INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS CORRENTE
CURSO ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

PEDRO HENRIQUE VOGADO MAIA

# RELATÓRIO DE IMPLEMENTAÇÃO DOS CÓDIGOS (KNN, NAIVE BAYES, ÁRVORE DE DECISÃO, PERCEPTRON, MLP)

CORRENTE

2024

# Sumário

# 1 INTRODUÇÃO

Neste projeto, exploramos a previsão de doenças cardíacas utilizando algoritmos de aprendizado de máquina, com o objetivo de identificar possíveis problemas de saúde antes que eles se manifestem. Utilizamos cinco métodos: K-Nearest Neighbors (KNN), Naive Bayes, Árvore de Decisão, Perceptron Simples, e Redes Neurais Multicamadas (MLP), comparando seus desempenhos para entender suas vantagens e limitações. Além de buscar um modelo preditivo eficiente, nosso trabalho visa também tornar a complexidade do aprendizado de máquina acessível, demonstrando como a inteligência artificial pode ser uma aliada na prevenção de doenças e na melhoria da qualidade de vida.

---

# 2 DESCRIÇÃO DO PROBLEMA

As doenças cardíacas são responsáveis por um grande número de mortes ao redor do mundo, tornando a sua detecção precoce essencial para salvar vidas. A complexidade dessa detecção reside na necessidade de analisar uma variedade de fatores de risco, como idade, pressão arterial, colesterol, e hábitos de vida, que podem interagir de maneiras complexas. O desafio é desenvolver modelos preditivos que possam, com precisão, estimar a probabilidade de uma pessoa desenvolver uma doença cardíaca, identificando padrões nos dados que não são imediatamente aparentes. Para enfrentar esse desafio, foram utilizados diferentes algoritmos de aprendizado de máquina, comparando suas habilidades em prever doenças cardíacas de forma eficiente e confiável, ajudando a determinar qual abordagem oferece os melhores resultados para essa tarefa crítica.

# 3 ALGORITMOS UTILIZADOS

## K-Nearest Neighbors (KNN)

O algoritmo K-Nearest Neighbors, ou KNN, é uma abordagem intuitiva para a classificação e regressão de dados. Imagine que você está em uma festa e quer saber quais pessoas provavelmente compartilham os mesmos interesses que você. A ideia do KNN é simples: você olha para as pessoas ao seu redor (seus "vizinhos") e observa o que elas gostam. Se a maioria dessas pessoas gostar de algo, há uma boa chance de que você também vá gostar. No contexto do aprendizado de máquina, o KNN funciona de maneira semelhante. Quando um novo dado é apresentado ao algoritmo, ele procura os "k" dados mais próximos no conjunto de dados de treinamento – ou seja, os dados que são mais semelhantes ao novo dado, com base em uma métrica de distância, como a distância euclidiana. A classe ou valor mais comum entre esses vizinhos próximos é então atribuído ao novo dado. Por exemplo, se a maioria dos vizinhos próximos a um paciente tiver uma doença cardíaca, o KNN pode prever que esse novo paciente também está em risco. Uma das grandes vantagens do KNN é sua simplicidade e facilidade de implementação. No entanto, essa simplicidade vem com alguns desafios, como a necessidade de armazenar todos os dados de treinamento, o que pode se tornar um problema em conjuntos de dados grandes. Além disso, o desempenho do KNN pode ser sensível ao valor de "k" escolhido, que precisa ser cuidadosamente ajustado para evitar classificações incorretas. A lista de referências, ilustrada a seguir, é a relação de todas as obras citadas na pesquisa, organizada em ordem alfabética de entrada (autores pessoais, entidades ou títulos). As referências devem ser indicadas em espaço simples (1,0), alinhadas à margem esquerda do texto e separadas entre si por uma linha em branco de espaço simples.

