e que foram elencados no capítulo anterior.

O modelo de regressão logística ajustado, nos resultou em melhores medidas de performance e uma minimização do erro que mais precisamos nos preocupar, o ”Erro 2”, que propõe um indivíduo ser saudável segundo os exames CTGs quando na verdade, o indivíduo é doente. É claro que precisamos mensurar o custo do ”Erro 1” e a assertividade do modelo em relação a esse tipo de erro. Contudo, como o ”Erro 2” pode levar ao óbito dos pacientes, nossos objetivos ao longo do trabalho envolviam minimizar ao máximo esse tipo de erro.

Embora o modelo de regressão logística tenha nos fornecido resultados mais satisfatórios conforme os objetivos da pesquisa, na análise de diagnóstico o modelo deixou a desejar, não cumprindo com as suposições básicas para declarar o modelo como satisfatório.

Buscando uma alternativa para contornar esse problema no diagnóstico envolvendo o modelo logístico, foi proposta uma árvore de decisão. Esse segundo modelo, também nos forneceu ótimos resultados no sentido de previsão. Contudo, aumentou ligeiramente o ”Erro 2” que estamos buscando evitar.

Dessa forma, concluímos que, embora nosso objetivo seja estabelecer um modelo preditivo e o modelo de Regressão Logística parece ter sido mais assertivo, o modelo de Árvore de decisão é mais estável. Assim, o ajuste a ser escolhido deve levar em consideração as premissas citadas anteriormente e, para o grupo que desenvolveu o trabalho, **foi decidido pelo modelo de Árvore de decisão por não violar nenhuma suposição inicial de ajuste do modelo**. Entretanto, é salientado que a preferência pelos modelos varia de acordo com o que o pesquisador preza mais ao executar a modelagem: O poder preditivo do ajuste ou a segurança no ajuste.

6 Códigos

```
library(dplyr)
library(MASS)
library(tidyverse)
library(DBI)
library(ROCR)
library(pROC)
library(ggpubr)
library(corrplot)
library(GGally)
library(car)
library(plotly)
library(gapminder)
library(TSA)
library(forecast)
library(xtable)
library(glmnet)
library(SMOTEWB)
library(ROSE)
library(caret)
library(DMwR2)
library(stats)
library(caTools)
library(rpart)
library(ggcorrplot)
library(vip)

View(dados5)

## DESCRITIVA

summary(dados5)
hist(dados5)

## histograma classes
dados5 %>%
  ggplot(aes(fill=fetal_health, x=fetal_health)) +
  geom_bar(position="dodge", stat="count") +
  ggtitle("Distribuição da variável resposta") +
  xlab("Estado de saúde do feto") +
  ylab("Contagem") +
```

```
  theme_bw() +
  scale_fill_brewer(palette = "Dark2")

##histograma todos
dados5 %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~ key, scales = "free") +
  geom_histogram()

## Nova variavel Y e tratamento de dados

str(dados5)
#dados5 <- lapply(dados5, as.numeric)
dados5$fetal_health <- as.factor(dados5$fetal_health)
dados5$Y <- ifelse(dados5$fetal_health==1,0,1)
dados5$Y <- as.factor(dados5$Y)

dados5$fetal_health <- NULL
dados5$histogram_tendency <- as.factor(dados5$histogram_tendency)

## histograma novo Y
dados5 %>%
  ggplot(aes(fill=Y, x=Y)) +
  geom_bar(position="dodge", stat="count") +
  ggtitle("Distribuição da nova variável resposta") +
  xlab("Estado de saúde do feto") +
  ylab("Contagem") +
  theme_bw() +
  scale_fill_brewer(palette = "Dark2")

## DF de risco e saudaveis
risco <- dados5 %>% filter(Y==1)
saudaveis <- dados5 %>% filter(Y==0)

risco_treino <- df_treino %>% filter(Y==1)
saud_treino  <- df_treino %>% filter(Y==0)


## zoom dos grafs por classe
```

```
h1 <- dados5 %>%
  ggplot(aes(x=accelerations)) +
  geom_bar(position="dodge", stat="count") +
  ggtitle("Distribuição da aceleração") +
  xlab("Aceleração") +
  ylab("Contagem") +
  theme_bw() +
  scale_fill_brewer(palette = "Dark2")

h2 <- dados5 %>%
  ggplot(aes(x=light_decelerations)) +
  geom_bar(position="dodge", stat="count") +
