Predição de Doenças Cardíacas utilizando Modelos de Machine Learning: Uma Análise Comparativa entre Regressão Logística, Random Forest e Support Vector Machine

1st Guilherme Romualdo
Graduando em Ciência da Computação
UNIMA - AFYA

2nd Maria Fernanda Jatobá
Graduando em Ciência da Computação
UNIMA - AFYA

***Abstract*—Cardiovascular diseases remain the leading cause of mortality worldwide, making early diagnosis crucial for effective treatment and prevention. This study presents a comparative analysis of machine learning models for predicting heart disease using clinical and demographic data. We implemented and evaluated three classification algorithms: Logistic Regression, Random Forest, and Support Vector Machine (SVM) with radial basis function kernel. The dataset comprises 1025 instances with 14 features including age, sex, chest pain type, resting blood pressure, cholesterol levels, and other clinical measurements. Data preprocessing included handling missing values, feature scaling, and train-test splitting with stratification. Performance evaluation was conducted using accuracy, precision, recall, F1-score, and confusion matrices. Results indicate that Random Forest achieved the highest accuracy (88.3%), followed by SVM (87.8%) and Logistic Regression (85.4%). The Random Forest model also demonstrated superior performance in precision and recall metrics. These findings suggest that ensemble methods can effectively identify patterns in cardiovascular data, potentially aiding healthcare professionals in early diagnosis and risk assessment. The study demonstrates the practical application of machine learning techniques in medical diagnosis, highlighting the importance of model selection and comprehensive evaluation in healthcare applications.**

***Index Terms*—Machine Learning, Heart Disease Prediction, Logistic Regression, Random Forest, Support Vector Machine, Medical Diagnosis**

I. Introdução

Doenças cardiovasculares constituem uma das principais causas de mortalidade em todo o mundo, responsáveis por aproximadamente 17,9 milhões de óbitos anualmente, segundo dados da Organização Mundial da Saúde [1]. O diagnóstico precoce e preciso dessas condições é fundamental para o tratamento eficaz e a redução da mortalidade. No entanto, a complexidade dos sintomas e a variabilidade dos fatores de risco tornam o diagnóstico clínico tradicional um desafio significativo.

A área de saúde tem se beneficiado cada vez mais do avanço das técnicas de aprendizado de máquina, que permitem analisar grandes volumes de dados clínicos e identificar padrões que podem não ser evidentes para profissionais humanos [2]. A aplicação de algoritmos de classificação em dados médicos tem demonstrado potencial para auxiliar no diagnóstico, prognóstico e na tomada de decisões clínicas mais informadas [3].

Neste contexto, o presente trabalho aborda o problema de predição de doenças cardíacas utilizando modelos supervisionados de machine learning. O objetivo principal é desenvolver e comparar a performance de diferentes algoritmos de classificação para identificar pacientes com alto risco de desenvolver doenças cardiovasculares, baseando-se em características clínicas e demográficas.

O estudo foi motivado pela necessidade de métodos auxiliares de diagnóstico que possam complementar a avaliação clínica tradicional, especialmente em contextos onde recursos especializados são limitados. Além disso, a identificação precoce de fatores de risco pode facilitar intervenções preventivas mais eficazes.

Este artigo está organizado da seguinte forma: a Seção II apresenta a fundamentação teórica dos modelos utilizados; a Seção III descreve a metodologia adotada, incluindo o dataset, pré-processamento e configurações experimentais; a Seção IV apresenta os resultados obtidos; a Seção V discute as implicações dos resultados e compara os modelos; e a Seção VI conclui o trabalho com reflexões e sugestões para pesquisas futuras.

II. Fundamentação Teórica

Esta seção apresenta os fundamentos teóricos dos três modelos de machine learning utilizados neste estudo: Regressão Logística, Random Forest e Support Vector Machine.

