UNIVERSIDADE FEDERAL
DE SANTA CATARINA

Modelo de Classificação da cardiotocografia do feto.

Ana P. Rocha, Gabryel L. Bernardes, Helena B. Daré e João V. Assis

4 de Maio de 2021

1 Entendimento do negócio

A cardiotocografia fetal é um exame realizado durante a gravidez para verificar os batimentos cardíacos e o bem estar do bebê feito com sensores ligados à barriga da gestante que coletam estas informações [1]. O seu objetivo é medir e registrar a atividade cardíaca do feto, observando, por exemplo, sua frequência cardíaca, seus movimentos e a contração uterina.

O objetivo do negócio é encontrar alguma possível alteração relacionada a oxigenação cerebral. Certas alterações podem estar associadas com a posição do feto, problemas na placenta ou cordão umbilical mal posicionado [2]. Dessa forma, é possível intervir o mais rápido possível quando alguma alteração no bem estar do feto é identificada e o obstetra pode proporcionar um melhor atendimento de acordo com a gestação e gravidade do caso.

O conjunto de dados [3] apresenta informações relacionadas ao sistema cardiovascular do feto, e a partir dessas informações classifica a cardiotocografia do feto em normal, suspeito ou patológico. O objetivo da mineração de dados é propor e validar um modelo de classificação para a predição da cardiotocografia do feto em normal, suspeito ou patológico, possibilitando assim uma intervenção rápida e que garanta a saúde do feto.

A organização do trabalho foi baseada na metodologia CRISP-DM (Processo Padrão de Vários Segmentos de Mercados para Mineração de Dados) com o objetivo de orientar os esforços na mineração dos dados [4]. Pretende-se realizar o entendimento dos dados, bem como uma análise exploratória e descritiva para visualizar possíveis relações entre a variável resposta com as variáveis explicativas. Além de instigar possíveis mudanças no conjunto de dados, como remoção de valores nulos ou duplicados e padronização das variáveis.

Neste trabalho também será modelado dois modelos de classificação e analisado qual obteve o melhor desempenho. Escolheu-se comparar o modelo de regressão logística com o modelo de floresta aleatória. Além disso, a partir da seleção de variáveis explicativas pelo método *SelectK-Best* da biblioteca scikit-learn será variado o número de variáveis explicativas para encontrar o modelo com maior poder de generalização. Sendo, por último, utilizado métricas como matriz de confusão e precisão para avaliar o desempenho dos modelos e escolher o melhor modelo para classificação da cardiotocografia do feto.

2 Entendimento dos dados

O conjunto de dados foi retirado de um repositório público no site Kaggle [3], esse conjunto contém 2.126 registros de características extraídas de exames de cardiotocografia [5], sendo a atividade cardíaca fetal classificada por três obstetras especialistas em três classes: Normal, suspeito e patológico [3], sendo essas classes referenciadas como 1, 2 e 3, respectivamente. O conjunto de dados contém 21 variáveis explicativas, que são apresentadas na Tabela 1.

Variável	Significado
Frequência cardíaca fetal	Frequência cardíaca fetal (batimentos por minuto)
Acelerações	Número de acelerações por segundo
Movimento fetal	Número de movimentos fetais por segundo
Contrações uterinas	Número de contrações uterinas por segundo
Desacelerações leves	Número de desacelerações leves por segundo
Desacelerações severas	Número de severas desacelerações por segundo
Desacelerações prolongadas	Número de desacelerações prolongadas por segundo
Variabilidade anormal de curto prazo	Porcentagem de tempo com variabilidade anormal de curto prazo
Valor médio da variabilidade de curto prazo	Valor médio da variabilidade de curto prazo
Porcentagem de tempo com variabilidade anormal de longo prazo	Porcentagem de tempo com variabilidade anormal de longo prazo
Valor médio da variabilidade de longo prazo	valor médio da variabilidade de longo prazo
Largura do histograma	Largura do histograma FHR
Valor mínimo (baixa frequência) do histograma FHR	Valor mínimo (baixa frequência) do histograma FHR
Valor máximo (alta frequência) do histograma FHR	Valor máximo (alta frequência) do histograma FHR
Número de picos do histograma	Número de picos do histograma
Número de zeros do histograma	Número de zeros do histograma
Moda do histograma	Moda do histograma
Média do histograma	Média do histograma
Mediana do histograma	Mediana do histograma
Variância do histograma	Variância do histograma
Tendência do histograma	Tendência do histograma

Tabela 1: Variáveis explicativas.

2.0.1 Análise exploratória

A Figura 1 apresenta o histograma da variável resposta saúde fetal. Além disso, a Tabela 2 apresenta a contagem de cada classe e evidencia que os dados estão desbalanceados. A consequência desse desequilíbrio é que o modelo terá um desempenho bom para as entradas da classe majoritária e um desempenho inferior para a classe minoritária.

No caso de classificar a saúde do feto, no qual o número de fetos normais é maior que as demais classes, o classificador tenderá a apresentar muitos 'falsos normais'. Em um *trade-off*, é preferível ter uma quantidade maior de 'falsos patológicos', visto que é importante identificar o mais rápido possível anomalias na frequência cardíaca do feto.

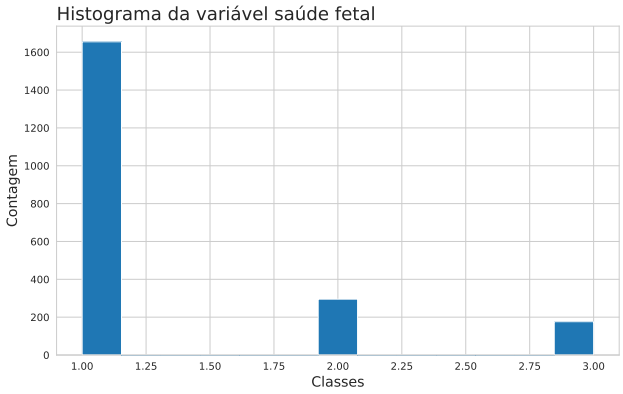


Figura 1: Histograma da variável resposta saúde fetal.

Classe	Contagem
Normal	1655
Suspeito	295
Patológico	176

Tabela 2: Contagem de amostras de cada classe.

Enquanto para uma análise exploratória mais efetiva das variáveis explicativas, primeiro será analisado a correlação de Pearson entre as variáveis explicativas, visto que variáveis com alto grau de correlação podem causar problemas no ajuste e interpretação do modelo. Assim, deve-se evitar construir um modelo com variáveis correlacionadas entre si [6].