  ggtitle("Distribuição das desacelerações leves") +
  xlab("Desacelerações leves") +
  ylab("Contagem") +
  theme_bw() +
  scale_fill_brewer(palette = "Dark2")

h3 <- dados5 %>%
  ggplot(aes(x=severe_decelerations)) +
  geom_bar(position="dodge", stat="count") +
  ggtitle("Distribuição das desacelerações severas") +
  xlab("Desacelerações severas") +
  ylab("Contagem") +
  theme_bw() +
  scale_fill_brewer(palette = "Dark2")

h4 <- dados5 %>%
  ggplot(aes(x=prolongued_decelerations)) +
  geom_bar(position="dodge", stat="count") +
  ggtitle("Distribuição das desacelerações prolongadas") +
  xlab("Desacelerações prolongadas") +
  ylab("Contagem") +
  theme_bw() +
  scale_fill_brewer(palette = "Dark2")

ggarrange(h1, h2, h3, h4, ncol = 2, nrow = 2)

## boxplots

b1 <- dados5 %>%
  ggplot( aes(x=Y, y=accelerations, fill=Y)) +
```

```
  geom_boxplot() +
  theme_bw() +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("BoxPlot do número de acelerações de batimentos") +
  xlab("Estado de saúde") + ylab("Acelerações") +
  scale_fill_brewer(palette = "Dark2")

b2 <- dados5 %>%
  ggplot( aes(x=Y, y=abnormal_short_term_variability, fill=Y)) +
  geom_boxplot() +
  theme_bw() +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("BoxPlot da Variação anormais de curto período") +
  xlab("Estado de saúde") + ylab("Variação anormais de curto período") +
  scale_fill_brewer(palette = "Dark2")

b3 <- dados5 %>%
  ggplot( aes(x=Y, y=baseline.value, fill=Y)) +
  geom_boxplot() +
  theme_bw() +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("BoxPlot da Frequência cardíaca basal") +
  xlab("Estado de saúde") + ylab("Frequência cardíaca basal") +
  scale_fill_brewer(palette = "Dark2")

b4 <- dados5 %>%
  ggplot( aes(x=Y, y=percentage_of_time_with_abnormal_long_term_variability
  geom_boxplot() +
  theme_bw() +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
```

```
  ggtitle("BoxPlot do % de tempo com variações anormais longas") +
  xlab("Estado de saúde") + ylab("% Tempo de variações anormais") +
  scale_fill_brewer(palette = "Dark2")

ggarrange(b1, b2, b3, b4, ncol = 2, nrow = 2)

dados5 %>%
  ggplot( aes(x=Y, y=uterine_contractions, fill=Y)) +
  geom_boxplot() +
  theme_bw() +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) + ggtitle("BoxPlot do número de contrações uterinas") +
  xlab("Estado de saúde") + ylab("Número de contrações uterinas") +
  scale_fill_brewer(palette = "Dark2")

# tabelas

# aceleracoes
summary(saudaveis$accelerations)
summary(risco$accelerations)

# FC basal
summary(saudaveis$baseline.value)
summary(risco$baseline.value)

# % tempo
summary(saudaveis$percentage_of_time_with_abnormal_long_term_variability)
summary(risco$percentage_of_time_with_abnormal_long_term_variability)

# var anormais
summary(saudaveis$abnormal_short_term_variability)
summary(risco$abnormal_short_term_variability)

#uterine contraction
summary(saudaveis$uterine_contractions)
summary(risco$uterine_contractions)


data.frame(summary(saudaveis$accelerations),summary(risco$accelerations))
```

```
## correlacao

quant <- data.frame(dados5$accelerations,
                    dados5$prolongued_decelerations,
                    dados5$severe_decelerations,
                    dados5$light_decelerations,
                    dados5$fetal_movement,
                    dados5$baseline.value,
                    dados5$uterine_contractions,
                    dados5$abnormal_short_term_variability,
                    dados5$mean_value_of_short_term_variability,
                    dados5$percentage_of_time_with_abnormal_long_term_varia
                    dados5$mean_value_of_long_term_variability,
                    dados5$histogram_width,dados5$histogram_min,
                    dados5$histogram_max,
                    dados5$histogram_number_of_peaks,