## Naive Bayes

O Naive Bayes é um algoritmo baseado na teoria da probabilidade, mais especificamente no Teorema de Bayes, que permite calcular a probabilidade de um evento ocorrer, dado o conhecimento prévio de outros eventos relacionados. Imagine que você quer prever se vai chover hoje. Você pode usar informações como a umidade do ar, a temperatura e a presença de nuvens para calcular essa probabilidade. O Naive Bayes faz algo parecido, mas com uma suposição muito importante: ele assume que todas as características que você está analisando são independentes umas das outras. Embora essa suposição de independência raramente seja verdadeira no mundo real, o Naive Bayes ainda funciona surpreendentemente bem em muitos casos práticos. No contexto da previsão de doenças cardíacas, o algoritmo calcularia a probabilidade de um paciente ter a doença com base em características como idade, níveis de colesterol e pressão arterial. Mesmo que essas características possam estar inter-relacionadas, o Naive Bayes assume que cada uma contribui de forma independente para a probabilidade final. Uma das grandes vantagens do Naive Bayes é que ele é rápido e eficiente, especialmente com conjuntos de dados grandes. Ele também lida bem com classes que tenham múltiplos atributos, tornando-o uma escolha popular para problemas de classificação. No entanto, sua principal limitação é a suposição de independência, que pode levar a previsões imprecisas quando as características são, na verdade, interdependentes.

## Árvore de Decisão

Imagine que você está planejando uma viagem. Você começa escolhendo o destino, depois pensa em como vai chegar lá e, por fim, decide onde vai se hospedar. Cada escolha que você faz abre novas possibilidades, e suas decisões vão moldando o resultado final da viagem. Uma Árvore de Decisão funciona de maneira parecida no contexto do aprendizado de máquina. Ela organiza as decisões de forma hierárquica, dividindo os dados em grupos com base nas características mais importantes para o problema em questão. No caso da previsão de doenças cardíacas, por exemplo, a árvore pode primeiro separar os pacientes pela idade, depois pela pressão arterial, e assim por diante, até que uma decisão final seja tomada sobre a probabilidade de o paciente ter ou não a doença. O que torna as Árvores de Decisão tão úteis é a clareza com que mostram como as decisões estão sendo tomadas. Você pode seguir cada ramificação da árvore e entender exatamente como o modelo chegou a uma conclusão. No entanto, como em qualquer ferramenta, existem desafios. Se a árvore for muito complexa, ela pode se ajustar demais aos dados que você usou para treiná-la, o que chamamos de "overfitting". Isso pode fazer com que o modelo não funcione tão bem quando enfrenta novos dados, diferentes daqueles que usou no treinamento.

## Perceptron Simples

O Perceptron Simples é um dos algoritmos mais antigos e fundamentais no campo do aprendizado de máquina, sendo a base para redes neurais mais complexas. Ele funciona de maneira similar a um neurônio biológico, recebendo entradas (sinais), processando-as, e gerando uma saída (resposta). Imagine que você está tentando decidir se vai para uma festa baseado em várias condições, como o clima, sua disposição, e a presença de amigos. Cada uma dessas condições tem um peso na sua decisão final, e o Perceptron Simples faz algo parecido. No contexto da previsão de doenças cardíacas, o Perceptron Simples recebe várias entradas, como idade, pressão arterial e colesterol, e as combina para fazer uma previsão. Se a combinação das entradas ultrapassar um certo limiar, o Perceptron “dispara” e classifica o dado em uma das duas classes: presença ou ausência de doença cardíaca. Uma das vantagens

do Perceptron Simples é a sua simplicidade e eficiência para resolver problemas de classificação linear. No entanto, ele tem limitações significativas, especialmente em problemas que não podem ser resolvidos com uma simples linha de separação entre as classes. É aqui que redes neurais mais complexas, como as Redes Neurais Multicamadas (MLP), entram em cena.

