Classificação da Saúde Fetal por Cardiotocografia: Uma Análise Comparativa de Algoritmos de Machine Learning

Gabriel Alves de Araújo Santos[1], **Maria Clara da Silva Ferreira**[1], **Sabrina Gabriele de Souza Freire Beserra**[1], **Yann Keven Jordão Leão**[1]

[1] Universidade Federal Rural de Pernambuco
Engenharia da Computação

{gabriel.alvesaraujo,mariaclara.ferreira,sabrina.gabriele,yann.kjleao}@ufrpe.br

Abstract. *Cardiotocography (CTG) analysis is important for monitoring fetal well-being; however, its interpretation is challenging and subjective. This paper presents a comparative study of five machine learning models—Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes, Logistic Regression, and Multilayer Perceptron (MLP)—for the automatic classification of fetal health into three classes: Normal, Suspect, and Pathological. Using a dataset containing 2,126 CTG exam records, the models underwent normalization pre-processing and were optimized using Grid Search with stratified 10-fold cross-validation. The results showed that the Decision Tree achieved the best overall performance, reaching an F1-Score of 0.9340 and outperforming the other approaches. Thus, it is observed that rule-based models offer high accuracy and interpretability, serving as promising tools for clinical decision support.*

Resumo. *A análise da Cardiotocografia (CTG) é importante para monitorar o bem-estar fetal, entretanto a sua interpretação é desafiadora e subjetiva. Este trabalho apresenta um estudo comparativo de cinco modelos de aprendizado de máquina — Árvore de Decisão, K-Vizinhos Mais Próximos (KNN), Naive Bayes, Regressão Logística e Multilayer Perceptron (MLP) — para a classificação automática de saúde fetal em três classes: Normal, Suspeito e Patológico. Utilizando uma base de dados com 2126 registros de exames CTG, os modelos foram submetidos a um pré-processamento de normalização e otimizados utilizando o Grid Search com validação cruzada estratificada (10-fold). Os resultados mostraram que a Árvore de Decisão obteve o melhor desempenho global, alcançando um F1-Score de 0.9340, superando as demais abordagens. Com isso, observa-se que modelos baseados em regras oferecem alta precisão e interpretabilidade, sendo ferramentas promissoras para o suporte à decisão clínica.*

1. Introdução

As complicações durante a gravidez representam riscos significativos tanto para a mulher quanto para o desenvolvimento fetal, tornando a identificação precoce dessas complicações imperativa para intervenções que salvam vidas [Salini et al., 2024]. Uma prática convencional entre obstetras é a análise manual de testes de cardiotocografia (CTG), um processo que é frequentemente intensivo em trabalho e sujeito a subjetividade e variabilidade interobservador [Salini et al., 2024].

A Cardiotocografia (CTG) é um exame não invasivo e de baixo custo que monitora o bem-estar fetal, registrando a frequência cardíaca fetal (FHR) e as contrações uterinas (UC) [Salini et al., 2024]. Tradicionalmente, a interpretação da CTG baseia-se na análise de características como padrões de FHR, acelerações e desacelerações, mas essa interpretação manual é complexa e pode levar à perda de sinais sutis de sofrimento fetal [Salini et al., 2024].

Para superar as limitações da análise manual, o desenvolvimento de modelos eficientes de classificação da saúde fetal baseados em aprendizado de máquina (Machine Learning - ML) é crucial para otimizar os recursos médicos e o tempo [Salini et al., 2024]. Os modelos de ML oferecem o potencial para melhorar significativamente a precisão e a eficiência da avaliação da saúde fetal, auxiliando os obstetras na tomada de decisões informadas [Salini et al., 2024].

Este trabalho propõe a utilização e comparação de cinco algoritmos de classificação em um conjunto de dados de CTG. Os algoritmos explorados incluem Árvore de Decisão, Vizinhos Mais Próximos (KNN), Naive Bayes, Regressão Logística e Redes Neurais MLP.