A Figura 2 apresenta a interpretação do coeficiente de correlação de Pearson. Assim, baseado na matriz de correlação da Figura 3, pode-se concluir que: valor de linha de base possui forte correlação (cores escuras) com a moda, média e mediana do histograma. A mediana do histograma também possui forte correlação com os valores da moda e média do histograma. Logo, as variáveis relacionadas ao histograma, como média, mediana e moda são candidatas a não serem incluídas no treinamento do modelo.

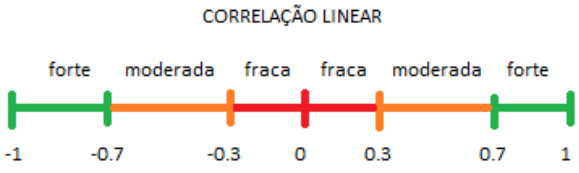


Figura 2: Correlação de Pearson

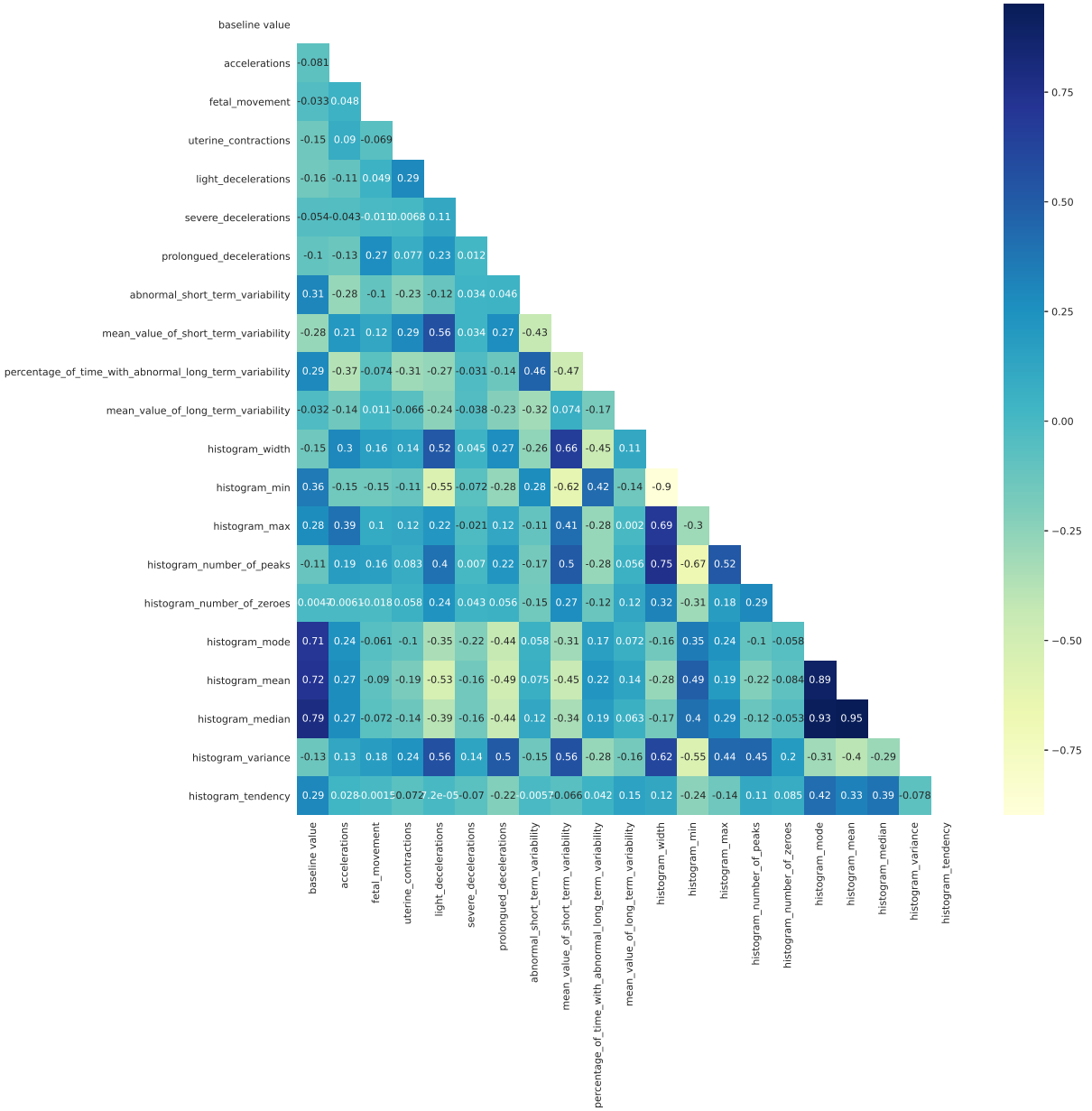


Figura 3: Matriz de correlação entre as variáveis explicativas.

A partir da Figura 3 apresenta-se os valores das variáveis com correlação de Pearson superior a 70%:

- A variável número de picos do histograma apresenta 75% de correlação com a variável largura do histograma.
- A variável frequência cardíaca fetal apresenta correlação de 71% com a variável moda do histograma, 72% com a variável média do histograma e 79% com a mediana do histograma.
- A variável moda do histograma apresenta correlação de 89% com a variável média do histograma e 93% de correlação com a variável mediana do histograma.
- A variável média do histograma apresenta correlação de 95% com a mediana do histograma.

Enquanto algumas das variáveis que apresentam correlação inferior a 1% são:

- A variável número de zeros do histograma apresenta correlação de -0,6% com a variável acelerações e -0,47% com a frequência cardíaca fetal.
- A variável tendência do histograma apresenta correlação de 0,15% com o movimento fetal, 0,0072% com desacelerações leves e 0,57% com variabilidade anormal de curto prazo.
- A variável desacelerações severas apresenta correlação de 0,67% com contrações uterinas e 0,7% com número de picos do histograma.

A Figura 4 apresenta a relação entre o movimento do feto e a atividade cardíaca fetal, percebe-se que fetos com atividade cardíaca normal se movimentam menos que fetos com patologias. Todavia, há uma quantidade considerável de fetos com movimentação alta e atividade cardíaca normal. Enquanto para a variável acelerações, fetos com atividade cardíaca normal possuem o número de acelerações por segundo superior aos demais, sendo consideravelmente bem distribuída.