## Redes Neurais Multicamadas (MLP)

As Redes Neurais Multicamadas (MLP) representam uma evolução significativa em relação ao Perceptron Simples. Enquanto o Perceptron Simples pode ser comparado a um único neurônio, uma MLP é como um cérebro completo, com múltiplas camadas de neurônios que podem processar informações de forma muito mais complexa. Cada camada em uma MLP transforma as entradas antes de passá-las para a próxima camada, permitindo ao modelo capturar padrões muito mais sofisticados nos dados. Imagine que você está tentando identificar um objeto em uma imagem. Uma rede neural multicamada não apenas olha para as cores e formas, mas também combina essas informações em níveis cada vez mais abstratos, permitindo a identificação precisa do objeto. No caso da previsão de doenças cardíacas, a MLP pode combinar múltiplas características do paciente de maneiras complexas para prever com maior precisão se o paciente está em risco. A MLP é poderosa porque pode resolver problemas não-lineares e capturar padrões muito complexos nos dados. No entanto, essa complexidade vem com um custo: as MLPs exigem mais dados e mais poder computacional para treinar, e são mais propensas a superajuste (overfitting) se não forem corretamente regularizadas.

# 4 RESULTADOS ALCANÇADOS

## Resultados do K-Nearest Neighbors (KNN)

O modelo K-Nearest Neighbors (KNN) foi testado para prever a presença de doenças cardíacas, e agora vamos explorar as principais métricas para entender como ele se desempenhou nessa tarefa.

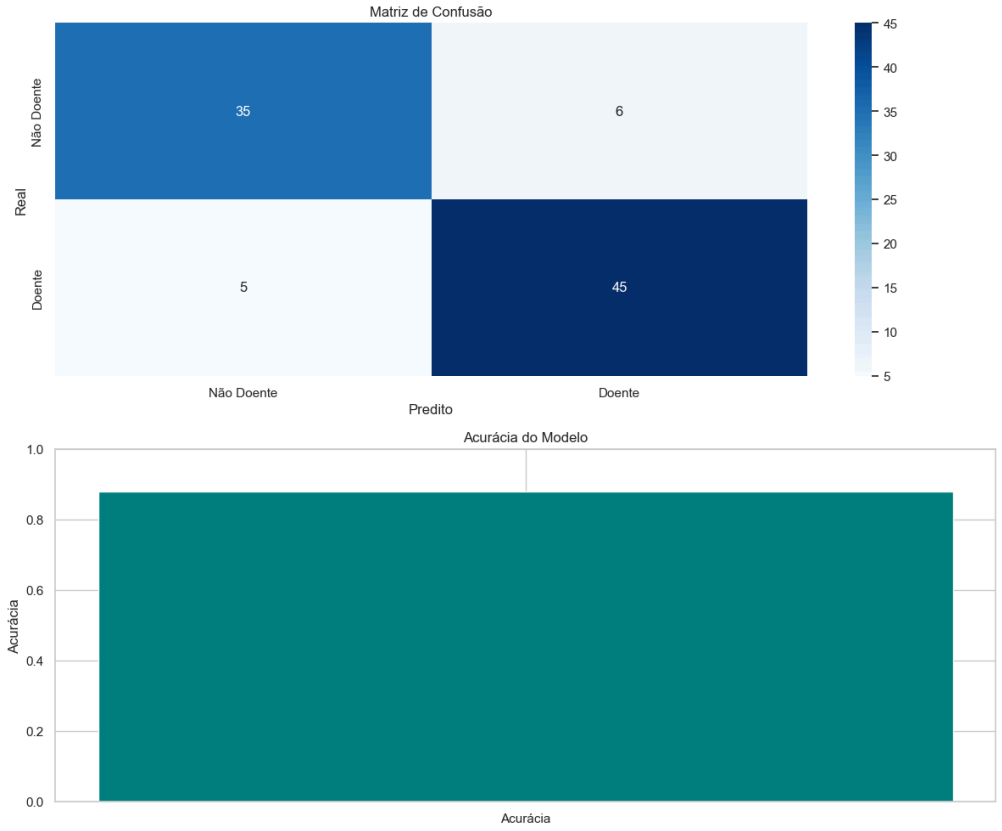
**Precisão:** Para a Classe 0 (não doente), o modelo teve uma precisão de 0.88, e para a Classe 1 (doente), também foi de 0.88. Isso significa que o KNN foi muito bom em prever corretamente se uma pessoa tinha ou não a doença, mantendo um alto nível de acerto em ambas as classes.