O restante deste artigo está organizado da seguinte forma: a Seção 2 discute trabalhos correlatos na classificação da saúde fetal usando ML. A Seção 3 detalha a metodologia experimental, incluindo a descrição do conjunto de dados, a análise exploratória e as etapas de pré-processamento e modelagem. Por fim, as Seções 4 e 5 apresentam os resultados e as considerações finais, respectivamente.

2. Trabalhos Relacionados

A aplicação de técnicas de aprendizado de máquina (ML) para a classificação da saúde fetal a partir de dados de Cardiotocografia (CTG) tem sido amplamente investigada, visando automatizar e aumentar a precisão do diagnóstico.

O estudo de Salini et al. [Salini et al., 2024] aborda a classificação da saúde fetal utilizando diversos modelos de ML (incluindo Random Forests, Regressão Logística, Árvores de Decisão, Classificadores de Vetores de Suporte, Classificadores por Votação e K-Nearest Neighbors) no conjunto de dados de CTG. Os resultados dessa pesquisa indicaram que os modelos de ML implementados alcançaram uma precisão notável de 93%, superando métodos anteriores [Salini et al., 2024, Salini et al., 2024]. Esse trabalho sugere que a integração de modelos de ML pode otimizar a alocação de recursos médicos e a eficiência do tempo na avaliação fetal [Salini et al., 2024]. Em sua comparação, o algoritmo Random Forest obteve o maior desempenho, com 93% de acurácia, enquanto o KNN e o Gradient Boosting Classifier atingiram 90% de acurácia [Salini et al., 2024].

Outro estudo relevante, realizado por Hoodbhoy et al. [Hoodbhoy et al., 2019], também utilizou o mesmo conjunto de dados de CTG (2126 registros, classificados por obstetras em Normal, Suspeito ou Patológico) para prever o risco fetal [Hoodbhoy et al., 2019, Hoodbhoy et al., 2019]. Os autores aplicaram dez modelos de classificação de ML e utilizaram a técnica de Sobreamostragem de Minorias Sintéticas (SMOTE) para lidar com o desbalanceamento dos dados (70% Normal, 20% Suspeito, 10% Patológico) [Hoodbhoy et al., 2019, Hoodbhoy et al., 2019]. Foi reportado que o modelo baseado em XGBoost obteve a maior precisão (93% de acurácia geral no teste)

para prever um desfecho fetal adverso, superando outros algoritmos testados, como Árvore de Decisão, Random Forest e KNN [**?**, Hoodbhoy et al., 2019]. Os autores ressaltam que, embora o XGBoost tenha alta precisão para o estado patológico (92%), sua precisão para o estado suspeito (73%) foi mais alta que a de outros modelos, sendo crucial ter alta precisão para ambas as classes de risco [Hoodbhoy et al., 2019].

Ambos os trabalhos demonstram a viabilidade da classificação automatizada de CTG e reforçam a importância da escolha e calibração adequadas dos modelos de ML para obter alta acurácia na identificação de fetos de alto risco [Salini et al., 2024, Hoodbhoy et al., 2019]. Tais estudos utilizam um procedimento experimental similar ao proposto neste projeto (comparação de múltiplos classificadores em dados de CTG) [**?**].

3. Metodologia Experimental

3.1. Descrição do Dataset

O conjunto de dados utilizado neste estudo é o *Fetal Health Classification*, que contém 2.126 registros derivados de exames cardiotocográficos (CTG). O dataset possui 21 *features* que representam medidas fisiológicas, como indicadores de batimentos cardíacos fetais (FHR), contrações uterinas, variabilidade cardíaca e variáveis estatísticas baseadas no histograma de FHR (e.g., *baseline value*, *accelerations*, *fetal_movement*, *histogram_width*, *histogram_mode*, etc.).

