Previsão de Insuficiência Cardíaca Utilizando Modelos de Machine Learning

Arthur Novaes
Afya Unima
Maceió, Brasil
Email: arthur.novaes@email.com

Igor Lima Laranjeiras
Afya Unima
Maceió, Brasil
Email: igor.lima@email.com

***Abstract*—The objective of this study is to predict the presence of heart disease using the "Heart Failure Prediction Dataset," which combines data from various renowned medical sources. To achieve this goal, machine learning models such as Naive Bayes, Random Forest, and Support Vector Machines (SVM) were applied, along with an ensemble approach using the Voting Classifier, which combines the predictions of individual models to improve accuracy. The data set used, sourced from recognized medical institutions, provides a robust foundation for predicting heart disease based on clinical features. The results demonstrate the potential of these approaches to accurately identify individuals at risk of developing heart-related issues.**

***Index Terms*—Machine Learning, Heart Disease Prediction, Naive Bayes, Random Forest, Voting Classifier**

I. Introdução

A insuficiência cardíaca é uma das condições médicas mais prevalentes e desafiadoras da atualidade, com impacto significativo na saúde pública global. Essa condição ocorre quando o coração não consegue bombear sangue suficiente para atender às necessidades metabólicas do corpo, frequentemente resultando de fatores como hipertensão, doenças coronarianas e diabetes. De acordo com estimativas, a insuficiência cardíaca afeta milhões de pessoas em todo o mundo, com taxas crescentes de incidência devido ao envelhecimento populacional e ao aumento de fatores de risco associados ao estilo de vida. Essa doença apresenta um desafio clínico considerável, dado que diagnósticos tardios podem levar a complicações graves e à redução da qualidade de vida dos pacientes [1].

Nesse contexto, o avanço da inteligência artificial e, mais especificamente, das técnicas de aprendizado de máquina, tem trazido novas possibilidades para o campo da medicina [2]. Essas tecnologias oferecem ferramentas poderosas para analisar grandes volumes de dados médicos e identificar padrões que podem escapar à observação humana. Em vez de depender exclusivamente de métodos tradicionais de diagnóstico, que frequentemente são demorados e sujeitos a erros, algoritmos de Machine Learning podem fornecer insights rápidos e precisos para apoiar a tomada de decisões clínicas [3].

A aplicação de Machine Learning em doenças cardiovasculares, como a insuficiência cardíaca, é particularmente promissora. Dados médicos rotineiros, incluindo idade, pressão arterial, colesterol, frequência cardíaca e características de eletrocardiogramas, podem ser usados como insumos para treinar modelos preditivos. Esses modelos ajudam não apenas a diagnosticar a condição, mas também a prever o risco de desenvolvimento de doenças em pacientes assintomáticos, facilitando intervenções precoces e personalizadas. Além disso, essas abordagens têm o potencial de reduzir custos no sistema de saúde, otimizando recursos ao concentrar esforços em pacientes com maior risco [4].

Este estudo aplica diferentes modelos de aprendizado de máquina para prever a presença ou ausência de doenças cardíacas, incluindo Naive Bayes, Random Forest e Support Vector Machines (SVM) [5]. Por fim, um ensemble learning, com a implementação de um Voting Classifier [6], é utilizado para combinar as previsões dos modelos individuais, buscando maximizar a precisão e a robustez das previsões. Os resultados evidenciam o impacto positivo do aprendizado de máquina no diagnóstico precoce de condições médicas críticas, como a insuficiência cardíaca, contribuindo para avanços significativos no cuidado à saúde.

II. Fundamentação Teórica

A. Naive Bayes

O classificador Naive Bayes é baseado no Teorema de Bayes e assume independência condicional entre as características. Para um conjunto de características $x = (x_1, x_2, ..., x_n)$ e uma classe C_k, o teorema de Bayes é formulado como:

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)} \tag{1}$$

A suposição "ingênua" de independência permite que a probabilidade conjunta seja calculada como o produto das probabilidades individuais:

$$P(x|C_k) = \prod_{i=1}^{n} P(x_i|C_k) \tag{2}$$

Esta abordagem é computacionalmente eficiente e funciona bem em diversos problemas de classificação, mesmo quando a suposição de independência não é estritamente válida [16].

B. Random Forest

Random Forest é um método de ensemble learning que constrói múltiplas árvores de decisão durante o treinamento. Cada árvore é treinada em uma amostra bootstrap do conjunto de dados original, e na divisão de cada nó, um subconjunto

aleatório de características é considerado. A predição final é obtida por votação majoritária (classificação) ou média (regressão) das predições das árvores individuais [12].

A aleatorização no processo de construção das árvores reduz a correlação entre elas, melhorando a generalização e reduzindo o overfitting. A importância das características pode ser calculada com base na redução média da impureza proporcionada por cada característica em todas as árvores [13].

C. Support Vector Machines (SVM)

SVM é um algoritmo de aprendizado supervisionado que encontra um hiperplano ótimo para separar classes em um espaço de características. Para dados linearmente separáveis, o hiperplano é definido como:

$$w \cdot x + b = 0 \tag{3}$$

onde w é o vetor de pesos e b é o viés. A otimização busca maximizar a margem entre as classes, que é inversamente proporcional a $||w||^2$. Para dados não linearmente separáveis, o SVM utiliza funções kernel para mapear os dados para um espaço de dimensão superior onde se tornam linearmente separáveis [20].

O kernel radial (RBF) é definido como:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \tag{4}$$

onde γ controla a influência de cada exemplo de treinamento [21].

D. Voting Classifier

O Voting Classifier é uma técnica de ensemble learning que combina as predições de múltiplos modelos base. No soft voting, as probabilidades preditas por cada classificador são averaged para produzir a predição final:

$$P(y = c) = \frac{1}{M} \sum_{m=1}^{M} P_m(y = c) \tag{5}$$

onde M é o número de classificadores e $P_m(y = c)$ é a probabilidade predita pelo m-ésimo classificador para a classe c [25].

III. Metodologia

A metodologia deste projeto foi desenvolvida com o objetivo de prever a presença de insuficiência cardíaca em pacientes com base nos dados clínicos disponíveis no Heart Failure Prediction Dataset.

A. Descrição do Conjunto de Dados

O conjunto de dados utilizado é o *Heart Failure Prediction Dataset*, disponível publicamente na plataforma Kaggle. Ele combina dados de cinco fontes médicas renomadas: Cleveland Clinic Foundation, Hungarian Institute of Cardiology, University Hospital (Zurique e Basel) e o V.A. Medical Center em Long Beach.

- **Total de observações:** 918.
- **Variáveis:** 12 (11 preditoras e 1 alvo: *HeartDisease*) [7].

1) Variáveis Preditivas:

- **Numéricas:** Idade (*Age*), pressão arterial em repouso (*RestingBP*), colesterol (*Cholesterol*), frequência cardíaca máxima (*MaxHR*) e depressão do segmento ST após exercício (*Oldpeak*).
- **Categóricas:** Sexo (*Sex*), tipo de dor no peito (*ChestPainType*), glicemia em jejum (*FastingBS*), resultados de eletrocardiograma (*RestingECG*), angina induzida por exercício (*ExerciseAngina*) e inclinação do segmento ST (*ST_Slope*).

B. Pré-processamento

- **Remoção de valores ausentes:** Valores foram tratados utilizando métodos como imputação pela média (variáveis numéricas).
- **Normalização:** Variáveis numéricas foram padronizadas utilizando o *StandardScaler*.
- **Codificação de variáveis categóricas:** Aplicado *OneHotEncoding*.
- **Balanceamento de classes:** O dataset apresentava distribuição balanceada (55% com doença, 45% sem doença), portanto técnicas de balanceamento não foram necessárias.

C. Ferramentas e Tecnologias

- **Software:** Python, com as bibliotecas *pandas*, *numpy*, *scikit-learn*, *seaborn* e *matplotlib*.
- **Ambiente:** Jupyter Notebook.

D. Modelos e Algoritmos

Três algoritmos principais foram aplicados:

1) **Naive Bayes (GaussianNB):** Modelo probabilístico baseado no Teorema de Bayes.
2) **Random Forest:** Modelo de árvores de decisão com 200 estimadores e profundidade máxima de 10.
3) **SVM (Support Vector Machines):** Kernel radial (*RBF*), regularização $C = 1$, $\gamma = 0.1$.

E. Estratégia de Treinamento e Validação

Foi utilizada validação cruzada com 5 *folds*, assegurando robustez e generalização. O conjunto de dados foi dividido em 80% para treinamento e 20% para teste.

F. Métricas de Avaliação

As métricas utilizadas para avaliação foram [9]:

- **Acurácia:** Proporção de previsões corretas.
- **Precisão:** Exatidão das previsões positivas.
- **Recall (Sensibilidade):** Proporção de verdadeiros positivos identificados corretamente.
- **F1-Score:** Média harmônica entre precisão e recall.
- **AUC-ROC:** Capacidade de discriminação dos modelos.

G. Ensemble Learning

Foi implementado um *Voting Classifier* utilizando *soft voting*. Este combinou as probabilidades preditivas de Naive Bayes, Random Forest e SVM, maximizando precisão e robustez [10].

IV. Proposta e Implementação

A implementação do experimento seguiu as seguintes etapas:

A. Análise Exploratória de Dados

Inicialmente, foi realizada uma análise exploratória para compreender a distribuição das variáveis, identificar correlações e detectar possíveis outliers. Esta etapa foi crucial para orientar as decisões de pré-processamento.

B. Seleção de Modelos

A seleção dos modelos (Naive Bayes, Random Forest e SVM) foi baseada em sua complementaridade: Naive Bayes como modelo probabilístico simples e rápido, Random Forest como método ensemble robusto, e SVM como classificador de margem máxima eficaz em espaços de alta dimensão.

C. Configuração de Parâmetros

Os parâmetros dos modelos foram configurados através de validação cruzada com busca em grid para otimização:

- **Random Forest:** n_estimators=200, max_depth=10
- **SVM:** kernel='rbf', C=1, gamma=0.1
- **Naive Bayes:** Parâmetros padrão do GaussianNB

D. Implementação do Ensemble

O Voting Classifier foi implementado utilizando a classe `VotingClassifier` do scikit-learn, com a estratégia de soft voting para combinar as probabilidades preditas pelos três modelos base.

V. Resultados

A análise dos resultados obtidos com os modelos de aprendizado de máquina forneceu insights valiosos sobre o desempenho dos algoritmos na previsão de doenças cardíacas. Foram utilizados gráficos, tabelas e métricas de avaliação para comparar os modelos e destacar o desempenho geral [11].

A. Desempenho Geral dos Modelos

Os resultados das métricas de avaliação foram organizados em uma tabela para facilitar a comparação, conforme mostrado na Tabela I.

Table I
Desempenho dos Modelos nas Métricas de Avaliação

Modelo	Acurácia (%)	Precisão (%)	Recall (%)	F1-Score (%)
Naive Bayes	85.87	90.63	84.11	87.24
Random Forest	87.50	89.17	89.72	89.44
SVM	85.33	87.18	88.79	87.97
Voting Classifier	**88.04**	**90.00**	**89.72**	**89.86**

B. Matriz de Confusão

As matrizes de confusão [24] foram geradas para cada modelo. Abaixo, é apresentada a matriz de confusão do *Voting Classifier*:

O *Voting Classifier* teve uma taxa reduzida de falsos negativos (12 casos), destacando sua eficácia na identificação de pacientes com doenças cardíacas.

Table II
Matriz de Confusão do Voting Classifier

	Predito: Sem Doença	Predito: Com Doença
Real: Sem Doença	67	10
Real: Com Doença	12	95

C. Curva ROC

A análise da curva ROC mostrou que o Voting Classifier obteve a maior área sob a curva (AUC = 0.94), seguido pelo Random Forest (AUC = 0.92), SVM (AUC = 0.89) e Naive Bayes (AUC = 0.87).

VI. Discussão

A. Interpretação dos Resultados

Os resultados demonstram a eficácia dos modelos de Machine Learning na predição de insuficiência cardíaca. O Voting Classifier obteve o melhor desempenho geral, com acurácia de 88.04% e F1-Score de 89.86%, confirmando a vantagem de combinar múltiplos algoritmos através de ensemble learning [25].

O Random Forest apresentou desempenho destacado como modelo individual, com alta sensibilidade (89.72%) que é crucial em aplicações médicas onde falsos negativos podem ter consequências graves. Esta performance pode ser atribuída à capacidade do algoritmo de capturar relações não-lineares e interações entre features [13].

O SVM, apesar de ser um algoritmo poderoso, teve desempenho ligeiramente inferior, possivelmente devido à sensibilidade aos parâmetros de regularização e à necessidade de ajuste fino do kernel [21].

O Naive Bayes, embora tenha apresentado a menor performance entre os modelos testados, mostrou-se computacionalmente eficiente e pode ser útil em aplicações que requerem predições rápidas com recursos limitados.

B. Limitações e Desafios

Algumas limitações do estudo incluem:

- O tamanho moderado do dataset (918 instâncias) pode limitar a generalização dos modelos
- A suposição de independência do Naive Bayes não se adequa perfeitamente aos dados médicos
- A interpretabilidade dos modelos ensemble é reduzida em comparação com modelos individuais

C. Impacto do Pré-processamento

O pré-processamento adequado, particularmente a padronização das features numéricas e a codificação one-hot das variáveis categóricas, foi essencial para o bom desempenho dos modelos, especialmente do SVM que é sensível à escala dos dados.

VII. Conclusão

Este trabalho apresentou uma abordagem sistemática para prever a presença de doenças cardíacas utilizando algoritmos de aprendizado de máquina aplicados ao *Heart Failure Prediction Dataset*.

Os resultados obtidos destacam a eficácia do *Voting Classifier*, que apresentou o melhor desempenho geral, com uma acurácia de 88.04% e um F1-Score de 89.86%. Esse modelo se beneficiou da combinação dos algoritmos *Naive Bayes*, *Random Forest* e *SVM*, evidenciando a importância de ensembles no aumento da robustez e precisão de sistemas preditivos. O *Random Forest* foi o modelo individual de maior destaque, apresentando resultados consistentes e equilibrados em todas as métricas avaliadas.

A análise aprofundada das métricas, como F1-Score e AUC-ROC, foi crucial para identificar os pontos fortes e limitações de cada modelo. Além disso, a matriz de confusão do *Voting Classifier* revelou sua baixa taxa de falsos negativos, um aspecto crítico no contexto médico, onde falhas em identificar casos positivos podem ter graves consequências para os pacientes [28].

A. Trabalhos Futuros

Como trabalhos futuros, sugere-se:

- Expansão do conjunto de dados com mais amostras e variáveis clínicas
- Exploração de técnicas avançadas como Redes Neurais Profundas e XGBoost
- Implementação de sistemas de explicação (Explainable AI) para aumentar a transparência dos modelos
- Validação dos modelos em ambientes clínicos reais
- Desenvolvimento de aplicações web ou móveis para disponibilizar as predições aos profissionais de saúde

Conclui-se que o aprendizado de máquina é uma ferramenta promissora no campo da saúde, com potencial significativo para melhorar o diagnóstico precoce e a personalização do tratamento. Este estudo reforça a relevância da integração entre tecnologia e medicina, pavimentando o caminho para inovações futuras que salvem vidas e aprimorem a qualidade do cuidado médico [30].

Agradecimentos

Gostaríamos de expressar nossa sincera gratidão às instituições e indivíduos que tornaram este estudo possível. O conjunto de dados utilizado nesta pesquisa, o Heart Failure Prediction Dataset, é uma compilação de dados fornecidos por renomadas instituições médicas e profissionais. Também reconhecemos a plataforma Kaggle por hospedar o conjunto de dados e promover a pesquisa colaborativa em aprendizado de máquina e saúde.

References

[1] M. S. Khan et al., "Global epidemiology of heart failure" Nature Reviews Cardiology, 2024.
[2] M. Shehab et al., "Machine learning in medical applications: A review of state-of-the-art methods," Computers in Biology and Medicine, 2022.
[3] A. M. Rahmani et al., "Machine Learning (ML) in Medicine: Review, Applications, and Challenges," Electronics, 2021.
[4] H. El-Sofany et al., "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," Scientific Reports, 2024.
[5] C. Boukhatem et al., "Heart Disease Prediction Using Machine Learning" 2022 Advances in Science and Engineering Technology International Conferences, 2022.
[6] A. da Cunha et al., "Boosting, Voting Classifiers and Randomized Sample Compression Schemes" Journal of Machine Learning Research, 2024.
[7] F. Kaliyadan et al., "Types of Variables, Descriptive Statistics, and Sample Size" Indian Dermatology Online Journal, 2019.
[8] V. Duarte et al., "Benchmarking machine-learning software and hardware for quantitative economics" Journal of Economic Dynamics and Control, 2020.
[9] O. Rainio et al., "Evaluation metrics and statistical tests for machine learning" Scientific Reports, 2024.
[10] C. Cornelio et al., "Voting with random classifiers (VORACE): theoretical and experimental analysis," Machine Learning, 2021.
[11] F. Xia et al., "Graph Learning: A Survey," IEEE Transactions on Artificial Intelligence, 2021.
[12] L. Breiman, "Random Forests," Machine Learning, 2001.
[13] M. Schonlau et al., "The random forest algorithm for statistical learning," The Stata Journal, 2020.
[14] K. Fawagreh et al., "Random forests: from early developments to recent advancements," Systems Science & Control Engineering, 2014.
[15] G. Biau, "Analysis of a random forests model," Journal of Machine Learning Research, 2012.
[16] V. Sheth et al., "A Comparative Analysis of Machine Learning Algorithms for Classification Purpose," Procedia Computer Science, 2022.
[17] S. U. Hassan et al., "Analytics of machine learning-based algorithms for text classification" Sustainable Operations and Computers, 2022.
[18] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," SN Computer Science, 2021.
[19] I. Wickramasinghe et al., "Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation," Knowledge and Information Systems, 2021.
[20] J. Cervantes et al., "A comprehensive survey on support vector machine classification: Applications, challenges and trends," Neurocomputing, 2020.
[21] G. Doran et al., "A theoretical and empirical analysis of support vector machine methods for multiple-instance classification," Machine Learning, 2013.
[22] R. Guido et al., "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," Diagnostics, 2024.
[23] V. Piccialli et al., "Nonlinear optimization and support vector machines," 4OR, 2022.
[24] K. M. Ting, "Confusion Matrix," Encyclopedia of Machine Learning, 2011.
[25] I. Gandhi et al., "Hybrid Ensemble of classifiers using voting," 2015 International Conference on Green Computing and Internet of Things, 2015.
[26] A. Mohammed et al., "A comprehensive review on ensemble deep learning: Opportunities and challenges," Journal of King Saud University - Computer and Information Sciences, 2023.
[27] M. Krichen et al., "Performance enhancement of artificial intelligence: A survey" Computer Science Review, 2024.
[28] R. Pugliese et al., "Machine learning-based approach: global trends, research directions, and regulatory standpoints," Data Science and Management, 2021.
[29] B. Gaye et al., "Improvement of Support Vector Machine Algorithm in Big Data Background," Mathematical Problems in Engineering, 2021.
[30] S. Liu et al., "Towards better analysis of machine learning models: A visual analytics perspective," Visual Informatics, 2017.