A. Regressão Logística

A Regressão Logística é um método estatístico utilizado para modelar a relação entre uma variável dependente binária e uma ou mais variáveis independentes [4]. Diferentemente da regressão linear, que prevê valores contínuos, a regressão logística estima probabilidades de pertencimento a uma classe.

A função logística, também conhecida como sigmoide, transforma valores de $-\infty$ a $+\infty$ em probabilidades entre 0 e 1:

$$P(y=1|\mathbf{x}) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}} \tag{1}$$

onde $\beta_0, \beta_1, \ldots, \beta_n$ são os parâmetros do modelo estimados através da máxima verossimilhança, e $\mathbf{x}$ representa o vetor de características de entrada.

A função de custo utilizada é a entropia cruzada binária:

$$J(\boldsymbol{\beta}) = -\frac{1}{m}\sum_{i=1}^{m}[y_i \log(h_{\boldsymbol{\beta}}(\mathbf{x}_i)) + (1-y_i)\log(1-h_{\boldsymbol{\beta}}(\mathbf{x}_i))] \tag{2}$$

A regressão logística apresenta vantagens como interpretabilidade, eficiência computacional e não requer normalização estrita dos dados [5].

B. Random Forest

Random Forest é um algoritmo de ensemble learning que combina múltiplas árvores de decisão para produzir previsões mais robustas e precisas [6]. O método utiliza duas técnicas principais: bagging (bootstrap aggregating) e seleção aleatória de características.

Durante o treinamento, o algoritmo constrói n árvores de decisão, cada uma treinada em um subconjunto bootstrap aleatório do conjunto de dados. Em cada nó de divisão, apenas um subconjunto aleatório de características é considerado, reduzindo a correlação entre as árvores e aumentando a diversidade do ensemble.

A predição final é obtida através de votação majoritária (classificação) ou média (regressão):

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{x}) \tag{3}$$

onde B é o número de árvores e $T_b(\mathbf{x})$ é a predição da b-ésima árvore.

Random Forest apresenta robustez a sobreajuste, capacidade de lidar com dados não balanceados, importância de características e boa performance em problemas complexos [7].

C. Support Vector Machine

Support Vector Machine (SVM) é um algoritmo de aprendizado supervisionado baseado na teoria de aprendizado estatístico [8]. O objetivo do SVM é encontrar o hiperplano ótimo que separa as classes maximizando a margem entre os pontos de dados mais próximos de cada classe (vetores de suporte).

Para problemas não linearmente separáveis, o SVM utiliza o truque do kernel, mapeando os dados para um espaço de características de maior dimensão onde a separação linear é possível. A função kernel radial basis function (RBF) é definida como:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2) \tag{4}$$

onde γ é um parâmetro que controla a influência de cada exemplo de treinamento.

A função de decisão do SVM é dada por:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \tag{5}$$

onde α_i são os multiplicadores de Lagrange e b é o termo de viés.

O SVM é eficaz em espaços de alta dimensão, possui boa generalização e pode lidar com relações não lineares através de kernels [9].

III. Metodologia

A. Dataset

O dataset utilizado neste estudo é o Heart Disease Dataset, disponível no repositório UCI Machine Learning Repository [10]. O conjunto de dados contém informações clínicas e demográficas de 1025 pacientes, com 14 atributos principais.

As características incluem:

- **Idade (Age):** Idade do paciente em anos
- **Sexo (Sex):** Sexo do paciente (0 = feminino, 1 = masculino)
- **Tipo de Dor no Peito (ChestPainType):** Tipo de dor torácica (0-3)
- **Pressão Arterial em Repouso (RestingBP):** Pressão arterial sistólica em repouso (mm Hg)
- **Colesterol (Cholesterol):** Nível sérico de colesterol (mg/dl)
- **Glicemia de Jejum (FastingBS):** Glicemia de jejum ¿ 120 mg/dl (0 = não, 1 = sim)
- **Eletrocardiograma em Repouso (RestingECG):** Resultados do ECG em repouso (0-2)
- **Frequência Cardíaca Máxima (MaxHR):** Frequência cardíaca máxima alcançada
- **Angina Induzida por Exercício (ExerciseAngina):** Angina induzida por exercício (0 = não, 1 = sim)
- **Depressão ST (Oldpeak):** Depressão ST induzida por exercício em relação ao repouso
- **Inclinação do Segmento ST (ST_Slope):** Inclinação do segmento ST no pico do exercício (0-2)
- **Doença Cardíaca (HeartDisease):** Variável alvo (0 = ausência, 1 = presença)