A variável alvo (*fetal_health*) é categórica e multinível, com três classes: 1 (Normal), 2 (Suspeito) e 3 (Patológico). Inicialmente, a coluna alvo foi convertida de tipo `float64` para `int64`.

3.2. Análise Exploratória de Dados (EDA)

A Análise Exploratória de Dados (EDA) e o pré-processamento inicial foram conduzidos com o objetivo de inspecionar a qualidade dos dados e preparar a base para a modelagem. As etapas de exploração incluíram:

3.2.1. Ausência de Valores Ausentes e Duplicados

O dataset não apresentou valores ausentes, eliminando a necessidade de imputation. No entanto, foram identificados 13 dados duplicados. Como esses dados derivam de exames CTG fisiológicos, as duplicatas não possuíam significado clínico e foram removidas para evitar o enviesamento de certos modelos, como KNN, Naive Bayes e Redes Neurais.

3.2.2. Distribuições e Outliers

Histogramas foram utilizados para visualizar a distribuição das variáveis numéricas, revelando comportamentos estatísticos distintos. Muitas *features*, como *fetal_movement* e *severe_decelerations*, mostraram alta concentração de registros próximos de zero.

A análise de *outliers* foi realizada visualmente por meio de *boxplots* e quantitativamente pela técnica do Intervalo Interquartil (IQR). Observou-se a presença de outliers em várias *features*. Decidiu-se manter esses valores extremos, pois eles não são considerados ruído, mas sim representações de eventos clínicos raros, como movimentos fetais

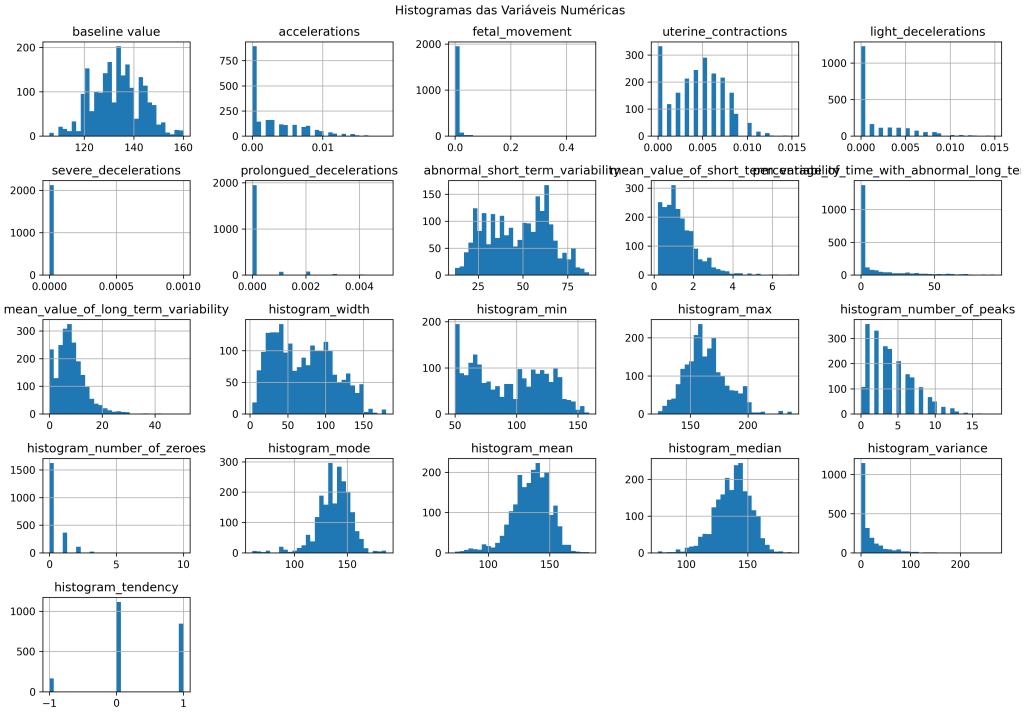

Figura 1. Distribuição de todas as variáveis preditoras no dataset (Histogramas).

intensos ou alterações abruptas no padrão de variabilidade cardíaca, sendo essenciais para uma modelagem fiel.

