resdite

# Classificação de exames de cardiotocografia usando modelagem por inteligência artificial

CLASSIFICATION OF CARDIOTOCOGRAPHY EXAMINATIONS USING ARTIFICIAL INTELLIGENCE MODELING

Larissa Favaro Redondo[1], Miguel Antonio do Nascimento Garcia[2], Juliana Oliveira de Meneses[3], Luciano Rodrigo Lopes[4], João Brainer Clares de Andrade[5], Ivan Torres Pisa[6]

[1] Graduanda em informática em saúde. Universidade Federal de São Paulo.
ORCID: https://orcid.org/0009-0008-5208-8119
**Email:** larissafredondo@gmail.com
[2] Graduando em informática em saúde. Universidade Federal de São Paulo.
ORCID: https://orcid.org/0000-0001-9777-283X
**Email:** miguel.garcia@unifesp.br
[3] Graduanda em informática em saúde. Universidade Federal de São Paulo.
ORCID:https://orcid.org/0009-0000-2337-8653
**Email:** jmeneses@unifesp.br
[4] Doutor em ciências. Universidade Federal de São Paulo.
ORCID: https://orcid.org/0000-0002-0284-2821
**Email:** luciano.lopes@unifesp.br
[5] Doutor em medicina. Universidade Federal de São Paulo.
ORCID: https://orcid.org/0000-0001-8768-7164
**Email:** joaobrainer@gmail.com
[6] Livre-docente em informática em saúde. Universidade Federal de São Paulo.
ORCID: https://orcid.org/0000-0002-5106-3904
**Email:** ivanpisa@unifesp.br

**Correspondência**: Larissa Favaro Redondo. Departamento de Informática em Saúde, Escola Paulista de Medicina. Rua Botucatu 862. Vila Clementino. 04023-062. São Paulo, SP.



**Conflito de interesses**: os autores declaram que não há conflito de interesses.

**Como citar este artigo**

Redondo LF, Garcia MAN, Meneses JO, Moreto AJ, Lopes LR, Andrade JBC, Pisa IT. Classificação de exames de cardiotocografia usando modelagem por inteligência artificial. Revista de Saúde Digital e Tecnologias Educacionais. [*online*], volume 9, n. 2. Editor responsável: Luiz Roberto de Oliveira. Fortaleza, out de 2024. Disponível em: http://periodicos.ufc.br/resdite/index. Acesso em "dia/mês/ano".

**Data de recebimento do artigo:** 04/12/2023
**Data de aprovação do artigo:** 28/02/2024
**Data de publicação:** 07/10/2024

## Resumo

**Introdução:** A cardiotocografia (CTG) avalia o bem-estar fetal durante o parto, mas sua interpretação pode gerar falsos positivos. Este estudo buscou aprimorar a interpretação da CTG usando técnicas de inteligência artificial (IA). Analisando 2.126 exames com 22 parâmetros, incluindo classificações de saúde fetal, técnicas de seleção de variáveis foram aplicadas. Modelos de IA, como árvore de decisão, random forest, regressão logística e naive bayes, foram implementados. **Métodos:** A análise exploratória identificou características cruciais para prever a saúde fetal, enfatizando a importância da avaliação sistemática das variáveis. Estratégias de correlação foram usadas na

criação de conjuntos de dados específicos, fortalecendo a base para comparar modelos. **Resultados:** A avaliação do desempenho, com métricas como acurácia e F2-score em diferentes conjuntos de variáveis, mostrou melhorias consistentes ao incorporar mais variáveis. Os modelos random forest destacaram-se, indicando eficácia na classificação dos resultados da CTG. **Conclusão:** A abordagem estratificada na seleção de variáveis foi ressaltada para otimizar a precisão dos modelos. Essa análise proporciona insights para interpretar dados cardiotocográficos, destacando a relevância clínica e a aplicabilidade dos modelos de IA na medicina obstétrica. Contribui para futuras pesquisas e desenvolvimentos na área.

**Palavras-chave:** Cardiotocografia, Obstetrícia, Inteligência artificial.

**Abstract**

***Introduction:*** *Cardiotocography (CTG) measures fetal well-being during labor, but this interpretation can lead to false positives. This study sought to improve CTG interpretation using artificial intelligence (AI) techniques. By analyzing 2,126 exams with 22 parameters, including fetal health classifications, variable selection techniques were applied. AI models such as decision tree, random forest, logistic regression and naive bayes were implemented.* ***Methods:*** *Exploratory analysis identified crucial characteristics for predicting fetal health, emphasizing the importance of systematic evaluation of variables. Correlation strategies were used to specific data sets, strengthening the basis for comparing models.* ***Results:*** *Performance evaluation, with metrics such as accuracy and F2-score on different sets of variables, showed consistent improvements when incorporating more variables. The random forest models stood out, indicating effectiveness in classifying CTG results.* ***Conclusion:*** *The stratified approach to variable selection was highlighted to optimize model accuracy. This analysis provides insights for interpreting cardiotocographic data, highlighting the clinical relevance and applicability of AI models in obstetric medicine. It contributes to future research and development in the area.*