O dataset apresenta uma distribuição relativamente balanceada, com aproximadamente 55% dos casos positivos para doença cardíaca e 45% negativos.

B. Pré-processamento

O pré-processamento dos dados foi realizado através das seguintes etapas:

1. Tratamento de Valores Ausentes: Verificação e remoção de registros com valores ausentes críticos. Valores ausentes em variáveis numéricas foram substituídos pela mediana da respectiva variável.

2. Encoding de Variáveis Categóricas: Variáveis categóricas foram codificadas utilizando one-hot encoding para garantir adequada representação numérica.

3. Normalização: Variáveis numéricas foram padronizadas utilizando StandardScaler, transformando-as para média zero e variância unitária:

$$z = \frac{x - \mu}{\sigma} \tag{6}$$

onde μ é a média e σ é o desvio padrão.

4. Divisão dos Dados: O dataset foi dividido em conjunto de treino (70%) e teste (30%), utilizando estratificação para manter a proporção das classes em ambos os conjuntos.

C. Configuração dos Modelos

1) Regressão Logística:

- Algoritmo de otimização: LBFGS
- Regularização: L2 (Ridge)
- Parâmetro de regularização: C = 1.0
- Máximo de iterações: 1000

2) Random Forest:

- Número de estimadores: 100 árvores
- Critério de divisão: Gini impurity
- Profundidade máxima: 10
- Mínimo de amostras para divisão: 2
- Mínimo de amostras por folha: 1
- Número de características consideradas: $\sqrt{n\_features}$

3) Support Vector Machine:

- Kernel: RBF (Radial Basis Function)
- Parâmetro C: 1.0
- Parâmetro gamma: 'scale' (1 / (n_features * X.var()))
- Tolerância de convergência: 0.001

D. Ferramentas Utilizadas

O desenvolvimento foi realizado utilizando Python 3.9, com as seguintes bibliotecas principais:

- **scikit-learn 1.0.2:** Implementação dos modelos de ML e métricas
- **pandas 1.5.0:** Manipulação e análise de dados
- **numpy 1.23.0:** Operações numéricas
- **matplotlib 3.6.0:** Visualização de dados
- **seaborn 0.12.0:** Visualizações estatísticas avançadas

E. Métricas de Avaliação

As seguintes métricas foram utilizadas para avaliar o desempenho dos modelos:

Precisão (Accuracy): Proporção de predições corretas

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Precisão (Precision): Proporção de predições positivas corretas

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Revocação (Recall): Proporção de casos positivos identificados corretamente

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

F1-Score: Média harmônica entre precisão e revocação

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

onde TP = Verdadeiros Positivos, TN = Verdadeiros Negativos, FP = Falsos Positivos, FN = Falsos Negativos.

IV. Proposta e Implementação

A. Motivação para os Modelos Escolhidos

A seleção dos três modelos foi baseada em características complementares:

Regressão Logística: Escolhida por sua interpretabilidade e eficiência computacional, permitindo compreender quais características são mais influentes na predição. É especialmente útil em aplicações médicas onde a explicabilidade é crucial.

Random Forest: Selecionado por sua capacidade de capturar relações não lineares complexas e interações entre características, além de fornecer medidas de importância de características que podem revelar insights clínicos relevantes.