3.2.3. Distribuição da Variável Alvo e Correlação

A distribuição da variável alvo (*fetal_health*) foi examinada por meio de um gráfico de barras, confirmando um forte desbalanceamento com a predominância da classe 1 (Normal). Esse desbalanceamento é comum em dados clínicos e será considerado na fase de modelagem.

A matriz de correlação (Figura 4) revelou um padrão complexo. Embora a maioria das correlações entre *features* seja baixa (variando entre ≈ -0.25 e $+0.25$), grupos de variáveis derivadas do histograma (*histogram_mean, histogram_median* e *histogram_mode*) mostraram correlação quase perfeita (≈ 1). Esse achado sugere que, embora o restante das variáveis seja amplamente independente (o que favorece algoritmos como Naive Bayes), há redundância dentro do grupo de variáveis do histograma, o que pode influenciar modelos sensíveis à multicolinearidade, como a Regressão Logística e Redes Neurais.

3.3. Pré-processamento

A etapa final de pré-processamento envolveu a normalização dos dados. Optou-se pelo **StandardScaler** para transformar cada *feature* para média 0 e desvio padrão 1.

Essa escolha é adequada para os modelos de ML que serão comparados, como KNN (baseado em distância), Regressão Logística (otimização mais estável) e Naive

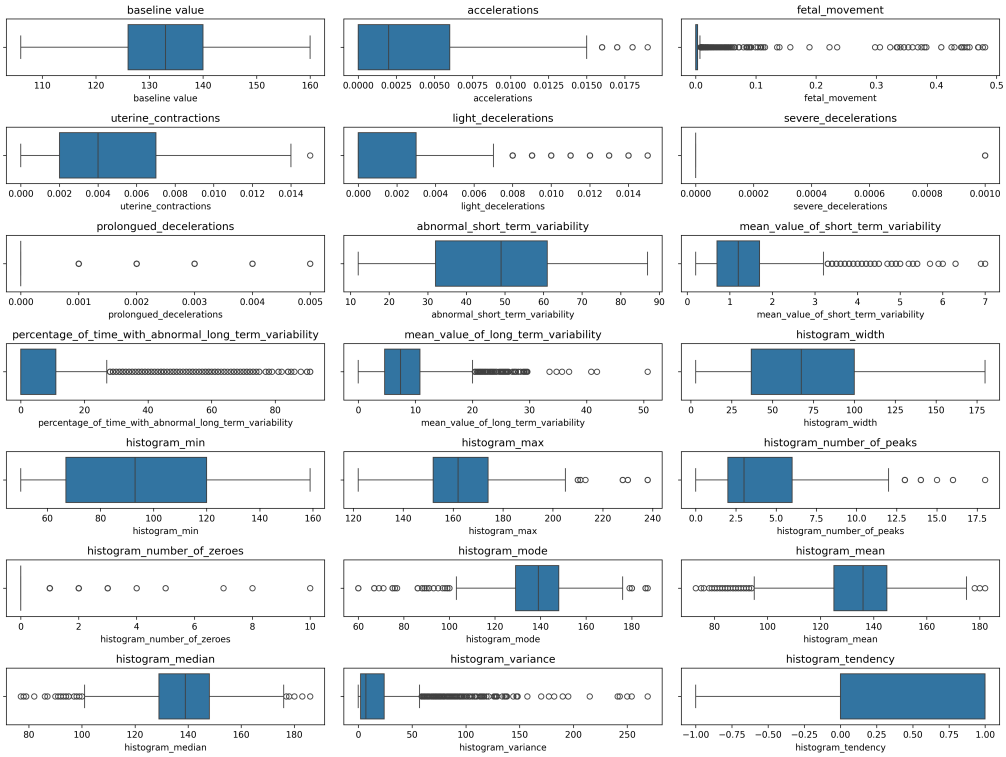

Figura 2. Boxplots das features, evidenciando a presença de outliers.