                    dados5$histogram_number_of_zeroes,
                    dados5$histogram_mode,
                    dados5$histogram_mean,
                    dados5$histogram_median,
                    dados5$histogram_variance,
                    as.numeric(dados5$Y))

colnames(quant) <- c("accelerations","prolongued_decel","severe_decel","lig
                     "fetal_mov","baseline_value", "uterine_contrac","anorm
                     "mean_STV","percent_of_time_STV","mean_LTV", "histogra
                     "histogram_min","histogram_max","histogram_n_peaks","h
                     "histogram_mode","histogram_mean","histogram_med","his
                     "Resultado do Exame")

cor <- cor(quant)

#ggpairs(dados5[,2:10])

ggcorrplot(cor, hc.order = TRUE, type = "lower", lab=TRUE,lab_size = 2.3,
           outline.col = "white",
           ggtheme = ggplot2::theme_bw,
           colors = c("#0a72ac", "white", "#ec6c04"))

cor_resp <- cor(quant$`Resultado do Exame`,quant)

## modelo
```

```
#treino e teste
prop = sample.split(Y = dados5$Y, SplitRatio = 0.7)
df_treino <- dados5[prop,]
treino_sat <- df_treino %>% filter(Y==0)
treino_insat <- df_treino %>% filter(Y==1)
#teste
df_teste = dados5[!prop,]

#arvore
arvore <- rpart(Y ~ ., data = df_treino, method = "class")
p_arvore <- predict(arvore, newdata=subset(df_teste), type="class")

## arvore 2

arvore2 <- rpart(Y ~ accelerations+
                   prolongued_decelerations+
                   severe_decelerations+
                   light_decelerations+
                   fetal_movement+
                   baseline.value+
                   uterine_contractions+
                   abnormal_short_term_variability+
                   mean_value_of_short_term_variability+
                   percentage_of_time_with_abnormal_long_term_variability+
                   mean_value_of_long_term_variability+
                   histogram_min+
                   histogram_number_of_peaks+
                   histogram_mode+
                   histogram_variance, data = df_treino, method = "class")
p_arvore2 <- predict(arvore2, newdata = subset(df_teste), type = "class")
table(p_arvore2,df_teste$Y)


#matriz de confusao
table(p_arvore,df_teste$Y)

#logistica

modelo <- glm(Y ~ .,family=binomial(link='logit'),data=df_treino)
```

```
modelo <- glm(Y ~ accelerations+
                  prolongued_decelerations+
                  severe_decelerations+
                  light_decelerations+
                  fetal_movement+
                  baseline.value+
                  uterine_contractions+
                  abnormal_short_term_variability+
                  mean_value_of_short_term_variability+
                  percentage_of_time_with_abnormal_long_term_variability+
                  mean_value_of_long_term_variability+
                  histogram_width+histogram_min+histogram_max+
                  histogram_number_of_peaks+histogram_number_of_zeroes+
                  histogram_mode+histogram_mean+histogram_median+
                  histogram_variance+histogram_tendency,
                  family=binomial(link='logit'),data=df_treino)

modelo2 <- glm(Y ~ accelerations+
                  prolongued_decelerations+
                  severe_decelerations+
                  light_decelerations+
                  fetal_movement+
                  baseline.value+
                  uterine_contractions+
                  abnormal_short_term_variability+
                  mean_value_of_short_term_variability+
                  percentage_of_time_with_abnormal_long_term_variability+
                  mean_value_of_long_term_variability+
                  histogram_min+
                  histogram_number_of_peaks+
                  histogram_mode+
                  histogram_variance,