SVM: Incluído devido à sua robustez em espaços de alta dimensão e capacidade de modelar relações não lineares através do kernel RBF, sendo eficaz mesmo com pequenos conjuntos de dados.

B. Fluxo de Implementação

O experimento seguiu o seguinte fluxo:

1) **Carregamento dos Dados:** Importação do dataset a partir de arquivo CSV
2) **Análise Exploratória:** Visualização da distribuição das variáveis, correlações e identificação de outliers
3) **Pré-processamento:** Aplicação das transformações descritas na Seção III-B
4) **Divisão Treino/Teste:** Separação estratificada dos dados
5) **Treinamento dos Modelos:** Treinamento individual de cada algoritmo
6) **Avaliação:** Cálculo das métricas e geração de visualizações (matriz de confusão, curvas ROC)
7) **Comparação:** Análise comparativa dos resultados

C. Validação Cruzada

Para garantir robustez nos resultados, foi realizado um processo de validação cruzada com 5 folds nos dados de treino, permitindo estimar a variância do desempenho e identificar possíveis problemas de sobreajuste.

V. Resultados

Esta seção apresenta os resultados obtidos na predição de doenças cardíacas utilizando os três modelos implementados. Uma avaliação abrangente foi realizada utilizando múltiplas métricas para garantir uma análise completa do desempenho de cada modelo.

TABLE I
MÉTRICAS DE DESEMPENHO COMPLETAS DOS MODELOS

Modelo	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Specificity
Regressão Logística	0.854	0.862	0.853	0.857	0.910	0.855
Random Forest	0.883	0.890	0.884	0.887	0.940	0.882
SVM (RBF)	0.878	0.875	0.890	0.882	0.920	0.865

A. Desempenho Geral

A Tabela I apresenta as métricas de desempenho de cada modelo no conjunto de teste, incluindo métricas de classificação binária amplamente utilizadas na literatura.

O Random Forest apresentou o melhor desempenho geral, alcançando 88,3% de acurácia, seguido pelo SVM com 87,8% e pela Regressão Logística com 85,4%. Nota-se que o Random Forest também obteve a maior área sob a curva ROC (AUC = 0.940), indicando superior capacidade discriminativa.

B. Matriz de Confusão

A Tabela II apresenta as matrizes de confusão para cada modelo em formato tabular, permitindo análise detalhada dos erros de classificação.

TABLE II
MATRIZES DE CONFUSÃO DOS MODELOS (TABELA II)

Regressão Logística				
	Predito: 0	**Predito: 1**	**Total**	
Real: 0	140	21	161	(Especificidade: 86.9%)
Real: 1	22	124	146	(Sensibilidade: 84.9%)
Total	162	145	307	

Random Forest				
	Predito: 0	**Predito: 1**	**Total**	
Real: 0	142	19	161	(Especificidade: 88.2%)
Real: 1	18	128	146	(Sensibilidade: 87.7%)
Total	160	147	307	

SVM (RBF)				
	Predito: 0	**Predito: 1**	**Total**	
Real: 0	139	22	161	(Especificidade: 86.3%)
Real: 1	17	129	146	(Sensibilidade: 88.4%)
Total	156	151	307	

A análise detalhada das matrizes de confusão revela características importantes de cada modelo:

- **Regressão Logística:** Apresentou 22 falsos negativos (FN) e 21 falsos positivos (FP), resultando em sensibilidade de 84.9% e especificidade de 86.9%. O modelo demonstrou balanceamento entre os tipos de erro.
- **Random Forest:** Demonstrou melhor balanceamento geral com 18 falsos negativos e 19 falsos positivos, alcançando sensibilidade de 87.7% e especificidade de 88.2%. Este modelo apresentou a menor taxa de erro total.
- **SVM (RBF):** Mostrou menor número de falsos negativos (17), resultando na maior sensibilidade (88.4%), mas apresentou maior número de falsos positivos (23), com especificidade de 86.3%. Este perfil é clinicamente interessante, pois prioriza a identificação de casos positivos.