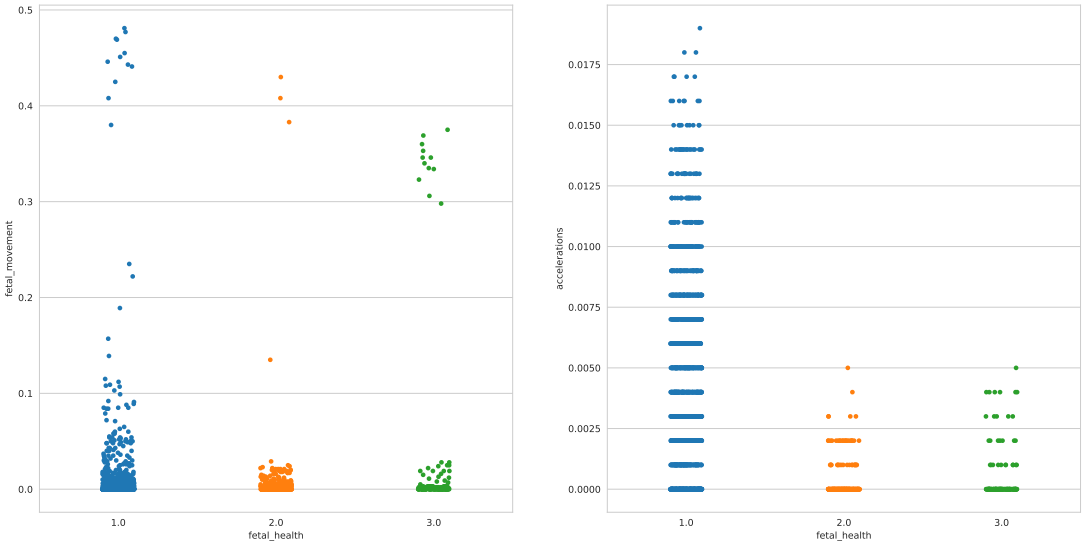


Figura 4: Relação entre a frequência cardíaca fetal e a movimentação do feto e a aceleração.

A Figura 5 apresenta os histogramas das variáveis explicativas, com exceção da variável resposta e das variáveis relacionadas ao histograma de média, mediana e moda. O comportamento que mais se aproxima de uma distribuição normal são das variáveis explicativas: frequência cardíaca fetal (único com distribuição simétrica), máximo da frequência cardíaca fetal, valor médio da variabilidade de curto prazo, valor médio da variabilidade de longo prazo e número de picos da frequência cardíaca fetal.

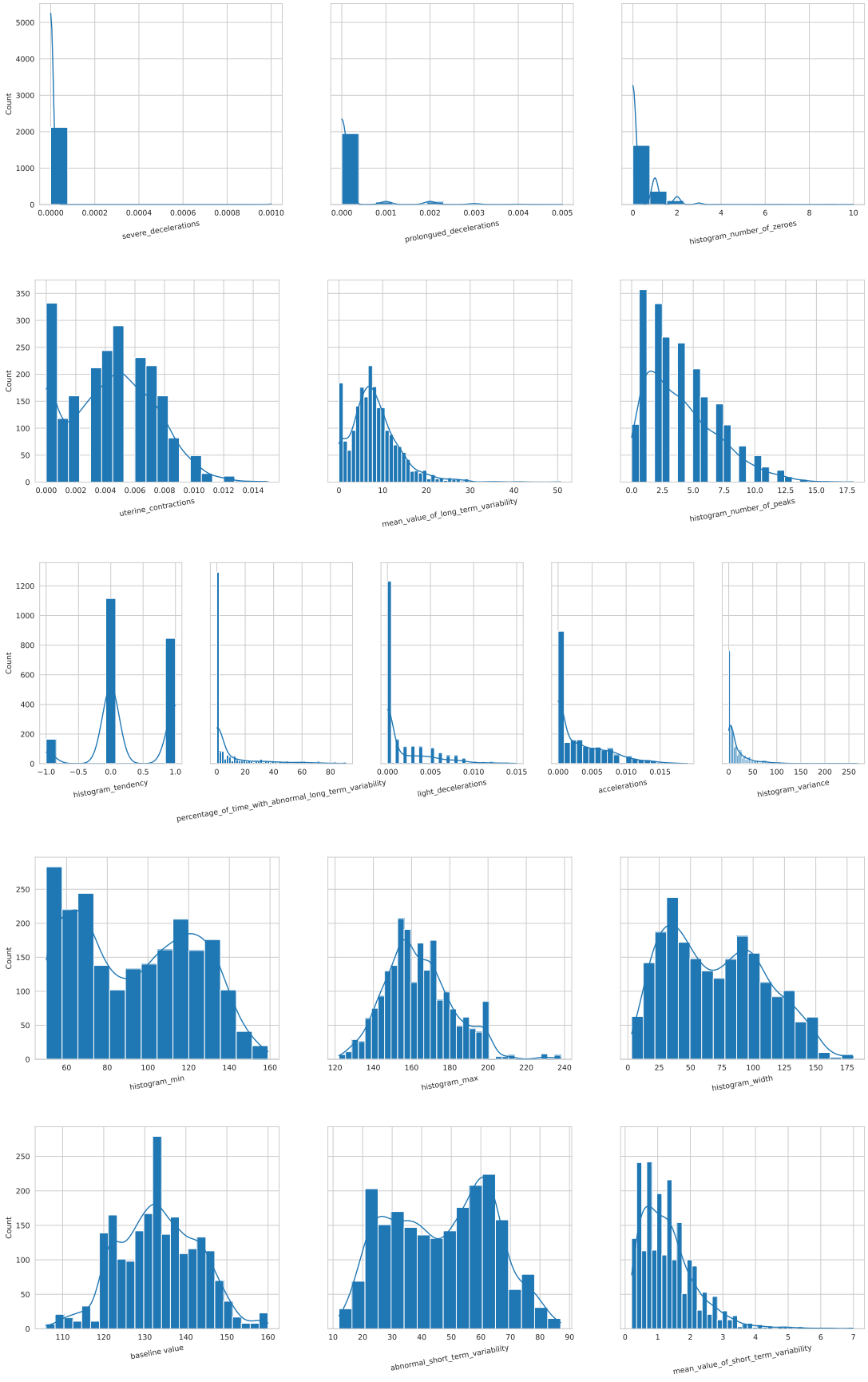


Figura 5: Histograma das variáveis explicativas.

Enquanto as Figuras 6, 7, 8 e 9 apresentam os diagramas de caixa entre as variáveis explicativas e a variável resposta. Note que os pontos no diagrama de caixa são considerados valores discrepantes, isto é, estão além das proporções do intervalo interquartílico inferior e superior [7].