                family=binomial(link='logit'),data=df_treino)

modelo3 <- glm(Y ~ accelerations+
                   prolongued_decelerations+
                   #severe_decelerations+
                   #light_decelerations+
                   #fetal_movement+
                   baseline.value+
                   uterine_contractions+
                   abnormal_short_term_variability+
```

```
                  mean_value_of_short_term_variability+
                  percentage_of_time_with_abnormal_long_term_variability+
                  mean_value_of_long_term_variability+
                  histogram_min+
                  histogram_number_of_peaks+
                  histogram_mode+
                  histogram_variance,
                family=binomial(link='logit'),data=df_treino)

p3 <- predict(modelo3, newdata = subset(df_teste), type = "response")
valores_preditos_GERAL3 <- ifelse(p3 > 0.284,1,0)
table(df_teste$Y,valores_preditos_GERAL3)



p2 <- predict(modelo2, newdata = subset(df_teste),type="response")
valores_preditos_GERAL2 <- ifelse(p2 > 0.284,1,0)
table(df_teste$Y,valores_preditos_GERAL2)

roc_score1=roc(df_teste$Y~p2,plot=FALSE)

plot(roc_score1,
     main ="Curva ROC para o modelo geral ajustado",
     xlab="Especificidade",
     ylab="Sensitividade",
     col = "Blue")

auc(df_teste$Y,p2)

stepAIC(modelo)

p_geral <- predict(modelo, newdata=subset(df_teste), type="response")

valores_preditos_GERAL <- ifelse(p_geral > 0.284,1,0)

table(df_teste$Y,valores_preditos_GERAL)

modelo_step <- glm(Y ~ accelerations +
                       prolongued_decelerations +
                       severe_decelerations +
                       uterine_contractions +
                       abnormal_short_term_variability +
```

```
                        percentage_of_time_with_abnormal_long_term_variabilit
                        histogram_width + histogram_min +
                        histogram_number_of_peaks + histogram_mean +
                        histogram_median + histogram_variance +
                        histogram_tendency,
                        family = binomial(link = "logit"), data = df_treino)

p_step <- predict(modelo_step, newdata=subset(df_teste), type="response")


## valor de corte p classificacao = proporcao de doentes
valores_preditos <- ifelse(p_step > 0.284,1,0)

## matriz de confusao

table(df_teste$Y,valores_preditos)

# auc

auc(df_teste$Y,p_step)

## SMOTE p oversampling

sample_SMOTE <- SMOTE(df_treino[,1:20],df_treino[,22],k=2)

df_SMOTE <- data.frame(sample_SMOTE$x_new,sample_SMOTE$y_new)


## modelo smote sem histogram tendency
modelo_SMOTE <- glm(sample_SMOTE$y_new ~ accelerations+
                        prolongued_decelerations+
                        severe_decelerations+
                        light_decelerations+
                        fetal_movement+
                        baseline.value+
                        uterine_contractions+
                        abnormal_short_term_variability+
                        mean_value_of_short_term_variability+
                        percentage_of_time_with_abnormal_long_term_variabilit
                        mean_value_of_long_term_variability+
                        histogram_width+histogram_min+histogram_max+
                        histogram_number_of_peaks+histogram_number_of_zeroes+
```

```
                    histogram_mode+histogram_mean+histogram_median+
                    histogram_variance,
                    #+histogram_tendency,
                  family=binomial(link='logit'),
                  data=df_SMOTE)

## modelo smote 2

modelo_SMOTE2 <- glm(sample_SMOTE$y_new ~ accelerations+
                       prolongued_decelerations+
                       severe_decelerations+
                       light_decelerations+
                       fetal_movement+
                       baseline.value+
                       uterine_contractions+
                       abnormal_short_term_variability+
                       mean_value_of_short_term_variability+
                       percentage_of_time_with_abnormal_long_term_variabili
                       mean_value_of_long_term_variability+
                       histogram_min+
                       histogram_number_of_peaks+
                       histogram_mode+
                       histogram_variance,
                     #+histogram_tendency,
                     family=binomial(link='logit'),
                     data=df_SMOTE)


p_smote <- predict(modelo_SMOTE, newdata=subset(df_teste), type="response")
valores_preditos_SMOTE <- ifelse(p_smote > 0.5,1,0)

# matriz de confusao SMOTE
table(df_teste$Y,valores_preditos_SMOTE)
auc(df_teste$Y,p_smote)

# modelo LASSO

modelo_lasso <- glmnet(df_treino[,1:20], df_treino[,22], family="binomial",
p_lasso <- predict(modelo_lasso, newx = as.matrix(subset(df_teste[,1:20])),
                   type = "response", s = 0.01)

valores_preditos_LASSO <- ifelse(p_lasso > 0.284,1,0)
```

```
table(df_teste$Y,valores_preditos_LASSO)
auc(df_teste$Y,p_lasso)

# modelo LASSO SMOTE - MELHOR MODELO NO SENTIDO DE ERRO MENOS CUSTOSO

modelo_lasso_SMOTE <- glmnet(df_SMOTE[,1:20], df_SMOTE[,21],
                             family="binomial", alpha=1)

p_lasso_SMOTE <- predict(modelo_lasso_SMOTE,
                         newx = as.matrix(subset(df_teste[,1:20])),
                         type = "response", s = 0.01)

valores_preditos_LASSO_SMOTE <- ifelse(p_lasso_SMOTE > 0.5,1,0)
table(df_teste$Y,valores_preditos_LASSO_SMOTE)
auc(df_teste$Y,p_lasso_SMOTE)

## arvore SMOTE

arvore_SMOTE <- rpart(sample_SMOTE.y_new ~ .,
                      data = df_SMOTE, method = "class")

p_arvore_SMOTE <- predict(arvore_SMOTE, newdata=subset(df_teste), type="cla

m_conf_ASMOTE <- table(df_teste$Y,p_arvore_SMOTE)
auc(df_teste$Y,as.numeric(p_arvore_SMOTE))

##PLOT MATRIZ CONFUSAO

ggplot(data =  as.data.frame(m_conf_ASMOTE),
       mapping = aes(x = Var1, y = p_arvore_SMOTE)) +
  geom_tile(aes(fill = Y), colour = "white") +
  geom_text(aes(label = sprintf("%1.0f", Y)), vjust = 1) +
  scale_fill_gradient(low = "blue", high = "red") +
  theme_bw() + theme(legend.position = "none")


# Identify non-zero coefficients
nonzero_coefs <- which(coef(modelo_lasso_SMOTE) != 0)

# Extract non-zero coefficients
coef(modelo_lasso_SMOTE)[nonzero_coefs]
```

```
# matriz de confusao SMOTE
table(df_teste$Y,valores_preditos_LASSO_SMOTE)
auc(df_teste$Y,p_smote)

## diagnostico

residuos <- p_lasso_SMOTE-as.numeric(df_teste$Y)
plot(p_lasso_SMOTE,residuos)
qresid(modelo_lasso_SMOTE)
qqPlot(residuals(modelo2), envelope = 0.95)

vif(modelo_lasso_SMOTE)


fit <- rpart(Y ~ accelerations+
               prolongued_decelerations+
               severe_decelerations+
               light_decelerations+
               fetal_movement+
               baseline.value+
               uterine_contractions+
               abnormal_short_term_variability+
               mean_value_of_short_term_variability+
               percentage_of_time_with_abnormal_long_term_variability+
               mean_value_of_long_term_variability+
               histogram_min+
               histogram_number_of_peaks+
               histogram_mode+
               histogram_variance, method = "class", data = df_treino)
melhor_cp <- fit$cptable[which.min(fit$cptable[, "xerror"]),
                         "CP"]
pfit <- rpart::prune(fit, cp = melhor_cp)
rpart.plot::rpart.plot(pfit, type = 4, extra = 106)

p_arvoreFIT <- predict(fit, newdata = subset(df_teste), type = "class")
table(p_arvoreFIT,df_teste$Y)

vi(fit)

table(p_arvore,df_teste$Y)
```