**Keywords:** *Cardiotocography, Obstetrics, Artificial intelligence.*

## 1. Introdução

A cardiotocografia (CTG) é uma técica diagnóstica amplamente usada para monitorar a frequência cardíaca fetal e avaliar o bem-estar do feto. Seu principal objetivo é detectar sinais de hipóxia intrauterina e acidose fetal, que podem estar relacionados a várias condições de saúde materna, crescimento fetal anormal, analgesia epidural, presença de mecônio ou atividade uterina excessiva[1]. A CTG foi introduzida no final da década de 1960 por Hon, Hammacher e Caldeyro-Barcia, após estudos que observaram padrões anormais de frequência cardíaca correlacionados com baixos escores de Apgar, acidose fetal e alta mortalidade neonatal[2,3].

Entretanto, a interpretação dos resultados da CTG é influenciada por vários fatores, levando a uma taxa de falsos positivos de até 60%[1]. Além disso, há variabilidade na interpretação, destacando a necessidade de um entendimento adequado da fisiologia fetal, experiência clínica acumulada e análise do contexto clínico durante o trabalho de parto para

uma classificação precisa[1,2]. Fatores de risco indesejáveis ou inevitáveis relacionados ao parto, como ruptura prolongada de membranas, prematuridade, restrição de crescimento intrauterino, infecção, presença de líquido amniótico com coloração de mecônio, uso de ocitocina e cicatriz uterina, também precisam ser considerados para a interpretação adequada da CTG[4].

A cardiotocografia depende de instrumentação composta principalmente de transdutores ou sensores, processadores de sinal, dispositivos de exibição, software de análise e dispositivos de ultrassom Doppler[2]. No entanto, essa técnica diagnóstica, cujo aprendizado e interpretação são complexos, pode levar a intervenções desnecessárias e até mesmo a práticas obstétricas inadequadas. Além disso, a CTG gera uma grande quantidade de dados, exigindo apoio computacional do equipamento. Nesse contexto, o desenvolvimento de técnicas rápidas e seguras se apresenta como uma vantagem significativa no campo da obstetrícia de emergência. O emprego de algoritmos e técnicas de inteligência artificial (IA) na área da saúde se apresenta como desenvolvimento importante para apoiar problemas dessa natureza[11].

O objetivo deste estudo é analisar os resultados da experimentação de modelagem de inteligência artificial para dados provenientes de exames de CTG para fornecer uma classificação automatizada.

## 2. Métodos

Os dados usados foram extraídos de um artigo de Ayres de Campos et al.[1] que apresenta resultados de uma análise automatizada de CTGs. Este trabalho foi considerado como pesquisa-base para o estudo aqui apresentado. A amostra de dados da pesquisa-base contém 2.126 registros de resultados de exames de CTGs coletados no ano de 2015 e foram respeitadas diretrizes desenvolvidas pela Federação Internacional de Ginecologia e Obstetrícia (FIGO) para orientar a prática clínica no monitoramento fetal intraparto.

Um CTG consiste em um exame médico que monitora a frequência cardíaca fetal e as contrações uterinas durante a gravidez. É um exame importante para avaliar a saúde do feto e a progressão do trabalho de parto. Esses sinais são registrados e podem ser visualizados em tempo real ou armazenados para posterior análise. Para garantir a qualidade do registro do CTG na pesquisa-base[1] foram adotados critérios de inclusão e exclusão. Foram incluídos apenas CTGs de boa qualidade técnica, sem artefatos significativos que pudessem comprometer a análise posterior. Foram excluídos CTGs com interrupções ou falhas na aquisição dos sinais, bem como aqueles com informações incompletas ou ilegíveis. Os CTGs foram digitalizados e pré-processados para possibilitar uma análise computacional. Foi

realizada a correção de eventuais artefatos e a normalização dos sinais. Em seguida, os registros foram segmentados em intervalos de tempo específicos, que correspondem a janelas de análise para extração das características relevantes. Informações clínicas adicionais foram coletadas para cada CTG, incluindo dados demográficos das gestantes, histórico médico, resultados de exames complementares e informações sobre complicações maternas ou fetais.

Os dados coletados dos CTGs na pesquisa-base[1] foram dispostos em um arquivo *comma-separated values* (CSV) contendo 22 parâmetros com informação do exame como também resultados de análises estatísticas disponibilizadas pelos autores do estudo original. As colunas incluem:

- Frequência cardíaca basal (*baseline heart rate*): Refere-se à taxa base da frequência cardíaca fetal (FCF) em batimentos por minuto (bpm). Esse dado é medido em tempo real através de um gráfico. A FCF é monitorada por pelo menos 10 minutos e os valores normais variam entre 110 e 160 bpm. Valores acima ou abaixo dessa faixa podem indicar possíveis patologias fetais.
- Acelerações (*accelerations*): Representa o número de acelerações cardíacas por segundo em resposta aos movimentos fetais. A maioria das acelerações coincide com os movimentos do feto e indica uma resposta neurológica adequada, indicando a ausência de hipóxia/acidose fetal.
- Desacelerações (*decelerations*): São quedas na frequência cardíaca fetal abaixo da linha de base, com amplitude superior a 15 bpm e duração superior a 15 segundos. As desacelerações leves, de curta duração e com variabilidade normal dentro delas, geralmente coincidem com as contrações uterinas e são causadas pela compressão da cabeça do feto, não indicando hipóxia/acidose fetal. Já as desacelerações prolongadas, com duração superior a 3 minutos, podem indicar hipoxemia e hipóxia/acidose fetal aguda, requerendo intervenção emergencial.
- Variabilidade (*variability*): Refere-se às variações no sinal da frequência cardíaca fetal (FCF) em relação aos valores de acelerações e desacelerações. É avaliada em amplitude de segmentos de um minuto do sinal do FCF.
- Número contrações uterinas por segundo: Durante o CTG, é monitorado o número de contrações uterinas por minuto. Uma média saudável de contrações uterinas é de cerca de 2 a 5 durante o trabalho de parto ativo, podendo variar de acordo com fatores como a fase do trabalho de parto, posição da mãe, posição do feto e intervenções médicas.
- Histograma de valores do exame: É uma representação gráfica das frequências das contrações uterinas e da frequência cardíaca fetal ao longo de um determinado período

de tempo. Os picos no histograma podem indicar eventos importantes para a avaliação da saúde fetal durante o trabalho de parto.

- Número de zeros (0) no histograma do exame: Representa os períodos em que a frequência cardíaca fetal não pôde ser detectada pelo equipamento de monitoramento. A ocorrência frequente ou prolongada de valores 0 pode indicar sofrimento fetal e requer atenção médica imediata.
- Mediana, moda, média e variância do histograma: Essas medidas estatísticas são utilizadas para descrever características específicas dos dados representados pelo histograma, como a distribuição dos valores, o valor mais frequente, a média dos dados e a dispersão em relação à média.
- Tendência do histograma: Refere-se à direção geral dos valores representados pelo histograma, podendo ser positiva (aumento), negativa (diminuição) ou neutra (sem tendência).
- Classificação ou comportamento do histograma saúde fetal (*fetal health*): Classifica o resultado do CTG em três categorias: normal, suspeito ou patológico, com base nas características observadas no histograma.

Com base na análise computacional exploratória foram identificadas as características mais relevantes para o estudo de predição da saúde fetal. Métodos estatísticos e de inteligência artificial (aprendizado de máquina) foram aplicados para selecionar as variáveis mais significativas conforme descrição a seguir. Foram considerados critérios como relevância clínica, correlações com a saúde fetal e interpretabilidade. Para o critério de classificação dos dados obtidos foi considerada uma classificação proposta pelos médicos especialistas, autores do artigo original[1], sendo normal (ID1), suspeito (ID2) e patológico (ID3).

Para avaliar a significância das variáveis com o propósito de realizar uma predição, bem como considerar quais desses valores eram observados em tempo real pelo médico no ato do exame e se esses mesmos estariam estatisticamente relacionados ao diagnóstico disponibilizado pelos autores, foi conduzido um levantamento empregando o MeanDecreaseGini[5] como medida-chave. Esta etapa possibilitou quantificar o impacto de cada variável na redução da impureza das divisões nas árvores de decisão, proporcionando uma avaliação sistemática da relevância de cada componente para o desempenho global do modelo. Variáveis com um MeanDecreaseGini mais elevado são consideradas mais importantes, indicando uma maior capacidade de informação e impacto na qualidade das decisões preditivas do modelo. Essa análise refinada da importância das variáveis não apenas aprimora a compreensão dos fatores determinantes na cardiotocografia, mas também oferece insights valiosos para otimização e interpretabilidade do modelo.

Posteriormente foi realizada a implementação de quatro modelos de inteligência artificial baseados nas técnicas árvore de decisão, random forest, regressão logística e naive bayes. Essa escolha baseia-se em sua comprovada eficácia no tratamento de problemas de classificação e predição. A técnica árvore de decisão[5] usa uma estrutura hierárquica de regras de decisão para classificar os dados. Cada nó da árvore representa uma decisão baseada em uma determinada característica do conjunto de dados, levando à classificação final do nó folha correspondente. Essa técnica é conhecida por sua interpretabilidade e capacidade de capturar relacionamentos não lineares entre as variáveis. A técnica random forest[6] usa um conjunto de árvores de decisão independentes para realizar a classificação. Cada árvore é construída utilizando uma amostra aleatória do conjunto de dados e, em seguida, a classificação final é obtida por meio de uma votação majoritária das árvores individuais. Essa técnica aproveita o conceito de *ensemble learning* para aumentar a precisão e a estabilidade das previsões. A técnica regressão logística[7] é amplamente usada em problemas de classificação binária. É capaz de estimar a probabilidade de pertencer a uma classe usando uma função logística que relaciona as variáveis de entrada com a variável dependente. Essa técnica está fundamentada na teoria da regressão e é capaz de lidar com dados categóricos e contínuos. Por fim, a técnica naive bayes[8] é frequentemente usada em diagnósticos médicos. É uma escolha popular quando a independência aproximada entre as variáveis preditoras é razoável, e é especialmente eficaz para conjuntos de dados grandes e de alta dimensão.

Os modelos de IA implementados foram treinados usando o conjunto de dados pré-processados. Foi empregada a técnica de validação cruzada[9], dividindo o conjunto de dados em partes para treinamento (80%) e teste (20%). A divisão do conjunto de dados em treinamento e teste foi realizada no código Python usando a função *train_test_split* da biblioteca *scikit-learn*. Adicionou-se um segundo parâmetro para garantir a reprodutibilidade do estudo, o *random_state=42*. Esse parâmetro é usado para garantir que a divisão seja reprodutível, com um valor 42 para a semente do gerador de números aleatórios. A escolha de uma divisão em 80% para treinamento e 20% para teste em um modelo de aprendizado de máquina é justificada com base na necessidade de equilibrar a capacidade de treinamento do modelo e a avaliação da sua capacidade de generalização. Essa divisão proporciona uma abordagem razoável para evitar o *overfitting* (ajuste excessivo) e para avaliar a capacidade de generalização do modelo, ajudando a garantir que ele seja eficaz na resolução de problemas práticos

Foram usadas métricas como acurácia e F-*measure*[10] para avaliar o desempenho dos modelos. A acurácia é uma métrica fundamental no campo da avaliação de modelos,

desempenhando um papel crucial na mensuração da eficácia de algoritmos na inteligência artificial. Essa métrica representa a proporção de predições corretas em relação ao total de predições feitas pelo modelo, oferecendo uma visão geral da precisão global do sistema. Já o F-*measure* fornece uma abordagem equilibrada para avaliar o desempenho do modelo. Essa métrica é particularmente valiosa em cenários nos quais tanto a precisão (*precision*) quanto a revocação (*recall*) são de igual importância, evitando a ênfase excessiva em uma métrica às custas da outra. Após uma discussão com um médico especialista chegou-se à conclusão de que é mais adequado usar F2-*score* já que, dentre os erros possíveis da classificação automatizada, é visto como preferível que uma mulher saudável faça uma cesariana do que permitir que uma gestante com um feto em situação patológica não a faça. O ponto crítico era dar prioridade à vida do feto. A saúde e a vida do feto são cruciais, e a acidose fetal pode ser uma condição potencialmente perigosa. Se um feto acidótico não for identificado a tempo e não receber uma cesariana, as consequências podem ser extremamente graves, incluindo lesões cerebrais, paralisia cerebral e até mesmo óbito. Portanto, minimizar a ocorrência de falsos negativos é fundamental para evitar essas tragédias. Focar na revocação mais alta, com o F2-*score*, significa que o modelo está otimizado para identificar a maioria dos casos reais de fetos em risco, minimizando o risco de falsos negativos.

Os resultados obtidos foram analisados criticamente considerando a interpretabilidade dos modelos, sua aplicabilidade clínica e possíveis limitações. Foram discutidas as vantagens e desafios de cada modelo de IA, bem como a possibilidade de implementação prática em ambiente clínico.

Por fim, com relação aos aspectos éticos do estudo, os CTGs da pesquisa-base[1] foram coletados em hospitais e clínicas obstétricas com a devida obtenção do consentimento informado das gestantes participantes. Os registros foram obtidos durante o trabalho de parto ou em momentos específicos da gestação, de acordo com as necessidades clínicas e os protocolos de monitoramento fetal. Os dados aqui estudados foram obtidos da pesquisa-base e possuem licença de dados abertos, sem qualquer identificação das gestantes.

