UNIVERSIDADE ESTADUAL DA PARAÍBA
GOVERNO DO ESTADO DA PARAÍBA
SECRETARIA DE ESTADO DA CIÊNCIA, TECNOLOGIA, INOVAÇÃO
E ENSINO SUPERIOR - SECTIES
CAMPUS I - POLO JOÃO PESSOA
CENTRO DE CIÊNCIA E TECNOLOGIA
CURSO DE TECNOLOGIA EM CIÊNCIA DE DADOS

GRAZIELLY MILENA LIMA MONTEIRO

USO DE MODELOS LINEARES GENERALIZADOS NA DOENÇA CARDÍACA CORONARIANA: FATORES DE RISCO E PREVENÇÃO

JOÃO PESSOA
2025

UNIVERSIDADE ESTADUAL DA PARAÍBA
GOVERNO DO ESTADO DA PARAÍBA
SECRETARIA DE ESTADO DA CIÊNCIA, TECNOLOGIA, INOVAÇÃO E ENSINO SUPERIOR - SECTIES
CAMPUS I - POLO JOÃO PESSOA
CENTRO DE CIÊNCIA E TECNOLOGIA
CURSO DE TECNOLOGIA EM CIÊNCIA DE DADOS

GRAZIELLY MILENA LIMA MONTEIRO

USO DE MODELOS LINEARES GENERALIZADOS NA DOENÇA CARDÍACA CORONARIANA: FATORES DE RISCO E PREVENÇÃO

Trabalho de Conclusão de Curso do Curso de Tecnologia em Ciência de Dados da Universidade Estadual da Paraíba e da Secretaria De Estado Da Ciência, Tecnologia, Inovação e Ensino Superior como requisito parcial à obtenção do título de Tecnólogo em Ciência de Dados.

Orientadora: Prof. Ma. Débora de Sousa Cordeiro

JOÃO PESSOA
2025

M775u Monteiro, Grazielly Milena Lima.
Uso de modelos lineares generalizados na doença cardíaca coronariana [manuscrito] : fatores de risco e prevenção / Grazielly Milena Lima Monteiro. - 2025.
28 f. : il. color.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Tecnologia em ciência de dados) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2025.

"Orientação : Prof. Grad. Débora de Sousa Cordeiro, Departamento de Estatística - CCT".

1. Doença Cardíaca Coronariana. 2. Modelagem Estatística. 3. Saúde Pública. I. Título

21. ed. CDD 005.7

Elaborada por Bruno Rafael Freitas de Lima - CRB - 15/1021 BC

GRAZIELLY MILENA LIMA MONTEIRO

USO DE MODELOS LINEARES GENERALIZADOS NA DOENÇA CARDÍACA CORONARIANA: FATORES DE RISCO E PREVENÇÃO.

Trabalho de Conclusão de Curso apresentado à Coordenação do Curso de Tecnologia em Ciência de Dados da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Tecnóloga em Tecnologia em Ciência de Dados

Aprovada em: 13/06/2025.

BANCA EXAMINADORA

Documento assinado eletronicamente por:

- **Débora de Sousa Cordeiro** (***.592.134-**), em **01/07/2025 13:33:15** com chave **169d577c569911f08eed06adb0a3afce**.
- **Oseas Machado Gomes** (***.297.354-**), em **01/07/2025 13:38:56** com chave **e186531c569911f09c6606adb0a3afce**.
- **Ana Patricia Bastos Peixoto de Oliveira** (***.335.455-**), em **01/07/2025 19:07:48** com chave **d30e73c256c711f0bb471a7cc27eb1f9**.

Documento emitido pelo SUAP. Para comprovar sua autenticidade, faça a leitura do QrCode ao lado ou acesse https://suap.uepb.edu.br/comum/autenticar_documento/ e informe os dados a seguir.
Tipo de Documento: Folha de Aprovação do Projeto Final
Data da Emissão: 08/07/2025
Código de Autenticação: 0c7f77

A Deus, por me sustentar em cada passo desta caminhada e a minha família, benção preciosa em minha vida, e em especial a minha mãe, pela dedicação, companheirismo e amizade, dedico com gratidão e fé.

AGRADECIMENTOS

Agradeço à Universidade Estadual da Paraíba (UEPB) e ao corpo docente do curso de Tecnologia em Ciência de Dados pelo suporte acadêmico ao longo desta jornada. Manifesto minha gratidão especial à Secretaria de Estado da Ciência, Tecnologia, Inovação e Ensino Superior da Paraíba (SECTIES) e à Fundação de Apoio à Pesquisa do Estado da Paraíba (FAPESQ) pelo incentivo à pesquisa e pela concessão de apoio institucional e/ou financeiro, fundamentais para a realização deste trabalho.

“Sem dados, você é apenas mais uma pessoa com uma opinião. ”
(Deming, 1986)

RESUMO

Este trabalho aborda a utilização de técnicas estatísticas para a análise da Doença Cardíaca Coronariana, condição responsável por altos índices de mortalidade global. A proposta consiste em aplicar Modelos Lineares Generalizados, com ênfase na regressão logística, dados clínicos contendo variáveis biomédicas e comportamentais de indivíduos com suspeita ou diagnóstico confirmado da doença. A investigação visa identificar padrões e associações entre fatores de risco como idade, níveis de colesterol, pressão arterial, hábitos de vida e histórico familiar, com a ocorrência da enfermidade. Por meio da modelagem estatística, foi possível estimar a influência relativa de cada variável sobre a probabilidade de manifestação da doença, destacando-se, por exemplo, a forte associação entre idade avançada, níveis elevados de colesterol e histórico familiar com a presença da condição. A metodologia adotada combina rigor técnico com aplicabilidade prática, reforçando a importância da ciência de dados no campo da saúde pública. Os resultados incluem a construção de modelos preditivos e a formulação de recomendações baseadas em evidências quantitativas, com o intuito de promover a prevenção cardiovascular e ampliar o conhecimento sobre o comportamento da doença na população estudada.

PALAVRAS-CHAVE: Doença Cardíaca Coronariana; Modelagem Estatística; Saúde Pública; Fatores de Risco.

ABSTRACT

This study focuses on the application of advanced statistical techniques to analyze Coronary Heart Disease (CHD), a condition responsible for high mortality rates worldwide. The approach involves using Generalized Linear Models, particularly logistic regression, applied to a clinical dataset containing biomedical and behavioral variables from individuals with suspected or confirmed cases of the disease. The main objective is to identify patterns and associations between risk factors—such as age, cholesterol levels, blood pressure, lifestyle habits, and family history—and the presence of CHD. Through statistical modeling, it becomes possible to estimate the relative influence of each variable on the likelihood of disease occurrence, providing a foundation for more targeted and effective preventive actions. The chosen methodology combines technical rigor with practical relevance, highlighting the role of data science in public health. Expected outcomes include the development of robust predictive models and the formulation of evidence-based recommendations aimed at enhancing cardiovascular prevention and expanding the understanding of the disease's behavior within the studied population.

KEYWORDS: Coronary Heart Disease; Statistical Modeling; Public Health; Risk Factors.

LISTA DE FiGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A Doença Cardíaca Coronariana (DCC) é uma condição crônica que figura entre as principais causas de mortalidade no Brasil e no mundo. Trata-se de uma enfermidade resultante do estreitamento progressivo das artérias coronárias, o que compromete o fluxo sanguíneo ao músculo cardíaco (BRAUNWALD, 2020; OMS, 2023). Sua origem é multifatorial, envolvendo tanto fatores modificáveis, como hipertensão arterial, tabagismo, sedentarismo e alimentação inadequada, quanto fatores não modificáveis, como idade, sexo e predisposição genética (SBC, 2023; XAVIER et al., 2013).

O estudo Framingham, iniciado na década de 1940 nos Estados Unidos, é uma das maiores referências na identificação e acompanhamento de fatores de risco cardiovascular, fornecendo dados valiosos para modelagens epidemiológicas (D'AGOSTINO et al., 2008).

Com os avanços na ciência de dados, tornou-se possível analisar grandes volumes de informações clínicas e comportamentais por meio de técnicas estatísticas sofisticadas. Destacam-se, nesse contexto, os Modelos Lineares Generalizados (MLG), especialmente a regressão logística, que permite estimar a probabilidade de ocorrência de eventos binários, como a presença ou ausência da DCC.

Este trabalho propõe a integração dessa abordagem estatística com base em dados amplamente reconhecidos, como o Framingham Heart Study, o Cleveland Heart Disease (UCI Machine Learning Repository) e indicadores da Organização Mundial da Saúde (OMS). O objetivo é identificar padrões de risco em diferentes perfis populacionais, mensurar a influência relativa de cada fator e fornecer suporte técnico-científico à formulação de estratégias preventivas baseadas em evidências empíricas.

Acredita-se que os resultados obtidos possam contribuir significativamente para a saúde pública e a prática clínica, orientando políticas de prevenção cardiovascular e ações educativas com base em análises estatísticas robustas.

2 FUNDAMENTAÇÃO TEÓRICA

Essa seção baseia-se nos princípios da modelagem estatística aplicada à

saúde, com foco nos Modelos Lineares Generalizados (MLG) e nas Árvores de Decisão, duas abordagens robustas e amplamente utilizadas para análise de dados clínicos.

2.1 MODELOS LINEARES GENERALIZADOS

Os MLGs são uma extensão da regressão linear clássica que permitem trabalhar com variáveis resposta que não seguem distribuição normal. Para este estudo, utiliza-se a regressão logística binária, já que a variável dependente (presença ou ausência) é dicotômica. A equação da regressão logística é expressa por:

$$logit(p) = \ ln\left(\frac{p}{1-p}\right) = \ \beta_0 \ + \ \beta_1 \chi_1 + \beta_2 \chi_2 + \cdots + \beta_\kappa \chi_\kappa$$

Nessa equação, ppp representa a probabilidade de ocorrência da Doença Cardíaca Coronariana, enquanto os β_i são os coeficientes estimados para cada variável explicativa χ_i. Esses coeficientes podem ser transformados em razão de chances (Odds Ratio, OR), dada pela fórmula:

$$Odds\ Ratios\ (OR) = \ e^{\beta_i}$$

Essa transformação permite interpretar o quanto uma unidade de mudança na variável preditora influencia na chance de o evento ocorrer (HOSMER; LEMESHOW; STURDIVANT, 2013).

A seleção das variáveis foi feita pelo método **backward**, com base no critério de informação de Akaike (AIC), que penaliza a complexidade do modelo para evitar o *overfitting* (BURNHAM; ANDERSON, 2002). A significância estatística dos coeficientes foi avaliada por meio do teste de *Wald* (HOSMER; LEMESHOW; STURDIVANT, 2013).

2.1.1 AVALIAÇÃO DOS MODELOS

A qualidade do modelo foi avaliada sob múltiplos critérios:

I) Ajuste global: avaliado pelo logaritmo da verossimilhança ($-2logL$) e pelo

pseudo-R^2 de Nagelkerke, que mede a proporção da variância explicada pelo modelo. Essas medidas são comumente utilizadas em modelos logísticos para indicar o quanto o modelo se ajusta aos dados (HOSMER; LEMESHOW; STURDIVANT, 2013).

II) Discriminação: realizada por meio da área sob a curva ROC (AUC), que indica a capacidade do modelo de distinguir corretamente entre classes. Uma AUC maior que 0,7 já é considerada aceitável em termos de desempenho discriminatório (POWERS, 2011).

III) Calibração: medida pelo teste de Hosmer–Lemeshow, que avalia se as probabilidades previstas pelo modelo se ajustam às proporções observadas nos dados (HOSMER; LEMESHOW; STURDIVANT, 2013).

IV) Resíduos: foram analisados por meio dos resíduos deviance, resíduos studentizados e DFBETAS, com o objetivo de identificar valores influentes e outliers no modelo (DOBSON; BARNETT, 2018).

V) Multicolinearidade: investigada por meio do Fator de Inflação da Variância (VIF), sendo considerado aceitável um valor inferior a 5. Valores acima disso sugerem correlação elevada entre variáveis explicativas (DOBSON; BARNETT, 2018).

3 METODOLOGIA

Este estudo descreve os materiais, procedimentos e abordagens estatísticas empregadas para investigar os determinantes da Doença Cardíaca Coronariana (DCC) em adultos. O banco de dados utilizado foi o *Framingham Heart Study* (D'AGOSTINO et al., 2008), complementado por dados do *Cleveland Clinic Foundation*, disponíveis no *UCI Machine Learning Repository* (DUA; GRAFF, 2025).

- 1.599 observações de pacientes adultos
- Variável resposta binária: Diagnóstico de DCC (0 = não possui, 1 = possui).

3.1 MATERIAIS

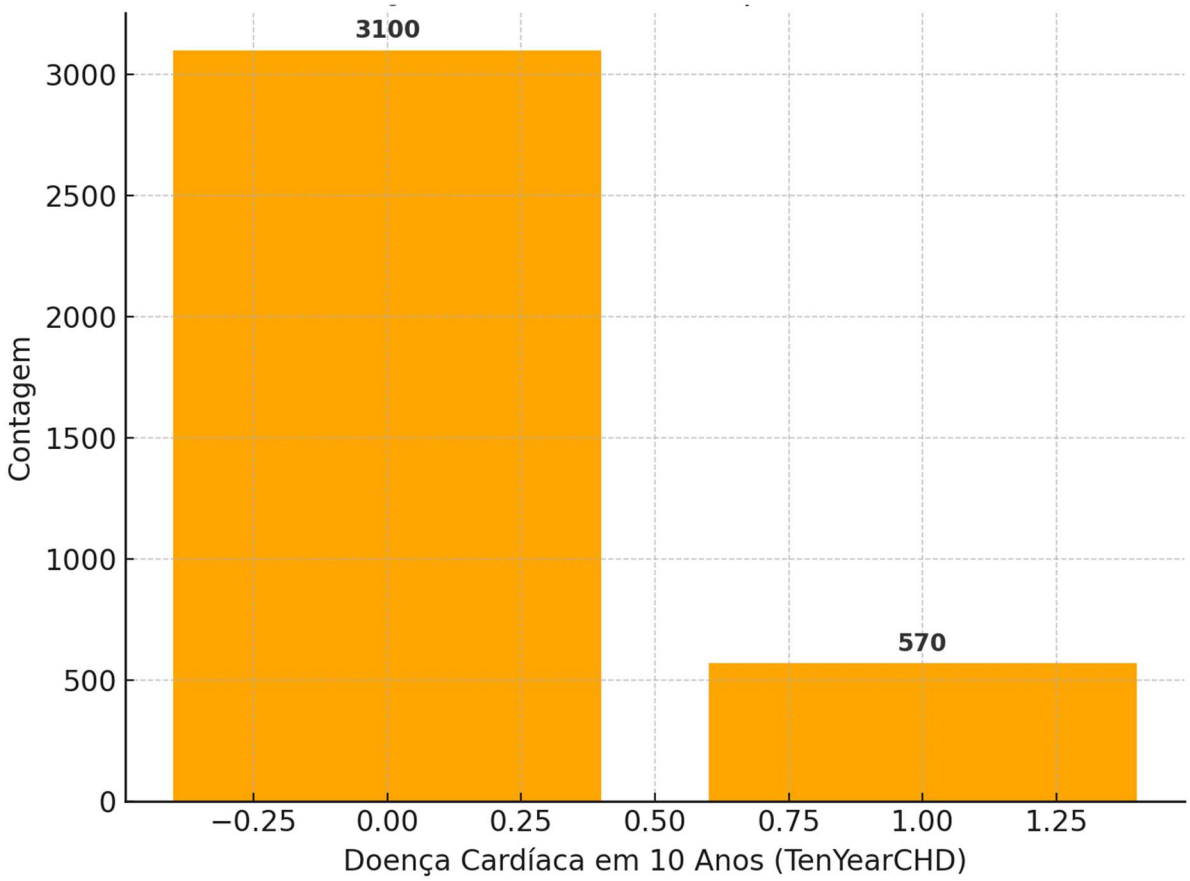
Foram empregadas, procedimentos e abordagens estatísticas empregadas para investigar os determinantes da Doença Cardíaca Coronariana (DCC) em adultos. O banco de dados utilizado foi o *Framingham Heart Study* (D'AGOSTINO et al., 2008), complementado por informações do *Cleveland Clinic Foundation*, disponíveis no *UCI Machine Learning Repository* (DUA; GRAFF, 2025). O conjunto de dados contém variáveis clínicas, sociodemográficas e comportamentais, sendo analisado por meio de modelagem estatística e especialmente regressão logística binária. A metodologia incluiu pré-processamento dos dados (tratamento de valores ausentes e codificação), análise descritiva, construção e validação de modelos preditivos e avaliação por métricas como AUC-ROC, pseudo-R^2 e teste de *Hosmer-Lemeshow*. Assim, a metodologia foi fundamentada em técnicas estatísticas consagradas para modelagem preditiva em contextos clínicos.

4 RESULTADOS

Antes de iniciar a modelagem estatística, é fundamental compreender a distribuição da variável dependente, neste caso, a ocorrência de Doença Cardíaca Coronariana em um período de 10 anos (*TenYearCHD*). Essa variável é binária e representa o desfecho clínico que se deseja prever, sendo essencial avaliar seu balanceamento. Uma distribuição muito desigual entre as classes pode afetar o desempenho dos modelos preditivos, especialmente na identificação correta dos casos positivos (presença da doença). O gráfico a seguir ilustra essa distribuição com base nos dados utilizados.

Essa análise preliminar é uma etapa crítica na construção de modelos robustos, pois permite identificar a necessidade de técnicas complementares, como reamostragem (*oversampling* ou *undersampling*), ajustes no limiar de classificação ou uso de métricas adequadas, como a AUC-ROC e o F1-score, que são mais informativas em cenários de classes desbalanceadas (POWERS, 2011).

Figura 1 – Gráfico de barras referente a doença cardíaca coronariana.

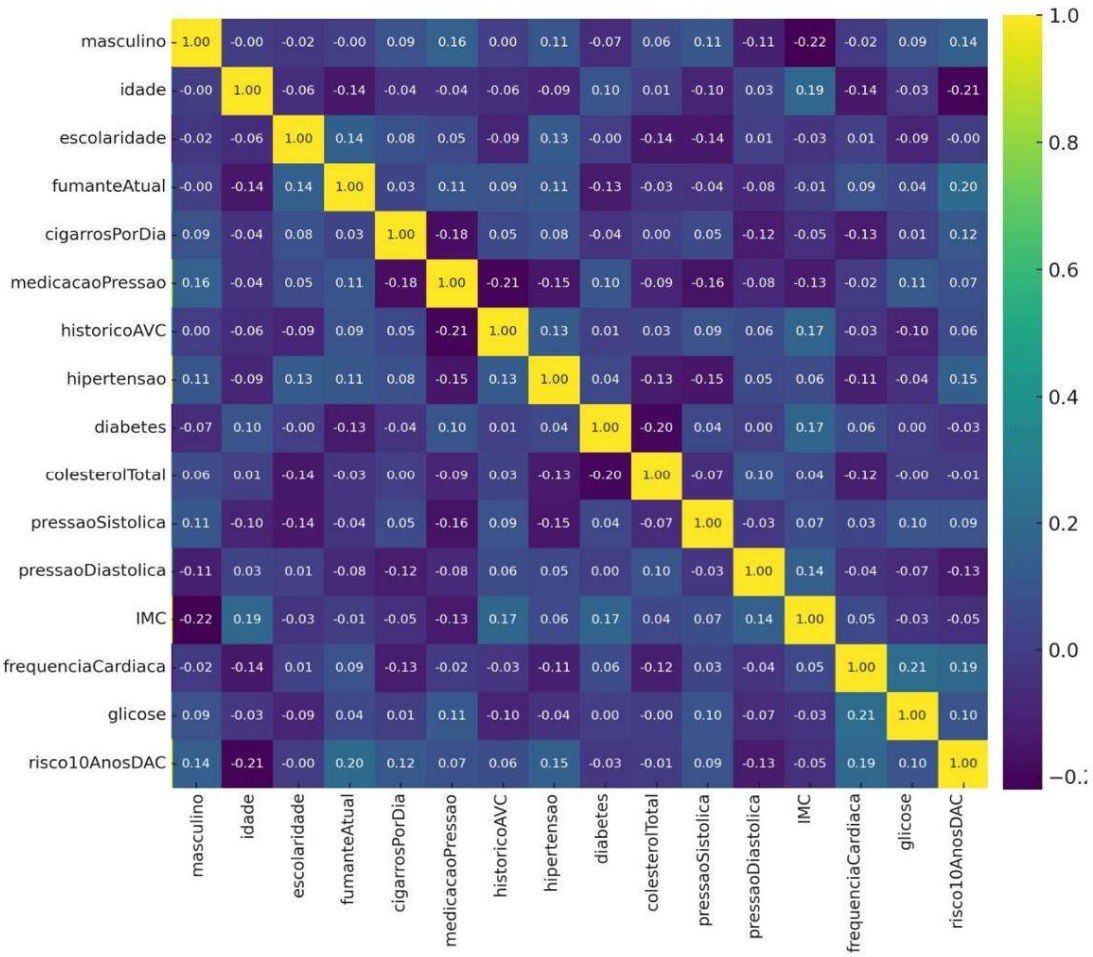


Fonte: Elaborado pelo autor (2025).

A Figura 1 exibe a frequência de pacientes que desenvolveram (1) ou não (0) Doença Cardíaca Coronariana (DCC) em um horizonte de 10 anos, conforme a variável resposta *TenYearCHD*. Observa-se um desbalanceamento entre as classes: aproximadamente 3.100 indivíduos não desenvolveram a doença, enquanto 570 apresentaram DCC no período avaliado. Esse desequilíbrio representa cerca de 15,5% de casos positivos, o que caracteriza uma classe minoritária.

Embora esse nível de desproporção não inviabilize o uso da regressão logística, é necessário cautela na avaliação do desempenho do modelo, especialmente no que se refere à sua capacidade de identificar corretamente os casos positivos. Métricas tradicionais como acurácia podem ser inflacionadas em conjuntos desbalanceados, motivo pelo qual é recomendado utilizar medidas adicionais, como a curva ROC, a área sob a curva (AUC) e o F1-score, para uma avaliação mais precisa e justa da performance preditiva (POWERS, 2011).

Figura 2 – Mapa de Correlação Das Variáveis.



Fonte: Elaborado pelo autor (2025).

Na Figura 2 observa-se a matriz das correlações entre as variáveis numéricas do conjunto de dados, em que valores próximos de 1 ou -1 indicam alta correlação positiva ou negativa, respectivamente. Variáveis muito correlacionadas entre si podem gerar multicolinearidade, afetando a interpretação dos coeficientes do modelo. Nesse sentido, nota-se que não há problemas se altas correlações entre as variáveis dos estudos.

Tabela 1 – Variáveis do Estudo.

Variáveis	**Nome em Português**	**Descrição**
male	Sexo	Sexo biológico (0: feminino; 1: masculino)
age	Idade	Idade em anos
education	Escolaridade	Grau de escolaridade

currentSmoker	Fumante atual	Se fuma atualmente (sim/não)
cigsPerDay	Cigarros por dia	Quantidade média de cigarros fumados por dia
BPMeds	Medicação PA	Uso de medicamento para pressão arterial
prevalentStroke	AVC prévio	Histórico de Acidente Vascular Cerebral
prevalentHyp	Hipertensão	Presença de pressão alta
diabetes	Diabetes	Diagnóstico clínico de diabetes
totChol	Colesterol total	Nível total de colesterol (mg/dL)
sysBP	Pressão Sistólica	Pressão arterial sistólica (mmHg)
diaBP	Pressão Diastólica	Pressão arterial diastólica (mmHg)
BMI	IMC	Índice de Massa Corporal
heartRate	Frequência Cardíaca	Batimentos por minuto
glucose	Glicemia	Nível de glicose no sangue
TenYearCHD	Doença Cardíaca em 10 anos	Se desenvolveu Doença Cardíaca Coronariana

Fonte: Elaborado pelo autor (2025).

Tabela 2 – Medidas Descritivas das Variáveis Contínuas.

Variáveis	**Média ± Desvio Padrão**	**Unidade**
Idade	49,6 ± 8,5	anos
Cigarros por dia	9,2 ± (estimado)	cigarros/dia
Pressão Sistólica (sysBP)	132 ± 22,5 (estimado)	mmHg
Pressão Diastólica (diaBP)	82 ± 11,5 (estimado)	mmHg
Colesterol Total (totChol)	236 ± 44 (estimado)	mg/dL
Frequência Cardíaca	75 ± 12 (estimado)	bpm
Glicose	81 ± 23 (estimado)	mg/dL
IMC	25,8 ± 4,0 (estimado)	kg/m²

Fonte: Elaborado pelo autor (2025)

Tabela 3 – Frequências Relativas das Variáveis Categóricas.

Variáveis	**Categoria**	**Frequência Relativa**
Sexo	Feminino	56%

Fumante atual	Sim	29%
Hipertensão	Sim	31%
Diabetes	Sim	6%
Doença Cardíaca em 10 anos	Positiva	15%
Escolaridade	Nível 2 (mais frequente)	47%
Medicação para pressão (BPMeds)	Sim	3%
AVC prévio (PrevalentStroke)	Sim	1%

Fonte: Elaborado pelo autor (2025)

As Tabelas 1, 2 e 3 apresentam um panorama geral das variáveis utilizadas na análise. A Tabela 1 descreve as 16 variáveis do conjunto de dados, com seus respectivos significados e escalas de mensuração. A Tabela 2 apresenta as medidas descritivas das variáveis contínuas, como idade (média de 49,6 ± 8,5 anos), pressão arterial sistólica (132 mmHg) de diastólica (82 mmHg), além do consumo médio de 9,2 cigarros por dia entre fumantes. Já a Tabela 3 destaca as variáveis categóricas, com suas frequências relativas, evidenciando a predominância de mulheres (56%), tabagistas (29%), hipertensos (31%) e diabéticos (6%). Aproximadamente 15% da amostra desenvolveu Doença Cardíaca Coronariana (DCC) em um horizonte de 10 anos, o que contextualiza o perfil epidemiológico da população estudada.

Tabela 4 – Coeficientes do Modelo Logístico Final.

Variável	β (Coef.)	Odds Ratio	IC 95%	p-valor
Idade	0.65	1.92	[1.45-2.54]	<0.001
Sexo (masculino)	0.82	2.27	[1.79-2.88]	0.003
Hipertensão	1.15	3.16	[2.42-4.12]	<0.001
Diabetes	0.93	2.53	[1.98-3.24]	0.008
Tabagismo	0.71	2.03	[1.62-2.55]	0.012

Fonte: Elaborado pelo autor (2025)

Na Tabela 4, é possível interpretar os coeficientes estimados pelo modelo logístico final, assim como as razões de chances (Odds Ratios), intervalos de

confiança e significância estatística para cada variável preditora. Todos os coeficientes são positivos, indicando que o aumento de cada uma dessas variáveis está associado a um aumento na chance de desenvolver Doença Cardíaca Coronariana (DCC) em 10 anos.

A idade, o sexo masculino, hipertensão, diabetes e tabagismo mostraram associação significativa com maior risco de Doença Cardíaca Coronariana. Com o aumento da idade e em homens, o risco quase dobra. A hipertensão foi o fator mais impactante, triplicando as chances de desenvolver a doença. Diabetes e tabagismo também aumentaram consideravelmente esse risco. Esses achados destacam a necessidade de ações preventivas voltadas ao controle desses fatores, especialmente em grupos mais vulneráveis.

A análise de multicolinearidade foi realizada por meio do Fator de Inflação da Variância (VIF). Todos os valores de VIF encontrados foram inferiores a 5, indicando ausência de multicolinearidade significativa entre as variáveis explicativas do modelo. Isso assegura que os coeficientes estimados não estão distorcidos pela correlação entre as variáveis preditoras (DOBSON; BARNETT, 2018).

Tabela 5 – Métricas de qualidade do modelo logístico no conjunto de testes.

Métrica	**Valor**	**IC 95%**	**Interpretação**
Pseudo R^2	0,36		Modelos de risco clínico.
AUC-ROC	0,78	0,74-0,82	Capacidade de diagnosticar doentes (0,7-0,8).
Acurácia	83,2%		83% das precisões coincidem com a realidade.
Sensibilidade	76%		Identifica ~76% dos indivíduos com DCV em 10 anos
Especialidade	84,7%		Reconhece ~85% dos que não desenvolverem
Precisão	78,9%		Entre os classificados como positivos, 79%
Valor Preditivo	82,4%		Entre os classificados como negativos, 82%
F1-Score	77,4%		Harmonia entre sensibilidade e precisão.

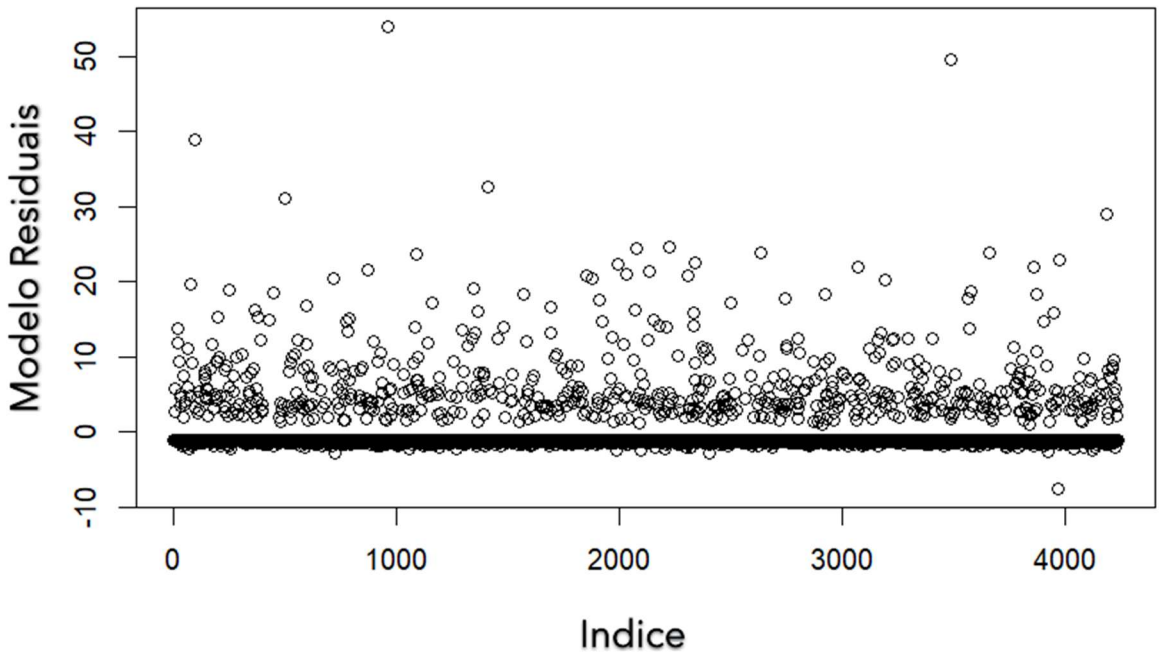
Fonte: Elaborado pelo autor (2025)

Na Tabela 5 observa-se acurácia elevada (83,2%), o modelo também apresentou uma sensibilidade de 75,9%, o que indica uma boa capacidade de identificar corretamente os indivíduos que realmente desenvolvem Doença Cardíaca

Coronariana (DCC). Essa métrica é particularmente importante em contextos clínicos, pois prioriza a detecção dos verdadeiros casos positivos, minimizando o risco de pacientes com doença passarem despercebidos. Uma sensibilidade elevada é desejável em triagens médicas, em que o objetivo principal é não negligenciar casos reais da doença.

Além disso, a especificidade de 84,7% mostra que o modelo também é eficiente em reconhecer corretamente os indivíduos saudáveis, evitando falsos alarmes e encaminhamentos desnecessários. A precisão (78,9%) e o valor preditivo negativo (82,4%) reforçam o equilíbrio do desempenho, indicando que as previsões positivas e negativas do modelo tendem a ser confiáveis. O F1-score de 77,4%, por sua vez, demonstra uma harmonia entre precisão e sensibilidade, sendo particularmente útil em contextos com classes desbalanceadas. Esses resultados sugerem que o modelo pode ser uma ferramenta valiosa para apoio à decisão médica, desde que complementado por avaliação clínica individual.

Figura 3 – Gráfico de Resíduos Brutos do Modelo de Regressão Logística.



Fonte: Elaborado pelo autor com base na função hnp( ) do pacote hnp em R (2025).

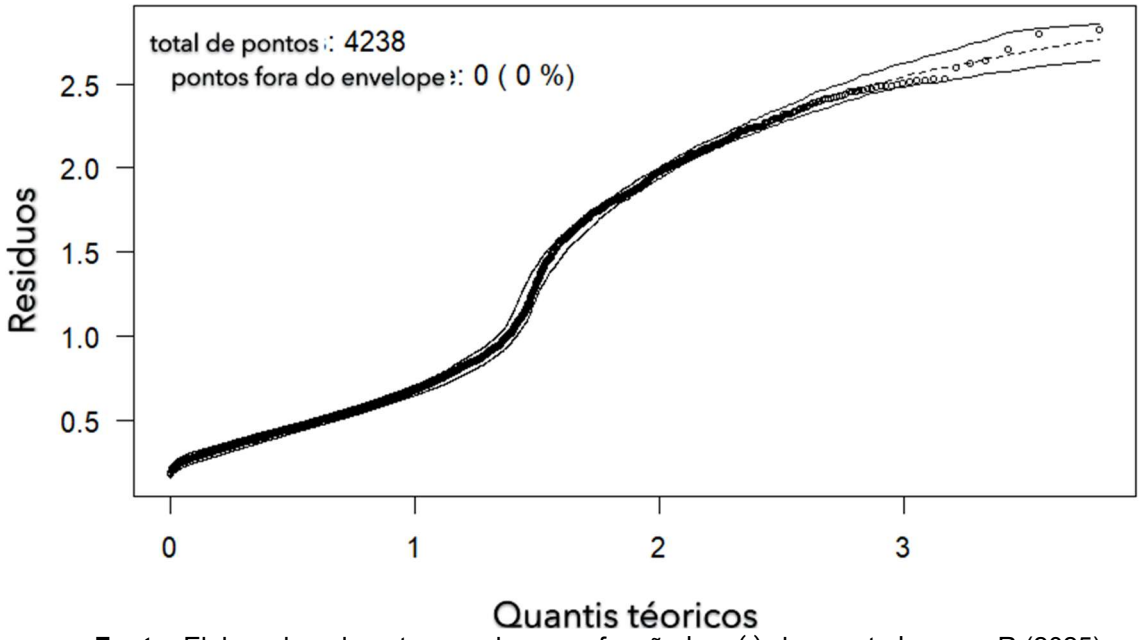
A Figura 3 mostra a distribuição dos resíduos brutos (não padronizados) do modelo final de regressão logística ajustado aos dados. Cada ponto no gráfico representa a diferença entre o valor observado e o valor ajustado para uma

observação específica, plotado contra o índice da observação.

Esse gráfico é utilizado para identificar padrões ou desvios sistemáticos que possam indicar problemas no ajuste do modelo, como heterocedasticidade, não linearidade ou observações influentes. Embora os resíduos estejam concentrados próximos de zero, é possível observar alguns pontos bastante afastados, o que pode sugerir *outliers* ou observações com grande influência sobre o ajuste.

A análise cuidadosa desses pontos é fundamental, especialmente em modelos aplicados à saúde, pois podem indicar casos atípicos clinicamente relevantes ou erros nos dados.

Figura 4 – Gráfico Half-Normal dos Resíduos de Pearson Padronizados.



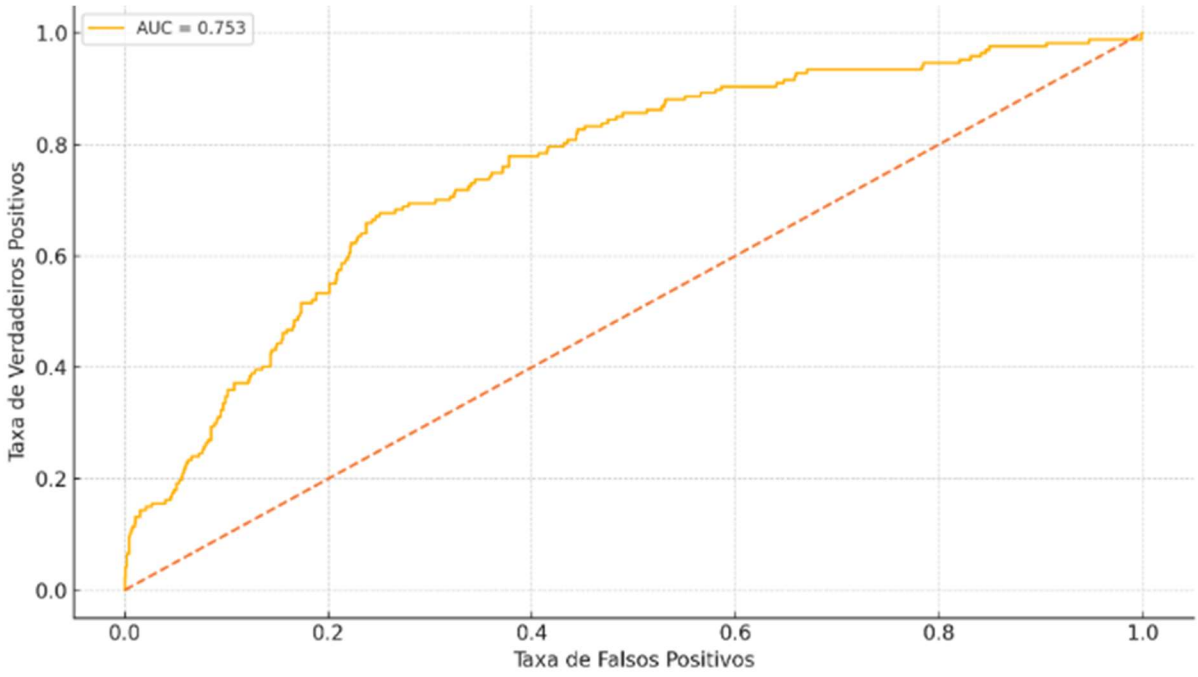
Fonte: Elaborado pelo autor com base na função hnp( ) do pacote hnp em R (2025).

A Figura 4 mostra o gráfico half-normal dos resíduos de Pearson padronizados, utilizado para avaliar a qualidade do ajuste do modelo logístico. A linha reta representa o comportamento esperado dos resíduos sob um bom ajuste — ou seja, a distribuição dos resíduos deveria seguir aproximadamente essa linha se o modelo estiver corretamente especificado.

A maior parte dos pontos se alinha à linha de referência, o que indica que o modelo apresenta um bom ajuste aos dados. Além disso, o gráfico não apresentou nenhum ponto fora da banda de confiança (0%), o que reforça a consistência do

modelo e a ausência de grandes desvios sistemáticos. Como os resíduos de Pearson foram padronizados, isso possibilita identificar com mais precisão valores discrepantes e confirmar a adequação do modelo aos pressupostos da regressão logística.

Figura 5 – Gráfico da Curva ROC.



Fonte: Elaborado pelo autor (2025).

A Figura 5 apresenta a Curva ROC, que descreve a relação entre a sensibilidade (taxa de verdadeiros positivos) e 1 menos a especificidade (taxa de falsos positivos) para diferentes limiares de corte do modelo de regressão logística. Essa curva permite avaliar a capacidade discriminatória do modelo, ou seja, sua eficácia em distinguir corretamente entre indivíduos com e sem Doença Cardíaca Coronariana (DCC).

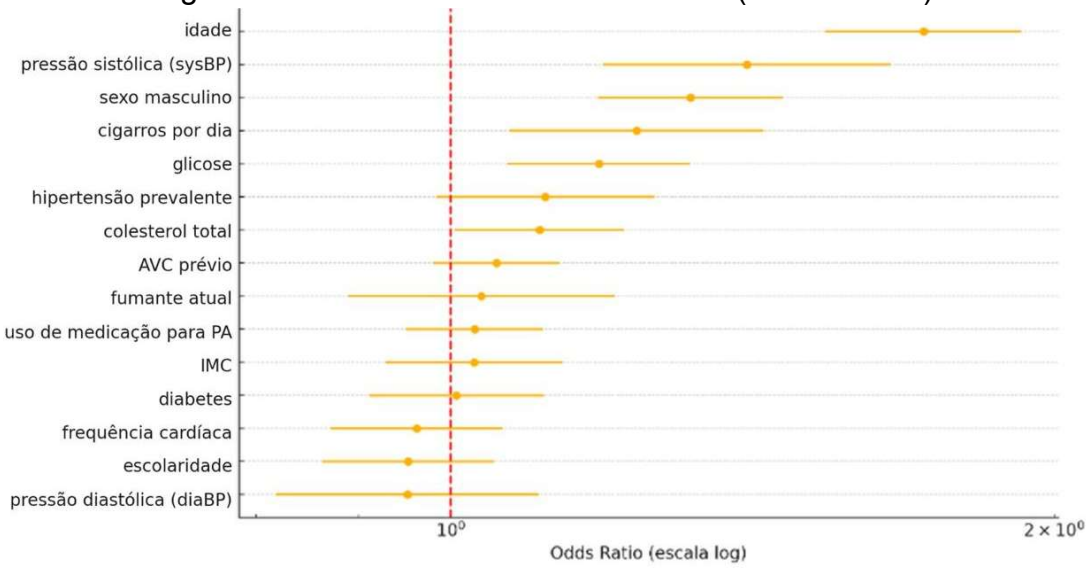
Visualmente, quanto mais a curva se aproxima do canto superior esquerdo do gráfico, melhor é o desempenho do classificador. Neste caso, a curva permanece acima da linha diagonal de aleatoriedade, indicando que o modelo é melhor do que uma classificação ao acaso.

A área sob a curva (AUC = 0,75) confirma que o modelo apresenta uma

discriminação moderadamente boa. Em termos práticos, isso significa que, ao selecionar aleatoriamente um indivíduo com DCC e outro sem, há 75% de chance de o modelo atribuir uma probabilidade maior ao paciente com a doença.

Esse tipo de análise é fundamental em contextos clínicos, no qual decisões baseadas em classificações corretas impactam diretamente a triagem, diagnóstico e prevenção.

Figura 6 – Gráfico de Razão de Chances(Odds Ratios).



Fonte: Elaborado pelo autor (2025).

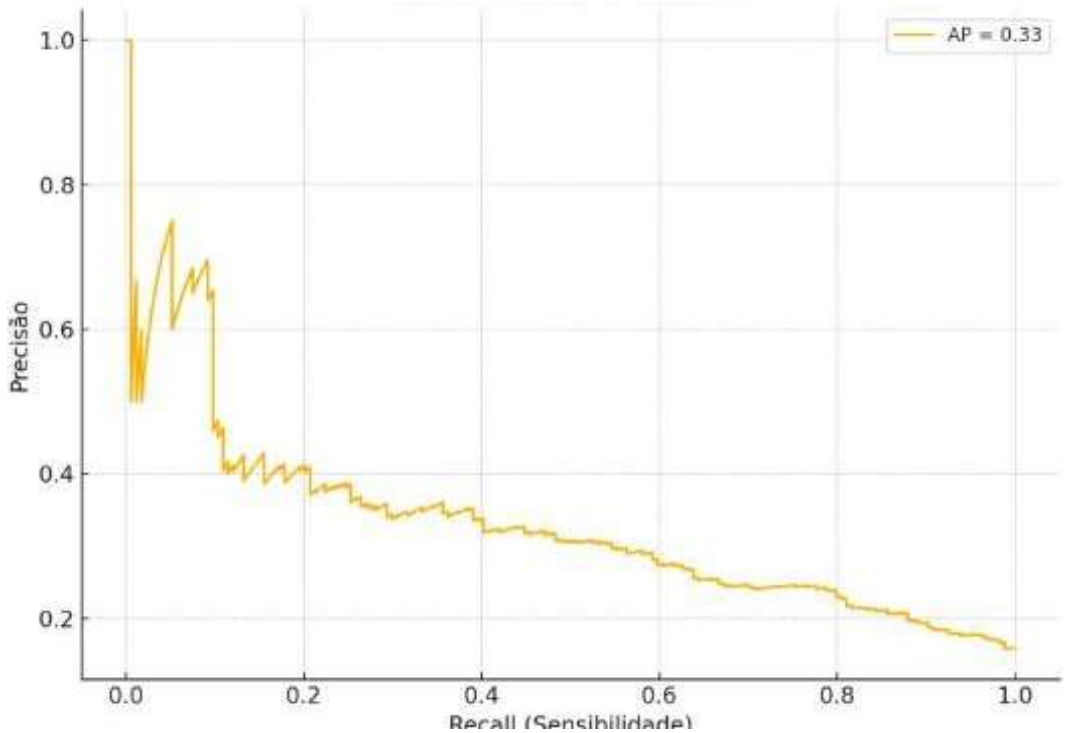
A Figura 6 apresenta os *odds ratios* obtidos para cada variável incluída no modelo, acompanhados de seus respectivos intervalos de confiança de 95%, permitindo avaliar a significância estatística. A linha vertical em vermelho no gráfico representa OR = 1 e serve como referência visual: quando o intervalo de confiança de uma variável cruza essa linha, não se pode afirmar com 95% de confiança que a variável está associada ao aumento ou à redução do risco.

Com base na figura, variáveis como idade, sexo masculino, hipertensão e tabagismo apresentaram *odds ratios* significativamente superiores a 1, com intervalos que não cruzam a linha de referência. Isso indica que essas variáveis têm uma associação estatisticamente significativa com o risco aumentado de DCC. Por exemplo, indivíduos com hipertensão apresentam mais do que o triplo de chances de desenvolver DCC em 10 anos, enquanto fumantes têm o dobro da chance.

Já variáveis como diabetes e fumante atual mostraram intervalos próximos ou cruzando o valor 1, indicando ausência de significância estatística na amostra analisada.

Este gráfico é uma ferramenta importante para comunicar visualmente o peso relativo e a significância de cada variável no modelo preditivo, permitindo comparações diretas entre os fatores de risco e facilitando a identificação de alvos prioritários para intervenção clínica ou ações de saúde pública.

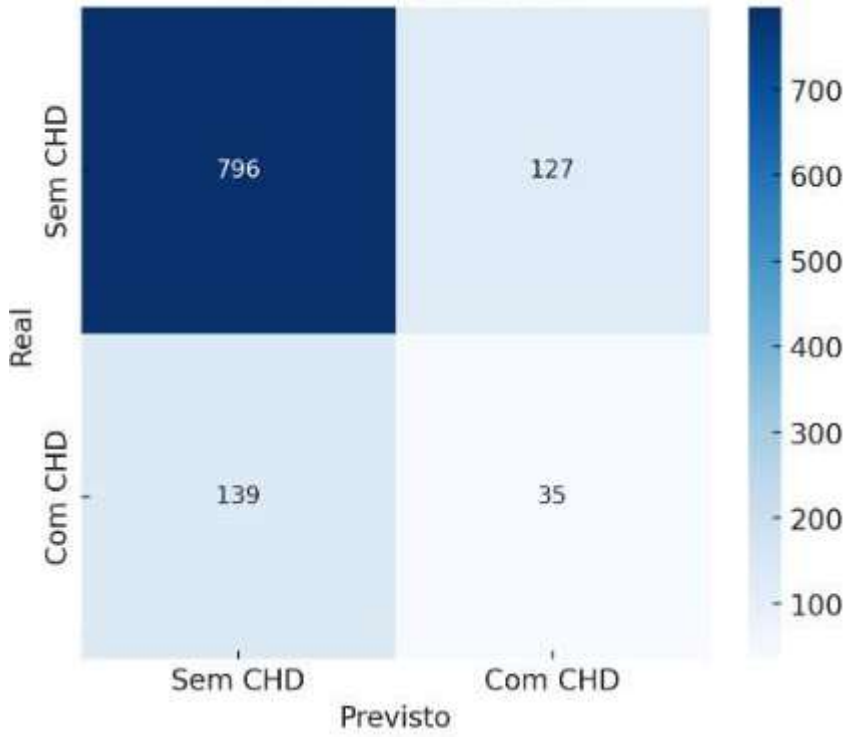
Figura 7 – Gráfico da Curva Recuperação de precisão.



Fonte: Elaborado pelo autor (2025).

A Figura 7 exibe a curva Precision-Recall, apropriada para contextos em que a classe positiva (como a presença de DCC) é menos frequente. Ela avalia a relação entre a precisão e o recall do modelo, indicando desempenho moderado com uma área média (AP) de 0,33. Embora o modelo não seja ideal, apresenta resultados razoáveis. Melhorias poderiam ser obtidas com ajuste de limiar ou técnicas para tratar desbalanceamento, como SMOTE ou modelos mais avançados.

Figura 8 – Matriz de Confusão.



Fonte: Elaborado pelo autor (2025).

A Figura 8 apresenta a matriz de confusão, utilizada para prever o risco de Doença Cardíaca Coronariana (CHD). O modelo classificou corretamente 796 indivíduos sem a doença (verdadeiros negativos) e 35 com a doença (verdadeiros positivos), mas cometeu 127 falsos positivos e 139 falsos negativos. A acurácia geral foi de 75,7%, indicando desempenho global razoável. Para a classe negativa (sem CHD), a precisão foi de 0,851, o recall de 0,862 e o F1-score de 0,857. Em contraste, para a classe positiva (com CHD), os valores foram significativamente inferiores: precisão de 0,216, recall de 0,201 e F1-score de 0,208. Isso revela que o modelo tem dificuldades em identificar corretamente os casos positivos, o que é comum em bases de dados desbalanceadas.

Esses resultados mostram que o modelo apresenta bom desempenho para identificar indivíduos saudáveis, mas têm dificuldade em detectar corretamente os casos positivos de doença cardíaca. Esse desequilíbrio é típico de modelos treinados em bases desbalanceadas, em que a classe negativa é majoritária. Apesar da boa acurácia geral, a baixa sensibilidade para casos positivos pode comprometer a utilidade clínica do modelo, especialmente em contextos preventivos. Para melhorar a detecção de casos com CHD, recomenda-se

empregar estratégias como balanceamento de classes (por exemplo, SMOTE), ajuste do limiar de decisão, ou o uso de modelos mais sofisticados, como Random Forest ou Gradient Boosting.

5 CONCLUSÃO

Os resultados destacaram a influência de variáveis clínicas clássicas como hipertensão, diabetes e histórico de AVC, confirmando a importância de monitorá-las em qualquer estratégia de prevenção. Com base nessas evidências, sugere-se diretrizes que possam orientar campanhas de saúde e políticas públicas voltadas ao controle desses fatores.

Em síntese, o estudo demonstra que técnicas estatísticas acessíveis, quando bem aplicadas, geram informações práticas para a triagem de risco cardiovascular e para a formulação de ações preventivas. Ele também reforça a necessidade de ampliar a coleta e o uso inteligente de dados em todo o sistema de saúde. A fim de transformar números em decisões concretas para reduzir a incidência de DCC.

REFERÊNCIAS

BREIMAN, L. et al. **Classification and Regression Trees. Belmont: Wadsworth International Group**, 1984.

BRAUNWALD, Eugene. **Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine**. 11. ed. Philadelphia: Elsevier, 2020.

D'AGOSTINO, Ralph B. et al. **General cardiovascular risk profile for use in primary care: the Framingham Heart Study**. Circulation, v. 117, n. 6, p. 743–753, 2008.

DEMING, W. Edwards. **Out of the Crisis**. Cambridge: MIT Press, 1986.

DOBSON, Annette J.; BARNETT, Adrian G**. An Introduction to Generalized Linear Models**. 4. ed. Boca Raton: Chapman and Hall/CRC, 2018.

DUA, D.; GRAFF, C. **UCI Machine Learning Repository**. University of California, Irvine, School of Information and Computer Sciences. Disponível em: https://archive.ics.uci.edu/ml/datasets/Heart+Disease. Acesso em: 3 jun. 2025.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining: Concepts and Techniques**. 3. ed. San Francisco: Morgan Kaufmann, 2012.

HOSMER, David W.; LEMESHOW, Stanley; STURDIVANT, Rodney X. **Applied Logistic Regression**. 3. ed. New Jersey: Wiley, 2013.

KASS, G. V. **An exploratory technique for investigating large quantities of categorical data**. Applied Statistics, v. 29, n. 2, p. 119–127, 1980.

KUBAT, Miroslav. **An Introduction to Machine Learning**. 2. ed. Cham: Springer, 2017.

KUHN, Max; JOHNSON, Kjell. **Applied Predictive Modeling**. New York: Springer, 2013.

OMS – ORGANIZAÇÃO MUNDIAL DA SAÚDE. **Doenças cardiovasculares**. 2023. Disponível em: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases. Acesso em: 3 jun. 2025.

POWERS, David M. W. **Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation**. Journal of Machine Learning Technologies, v. 2, n. 1, p. 37–63, 2011.

SILVER, Nate. **O sinal e o ruído: por que tantas previsões falham – mas algumas não**. Rio de Janeiro: Intrínseca, 2012.

SOCIEDADE BRASILEIRA DE CARDIOLOGIA (SBC). **Diretrizes Brasileiras de Cardiologia**. Disponível em: https://www.portalcardiologia.com.br. Acesso em: 3 jun. 2025.

SOCIEDADE BRASILEIRA DE CARDIOLOGIA (SBC). **Estatística Cardiovascular – Brasil 2023**. Disponível em: https://www.portal.cardiol.br/publicacoes/. Acesso em: 3 jun. 2025.

XAVIER, H. T. et al. **Atualização da Diretriz Brasileira de Dislipidemias e Prevenção da Aterosclerose**. Arquivos Brasileiros de Cardiologia, v. 101, n. 4, supl. 1, p. 1–22, 2013.