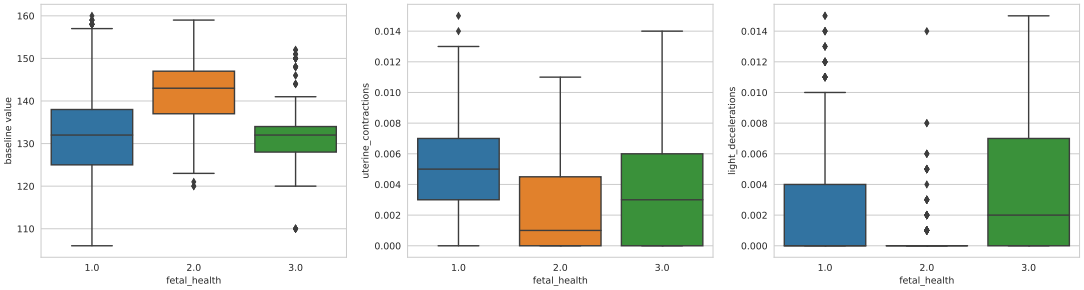


Figura 6: Diagrama de caixa das variáveis explicativas frequência cardíaca fetal, contrações uterinas e desacelerações leves.

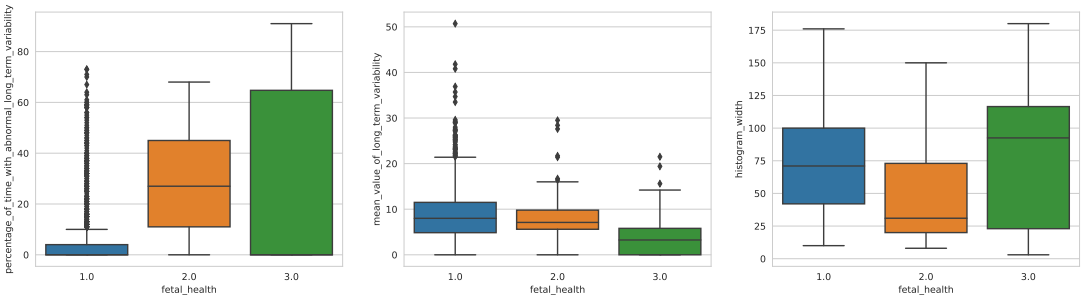


Figura 7: Diagrama de caixa das variáveis explicativas porcentagem de tempo com variabilidade anormal de longo prazo, valor médio da variabilidade de longo prazo, largura do histograma.

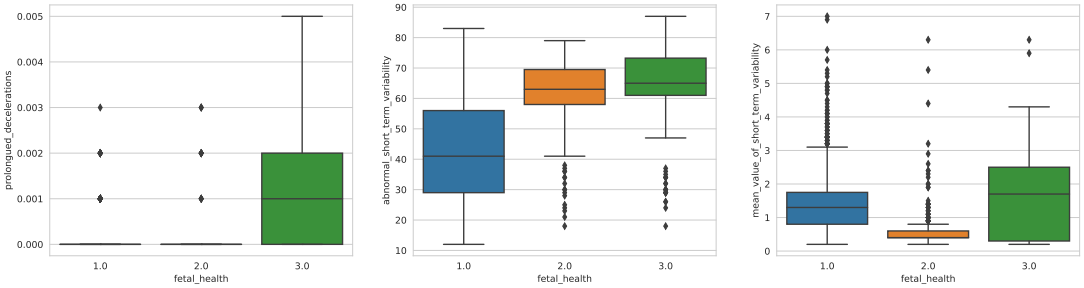


Figura 8: Diagrama de caixa das variáveis explicativas desacelerações prolongadas, variabilidade anormal de curto prazo e valor médio da variabilidade de curto prazo.

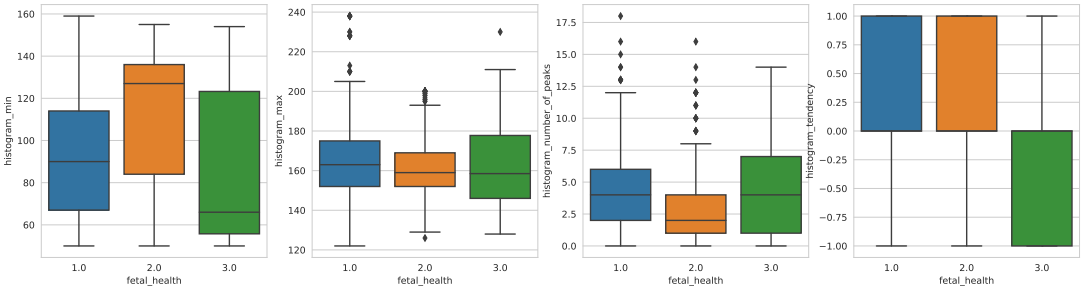


Figura 9: Diagrama de caixa das variáveis explicativas valor mínimo (baixa frequência) e máximo (alta frequência) do histograma FHR, número de picos do histograma e tendência do histograma.

É possível notar que grande parte das variáveis apresentadas possui valores discrepantes. Ressalta-se que esses valores são significativos para a análise de identificar quais padrões resultam uma atividade cardíaca fetal suspeita ou patológica, ou seja, quais características indicam que o feto possui problemas de saúde. Todavia, esses valores discrepantes podem influenciar negativamente o modelo para a classificação do feto entre suspeito, normal ou patológico.

Pela Figura 8 nota-se que a classe 3 (frequência cardíaca fetal patológica) apresenta altos valores de desacelerações prolongadas, isso pode ser causado pela compressão do cordão umbilical e pode refletir a hipóxia fetal (privação de oxigênio) [8]. Assim, é uma variável significativa para a classificação da saúde fetal. O mesmo não ocorre para a variável tendência do histograma, apresentada na Figura 9, as classes 1 e 2 não são significativamente diferentes, logo não seria uma variável interessante para diferenciar as classes. Ainda, é possível perceber um comportamento anômalo na classe 3, visto que as classes 1 e 2 apresentam valores em sua maioria acima de zero, enquanto a classe 3 apresenta valores e sua maioria abaixo de zero.

2.0.2 Análise descritiva

	Frequência cardíaca fetal (bpm)	Desacelerações severas	Desacelerações prolongadas	Variabilidade anormal de curto prazo	Moda do histograma
média	133,0	$3{,}0{\cdot}10^{-5}$	$15{,}9{\cdot}10^{-5}$	46,9	137,4
desvio padrão	9,8	$5{,}7{\cdot}10^{-5}$	$5{,}9{\cdot}10^{-4}$	17,2	16,4
moda	133,0	0,0	0,0	60,0	133,0
25%	126,0	0,0	0,0	32,0	129,0
75%	140,0	0,0	0,0	61,0	139,0
mín	106,0	0,0	0,0	12,0	60,0
máx	160,0	$1{,}0 \cdot 10^{3}$	$5{,}0{\cdot}10^{-3}$	87,0	187,0

Tabela 3: Análise descritiva das variáveis explicativas.

Selecionou-se 5 das 21 variáveis explicativas (todas contínuas) para a realização da análise descritiva, como vistas na tabela 3 para início da investigação das *features*. A Frequência Cardíaca Fetal (FCF) acima de 160 batimentos por minuto, que permanece assim, indica taquicardia fetal e FCF abaixo de 110 batimentos por minuto indica bradicardia fetal. A moda, ou seja, o valor mais frequência no conjunto observado, permite inferir que a maioria dos fetos analisados possui uma frequência cardíaca dentro do desejável.

As desacelerações também descritas no *dataset* são sinais de hipoxia fetal, que corresponde a ausência do oxigênio que deve ser recebido pelo feto através da placenta. São descritas em três gravidades: leve, moderada e grave. Neste caso, descreve-se as desacelerações da FCF

que podem ser variáveis no tempo e tamanho. Percebe-se que em grande parte dos casos, os fetos não apresentam desacelerações severas, porém em alguns casos o aumento do número de desacelerações por segundo, principalmente por intervalos superiores a 3 minutos carregam um alerta aos pais [9].

Os histogramas da frequência cardíaca fetal também podem ser possíveis atributos a serem considerados na construção do modelo, sendo que o conjunto dispõe de 6 medidas distintas do histograma cardíaco. Além disso, tem-se a variabilidade de curto prazo da frequência cardíaca fetal, a referência para a normalidade deve ser a duração superior a 3ms. Desse modo, é possível computar o tempo desta variabilidade, que por sua vez, auxilia no monitoramento fetal em gestações normais e de risco, evitando maior número de morte-fetal.

3 Preparação dos dados

3.1 Limpeza dos dados

Foi encontrado 12 linhas duplicadas no conjunto de dados. Esses dados duplicados foram retirados, visto que não estão agregando informação ao modelo, pois é um mesmo registro. Além de que os dados duplicados também podem atrapalhar nas pré-análises.

Ao procurar dados nulos [10] notou-se que no conjunto de dados escolhido não havia nenhuma linha nula ou valor nulo. A Figura 10 mostra que não há dados faltantes em nenhuma das colunas do conjunto de dados.

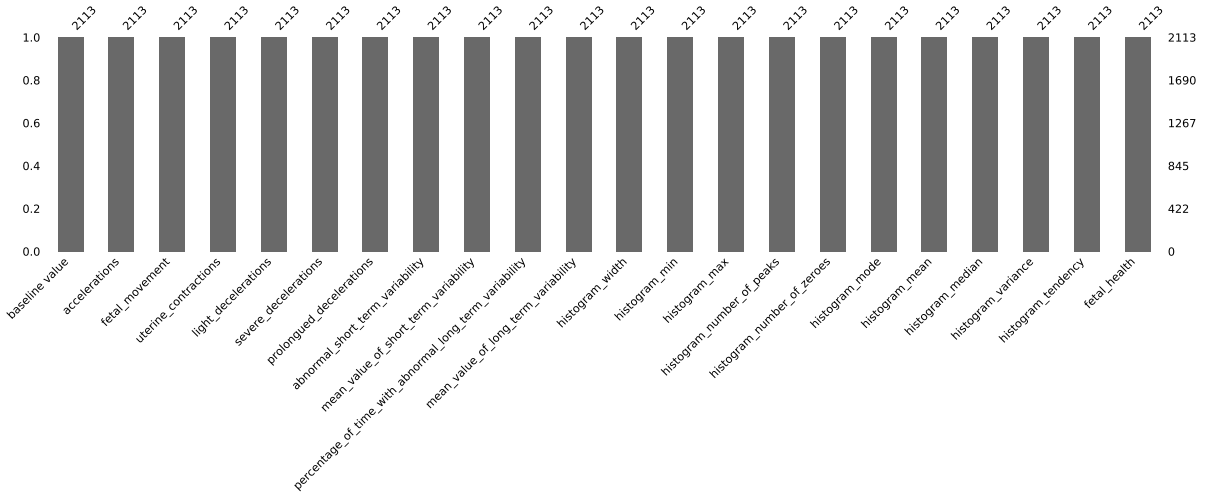


Figura 10: Dados faltantes.

3.2 Padronização dos dados

A padronização dos dados possui como objetivo transformar os dados para a mesma ordem de grandeza. A equação 1 apresenta a padronização utilizada nos dados.

$$z_i^j = \frac{x_i^j - \mu_j}{\sigma_j}. \tag{1}$$

Ao padronizar os dados, as variáveis resultarão em uma média igual a 0 e um desvio padrão igual a 1. Todas as variáveis explicativas do conjunto de dados foram padronizadas, visto que as variáveis possuem escalas diferentes e essa diferença pode aumentar a dificuldade de classificação da cardiotocografia do feto.

3.3 Extração das variáveis respostas

Para a seleção das variáveis respostas foi aplicado o método *SelectKBest* da biblioteca scikit-learn, na qual avalia para um problema de classificação quais são as variáveis respostas mais importantes para a classificação final.

A Figura 11 apresenta a pontuação dada para cada variável explicativa. Para compreender a influência das variáveis explicativas no desempenho do modelo será considerado diferentes quantidades de variáveis explicativas e analisado o desempenho do modelo. As quantidades escolhidas são:

- As 5 variáveis com maior pontuação: Desacelerações prolongadas, variabilidade anormal de curto prazo, porcentagem de tempo com variabilidade anormal de longo prazo, média e moda do histograma.
- As 10 variáveis com maior pontuação: Desacelerações prolongadas, variabilidade anormal de curto prazo, porcentagem de tempo com variabilidade anormal de longo prazo, média, moda e mediana do histograma, acelerações, variância do histograma, frequência cardíaca fetal e valor médio da variabilidade de curto prazo
- As 15 variáveis com maior pontuação: Desacelerações prolongadas, variabilidade anormal de curto prazo, porcentagem de tempo com variabilidade anormal de longo prazo, média, moda e mediana do histograma, acelerações, variância do histograma, frequência cardíaca fetal, valor médio da variabilidade de curto prazo, contrações uterinas, valor mínimo (baixa frequência) do histograma FHR, valor médio da variabilidade de longo prazo, desacelerações leves e largura do histograma.