**Revocação:** A revocação foi de 0.85 para a Classe 0 e 0.90 para a Classe 1. Esses números indicam que o modelo conseguiu identificar bem os pacientes que realmente tinham a doença e também foi eficaz em detectar aqueles que não tinham. Em outras palavras, o KNN fez um ótimo trabalho encontrando tanto os casos positivos quanto os negativos.

**F1-Score:** O F1-Score combina precisão e revocação para oferecer uma visão equilibrada do desempenho do modelo. Com 0.86 para a Classe 0 e 0.89 para a Classe 1, o KNN mostrou um desempenho sólido, refletindo uma boa harmonia entre a identificação correta dos casos e a minimização de erros.

**Acurácia Geral:** O modelo alcançou uma acurácia de 0.88, o que significa que acertou 88% das previsões feitas. Esse é um excelente resultado, demonstrando que o KNN foi altamente preciso em geral.

Além disso, as médias macro e ponderadas de todas as métricas foram de 0.88, indicando que o modelo manteve um desempenho consistente e equilibrado em todas as classes.O gráfico a seguir ilustra o desempenho do modelo:




## Resultados do Naive Bayes

O gráfico a seguir apresenta o desempenho do modelo Naive Bayes na previsão de doenças cardíacas. Vamos analisar as principais métricas para compreender como o modelo se saiu nessa tarefa.

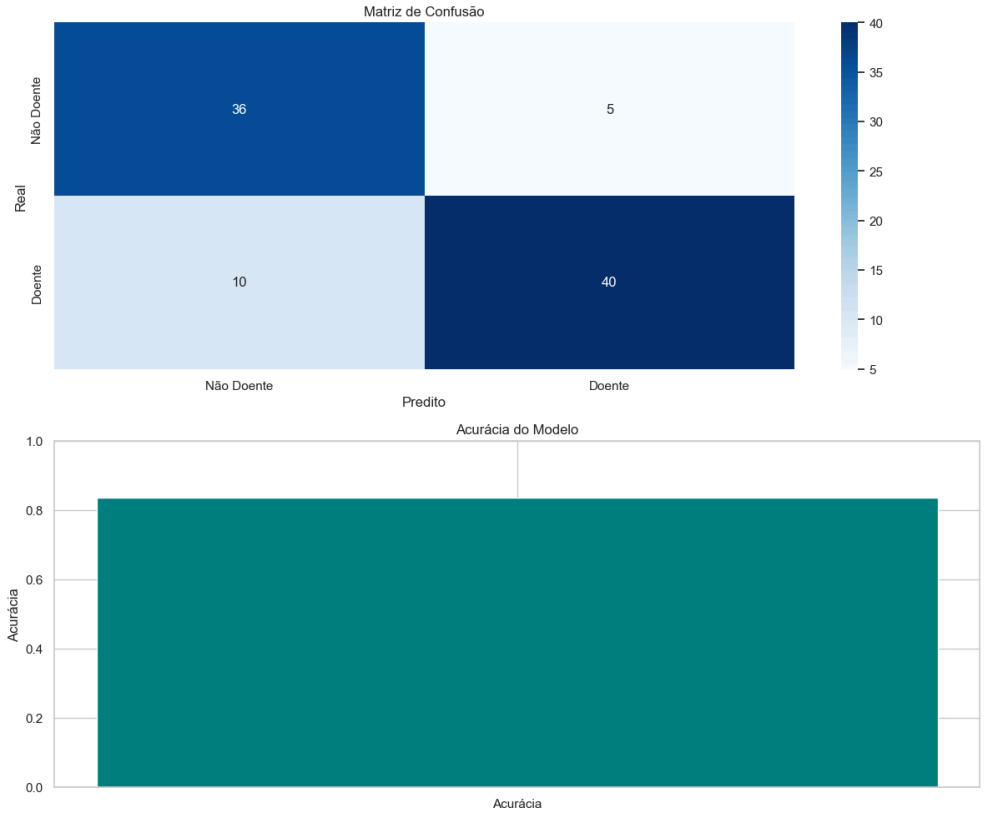
**Precisão:** Para a Classe 0 (não doente), a precisão foi de 0.78, e para a Classe 1 (doente), foi de 0.89. Isso indica que o Naive Bayes teve uma boa taxa de acerto ao prever se um paciente tinha a doença ou não, sendo especialmente eficaz em identificar corretamente a presença da doença.