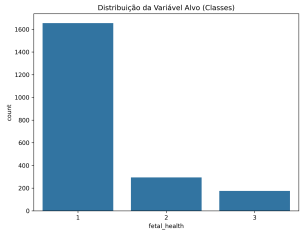

Figura 3. Distribuição das classes da variável alvo (*fetal_health*).

Bayes Gaussiano (que assume distribuição normal), além de ser mais robusta em cenários com a presença de *outliers* quando comparada ao Min-Max Scaling. O *scaler* é ajustado apenas nos dados de treino, dentro do procedimento de validação cruzada, para evitar *data leakage*, e posteriormente é salvo para uso futuro na fase de modelagem e avaliação.

3.4. Modelos de Classificação

Nesta etapa do trabalho, descrevem-se os algoritmos de classificação selecionados para a primeira fase de modelagem: Árvore de Decisão, K-Vizinhos Mais Próximos (KNN), Gaussian Naive Bayes, Regressão Logística Multinomial e Multilayer Perceptron (MLP).

3.4.1. Árvore de Decisão (Decision Tree)

A Árvore de Decisão é um modelo que representa uma função que recebe como entrada uma lista de valores de atributos para uma única decisão [Russell and Norvig, 2021]. Ela

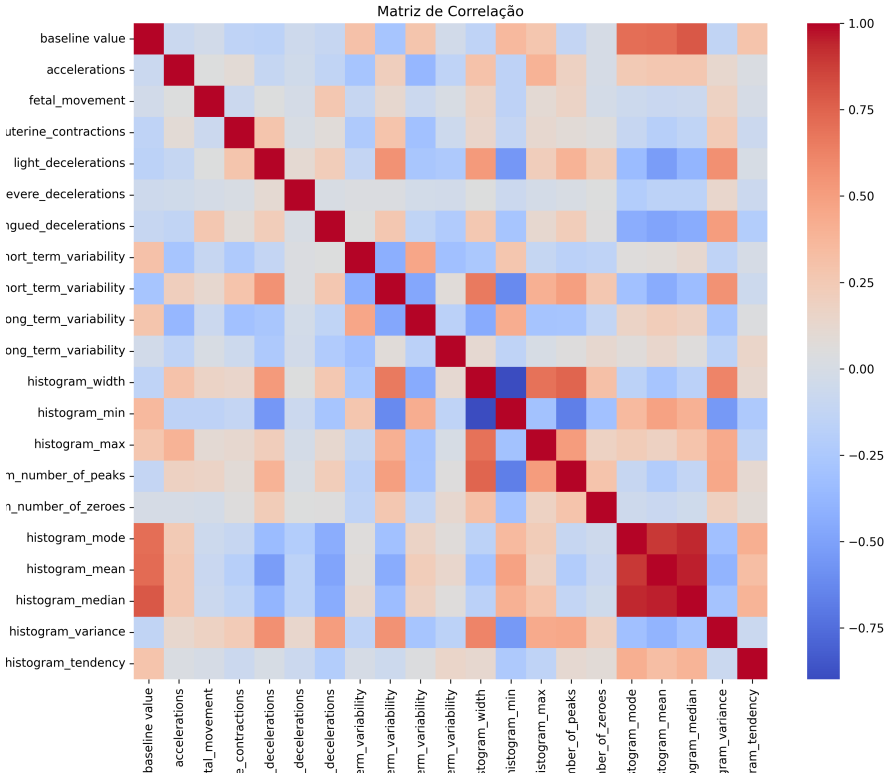

Figura 4. Matriz de Correlação entre as features.

começa da raiz seguindo o ramo até que uma folha seja alcançada. No presente trabalho, foi utilizado um algoritmo de predição chamado CART (*Classification And Regression Tree*) implementado de forma padrão pela biblioteca *Scikit-Learn*.

Com o intuito de garantir o desempenho do modelo e evitar o *overfitting*, houve uma variação dos seguintes hiperparâmetros durante a experimentação:

- **Critério de Divisão:** O objetivo é verificar qual métrica consegue dividir melhor as classes de forma mais eficiente para o conjunto de dados. Para isso foi comparado o índice Gini com a Entropia.
- **Profundidade Máxima:** Para impedir que o modelo apenas memorize ruídos do conjunto de treino foram testadas as profundidades (10, 20, 30) e a ausência de limite.
- **Mínimo de Amostras para Divisão:** Variou-se a quantidade mínima de exemplos para a divisão de um nó interno (2, 10, 20). Os valores maiores tendem a forçar o modelo para criar regras mais gerais.

Melhor Configuração: Após a otimização via *Grid Search*, a melhor combinação de hiperparâmetros obtida foi: critério **Entropia**, profundidade máxima **sem limite (None)** e mínimo de amostras para divisão igual a **20**.

3.4.2. K-Vizinhos Mais Próximos (KNN)

O algoritmo KNN consiste na predição de uma classe com base na votação majoritária de exemplos similares armazenados em um conjunto de treinamento. Ele utiliza seus k vizinhos mais próximos com base em instâncias (*lazy learning*).

Com base nisso a escolha de hiperparâmetros se torna determinante. Foram investigadas as seguintes configurações:

- **Número de Vizinhos:** Foram testados valores ímpares para que não houvesse empate na votação. Valores menores foram avaliados para capturar estruturas locais complexas, enquanto valores maiores visaram verificar o efeito da suavização nas fronteiras de decisão.
- **Ponderação:** Avaliou-se o desempenho entre a votação uniforme e a ponderação pela distância, visto que esta atribui maior relevância aos vizinhos mais próximos em comparação com aqueles que estão no limite da fronteira k.
- **Métrica de Distância:** Foram comparadas as distâncias Euclidiana e Manhattan para averiguar qual medida se adapta melhor à distribuição espacial dos atributos pré-processados.

Melhor Configuração: Dentre as combinações avaliadas, o modelo que apresentou melhor capacidade de generalização utilizou $k = \mathbf{7}$ vizinhos, ponderação **pela distância (distance)** e a métrica de distância **Manhattan**.

3.4.3. Gaussian Naive Bayes

O classificador Naive Bayes fundamenta-se na aplicação direta do Teorema de Bayes, assumindo a premissa de que as características (*features*) são condicionalmente independentes entre si dado o valor da classe alvo. Embora essa suposição de "ingenuidade" raramente se sustente em dados biológicos complexos — onde correlações fisiológicas são esperadas —, o modelo é amplamente utilizado devido à sua eficiência computacional e robustez em altas dimensões.

Dada a natureza contínua das variáveis do exame de CTG, adotou-se a variante Gaussian Naive Bayes, que assume que a verossimilhança das características segue uma distribuição normal (Gaussiana).

Parâmetros Utilizados: O principal ajuste realizado foi no parâmetro de suavização da variância (*var_smoothing*). Este hiperparâmetro adiciona uma pequena constante à variância de cada característica para garantir a estabilidade numérica e evitar divisões por zero em probabilidades. Na busca em grade, variou-se este valor entre 10^{-9} e 10^{-6}, permitindo ao modelo lidar melhor com irregularidades nas distribuições dos dados.

Melhor Configuração: O valor ideal encontrado para o parâmetro de suavização da variância (*var_smoothing*) que maximizou o desempenho do modelo foi de 10^{-9}.