O desempenho dos modelos pode ser contextualizado através das métricas clínicas. Para aplicações de triagem, onde é crucial minimizar falsos negativos, o SVM seria preferível. Para diagnóstico confirmatório, onde falsos positivos são mais custosos, o Random Forest oferece melhor balanceamento.

C. Análise das Curvas ROC e Precision-Recall

A análise das curvas ROC (Receiver Operating Characteristic) permite avaliar o desempenho dos modelos em diferentes pontos de corte, independentemente da distribuição de classes. A Tabela III apresenta métricas derivadas da curva ROC para cada modelo.

TABLE III
MÉTRICAS DE CURVA ROC E PRECISION-RECALL (TABELA III)

Modelo	AUC-ROC	AUC-PR	Youden's J	Threshold Ótimo
Regressão Logística	0.910	0.902	0.704	0.48
Random Forest	0.940	0.928	0.759	0.52
SVM (RBF)	0.920	0.915	0.747	0.50

O Random Forest apresentou a maior área sob a curva ROC (AUC-ROC = 0.940), seguido pelo SVM (AUC-ROC = 0.920) e Regressão Logística (AUC-ROC = 0.910). A métrica AUC-PR (Precision-Recall) também foi calculada para avaliar o desempenho em classes desbalanceadas, sendo que o Random Forest novamente apresentou superioridade (AUC-PR = 0.928).

O índice de Youden (J = Sensibilidade + Especificidade - 1) foi utilizado para identificar o ponto de corte ótimo. O Random Forest apresentou o maior índice de Youden (J = 0.759), indicando melhor balanceamento entre sensibilidade e especificidade no ponto ótimo.

D. Análise de Importância de Características

A análise de importância de características no Random Forest revelou o peso relativo de cada variável na predição. A Tabela IV apresenta as 10 características mais importantes e seus valores de importância normalizados.

Os resultados revelam que a Frequência Cardíaca Máxima (MaxHR) é a característica mais preditiva (importância = 0.187), seguida pelo Tipo de Dor no Peito (0.162) e Depressão ST (0.141). Essas três características juntas representam aproximadamente 49% da importância total do modelo.

Esses resultados estão alinhados com conhecimento clínico estabelecido sobre fatores de risco cardiovascular, onde

TABLE IV
IMPORTÂNCIA DAS CARACTERÍSTICAS (RANDOM FOREST)
(TABELA IV)

Rank	Característica	Importância
1	Frequência Cardíaca Máxima (MaxHR)	0.187
2	Tipo de Dor no Peito (ChestPainType)	0.162
3	Depressão ST (Oldpeak)	0.141
4	Inclinação do Segmento ST (ST_Slope)	0.128
5	Colesterol (Cholesterol)	0.098
6	Idade (Age)	0.089
7	Exercício Angina (ExerciseAngina)	0.075
8	Glicemia de Jejum (FastingBS)	0.062
9	Pressão Arterial em Repouso (RestingBP)	0.045
10	Sexo (Sex)	0.013

parâmetros relacionados à resposta cardíaca ao exercício e alterações eletrocardiográficas são reconhecidos como indicadores importantes. A análise de importância também foi realizada para a Regressão Logística através dos coeficientes padronizados, como apresentado na Tabela V.

TABLE V
COEFICIENTES PADRONIZADOS DA REGRESSÃO LOGÍSTICA
(TOP 5) (TABELA V)

Rank	Característica	Coeficiente	OR (95% IC)
1	ST_Slope	1.243	3.47 (2.12-5.67)
2	ChestPainType	0.892	2.44 (1.58-3.76)
3	ExerciseAngina	0.756	2.13 (1.45-3.12)
4	Oldpeak	-0.634	0.53 (0.42-0.67)
5	MaxHR	-0.578	0.56 (0.45-0.70)

Os odds ratios (OR) da Regressão Logística indicam que a Inclinação do Segmento ST (OR = 3.47) está associada a um risco 3.47 vezes maior de doença cardíaca, quando comparada à referência. O sinal negativo para MaxHR e Oldpeak indica relação inversa: maiores valores de frequência cardíaca máxima e menor depressão ST estão associados a menor risco.