**Revocação:** A revocação para a Classe 0 foi de 0.88 e para a Classe 1 foi de 0.80. Esses valores mostram que o modelo foi muito eficiente em identificar pacientes que realmente não tinham a doença, com uma boa taxa de detecção também para os pacientes com a doença.

**F1-Score:** O F1-Score, que combina precisão e revocação, foi de 0.83 para a Classe 0 e 0.84 para a Classe 1. Esses resultados refletem um desempenho equilibrado, com o Naive Bayes conseguindo um bom equilíbrio entre identificar corretamente os casos e minimizar erros.

**Acurácia Geral:** O modelo alcançou uma acurácia de 0.84, o que significa que ele fez previsões corretas em 84% dos casos. Esse resultado é bastante sólido, demonstrando que o Naive Bayes foi eficaz na classificação geral.

As médias macro e ponderadas das métricas também foram de 0.84, indicando que o modelo manteve um desempenho consistente e equilibrado ao longo das diferentes classes.




## Resultados da Árvore de Decisão

O gráfico a seguir mostra o desempenho do modelo de Árvore de Decisão na tarefa de prever doenças cardíacas. Vamos analisar as principais métricas para entender como o modelo se comportou.

**Precisão:** Para a Classe 0 (não doente), a precisão foi de 0.68, e para a Classe 1 (doente), foi de 0.80. Isso significa que o modelo foi um pouco menos preciso ao prever a ausência da doença, mas teve uma taxa de acerto relativamente boa ao identificar a presença da doença.

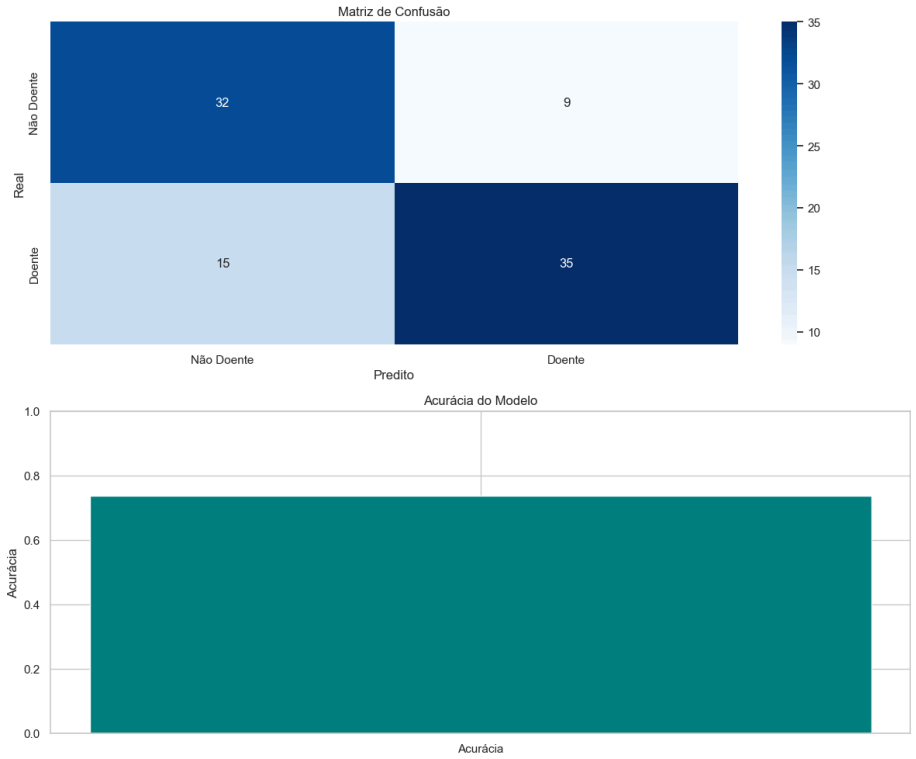
**Revocação:** A revocação foi de 0.78 para a Classe 0 e de 0.70 para a Classe 1. Esses números indicam que a Árvore de Decisão foi bastante eficaz em identificar pacientes que realmente não tinham a doença, embora sua capacidade de identificar todos os pacientes com a doença pudesse ser melhorada.

**F1-Score:** O F1-Score, que equilibra precisão e revocação, foi de 0.73 para a Classe 0 e de 0.74 para a Classe 1. Isso mostra que, apesar de alguns desafios na precisão e na revocação, a

Árvore de Decisão conseguiu manter um desempenho relativamente equilibrado para ambas as classes.

**Acurácia Geral:** O modelo obteve uma acurácia de 0.74, o que significa que acertou 74% das previsões feitas. Esse resultado é sólido, indicando que a Árvore de Decisão é uma ferramenta útil para a tarefa, apesar de algumas limitações em termos de precisão e revocação.

As médias macro e ponderadas das métricas foram todas de 0.74, sugerindo que o modelo manteve um desempenho consistente e equilibrado em todas as métricas, mesmo com algumas variações entre as classes.




**Desempenho do Perceptron Multicamadas (MLP)**

O modelo de Perceptron Multicamadas (MLP) foi avaliado com base em várias métricas para entender sua eficácia na previsão de doenças cardíacas. Aqui estão os principais resultados obtidos:

**1. Precisão (Precision):** A precisão do MLP foi de 0.72 para a classe de pacientes sem a doença (Classe 0) e de 0.82 para a classe de pacientes com a doença (Classe 1). Isso significa que, quando o modelo previu que um paciente não tinha a doença, ele estava correto 72% das vezes. Em contraste, quando o modelo indicou que um paciente tinha a doença, ele acertou 82% das vezes. Esses números refletem a capacidade do MLP em identificar corretamente pacientes com e sem a doença, com uma leve vantagem na identificação dos casos positivos.

**2. Revocação (Recall):** A revocação, que mede a capacidade do modelo de identificar todos os casos positivos reais, foi de 0.80 para a Classe 0 e 0.74 para a Classe 1. Isso indica que o MLP conseguiu identificar 80% dos pacientes que realmente não têm a doença e 74% dos pacientes que realmente têm a doença. Esses valores mostram que o modelo é bastante eficaz em detectar pacientes com a condição, embora haja espaço para melhorar a detecção de alguns casos.

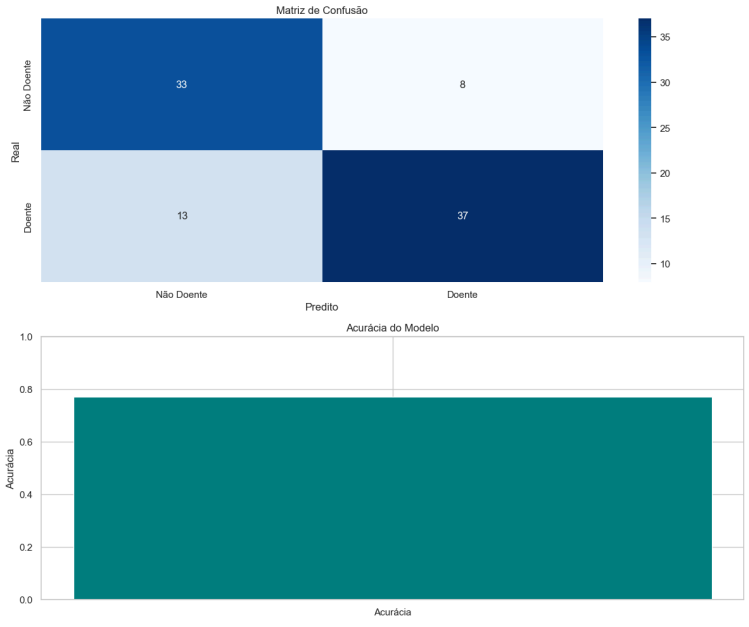
**3. F1-Score:** O F1-Score, que combina precisão e revocação em uma única métrica, foi de 0.76 para a Classe 0 e 0.78 para a Classe 1. O F1-Score de 0.76 para a Classe 0 indica um bom equilíbrio entre identificar corretamente os pacientes sem a doença e evitar classificações incorretas. O F1-Score de 0.78 para a Classe 1, por sua vez, mostra um equilíbrio ligeiramente melhor para a detecção de pacientes com a doença, refletindo um desempenho sólido na identificação de casos positivos.