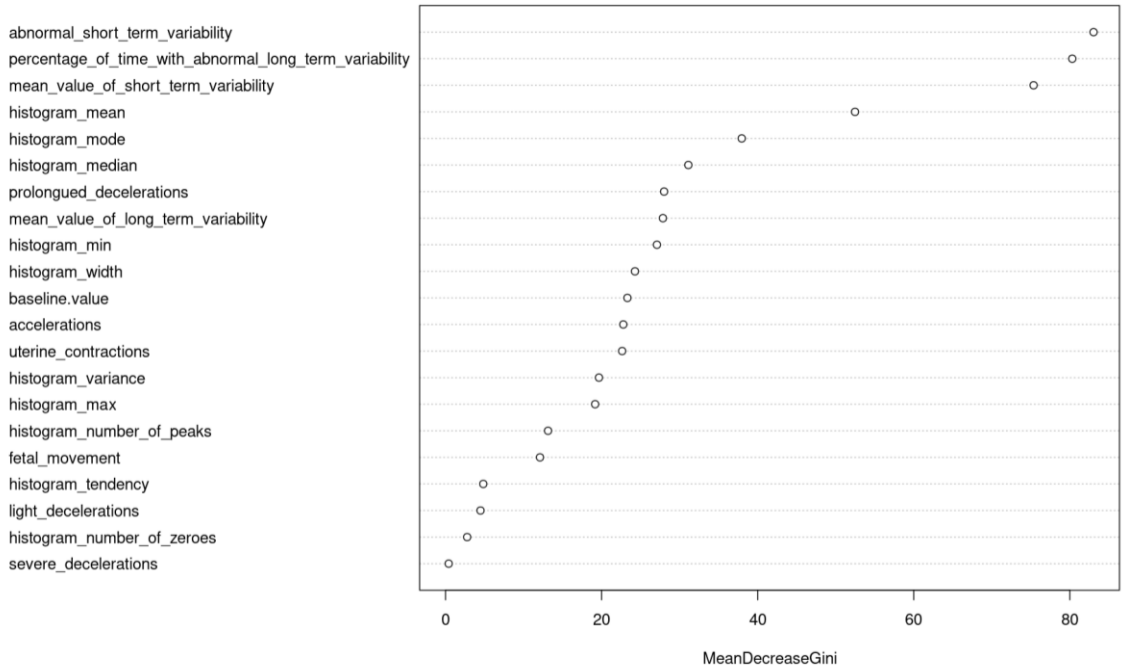
## 3. Resultados

A previsão da saúde fetal é um desafio importante no campo da medicina obstétrica, visando identificar precocemente possíveis complicações e tomar medidas preventivas. Neste estudo realizou-se uma análise exploratória para identificar as características mais relevantes para a predição da saúde fetal usando métodos estatísticos e técnicas de inteligência artificial (aprendizado de máquina). Considerou-se critérios como relevância

clínica, correlações com a saúde fetal e interpretabilidade para selecionar as variáveis mais significativas.

Ao trabalhar com 2.126 registros distintos, com 22 parâmetros cada, foi necessário avaliar quais tinham a maior significância para realizar a predição. O resultado está representado na Figura 1, proporcionando uma avaliação sistemática da relevância de cada componente para o desempenho global do modelo. Nesta figura, é apresentada a plotagem da distribuição dos parâmetros presentes no conjunto de dados originais.

**Figura 1 -** Gráfico de MeanDecreaseGini dos 22 parâmetros encontrados no conjunto de dados originais.



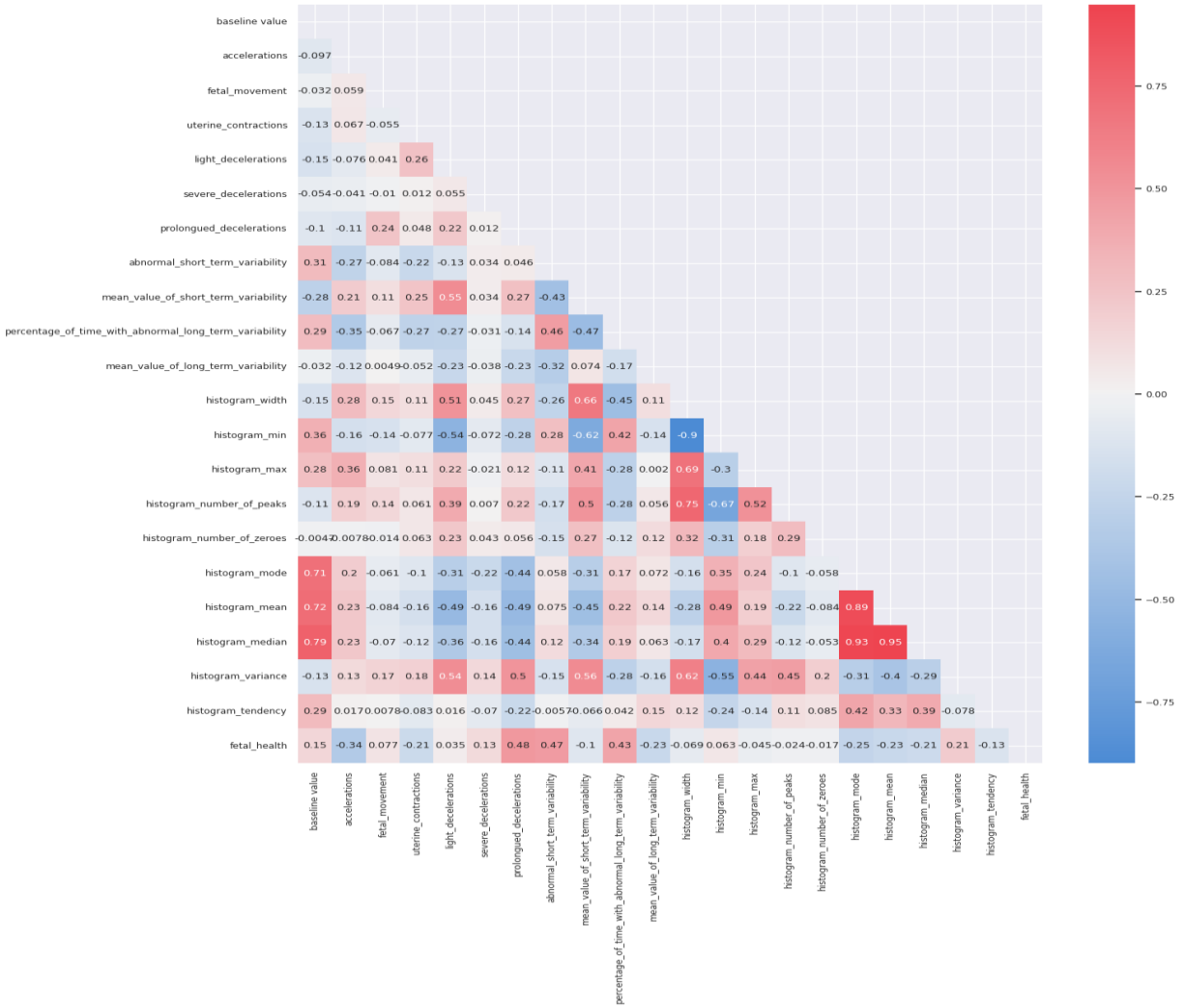

**Fonte:** Autoria própria (2023).

Levando em consideração o grande volume de parâmetros identificados, e como alguns deles não são necessariamente relacionados ao diagnóstico no momento do exame, foi elaborada uma matriz de correlação estatística para determinar quais parâmetros estavam estatisticamente relacionados ao parâmetro de *fetal health*, que determina o diagnóstico em si (ID1, ID2 ou ID3). Alguns parâmetros, como o de desacelerações prolongadas, apresentam uma forte correlação estatística, uma vez que esse dado é também analisado de forma visual por meio do gráfico do exame pelo profissional da saúde responsável.

Por meio da representação dessa análise em um mapa de calor (*heatmap)* triangular (Figura 2) foi possível averiguar quais são esses parâmetros. Pode-se observar uma forte correlação entre o gráfico de MeanDecreaseGini e o mapa de calor. Na Figura 2, quanto mais

próximo à cor vermelha intensa, maior correlação existe entre a variável e a predição de *fetal_health.* Em síntese, uma análise da distribuição bidimensional dos dados foi útil para identificar esses padrões. Nesta figura, os tons das cores indicam o grau de correlação entre os parâmetros. Tons mais intensos, como o vermelho escuro, indicam uma correlação positiva forte. Isso significa que, quando esses parâmetros aumentam ou diminuem, há uma tendência correspondente na variação da saúde fetal. Por outro lado, tons mais suaves, como o azul claro, representam uma correlação negativa fraca ou nula.

**Figura 2.** *Heatmap* (mapa de calor) mostrando a correlação estatística entre diferentes parâmetros e o parâmetro de diagnóstico *fetal_health*.




**Fonte:** Autoria própria (2023).

Após essa análise, assumindo *cutoff* de 0,10 - que representa todas as medidas avermelhadas do mapa de calor -, todos os valores correspondentes aos parâmetros que passaram nesse *cutoff* foram levados em consideração para criar novos conjuntos de dados,

esses contendo dinâmicas da correlação estatística vistas na Figura 1 e Figura 2. Esses conjuntos de dados foram então usados para fazer a comparação entre os modelos de inteligência artificial.

Foram empregados cinco conjuntos de dados distintos na análise preditiva e avaliação estatística de acurácia e F-*score*. Os conjuntos, denominados Gini 0, Gini 1, Gini 2, Gini 3 e Gini 4, foram compostos arbitrariamente a partir da ordem de correlação identificada na etapa anterior. No Gini 0 foram usadas apenas as três variáveis consideradas mais importantes pelo modelo gerado, a saber, *abnormal short term variability, percentage_of_time_with abnormal long term variability, e mean_value_of_short_term_variability*. O Gini 1 incorporou as três primeiras variáveis mais significativas e adicionou *histogram_mean, histogram_mode* e *histogram_median*. O Gini 2 compreendeu as treze primeiras variáveis mais relevantes, enquanto o Gini 3 ampliou esse conjunto para as quinze primeiras variáveis. Por fim, o Gini 4 abrangeu todas as 21 variáveis disponíveis, excluindo apenas o diagnóstico (ID1, ID2 ou ID3). Essa abordagem estratificada possibilitou uma análise detalhada do impacto progressivo das variáveis na qualidade das previsões.

Os modelos de inteligência artificial foram treinados usando um conjunto de dados pré-processados, incluindo conferência e recodificação, e empregou-se a técnica de validação cruzada para avaliar sua capacidade de generalização. Durante o treinamento, os modelos ajustaram seus parâmetros para minimizar a função de perda, medindo a diferença entre as classificações previstas e os rótulos verdadeiros do conjunto de dados de treinamento.

Os resultados das análises, incluindo as métricas de acurácia e F2-*score*, foram sistematicamente compilados e organizados em uma tabela (Tabela 1) para proporcionar uma visão consolidada do desempenho dos diferentes conjuntos de dados Gini. A acurácia foi utilizada como medida global da precisão do modelo, enquanto o F2-*score*, que pondera de maneira específica os falsos negativos em relação aos falsos positivos, ofereceu uma perspectiva para avaliar o desempenho em cenários de desequilíbrio de classes. A disposição tabular desses resultados possibilitou uma análise comparativa eficaz, destacando as nuances nas diferentes estratégias de seleção de variáveis e seu impacto nas métricas de avaliação.

No Gini 0, composto pelas três variáveis mais importantes, observa-se desempenhos variados nos modelos, destacando-se a árvore de decisão e o random forest com acurácia superior a 86,62% e 87,09%, respectivamente, e F2-*score* acima de 86%. No Gini 1, que incorpora seis variáveis essenciais, os modelos de árvore de decisão e random forest apresentam melhorias notáveis, com acurácia 89,90% e 93,19%, respectivamente, e F2-*score* acima de 89%. O Gini 2, compreendendo as treze variáveis mais importantes, revela um

aumento consistente na acurácia, atingindo 92,72% e 93,19% respectivamente, e F2-*score* acima de 92%. No Gini 3, com quinze variáveis, observa-se um desempenho robusto, destacando-se a árvore de decisão e o random forest com acurácia acima de 93% e F2-*score* superior a 93%. O Gini 4, usando todas as 22 variáveis, mostra resultados notáveis, com o modelo de random forest alcançando acurácia de 93,66% e F2-*score* de 93,60%.

**Tabela 1.** Tabela da acurácia e F2-*score* dos diferentes conjuntos de dados e modelos de predição.

| | Regressão Logística | | Árvore de Decisão | | Random Forest | | Naive Bayes | |
|---|---|---|---|---|---|---|---|---|
| | Acurácia | F2-Score | Acurácia | F2-Score | Acurácia | F2-Score | Acurácia | F2-Score |
| Gini 0 | 0.80281 | 0.79139 | 0.86619 | 0.86538 | 0.87089 | 0.86940 | 0.81924 | 0.66066 |
| Gini 1 | 0.83333 | 0.82653 | 0.89906 | 0.89912 | 0.93192 | 0.93040 | 0.80985 | 0.72022 |
| Gini 2 | 0.86150 | 0.85856 | 0.92723 | 0.92646 | 0.93192 | 0.93119 | 0.81924 | 0.75471 |
| Gini 3 | 0.85915 | 0.85730 | 0.92957 | 0.92829 | 0.93427 | 0.93374 | 0.81220 | 0.75596 |
| Gini 4 | 0.85446 | 0.85291 | 0.93192 | 0.93145 | 0.93661 | 0.93598 | 0.96244 | 0.94879 |

**Fonte:** Autoria própria (2023).

Os resultados obtidos nas análises dos diferentes conjuntos de variáveis Gini oferecem descobertas cruciais para a aplicação da cardiotocografia e destacam a importância da seleção de variáveis na precisão dos modelos preditivos. Notavelmente, a progressiva inclusão de variáveis nos conjuntos Gini resultou em melhorias consistentes nas métricas de acurácia e F2-*score*, sugerindo que a consideração de um conjunto mais abrangente de características contribui para aprimorar a capacidade preditiva dos modelos.

Os modelos de random forest se destacaram consistentemente, alcançando os melhores resultados em acurácia e F2-*score* em praticamente todos os conjuntos Gini. Isso sugere que a natureza de *ensemble* do random forest, que integra múltiplas árvores de decisão, contribuiu significativamente para a robustez e eficácia preditiva, especialmente quando mais variáveis são consideradas.

A análise dos resultados também destacou a importância de ponderar falsos negativos e falsos positivos, uma vez que o F2-*score* é particularmente sensível a essas métricas em cenários de desequilíbrio de classes, como na cardiotocografia. Nos conjuntos Gini 2, Gini 3 e

Gini 4 observa-se uma melhoria notável no F2-*score*, indicando uma redução na frequência de falsos negativos em relação aos falsos positivos.

Em suma, os resultados desta análise reforçam a necessidade de uma abordagem criteriosa na seleção de variáveis para modelos de cardiotocografia, com ênfase na inclusão de informações relevantes para otimizar a capacidade preditiva e garantir resultados mais precisos e clinicamente úteis. Essas descobertas são valiosas para profissionais de saúde que buscam aprimorar a interpretabilidade e confiabilidade dos modelos utilizados na interpretação de dados cardiotocográficos.

## 4. Discussão

A discussão desta análise destaca a relevância e complexidade associadas à cardiotocografia na predição da saúde fetal. A distribuição dos valores no conjunto de dados original revelou a diversidade e amplitude dos parâmetros cardiotocográficos, reforçando a necessidade de uma abordagem cuidadosa na seleção de variáveis para a construção de modelos preditivos. Diante do conjunto de 22 parâmetros, o uso do MeanDecreaseGini possibilitou uma avaliação sistemática da importância de cada variável, contribuindo para a identificação das características mais importantes na previsão da saúde fetal.

A análise estatística de correlação entre variáveis ofereceu revisões adicionais sobre a inter-relação dos parâmetros e sua associação com o diagnóstico de saúde fetal. A forte correlação entre certos parâmetros, como desacelerações prolongadas, tanto visualmente quanto estatisticamente, destaca a importância desses indicadores no contexto da cardiotocografia. A estratégia de definir um *cutoff* de correlação de 0,10 e criar conjuntos de dados específicos baseados nesse critério proporcionou uma base sólida para a comparação de modelos.

A avaliação da acurácia e F2-*score* em diferentes conjuntos de variáveis Gini revelou tendências consistentes de melhoria de desempenho à medida que mais variáveis foram incorporadas. Os modelos baseados em random forest destacaram-se em praticamente todos os conjuntos, demonstrando a eficácia desse método, especialmente quando todas as variáveis foram consideradas na modelagem.

Ressalta-se a importância da abordagem estratificada na seleção de variáveis para otimizar a precisão dos modelos de predição. A comparação detalhada dos conjuntos Gini evidenciou que a inclusão de mais variáveis não apenas aprimorou as métricas de desempenho, mas também proporcionou uma compreensão mais abrangente dos fatores determinantes na cardiotocografia.

## 5. Conclusão

Esta análise destacou a importância da cardiotocografia na predição da saúde fetal, revelando a complexidade e diversidade dos 22 parâmetros considerados. O uso do MeanDecreaseGini possibilitou uma avaliação sistemática da importância de cada variável, identificando características cruciais para a previsão da saúde fetal. A análise de correlação entre variáveis proporcionou esclarecimentos sobre a inter-relação dos parâmetros, especialmente evidenciando a relevância de indicadores como desacelerações prolongadas. Os modelos baseados em random forest demonstraram eficácia, sugerindo que a inclusão de mais variáveis aprimora consistentemente o desempenho, enfatizando a importância da abordagem estratificada na seleção de variáveis para otimizar a precisão dos modelos.

Esses resultados têm implicações significativas para a prática clínica, fornecendo orientações valiosas para a interpretação de dados cardiotocográficos e contribuindo para avanços na aplicação da inteligência artificial na medicina obstétrica. De fato, a análise não apenas ressalta a relevância das variáveis na cardiotocografia para a predição da saúde fetal, mas também enfatiza a eficácia dos modelos preditivos, na interpretabilidade e confiabilidade clínica. Essa compreensão refinada das variáveis relevantes não só aprimora a precisão diagnóstica, mas também impacta positivamente na tomada de decisões clínicas. Este estudo oferece resultados promissores para o avanço da inteligência artificial na medicina obstétrica, contribuindo para futuras investigações e desenvolvimentos clínicos na área.

### Agradecimentos

Os autores agradecem a colaboração inestimável de Gustavo Mitsuo Suguimoto, Felipe Chen e de Osvaldo Borges de Miranda, estudantes de graduação do Curso Superior de Tecnologia em Informática em Saúde da UNIFESP, nas discussões dos resultados; ao Departamento de Informática em Saúde, EPM, UNIFESP, pelo apoio na condução do estudo.

## Referências

1. Ayres-de-Campos D, et al. FIGO Intrapartum Fetal Monitoring Expert Consensus Panel. FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography. Int J Gynaecol Obstet. 2015 Jan;131(1):13-24. Available from: https://doi.org/10.1016/j.ijgo.2015.06.020. Accessed May 23, 2023.

2. Hon EH, Quilligan EJ. Electronic evaluation of the fetal heart rate. VII. Patterns preceding fetal death, further observations. Am J Obstet Gynecol. 1962;83:1359-1374.
3. Caldeyro-Barcia R, Poseiro JJ. Monitoring of fetal heart rate during labor. Am J Obstet Gynecol. 1962;84:1-12.
4. Murray H, Goldberg J. Malpractice issues in electronic fetal monitoring. Clin Obstet Gynecol. 2002 Dec;45(4):1022-1030. Available from: https://doi.org/10.1097/00003081-200212000-00027. Accessed May 12, 2023.
5. Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. CRC Press; 1984.
6. Breiman L. Random forests. Machine Learning. 2001 Oct;45(1):5-32.
7. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. John Wiley & Sons; 2013.
8. Kelleher JD, Mac Namee B, D'Arcy A. Fundamentals of Machine Learning for Predictive Data Analytics. MIT Press; 2015.
9. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence; 1995. p. 1137-1143.
10. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45(4):427-437.
11. Kim HY, Cho GJ, Kwon HS. Applications of artificial intelligence in obstetrics. Ultrasonography. 2023 Jan;42(1):2-9. doi: 10.14366/usg.22063. Epub 2022 Jul 20. PMID: 36588179; PMCID: PMC9816710.