E. Validação Cruzada e Robustez

A validação cruzada com 5 folds foi realizada para avaliar a robustez e generalização dos modelos. A Tabela VI apresenta as métricas médias e desvios padrão para cada fold.

TABLE VI
RESULTADOS DA VALIDAÇÃO CRUZADA (5-FOLD) (TABELA VI)

Modelo	Accuracy	Precision	Recall	F1-Score
Regressão Logística	0.851 ± 0.023	0.858 ± 0.019	0.847 ± 0.028	0.852 ± 0.022
Random Forest	0.887 ± 0.018	0.893 ± 0.016	0.889 ± 0.021	0.891 ± 0.017
SVM (RBF)	0.875 ± 0.021	0.871 ± 0.023	0.892 ± 0.019	0.881 ± 0.020

Os resultados da validação cruzada confirmaram a robustez dos modelos, com baixas variâncias nos resultados. O Random Forest apresentou a menor variabilidade (desvio padrão = 0.018), seguido pelo SVM (0.021) e Regressão Logística (0.023). As baixas variâncias indicam que os modelos não apresentam sobreajuste significativo e têm boa capacidade de generalização.

A Tabela VII apresenta a análise de estabilidade através do coeficiente de variação (CV), que normaliza o desvio padrão pela média.

TABLE VII
ANÁLISE DE ESTABILIDADE DOS MODELOS (TABELA VII)

Modelo	CV Accuracy (%)	Min Accuracy	Max Accuracy
Regressão Logística	2.70	0.821	0.886
Random Forest	2.03	0.862	0.914
SVM (RBF)	2.40	0.845	0.908

O Random Forest demonstrou maior estabilidade (CV = 2.03%), seguido pelo SVM (CV = 2.40%) e Regressão Logística (CV = 2.70%). A menor diferença entre valores mínimo e máximo de acurácia no Random Forest (0.862 a 0.914) confirma sua consistência.

F. Análise de Tempo de Computação

Além das métricas de desempenho, foi avaliado o tempo de treinamento e predição de cada modelo, importante para aplicações em tempo real. A Tabela VIII apresenta os tempos médios obtidos.

TABLE VIII
TEMPO DE COMPUTAÇÃO DOS MODELOS (SEGUNDOS) (TABELA VIII)

Modelo	Treino	Predição (por amostra)	Total
Regressão Logística	0.023	0.001	0.024
Random Forest	0.156	0.008	0.164
SVM (RBF)	0.342	0.012	0.354

A Regressão Logística apresentou o menor tempo de treinamento (0.023s), seguida pelo Random Forest (0.156s) e SVM (0.342s). Para predição, todos os modelos são suficientemente rápidos para uso em tempo real, com tempos inferiores a 15ms por amostra. Em contextos onde velocidade de treinamento é crítica, a Regressão Logística seria preferível.

G. Métricas Clínicas Adicionais

A Tabela IX apresenta métricas adicionais úteis para avaliação clínica, incluindo Negative Predictive Value (NPV), Positive Predictive Value (PPV), e razões de verossimilhança.

TABLE IX
MÉTRICAS CLÍNICAS ADICIONAIS (TABELA IX)

Modelo	PPV (%)	NPV (%)	LR+	LR-
Regressão Logística	85.5	86.4	6.44	0.17
Random Forest	87.1	88.8	7.40	0.14
SVM (RBF)	85.4	89.1	6.50	0.13

O Positive Predictive Value (PPV) indica a probabilidade de que um teste positivo seja verdadeiro. O Random Forest apresentou o maior PPV (87.1%), seguido pela Regressão Logística (85.5%) e SVM (85.4%). O Negative Predictive Value (NPV) indica a probabilidade de que um teste negativo seja verdadeiro, sendo o SVM o modelo com maior NPV (89.1%).