3.4.4. Regressão Logística Multinomial

Diferente da regressão linear, a Regressão Logística modela a probabilidade de uma instância pertencer a uma classe específica utilizando a função logística (sigmoide). Para o

cenário de saúde fetal, que possui três classes (Normal, Suspeito, Patológico), o algoritmo foi generalizado para o caso multinomial, utilizando a função Softmax para distribuir as probabilidades entre as categorias.

A presença de multicolinearidade no *dataset* (observada entre as variáveis de histograma) pode desestabilizar os coeficientes do modelo. Para mitigar isso, a regularização é indispensável.

- **Tipo de Penalidade (penalty):** Foram comparadas a regularização L1 (Lasso), que promove a seleção de atributos ao zerar coeficientes de variáveis redundantes, e a L2 (Ridge), que penaliza pesos elevados para reduzir a variância do modelo sem eliminar variáveis.
- **Inverso da Regularização (C):** Este parâmetro controla a intensidade da penalidade. Foram testados os valores 0.1, 1.0 e 10.0, onde valores menores indicam uma regularização mais forte (maior restrição aos coeficientes) para combater o *overfitting*.

Melhor Configuração: O melhor ajuste do modelo utilizou o tipo de penalidade **L2 (Ridge)** combinada com o inverso da regularização $C = \mathbf{10.0}$.

3.4.5. Multilayer Perceptron (MLP)

As Redes Neurais Artificiais do tipo Multilayer Perceptron (MLP) são modelos inspirados no funcionamento biológico, compostos por camadas de neurônios interconectados que aprendem através do algoritmo de *Backpropagation*. Sua principal vantagem reside na capacidade de modelar fronteiras de decisão não-lineares complexas, o que é ideal para capturar padrões sutis de sofrimento fetal que modelos lineares podem perder.

O funcionamento do MLP depende criticamente de sua topologia e da forma como os neurônios processam os sinais de entrada.

- **Arquitetura da Rede (hidden_layer_sizes):** Testou-se a profundidade da rede comparando uma estrutura mais rasa, com uma única camada oculta de 100 neurônios $(100,)$, contra uma estrutura mais profunda composta por duas camadas de 50 neurônios $(50, 50)$.
- **Função de Ativação (activation):** Para introduzir a não-linearidade necessária para separar as classes de risco, alternou-se entre a função Tangente Hiperbólica (*tanh*) e a Unidade Linear Retificada (*relu*).
- **Termo de Regularização L2 (alpha):** Para prevenir que a rede "memorize" o ruído dos dados de treino, aplicou-se uma penalidade nos pesos (*alpha*) com valores de 0.0001, 0.001 e 0.01.

Melhor Configuração: A arquitetura neural mais eficiente para este problema foi composta por **uma camada oculta de 100 neurônios (100,)**, utilizando a função de ativação **tanh** (Tangente Hiperbólica) e termo de regularização alpha de $\mathbf{0.01}$.

4. Resultados e Discussão

Para avaliar a eficácia dos modelos propostos na classificação da saúde fetal, foram realizados experimentos utilizando validação cruzada(*k-fold cross-validation*), visando garantir a robustez estatística das métricas obtidas. Como discutido na metodologia, o conjunto

de dados apresenta desbalanceamento de classes (predominância da classe "Normal"), o que torna a Acurácia uma métrica insuficiente isoladamente. Portanto, a análise priorizou o F1-Score (*Weighted*) e o Recall, dado que a não identificação de um feto em estado patológico (Falso Negativo) acarreta riscos severos à vida.

A Tabela abaixo resume o desempenho dos cinco classificadores otimizados via *Grid Search*.

Tabela 1. Desempenho dos classificadores otimizados via Grid Search.

Modelo	Configuração Principal	Acurácia	F1-Score	Recall
Decision Tree	Entropy, Depth=30	93.42%	0.9340	0.9342
KNN	Manhattan, k=7	91.19%	0.9077	0.9119
MLP (Neural Net)	Tanh, (100,)	89.53%	0.8911	0.8953
Regressão Logística	L2, C=10.0	88.11%	0.8724	0.8811
Naive Bayes	Var Smoothing=10^{-9}	81.35%	0.8299	0.8135

Observa-se que a Árvore de Decisão obteve o melhor desempenho global, alcançando um F1-Score de 0.934. Este resultado é consistente com a natureza dos dados de cardiotocografia (CTG), que frequentemente envolvem limites de decisão rígidos (ex: "se batimento < 110 bpm, então risco"), favorecendo modelos baseados em regras de particionamento em detrimento de modelos lineares ou probabilísticos simples.

O desempenho da Árvore de Decisão neste estudo alinha-se aos resultados do estado da arte mencionados nos trabalhos relacionados. Salini et al. [**?**] reportaram acurácia de 93% utilizando *Random Forest*, e Hoodbhoy et al. [**?**] obtiveram desempenho similar com XGBoost. O fato de uma única Árvore de Decisão (com poda otimizada) atingir resultados competitivos com métodos de *Ensemble* mais complexos sugere que as *features* extraídas do exame CTG possuem alto poder discriminativo quando modeladas de forma não-linear.

Em contrapartida, modelos lineares como a Regressão Logística (88.1%) e probabilísticos como Naive Bayes (81.3%) apresentaram desempenho inferior. A baixa performance do Naive Bayes pode ser explicada pela forte correlação identificada entre as variáveis de histograma na análise exploratória, violando a premissa de independência entre as variáveis preditoras exigida pelo algoritmo.

5. Considerações Finais

Este estudo demonstrou que a aplicação de aprendizado de máquina na análise de cardiotocografia (CTG) é uma estratégia eficaz para auxiliar no diagnóstico precoce de riscos à saúde fetal. Após comparar cinco algoritmos distintos, a Árvore de Decisão consolidou-se como o modelo mais competente para esta tarefa, alcançando um F1-Score de 93.4% e superando abordagens como Redes Neurais e Naive Bayes.

A superioridade da Árvore de Decisão neste cenário não foi acidental. A análise dos dados revelou que exames de CTG possuem distribuições complexas e eventos clínicos extremos (como desacelerações abruptas) que atuam como *outliers*. Enquanto modelos baseados em estatística pura ou distância tiveram dificuldade com essas características, a estrutura hierárquica da Árvore conseguiu isolar e interpretar esses padrões

críticos com alta precisão. Além da performance, este modelo oferece a vantagem crucial da interpretabilidade, permitindo que a equipe médica visualize as regras lógicas de decisão, o que aumenta a confiança no diagnóstico assistido por computador.

Em contrapartida, modelos que assumem a independência entre variáveis, como o Naive Bayes, apresentaram desempenho inferior, confirmando a existência de correlações complexas entre os parâmetros fisiológicos do feto que não podem ser ignoradas.

Para trabalhos futuros, os resultados aqui obtidos sugerem um caminho promissor para a integração destes algoritmos em sistemas de monitoramento hospitalar em tempo real, atuando como alertas automáticos para a equipe de obstetrícia. Além disso, investigações futuras podem focar na otimização de modelos para dispositivos portáteis e de baixo custo, democratizando o acesso a diagnósticos de alta precisão em regiões com escassez de especialistas.

Referências

Hoodbhoy, Z., Noman, M., Shafique, A., Nasim, A., Chowdhury, D., and Hasan, B. (2019). Use of machine learning algorithms for prediction of fetal risk using cardiotocographic data. *International Journal of Applied and Basic Medical Research*, 9:226–230.

Russell, S. J. and Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*. Pearson Education Limited, Harlow, 4th global edition edition.

Salini, Y., Mohanty, S. N., Ramesh, J. V. N., Yang, M., and Chalapathi, M. M. V. (2024). Cardiotocography data analysis for fetal health classification using machine learning models. *IEEE Access*, 12:3364755.