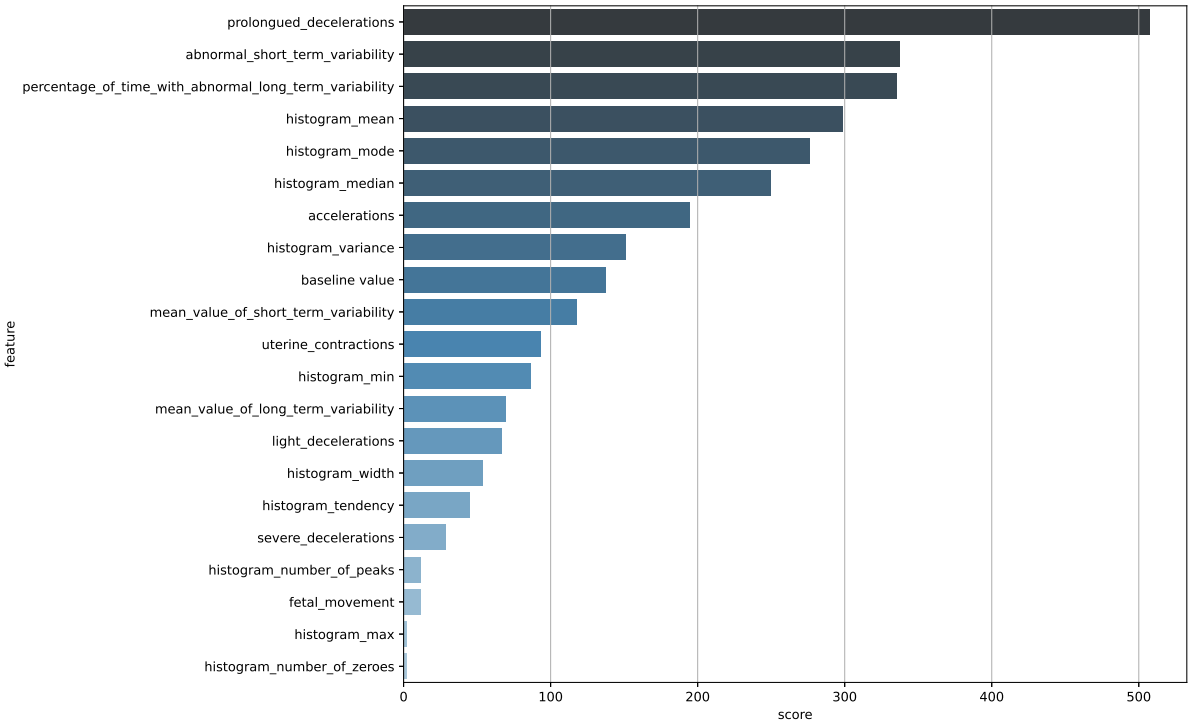


Figura 11: Pontuação das variáveis respostas com *SelectKBest*.

3.4 Balanceamento dos dados

Para realizar o balanceamento dos dados será utilizado o método de *Oversampling* que consiste em gerar dados sintéticos (não duplicados) da classe minoritária a partir de vizinhos [11]. Para isso, utilizou-se a técnica SMOTE (Synthetic Minority Over-sampling Technique) da biblioteca *imblearn*, na qual é re-amostrado as classes suspeito e patológico (classes minoritárias) e é considerado 5 vizinhos mais próximos para a construção das amostras sintéticas. Assim, foi obtido que todas as classes possuem 1646 amostras.

4 Modelos

Foi escolhido dois algoritmos para desenvolvimento do modelo de classificação da cardiotocografia do feto. Os algoritmos escolhidos foram regressão logística e floresta aleatória.

4.1 Regressão Logística

A regressão logística é um algoritmo de classificação usado quando a variável resposta é categórica. A ideia da regressão logística é encontrar uma relação entre as características e a probabilidade de um resultado específico. A curva de regressão logística é construída usando o logaritmo natural das "probabilidades" da variável alvo, ao invés da probabilidade. Além disso, os preditores não precisam ser normalmente distribuídos ou ter variâncias iguais em cada grupo.

Assim como os demais modelos de *data mining*, a regressão logística possui hiperparâmetros que devem ser finitos para o treinamento do modelo. Para a aplicação do modelo, foi utilizado

a biblioteca *sklearn* do python. Os hiperparâmetros ajustados do modelo de regressão foram: A quantidade máxima de iterações (150) e o algoritmo utilizado para a otimização (lbfgs).

4.2 Floresta aleatória

O algoritmo floresta aleatória é composto por diferentes árvores de decisão, cada uma com os mesmos nós, mas usando dados diferentes que levam a diferentes folhas. Ele mescla as decisões de várias árvores de decisão para encontrar uma resposta que representa a média de todas essas árvores de decisão. O algoritmo floresta aleatória é usado tanto para resolver problemas de regressão quanto classificação.

5 Validação dos modelos

O objetivo é medir quão distante o modelo está da classificação perfeita, todavia cada métrica faz de uma maneira diferente [12]. Para a análise dos modelos foi utilizado a matriz de confusão e outras métricas avaliativas, como precisão, sensibilidade e métrica F1. A precisão é a razão entre a quantidade de exemplos classificados corretamente como positivos e o total de exemplos classificados como positivo. Essa métrica dá maior ênfase para os erros por falso positivo [12].

A sensibilidade é a razão entre a quantidade de exemplos classificados corretamente como positivos e a quantidade de exemplos que são de fato positivos. Essa métrica dá maior ênfase para os erros por falso negativo. Enquanto a métrica F1 é uma média harmônica entre a precisão e a sensibilidade [12].

Para a cardiotocografia do feto a métrica mais relevante seria em relação ao erro por falso negativo, visto que não identificar uma anomalia quando ela de fato existe impacta negativamente a saúde do feto. Desta forma, a métrica com maior peso para a análise de desempenho do modelo será a sensibilidade.

5.1 Regressão Logística

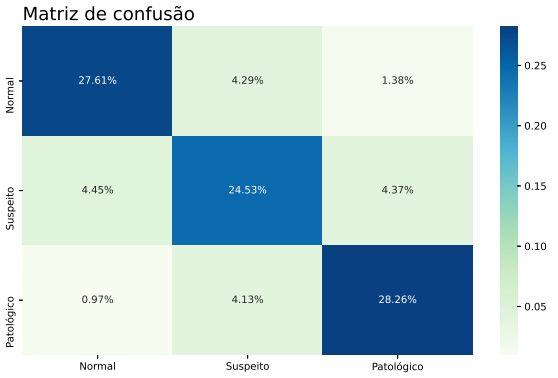
A Tabela 4 apresenta as métricas avaliativas para o modelo de regressão logística, sendo os valores na cor branca para 5 variáveis explicativas, na cor cinza para 10 variáveis explicativas e na cor ciano para 15 variáveis explicativas. Para o modelo com 15 variáveis explicativas a precisão foi de 97% para fetos com cardiotocografia normal, ou seja, a cada 100 fetos classificados como normais, é esperado que apenas 97 estejam de fato com a cardiotocografia normal. Enquanto a sensibilidade foi de 84%, ou seja, a cada 100 fetos que de fato possuem a cardiotocografia normal, é esperado que apenas 84 sejam corretamente identificados com cardiotocografia não normal.

Além disso, foi obtido 90% para a métrica F1 que é a combinação entre a precisão e a sensibilidade. A mesma lógica pode ser aplicada para o entendimento das métricas das outras classes. Ainda a partir da Tabela 4 nota-se que o modelo com 15 variáveis explicativas teve uma melhor precisão em relação aos outros modelos. A classe suspeito para o modelo com 5 variáveis explicativas teve o pior desempenho se comparado com os outros modelos, sendo que os valores de precisão, sensibilidade e métrica F1 ficaram na faixa de 70%.

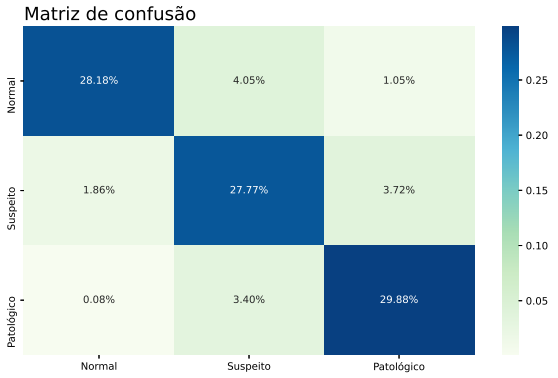
	Precisão	Sensibilidade	Métrica F1
Normal	0,84	0,83	0,83
	0,93	0,85	0,89
	0,97	0,98	0,90
Suspeito	0,74	0,73	0,74
	0,79	0,83	0,81
	0,80	0,88	0,84
Patológico	0,84	0,83	0,83
	0,86	0,89	0,88
	0,87	0,90	0,89

Tabela 4: Métricas avaliativas para o modelo de regressão logística.

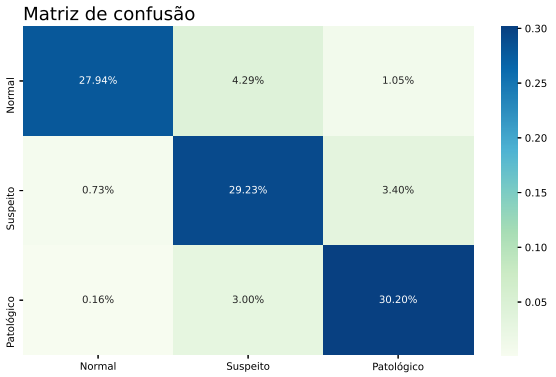
A Figura 12 apresenta as matrizes de confusão para cada modelo de regressão logística. A diagonal principal apresenta a porcentagem de exemplos que foram classificados corretamente para cada classe. O modelo com 15 variáveis explicativas (Figura 12c) apresentou os valores mais próximos de 33,3% (que seria o ideal), sendo que 27,94% dos exemplos que eram normais foram classificados como normais, 29,23% dos exemplos que eram suspeitos foram classificados como suspeito e 30,20% dos exemplos que eram patológicos foram classificados como patológico.



(a) 5 variáveis explicativas



(b) 10 variáveis explicativas



(c) 15 variáveis explicativas

Figura 12: Matrizes de confusão para o modelo de regressão logística.

O pior caso seria a cardiotocografia do feto ser classificada como normal, mas ser na verdade patológico. Essa situação teve uma porcentagem abaixo de 1% para todos os modelos, sendo o melhor desempenho para o modelo com 10 variáveis explicativas (Figura 12b), o qual obteve uma porcentagem de 0,08%. Além disso, de acordo com a Tabela 5 a acurácia dos modelos foi em torno de 80%, ou seja, a cada 100 fetos 80 tiveram a cardiotocografia classificada corretamente.

	5 variáveis explicativas	10 variáveis explicativas	15 variáveis explicativas
Acurácia	0,8	0,86	0,88

Tabela 5: Acurácia dos modelos para diferentes quantidades de variáveis explicativas.

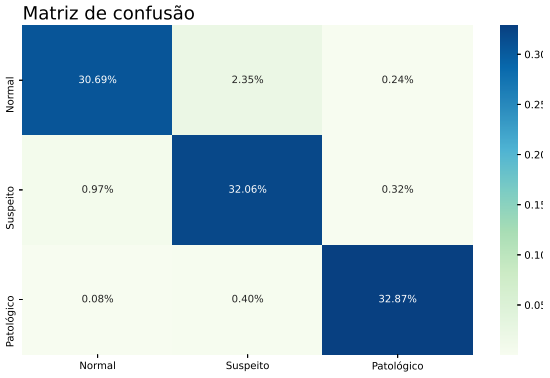
5.2 Floresta aleatória

O segundo modelo proposto para classificar a saúde do feto foi a floresta aleatória. Semelhante a análise aplicada na regressão logística, analisou-se a capacidade de generalização do modelo através das métricas avaliativas de precisão, sensibilidade e F1 como demonstrado na Tabela 6.

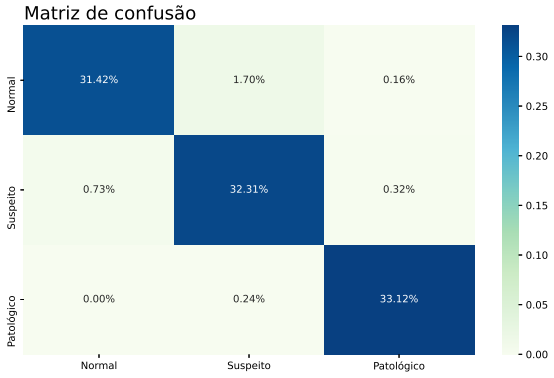
	Precisão	Sensibilidade	Métrica F1
Normal	0,97	0,92	0,94
	0,97	0,94	0,96
	0,98	0,94	0,96
Suspeito	0,92	0,96	0,94
	0,94	0,97	0,95
	0,94	0,97	0,96
Patológico	0,98	0,98	0,98
	0,98	0,99	0,99
	0,98	0,99	0,98

Tabela 6: Métricas avaliativas para a Floresta aleatória.

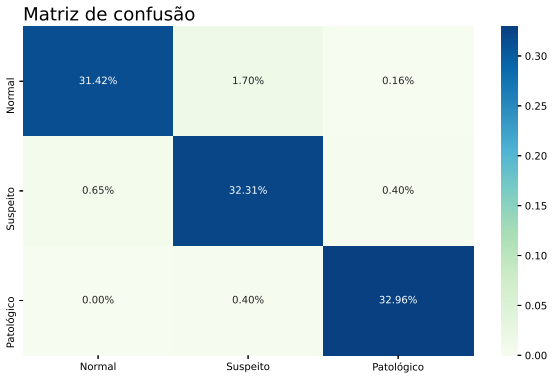
Observa-se que as métricas apresentam valores próximos de 1 para 5, 10 e 15 variáveis explicativas nas 3 métricas. Essa consideração aponta que os modelos de classificação são satisfatórios, porém a análise da matriz de confusão torna-se imprescindível para uma análise mais efetiva do modelo de floresta aleatória.



(a) 5 variáveis explicativas



(b) 10 variáveis explicativas



(c) 15 variáveis explicativas

Figura 13: Matrizes de confusão para o modelo de floresta aleatória.

A partir das matrizes de confusão, Figura 13, obtidas e analisando a performance geral para conjunto de *features* (Tabela 7), foi possível inferir que o modelos de floresta aleatória com 10 e 15 variáveis explicativas apresentam melhores resultados.

Na matriz de confusão para 10 variáveis tem-se um valor de 32,87% como indicação dos resultados de exames patológicos classificados como patológicos, enquanto que nas outras matrizes essa observação apresenta valores menores. Além disso, a métrica de maior interesse (sensibilidade) apresenta aproximadamente 99%. Esse valor indica que a cada 100 cardiotocografias classificadas como patológico em 99 delas era realmente um exame anômalo.

	5 variáveis explicativas	10 variáveis explicativas	15 variáveis explicativas
Acurácia	0,96	0,97	0,97

Tabela 7: Acurácia dos modelos para diferentes quantidades de variáveis explicativas.

6 Considerações Finais

Para o modelo de regressão logística através da análise da avaliação do desempenho do modelo conclui-se que o melhor resultado foi obtido com 15 variáveis explicativas. Enquanto para o modelo de floresta aleatória o modelo que classificou com efetividade e com menor complexidade foi o que apresentava 10 variáveis explicativas.

Além do mais, buscou-se outras análises com o mesmo *dataset* utilizado neste estudo. A exploração proposta em [13] apresenta o *RandomForestClassifier* como modelo de classificação, além da construção de um estimador de hiperparâmetros *GridSearchCV* da biblioteca *sklearn*. Porém os resultados para acurácia, sensibilidade, precisão e métrica F1 obtidos foram ligeiramente menores que os alcançados pelos modelos apresentados neste trabalho. A justificativa para essa diferença de desempenho pode ser encontrada na etapa de preparação dos dados, visto que o presente trabalho incluiu extração de $features$ e balanceamento das classes, o que não foi realizado pela autora em [13].

Portanto, o uso de um modelo de *machine learning* como o abordado neste trabalho oportuniza com confiança a obtenção de resultados precoces em relação a cardiotocografia do feto. Assim, auxiliando os especialistas na tomada de decisão para posterior tratamento do feto com o objetivo de garantir a saúde do feto e da mãe.

7 Referências

[1] SEDICIAS, S. *Como é feita a cardiotocografia fetal*. Acesso em 12 de março de 2021. Disponível em: <https://www.tuasaude.com/cardiotocografia-fetal/>.

[2] MELLO, H. C. *Cardiotocografia: entenda a importância para a saúde do bebê*. 2019. Acesso em 14 de março de 2021. Disponível em: <https://blog.medicalway.com.br/cardiotocografia-entenda-a-importancia-para-a-saude-do-bebe/>.

[3] LARXEL. *Fetal Health Classification*. Acesso em 12 de março de 2021. Disponível em: <https://www.kaggle.com/andrewmvd/fetal-health-classification>.

[4] IBM. *Guia do IBM SPSS Modeler CRISP-DM*. Acesso em 12 de março de 2021. Disponível em: <ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/17.1/br_po/ModelerCl

[5] NUNES, P. D. I. *Cardiotocografia*. Acesso em 12 de março de 2021. Disponível em: <https://www.csaudeboavista.com/cardiotocografia-o-que-e-como-funciona-e-para-que-serve/>.

[6] KAISER, M. *How to Tackle Multicollinearity*. 2020. Acesso em 14 de março de 2021. Disponível em: <https://kaiserm.medium.com/how-to-tackle-multicollinearity-79afe58e9479>.

[7] BOXPLOT - seaborn. Acesso em 14 de março de 2021. Disponível em: <http://seaborn.pydata.org/generated/seaborn.boxplot.html>.

[8] TODD DR MATTHEW RUCKLIDGE, M. T. K. D. C. *TUTORIAL DE ANESTESIA DA SEMANA MONITORIZAÇÃO DOS BATIMENTOS CARDÍACOS FETAIS – PRINCÍPIOS DA INTERPRETAÇÃO DA CARDIOTOCOGRAFIA*. Acesso em 1 de maio de 2021. Disponível em: <https://tutoriaisdeanestesia.paginas.ufsc.br/files/2013/11/Monitoriza%C3%A7%C3%A3o-fetal-Parte-2.pdf>.

[9] GONZALES, M. D. O. *Fatores que influenciam a cardiotocografia computadorizada em gestantes hipertensas*. 2019. Acesso em 14 de março de 2021. Disponível em: <https://www.teses.usp.br/teses/disponiveis/5/5139/tde-07012020-134808/publico/MarinadeOliveiraGonzales.pdf>.

[10] YıLDıRıM, S. *Visualize Missing Values with Missingno*. 2020. Acesso em 14 de março de 2021. Disponível em: <https://towardsdatascience.com/visualize-missing-values-with-missingno-ad4d938b00a1>.

[11] SANTANA, R. *Lidando com Classes Desbalanceadas – Machine Learning*. Acesso em 2 de maio de 2021. Disponível em: <https://minerandodados.com.br/lidando-com-classes-desbalanceadas-machine-learning/>.

[12] KUNUMI. *Métricas de Avaliação em Machine Learning: Classificação*. 2020. Acesso em 2 de maio de 2021. Disponível em: <https://medium.com/kunumi/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-em-machine-learning-classifica%C3%A7%C3%A3o-49340dcdb198>.

[13] KAPOOR, K. *Fetal Health Classification*. 2021. Acesso em 2 de maio de 2021. Disponível em: <https://www.kaggle.com/karnikakapoor/fetal-health-classification>.