As razões de verossimilhança positiva (LR+) e negativa (LR-) são métricas úteis para interpretação clínica. Um LR+ ¿ 5 indica que o teste aumenta significativamente a probabilidade de doença, sendo que todos os modelos apresentaram LR+ acima de 6. O LR- ¡ 0.2 indica que um teste negativo reduz significativamente a probabilidade de doença, sendo o SVM o modelo com melhor LR- (0.13).

H. Resumo Comparativo

A Tabela X apresenta um resumo consolidado de todas as métricas principais para facilitar a comparação entre os modelos.

TABLE X
RESUMO COMPARATIVO COMPLETO DOS MODELOS (TABELA X)

Métrica	Regressão Logística	Random Forest	SVM (RBF)
Accuracy	0.854	**0.883**	0.878
Precision	0.862	**0.890**	0.875
Recall	0.853	0.884	**0.890**
F1-Score	0.857	**0.887**	0.882
AUC-ROC	0.910	**0.940**	0.920
Specificity	0.855	**0.882**	0.865
Tempo Treino (s)	**0.023**	0.156	0.342
Estabilidade (CV)	2.70%	**2.03%**	2.40%
Interpretabilidade	**Alta**	Média	Baixa

O resumo comparativo destaca que o Random Forest apresenta superioridade na maioria das métricas de desempenho, enquanto a Regressão Logística oferece vantagens em velocidade e interpretabilidade. O SVM destaca-se pela maior sensibilidade (recall), sendo preferível em cenários onde minimizar falsos negativos é crítico.

VI. Discussão

A. Interpretação dos Resultados

Os resultados demonstram que todos os três modelos apresentam desempenho satisfatório na predição de doenças cardíacas, com acurácias superiores a 85%. O Random Forest emergiu como o modelo de melhor desempenho, o que pode ser atribuído à sua capacidade de capturar interações complexas entre características através do ensemble de árvores de decisão.

A Regressão Logística, apesar de apresentar a menor acurácia entre os modelos testados, oferece vantagens significativas em termos de interpretabilidade. Os coeficientes do modelo podem ser diretamente interpretados como odds ratios, fornecendo insights clínicos valiosos sobre a relação entre características e probabilidade de doença cardíaca.

O SVM com kernel RBF demonstrou competência comparável ao Random Forest, especialmente na minimização de falsos negativos, o que é clinicamente relevante, pois reduz o risco de pacientes doentes não serem identificados.

B. Comparação entre Modelos

A comparação entre os modelos revela trade-offs importantes:

Interpretabilidade vs. Performance: A Regressão Logística oferece maior interpretabilidade, enquanto Random Forest e SVM priorizam performance preditiva.

Velocidade Computacional: A Regressão Logística é mais rápida no treinamento, seguida pelo Random Forest. O SVM pode ser computacionalmente mais custoso, especialmente com grandes volumes de dados.

Robustez a Ruído: Random Forest demonstra maior robustez a outliers e ruído devido à natureza do ensemble. SVM também apresenta boa robustez através do conceito de margem.

Balanceamento de Classes: Todos os modelos lidaram adequadamente com a distribuição relativamente balanceada do dataset. Em casos de desbalanceamento severo, técnicas adicionais como SMOTE ou ajuste de pesos seriam recomendadas.

C. Limitações e Observações

Algumas limitações do presente estudo devem ser consideradas:

1) **Dataset Limitado:** O tamanho do dataset (1025 instâncias) pode limitar a generalização dos resultados para populações diversas.
2) **Variáveis Consideradas:** O modelo utiliza apenas características clínicas e demográficas disponíveis no dataset, não incluindo outros fatores de risco potencialmente relevantes como histórico familiar detalhado ou fatores socioeconômicos.
3) **Contexto Clínico:** Os modelos foram desenvolvidos para auxiliar no diagnóstico, não para substituir a avaliação clínica profissional. A interpretação dos resultados deve sempre considerar o contexto clínico completo do paciente.
4) **Externalização:** A validação foi realizada apenas no dataset original. Validação externa em datasets independentes seria necessária para confirmar a generalização dos resultados.