**4. Acurácia Geral:** O modelo alcançou uma acurácia geral de 0.77, o que significa que ele fez previsões corretas em 77% das vezes. Esse resultado indica que o MLP é eficaz no geral, com um bom desempenho na previsão tanto de pacientes com a doença quanto daqueles sem.

**5. Médias Macro e Ponderadas:** As médias macro e ponderadas das métricas foram todas de 0.77. A média macro considera cada classe igualmente, enquanto a média ponderada leva em conta o número de amostras em cada classe. Ambas as médias sugerem que o modelo tem um desempenho consistente e equilibrado ao longo das diferentes métricas, evidenciando uma abordagem robusta para a tarefa de previsão.

Em resumo, o Perceptron Multicamadas demonstrou ser uma ferramenta eficaz para prever a presença de doenças cardíacas, com um desempenho sólido em termos de precisão, revocação e acurácia. Embora haja áreas para melhorias, especialmente na detecção de alguns casos positivos, o modelo oferece uma boa base para avaliações e pode ser uma peça-chave na análise preditiva para a saúde cardíaca.

O gráfico a seguir ilustra o desempenho do modelo de Perceptron Multicamadas (MLP) na previsão de doenças cardíacas. As métricas destacadas incluem precisão, revocação e F1-Score para cada classe, além da acurácia geral do modelo.



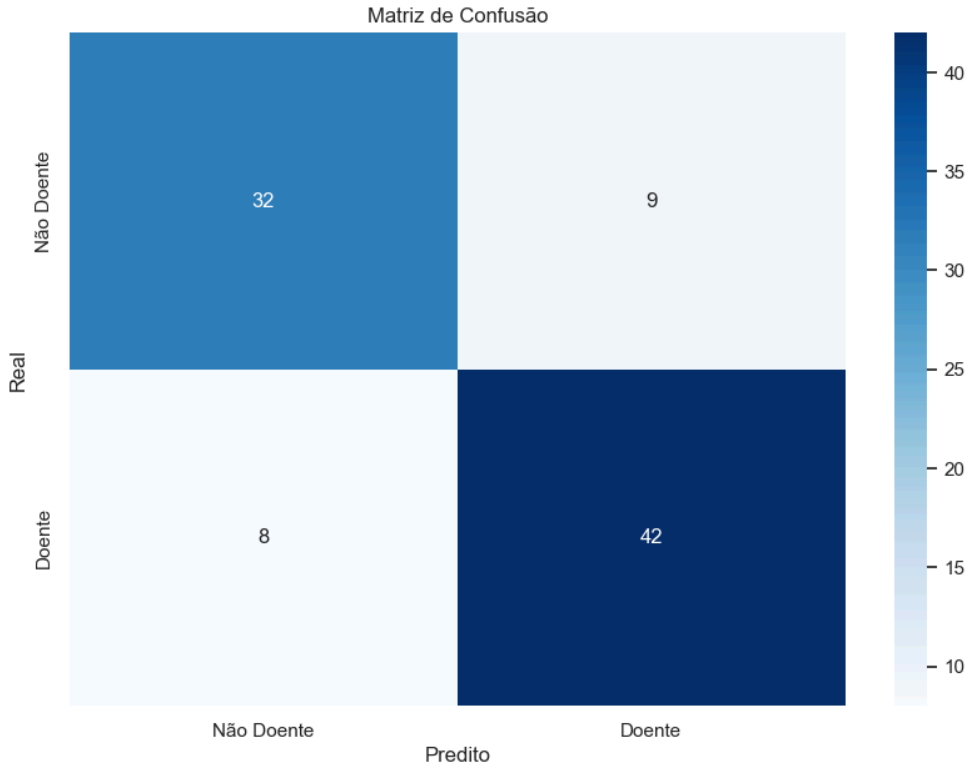
## Resultados do Perceptron Simples

O gráfico a seguir mostra o desempenho do Perceptron Simples As métricas apresentadas incluem precisão, revocação e F1-Score para cada classe, bem como a acurácia geral do modelo.

**Explicação dos Resultados:**

- **Precisão:** Mede a proporção de previsões corretas entre os casos classificados como positivos. Para a Classe 0 (não doente), a precisão foi de 0.80, e para a Classe 1 (doente), foi de 0.82. Isso indica que o modelo foi eficaz em prever corretamente tanto a ausência quanto a presença da doença.
- **Revocação:** Avalia a capacidade do modelo de identificar todos os casos positivos reais. A revocação para a Classe 0 foi de 0.78 e para a Classe 1 foi de 0.84, mostrando que o modelo detectou a maioria dos pacientes com a doença e teve uma boa taxa de detecção para os casos negativos.
- **F1-Score:** Combina precisão e revocação em uma única métrica equilibrada. O F1-Score para a Classe 0 foi de 0.79 e para a Classe 1 foi de 0.83, refletindo um bom equilíbrio entre a identificação correta dos casos e a minimização de falsos positivos.
- **Acurácia Geral:** A acurácia do modelo foi de 0.81, indicando que ele fez previsões corretas em 81% dos casos, o que é um excelente desempenho geral.

As médias macro e ponderadas das métricas foram todas de 0.81, mostrando que o modelo tem um desempenho equilibrado e consistente em todas as classes.O gráfico a seguir ilustra o desempenho do modelo:



## CONCLUSÃO

Neste trabalho, analisamos a eficácia de vários algoritmos de aprendizado de máquina para prever doenças cardíacas, usando um conjunto de dados específico. Avaliamos cinco modelos: Perceptron Multicamadas (MLP), Perceptron Simples, K-Nearest Neighbors (KNN), Naive Bayes e Árvore de Decisão, com base em métricas como precisão, revocação, F1-Score e acurácia geral.

O **Perceptron Multicamadas (MLP)** apresentou uma acurácia de 77%, mostrando um desempenho consistente, especialmente na identificação de casos positivos, mas com espaço para melhorar a detecção de casos negativos.

O **Perceptron Simples** obteve uma acurácia de 81%, com um bom equilíbrio entre precisão e revocação, demonstrando uma sólida capacidade de diferenciar entre pacientes com e sem a doença.

O **K-Nearest Neighbors (KNN)** foi o destaque, com uma acurácia de 88%. Este modelo teve excelente desempenho em todas as métricas, sendo o mais equilibrado e eficaz na previsão tanto da presença quanto da ausência da doença.

O **Naive Bayes** alcançou uma acurácia de 84%, com boas métricas gerais, sendo particularmente eficiente em identificar a presença da doença, embora pudesse melhorar na detecção de alguns casos específicos.

A **Árvore de Decisão** obteve uma acurácia de 74%, mostrando boa capacidade de identificar pacientes sem a doença, mas com algumas limitações na identificação dos casos positivos.

Em resumo, o KNN se destacou como o modelo mais eficaz para esta tarefa, oferecendo a melhor combinação de precisão e acurácia. Cada algoritmo mostrou seu valor e potencial, destacando a importância de escolher o modelo certo para diferentes necessidades e contextos. Esses resultados fornecem uma base sólida para futuras pesquisas e aprimoramentos na detecção de doenças cardíacas.