D. Pontos Fortes e Fracos

Pontos Fortes:

- Implementação rigorosa de múltiplos algoritmos de ML
- Uso de métricas múltiplas para avaliação abrangente
- Validação cruzada para garantir robustez
- Análise de importância de características fornecendo insights clínicos
- Comparação sistemática entre modelos com diferentes características

Pontos Fracos:

- Ausência de validação externa em datasets independentes
- Não exploração de técnicas avançadas como deep learning
- Ausência de análise de custo-benefício clínico
- Não consideração de técnicas de seleção de características mais sofisticadas

VII. Conclusão

Este trabalho apresentou uma análise comparativa de três modelos de machine learning para predição de doenças cardíacas: Regressão Logística, Random Forest e Support Vector Machine. Os resultados demonstram que todos os modelos são capazes de realizar predições com acurácia superior a 85%, com o Random Forest apresentando o melhor desempenho geral (88,3% de acurácia).

A aplicação prática desses modelos pode auxiliar profissionais de saúde no processo de diagnóstico, especialmente em contextos onde recursos especializados são limitados. A análise de importância de características revelou que frequência cardíaca máxima, tipo de dor no peito e depressão ST são os fatores mais preditivos, alinhados com conhecimento clínico estabelecido.

Durante o desenvolvimento deste trabalho, foram aprendidas práticas importantes sobre pré-processamento de dados médicos, seleção e ajuste de modelos, e interpretação de resultados em contextos clínicos. A importância da validação cruzada e da avaliação através de múltiplas métricas foi evidenciada, garantindo robustez e confiabilidade nos resultados.

Para trabalhos futuros, sugere-se:

- Investigação de modelos mais complexos, como redes neurais profundas ou ensemble híbridos
- Incorporação de características adicionais, incluindo dados de imagens médicas ou marcadores genéticos
- Validação externa em datasets de diferentes populações e contextos geográficos
- Desenvolvimento de sistemas de apoio à decisão clínica integrando os modelos desenvolvidos
- Análise de custo-efetividade da implementação dessas ferramentas em contextos clínicos reais

- Exploração de técnicas de interpretabilidade avançada, como SHAP values, para melhor compreensão das decisões dos modelos

Os resultados deste estudo contribuem para o crescente corpo de conhecimento sobre aplicação de machine learning em diagnóstico médico, demonstrando o potencial dessas técnicas como ferramentas auxiliares na prática clínica.

AGRADECIMENTOS

Os autores agradecem ao UCI Machine Learning Repository por disponibilizar o dataset utilizado neste estudo e à comunidade científica que desenvolve e mantém as bibliotecas open-source utilizadas.

REFERENCES

[1] World Health Organization, "Cardiovascular diseases (CVDs)," 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, 2019.

[3] K. Shameer, K. W. Johnson, B. S. Glicksberg, J. T. Dudley, and P. P. Sengupta, "Machine learning in cardiovascular medicine: are we there yet?," Heart, vol. 104, no. 14, pp. 1156–1164, 2018.

[4] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY: Springer, 2009.

[5] C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," The Journal of Educational Research, vol. 96, no. 1, pp. 3–14, 2002.

[6] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[7] A. Liaw and M. Wiener, "Classification and regression by randomForest," R News, vol. 2, no. 3, pp. 18–22, 2002.

[8] V. N. Vapnik, The Nature of Statistical Learning Theory. New York, NY: Springer-Verlag, 1995.

[9] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121–167, 1998.

[10] UCI Machine Learning Repository, "Heart Disease Dataset," 2021. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease