Auxílio ao Diagnóstico para Predição de Morte Súbita em Pacientes Chagásicos a Partir de Dados Clínicos: uma Abordagem baseada em Aprendizagem de Máquina

Pedro E.O. Primo[1], Weslley L. Caldas[1], Gabriel S. Almeida[1], Luan P.L. Brasil[2], Carlos H.L. Cavalcante[3], João P.V. Madeiro[1], Danielo G. Gomes[2], Roberto C. Pedrosa[4]

[1]Departamento de Computação – Universidade Federal do Ceará (UFC)
Campus do Pici – 60020-181 – Fortaleza – CE – Brasil

[2]Departamento de Engenharia de Teleinformática – Universidade Federal do Ceará (UFC)
Campus do Pici – 60455-970 – Fortaleza – CE – Brasil

[3]Instituto Federal de Educação e Tecnologia do Ceará (IFCE)
Campus de Maracanaú – 61939-140 – Maracanaú – CE – Brasil

[4]Hospital Universitário Clementino Fraga Filho (HUCFF)
Universidade Federal do Rio de Janeiro (UFRJ)
21941-617 – Rio de Janeiro - RJ – Brasil

{pedroernesto2406, gabrielsuassuna, luanbrasil}@alu.ufc.br,

weslleylc@lia.ufc.br, jpaulo.vale@dc.ufc.br, danielo@ufc.br,

henriqueleitao@ifce.edu.br, coury@hucff.ufrj.br

Abstract. *Chagas Disease (CD) affects about 7 million people worldwide and one of the main adverse outcomes is the sudden cardiac death (SCD) caused by cardiomyopathy, whose evolution can be controlled with an early diagnosis. At this paper, we use 7 machine learning algorithms over a specific dataset with clinic data from chagasic patients aiming at discrimination among patients with high and patients with low predisposition for SCD, applying feature selection and resampling methods. K-Nearest Neighbors showed the best performance, with AUC:85.35 and F1:75.79. Due to their high weights in the machine learning classifiers, we suggest Non-Sustained Ventricular Tachycardia and Total Ventricular Extrasystoles as important features to identify SCD.*

Resumo. *A doença de Chagas (DC) afeta cerca de 7 milhões de pessoas no mundo e pode levar à Morte Súbita Cardíaca (MSC) do paciente por cardiomiopatia, cuja evolução pode ser controlada com diagnóstico precoce. Neste artigo, foram utilizados 7 algoritmos de aprendizagem de máquina com uma base de dados clínicos de pacientes chagásicos, objetivando a classificação em alta ou baixa predisposição do paciente à MSC, com seleção de atributos e balanceamento dos dados. Os melhores resultados indicam AUC:85.35 e F1:75.79 para o algoritmo K-Vizinhos Mais Próximos. Devido ao forte impacto nos modelos de aprendizagem de máquina, sugerimos o uso da Taquicardia Ventricular Não Sustentada e Extrassístole Ventricular Total como indicadores de MSC iminente.*

1. Introdução

A doença de Chagas (DC), também conhecida como tripanossomíase americana, ocorre em mais de 21 países, principalmente na América Latina (no Brasil e no México). O número de infectados gira de 6 a 7 milhões de pessoas [Silva et al. 2021, WHO 2021]. A disseminação da doença geralmente ocorre quando as fezes dos insetos da subfamília *Triatominae*, portadores do protozoário *Trypanosoma cruzi*, entram em contato com o sangue do paciente [Coura and Viñas 2010]. A infecção, então, ocorre em duas fases: aguda e crônica. O período agudo dura dois meses após a infecção, mas os sintomas estão ausentes na maioria dos casos [Guedes et al. 2012, Li et al. 2015]. No entanto, de acordo com a Organização Mundial de Saúde (OMS), durante a fase crônica, 30% dos pacientes desenvolvem algum tipo de cardiomiopatia, como arritmias cardíacas (normalmente taquicardia ventricular e fibrilação ventricular) ou insuficiência cardíaca progressiva, que, se não tratadas, podem levar o paciente chagásico à Morte Súbita Cardíaca (MSC).

A MSC, por ser um dos principais resultados adversos causados pela doença de Chagas, tem chamado a atenção dos pesquisadores nos últimos anos. Diversos estudos foram propostos para a identificação de variáveis clínicas e laboratoriais que auxiliem na detecção precoce de uma possível alta propensão à MSC, como Rassi score [Rassi Jr et al. 2006], análise multivariada [de Souza et al. 2015] e análises sobre a variabilidade da frequência cardíaca (HRV) [Alberto et al. 2017], [Alberto et al. 2020].

Rassi Jr et al. (2006) propuseram um escore para predizer óbito em pacientes com doença de Chagas crônica. No entanto, foi uma pontuação para a morte geral e não considerou o modo de morte. O artigo escrito por de Sousa et al. (2015) usou o modelo de risco proporcional Cox para avaliar a relação entre fatores de risco para cardiopatia crônica chagásica e morte súbita em pacientes chagásicos, determinando-se um escore de risco, analisando-se as curvas ROC e curvas de sobrevida de Kaplan Meier para avaliar a performance preditiva dos escores. Apesar de terem uma amostragem de 373 exemplos (43 exemplos para MSC), não aplicaram nenhuma técnica de balanceamento dos dados nem forneceram detalhes sobre sensibilidade e especificidade.

Alberto et al. (2017) extraíram atributos da Turbulência da Frequência Cardíaca (TFC) e parâmetros da Variabilidade da Frequência Cardíaca (HRV) no domínio do tempo de sinais ECG divididos em dois períodos de 12h (dia e noite). Esses parâmetros foram usados como entrada para dois modelos lineares multivariados - regressão logística (LR) e discriminante linear de Fisher (LDA). No entanto, apesar de ser estratificado, o estudo contou com uma amostragem limitada a 22 exemplos.

Finalmente, em outra abordagem, Alberto et al. (2020) aplicaram técnicas de TFC e HRV para extrair atributos de sinais de Holter ECG e investigar possíveis associações com a ocorrência de morte súbita cardíaca, considerando 3 cenários diferentes: um sinal completo de 24h, apenas 12 horas de dia claro, e as 12 horas restantes; foram utilizados os métodos de seleção de atributos *forward* e *backward* para reduzir a quantidade de parâmetros, K-Vizinhos Mais Próximos para realizar a classificação e o método *leave-one-out* para a validação cruzada. Entretanto, o escopo desse trabalho se difere do nosso por usar uma amostragem limitada a 82 pacientes (20 positivos para MSC), bem inferior à amostragem de 310 pacientes (78 positivos para MSC) aqui proposta.

As abordagens mencionadas anteriormente utilizaram uma amostragem de dados

limitada, ou não apresentaram nenhum método em especial para lidar com o desbalanceamento dos dados, necessitando-se de informações sobre sensibilidade e especificidade. Para mitigar este problema, apresentamos os resultados de uma abordagem baseada em aprendizagem de máquina, utilizando algoritmos do estado da arte, bem como técnicas de subamostragem e sobreamostragem, para lidar com o problema do desbalanceamento. Estes algoritmos foram otimizados sobre uma amostragem de 310 pacientes, dos quais 78 são positivos para MSC, a fim de obter um modelo de diagnóstico auxiliado por computador para a identificação precoce de alta propensão à MSC em pacientes chagásicos. Em nossos achados, constatamos a importância das variáveis Taquicardia Ventricular Sustentada e Extrassístole Ventricular Total para a predição da MSC.

2. Material e Métodos

2.1. Base de Dados

Foram coletados dados clínicos e laboratoriais de 314 pacientes com doença de Chagas do Hospital Universitário Clementino Fraga Filho, HUCFF, da Universidade Federal do Rio de Janeiro (UFRJ) ao longo de vinte e seis anos, entre 1990 e 2016. Cerca de 160 pacientes possuem dois ou mais registros médicos. Dessa forma, a base de dados contém 550 amostras, dentre as quais 232 são de pacientes do gênero masculino e 318 de pacientes do gênero feminino. Cerca de 14,7% dos registros (81) são de pacientes que apresentaram morte súbita cardíaca em decorrência da doença de Chagas. Para a obtenção da base de dados utilizada neste trabalho, o protocolo foi apreciado pelo comitê de ética do HUCFF-UFRJ, o qual abdicou da necessidade de permissão por escrito sob o número 45360915.1.1001.5262, de acordo com as normas atualmente aplicadas pelo Comitê Nacional de Ética em Pesquisa e com os princípios descritos na Declaração de Helsinque.

O escopo deste trabalho se limita à classificação de pacientes chagásicos em duas classes: morte súbita cardíaca e não morte súbita cardíaca (MSC e não-MSC). Como consequência, algumas considerações são feitas quanto ao uso da base de dados original. Primeiramente, para a realização dos experimentos, os pacientes que vieram a óbito por meio de outras causas que não a morte súbita cardíaca (MSC), como causas naturais ou provindas de outras doenças, foram alocados na mesma categoria dos pacientes que não sofreram MSC e ainda estão vivos. Por fim, adotamos o registro mais recente de cada paciente, totalizando 310 registros únicos de pacientes, dos quais 78 (25,16%) faleceram por MSC e 232 (74,84%) permaneceram vivos até a última checagem ou faleceram por outros motivos. Além disso, foram considerados somente 37 atributos, referentes ao histórico médico prévio e a dados dos exames de eletrocardiograma, ecocardiograma e holter. Os demais atributos, como utilização de medicamentos, foram descartados para evitar qualquer tipo de viés nos dados. A Tabela 1 sumariza os atributos utilizados neste trabalho.

É importante ressaltar que esta é uma base que está sendo atualizada constantemente, apesar dos dados clínicos terem sido coletados até 2016. Informações como data e ocorrência de óbito são coletadas anualmente. Neste trabalho, contamos com atualização dos dados até o dia 10/02/2020, de forma que até essa data, não houve novos óbitos.

2.2. Metodologia

A metodologia proposta pode ser dividida em 4 passos: balanceamento dos dados, normalização dos atributos, seleção dos atributos e por fim classificação. Inicialmente,

Tabela 1. Atributos utilizados para experimentos

Grupo de Atributos	Variável	Tipo
Dados Paciente	Sexo	Categórica
	Índice de Massa Corporal	Quantitativa
Histórico Clínico	Hipertensão Arterial Sistêmica	Categórica
	Diabetes Melitus Tipo 2	Categórica
	Outras Cardiopatias	Categórica
	Marcapasso	Categórica
	Síncope	Categórica
	Fibrilação/Flutter Atrial	Categórica
	Insuficiência Renal Crônica	Categórica
	Embolia Pulmonar	Categórica
	Insuficiência Cardíaca	Categórica
	Derivação Ventrículo-Peritoneal	Categórica
	Tabagismo	Categórica
	Alcoolismo	Categórica
	Sedentarismo	Categórica
ECG	Área Elétrica Inativa	Categórica
	Extrassístole Ventricular	Categórica
	Extrassístole Supraventricular	Categórica
	Taquicardia Ventricular Não Sustentada	Categórica
	Pausa $>$ 3s	Categórica
	Alteração Primária	Categórica
	Distúrbio na Condução Interventricular	Categórica
	Distúrbio na Condução Atrioventricular	Categórica
ECO	Disfunção Diastólica	Categórica
	Diâmetro do Átrio Esquerdo	Quantitativa
	Diástole do Ventrículo Esquerdo	Quantitativa
	Sístole do Ventrículo Esquerdo	Quantitativa
	Déficit Segmentar	Categórica
Holter	Fibrilação/Flutter Atrial	Categórica
	Frequência Cardíaca Media	Quantitativa
	Disfunção do Nódulo Sinusal	Categórica
	Taquicardia Ventricular Sustentada	Categórica
	Taquicardia Ventricular Não Sustentada	Categórica
	Extrassístole Ventricular	Categórica
	Extrassístole Ventricular Total	Quantitativa
	Distúrbio na Condução Átrioventricular	Categórica

o balanceamento dos dados é necessário devido à grande diferença de amostras rotuladas como MSC e não-MSC. Para isso, foram aplicadas tanto técnicas de sobre amostragem quanto subamostragem nos dados a fim de melhorar o poder de predição do classificador final [Hernandez et al. 2013]. Para sobre amostragem, foi adotado o *Synthetic Minority Oversampling Technique* (SMOTE)[Chawla et al. 2002], que gera amostras sintéticas para a classe minoritária a partir das amostras existentes. Enquanto que para a subamos-

tragem, foi feito um simples sorteio para remoção das amostras da classe majoritária de forma aleatória.

Após isso, todos os atributos são reescalonados entre 0 e 1 para garantir que não haja discrepância de magnitudes. São então removidos os atributos redundantes ou não necessários, que podem prejudicar a interpretação e os resultados dos modelos. Para tanto, foi aplicado o método *K-Best* [Jović et al. 2015] de seleção de atributos. O *K-Best* seleciona os K melhores atributos conforme um *score* proveniente de uma determinada métrica. Neste trabalho foram escolhidas as seguintes métricas:

1. Chi2: O Qui-quadrado, ou Chi2, é calculado pela fórmula:

$$x^2 = \sum \left(\frac{(O - E)^2}{E} \right),$$

   onde O é a frequência observada, e E é a frequência esperada de uma determinada categoria.

2. f_classif: Este método calcula o *F-value*, baseado na análise de variância (ANOVA). O cálculo é feito pela divisão da variância entre os grupos pela variância interna dos grupos.

Por fim, os atributos restantes são utilizados como entrada para o algoritmo de classificação. Neste trabalho, os algoritmos selecionados foram: K-Vizinhos Mais Próximos (KNN), *GradientBoosting* (GB), Regressão Logística (LR), Naive Bayes (NB), Máquina de Vetores de Suporte (SVM), Floresta Aleatória Balanceada (BRF) e Perceptron Multicamadas (MLP).

Uma visão geral da sequência de passos proposta está ilustrada na Figura 1, que representa uma implementação desta metodologia usando a biblioteca *scikit-learn* do *Python*.

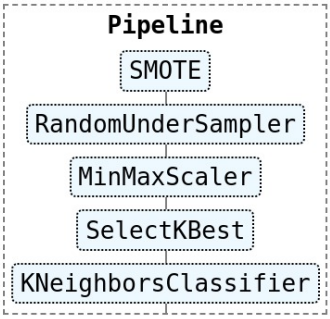

Figura 1. Exemplo da Pipeline para KNN usando-se a biblioteca scikit-learn.

3. Experimentos e Resultados

Para avaliar o desempenho do metodologia proposta, 3 cenários de experimentos foram conduzidos com uma série de algoritmos do estado da arte em aprendizagem de máquina: para o Cenário 1, não houve seleção de atributos nem balanceamento dos dados; para o Cenário 2, houve balanceamento, mas não houve seleção de atributos; por fim, para

Tabela 2. Cenários escolhidos

	Seleção de Atributos	Rebalanceamento	Algoritmo
Cenário 1	Não	Não	BRF
Cenário 2	Não	Sim	KNN
Cenário 3	Sim	Sim	BRF

o Cenário 3, utilizou-se tanto rebalanceamento quanto seleção de atributos. A Tabela 1 sumariza cada um dos cenários e o algoritmo vencedor correspondente.

Sobre a metodologia de treinamento, foram selecionadas aleatoriamente 80% das amostras para o conjunto de treino e 20% para o conjunto de testes de forma estratificada, garantindo-se a mesma proporção das classes nos conjuntos de treino e teste. Dentro do conjunto de treinamento, foi utilizado o método 5-fold estratificado para estimar os hiper-parâmetros, dentre os citados abaixo.

1. *GradientBoosting*
    (a) Taxa de aprendizagem: 0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2
    (b) Número de estimadores: 10, 30, 70, 100
    (c) Amostras mínimas para divisão de um nó: 12 amostras igualmente espaçadas no intervalo de 0.1 a 0.5
    (d) Amostras mínimas para um nó folha: 12 amostras igualmente espaçadas no intervalo de 0.1 a 0.5
    (e) Número máximo de features: logaritmo binário, raiz quadrada
    (f) Profundidade máxima: 3, 5, 8
    (g) Critério de avaliação de um *split*: erro médio absoluto, erro médio quadrático de Friedman
    (h) Subamostras de ajuste: 0.5, 0.618, 0.8, 0.85, 0.9, 0.95, 1
2. Floresta Aleatória Balanceada
    (a) Número de estimadores: 100 amostras igualmente espaçadas no intervalo de 151 a 1200
    (b) Amostras mínimas para divisão de um nó: 5, 7, 10, 14
    (c) Amostras mínimas para um nó folha: 4, 6, 8, 12
    (d) Número máximo de features: logaritmo binário, raiz quadrada, todas as features
    (e) Profundidade máxima: 10 amostras igualmente espaçadas no intervalo de 10 a 1200
    (f) Critério de avaliação de um *split*: índice Gini, ganho de informação
3. Perceptron Multicamadas
    (a) Taxa de aprendizagem: constante, adaptativa
    (b) Tamanho das camadas ocultas: (200, 50, 30), (100, 50, 10), (100, 50), (200, 100), (500, 250), (20,), (50,), (100,), (10,), (200,)
    (c) Função de ativação: tangente hiperbólica, unidade linear retificada
    (d) Solucionador de otimização de pesos: gradiente descendente estocástico, Adam
    (e) Parâmetro de regularização: 0.0001, 0.005, 0.05
4. Regressão Logística
    (a) Parâmetro de regularização: 0, 0.01, 0.1, 1.0, 10, 100

5. K-Vizinhos Mais Próximos
    (a) Número de vizinhos: 3, 5, 7, 9, 11
6. Máquina de Vetores de Suporte
    (a) Kernel: RBF, linear
    (b) Gamma (apenas para RBF): 2^{i-15}, para i de 0 a 19, com passo 2
    (c) Parâmetro de regularização: 2^{i-5}, para i de 0 a 21, com passo 2
7. Naive Bayes Gaussiana: Sem hiper-parâmetros

É importante notar que as técnicas de rebalanceamento, seleção de features e a normalização são aplicadas utilizando-se como base somente os dados no conjunto de treino, i.e não existe balanceamento no teste, e os valores usados para a normalização do conjunto de teste são extraídos do conjunto de treinamento. Para as técnicas de balanceamento dos dados, os hiper-parâmetros usados foram:

1. SMOTE
    (a) Taxa de reamostragem final da classe minoritária sobre a majoritária: 30%, 40%, 50%, 60%
2. Sub Amostragem Aleatória
    (a) Taxa de reamostragem final da classe majoritária sobre a minoritária: 130%, 120%, 110%, 100%

Após a realização dos experimentos, os resultados foram agrupados e tabulados. A Tabela 3 apresenta a média e o desvio padrão da Acurácia, Curva ROC-AUC, F1-score, Precisão e Sensibilidade para cada algoritmo, após 30 execuções sobre o conjunto de dados. Conforme pode ser observado, todos os algoritmos tiveram uma taxa de acurácia superior a 80%, porém, dado o desbalanceamento das classes, a sensibilidade foi afetada na maioria dos classificadores, com exceção do NB e do BRF, os quais obtiveram respectivamente $80.62 \pm 9.49\%$ e $83.96 \pm 9.67\%$. No entanto, nota-se que o BRF teve uma acurácia $82.1 \pm 4.89\%$ e F1 $70.94 \pm 6.1\%$ muito superiores ao NB, ocorrido provavelmente devido ao BRF ser um algoritmo que já leva em conta um tipo de balanceamento dos dados.

Tabela 3. Cenário 1: Sem balanceamento e sem seleção de features

	Acurácia		Curva ROC		F1		Precisão		Sensibilidade	
	Média	DP	Média	DP	Média	DP	Média	DP	Média	DP
BRF	**82.1%**	**± 4.89%**	**82.7%**	**± 4.7%**	**70.94%**	**± 6.1%**	**62.29%**	**± 7.8%**	**83.96%**	**± 9.67%**
NB	68.6%	± 9.39%	72.52%	± 7.38%	57.6%	± 7.75%	45.4%	± 8.22%	80.62%	± 9.48%
GBC	84.52%	± 3.4%	78.56%	± 4.93%	68.68%	± 6.97%	72.51%	± 9.02%	66.25%	± 9.79%
KNN	84.03%	± 3.96%	78.98%	± 5.86%	68.67%	± 8.47%	69.68%	± 8.51%	68.54%	± 11.19%
LR	82.8%	± 4.37%	76.45%	± 6.28%	65.26%	± 9.5%	68.39%	± 9.96%	63.33%	± 11.69%
MLP	82.47%	± 4.27%	75.35%	± 5.46%	63.94%	± 8.61%	69.7%	± 11.81%	60.62%	± 10.91%
SVM	82.85%	± 4.31%	77.1%	± 6.59%	65.85%	± 9.98%	68.03%	± 11.18%	65.21%	± 13.1%

Já no cenário 2, conforme a Tabela 4, a adição das técnicas de subamostragem e sobreamostragem resultaram na melhoria significativa de todos os algoritmos com exceção do BRF, que teve uma redução na sensibilidade, provavelmente por alguma interferência em sua própria metodologia interna de balanceamento. Nota-se, no entanto, que SVM, LR e KNN obtiveram sensibilidade superior a 80%. Além disso, o KNN se destacou por apresentar a melhor acurácia com $86.13 \pm 3.71\%$ e uma taxa f1 de $75.79 \pm 5.99\%$.

Tabela 4. Cenário 2: Com balanceamento e sem seleção de features

	Acurácia		Curva ROC		F1		Precisão		Sensibilidade	
	Média	DP	Média	DP	Média	DP	Média	DP	Média	DP
BRF	84.52%	± 2.7%	82.23%	± 2.37%	72.19%	± 3.4%	68.6%	± 9.22%	77.5%	± 7.13%
NB	78.71%	± 2.65%	77.09%	± 3.17%	64.08%	± 3.7%	57.11%	± 4.29%	73.75%	± 8.15%
GBC	83.23%	± 6.41%	81.36%	± 3.55%	71.1%	± 6.33%	67.65%	± 13.17%	77.5%	± 8.39%
KNN	**86.13%**	**± 3.71%**	**85.35%**	**± 4.3%**	**75.79%**	**± 5.99%**	**69.55%**	**± 7.77%**	**83.75%**	**± 7.13%**
LR	85.21%	± 4.36%	82.85%	± 2.97%	74.062%	± 5.98%	68.80%	± 13.26%	81.45%	± 5.59%
MLP	84.19%	± 4.02%	82.42%	± 4.56%	72.09%	± 6.18%	67.06%	± 8.42%	78.75%	± 8.39%
SVM	83.23%	± 4.05%	82.99%	± 2.23%	72.05%	± 4.56%	65.36%	± 11.61%	82.5%	± 8.15%

Por fim, a Tabela 5 apresenta os resultados do último cenário. Com exceção do classificador NB, que teve uma melhora significativa em todas as métricas (provavelmente porque a seleção de variáveis ajuda a aumentar a independência das mesmas, hipótese assumida pelo NB), não foi observada nenhuma melhoria significativa na média em nenhuma das métricas usadas para os demais classificadores; os resultados do segundo cenário apresentaram melhores taxas tanto em precisão, quanto sensibilidade. Para este grupo de experimentos, o melhor algoritmo foi o BRF, com um resultado de $70.68 \pm 2.88\%$, muito semelhante ao obtido no primeiro cenário de $70.94 \pm 6.61\%$. Porém, nota-se que o desvio padrão diminuiu drasticamente para todas as métricas utilizadas. Uma hipótese para isso é que a redução de atributos possa ajudar na redução da variância dos dados, resultando numa maior consistência do modelo. Resultados semelhantes podem ser observados para os algoritmos restantes.

Tabela 5. Cenário 3: Com balanceamento e seleção de features

	Acurácia		Curva ROC		F1		Precisão		Sensibilidade	
	Média	DP	Média	DP	Média	DP	Média	DP	Média	DP
BRF	**84.19%**	**± 1.77%**	**80.79%**	**± 1.89%**	**70.68%**	**± 2.88%**	**67.93%**	**± 3.69%**	**73.75%**	**± 2.8%**
NB	80.65%	± 1.14%	78.4%	± 2.11%	66.21%	± 2.23%	60.44%	± 2.62%	73.75%	± 6.85%
GBC	82.58%	± 2.39%	79.7%	± 1.59%	68.7%	± 2.78%	64.56%	± 5.31%	73.75%	± 2.8%
KNN	81.94%	± 2.65%	77.64%	± 2.93%	66.31%	± 4.32%	64.18%	± 5.39%	68.75%	± 4.42%
LR	83.23%	± 2.16%	80.14%	± 3.0%	69.39%	± 4.04%	65.6%	± 3.86%	73.75%	± 5.23%
MLP	82.58%	± 2.39%	78.48%	± 2.08%	67.55%	± 3.34%	65.46%	± 5.52%	70.0%	± 2.8%
SVM	80.32%	± 4.33%	78.99%	± 2.01%	66.95%	± 3.63%	60.16%	± 6.71%	76.25%	± 2.8%

Dentre os 3 cenários apresentados, podemos definir que os melhores classificadores foram o BRF do terceiro cenário e o KNN, SVM e LR no segundo cenário. O BRF no cenário 1 pode ser descartado por apresentar resultados semelhantes porém menos consistentes que sua versão no cenário 3.

Aqui nos deparamos com o *trade-off* de interpretabilidade/acurácia dos resultados. O BRF contém resultados mais consistentes, enquanto que o KNN e SVM resultados superiores. Já a LR, apesar de conter resultados inferiores aos demais classificadores, apresenta uma melhor interpretação e explicabilidade. Os coeficientes extraídos do SVM linear, LR e BRF estão presente na Tabela 6. É interessante notar, que a variável Síncope apresenta um grande peso no modelo de regressão linear, similar a resultados obtidos em [de Souza et al. 2015]. Porém, observa-se uma maior importância das variáveis Taquicardia Ventricular Não Sustentada e Extrassístole Ventricular Total em praticamente todos os classificadores e cenários. Podemos, portanto, somente hipotetizar que a maior amostragem de dados bem como o balanceamento forneceram diferentes perspectivas sobre a importância das variáveis clínicas e laboratoriais.

Tabela 6. Coeficientes de peso dos atributos para os principais modelos obtidos

	Cenário 1	Cenário 2		Cenário 3
Atributo	**BRF**	**SVM**	**LR**	**BRF**
Sexo	0.0	-0.48	0.18	0.0
Índice de Massa Corporal	0.02	-0.19	0.0	Não selecionada
Câncer	0.0	0.32	-0.0	Não selecionada
Hipertensão Arterial Sistêmica	0.0	0.13	0.17	0.0
Diabetes Melitus Tipo 2	0.0	0.6	0.11	0.01
Outras Cardiopatias	0.0	-0.12	0.06	0.0
Marcapasso	0.0	-0.58	-0.14	Não selecionada
Síncope	0.0	-0.1	0.63	0.03
Histórico - Fibrilação/Flutter Atrial	0.0	0.63	0.07	Não selecionada
Insuficiência Renal Crônica	0.01	-0.49	-0.19	0.0
Embolia Pulmonar	0.0	0.0	-0.0	0.0
Insuficiência Cardíaca	0.01	-0.5	-0.08	Não selecionada
Derivação Ventrículo-Peritoneal	0.0	0.59	0.14	Não selecionada
Tabagismo	0.0	-0.4	0.06	Não selecionada
Alcoolismo	0.0	-0.12	0.02	Não selecionada
Sedentarismo	0.01	-0.54	-0.33	0.01
Alteração Primária	0.01	-0.38	-0.3	0.01
Distúrbio na Condunção Interventricular	0.03	1.2	-0.15	0.01
ECG - Distúrbio na Condução Atrioventricular	0.0	0.78	0.02	0.0
Pausa $>$ 3s	0.0	0.09	0.06	Não selecionada
Extrassístole Supraventricular	0.03	-0.11	-0.39	0.05
ECG - Extrassístole Ventricular	0.02	-1.33	-0.26	0.01
ECG - Taquicardia Ventricular Não Sustentada	0.0	-0.21	-0.16	0.0
Área Elétrica Inativa	0.11	-0.61	-0.5	0.2
Diâmetro Átrio Esquerdo	0.04	0.33	0.18	0.05
Diástole do Ventrículo Esquerdo	0.02	1.0	0.3	0.02
Sístole do Ventrículo Esquerdo	0.03	-0.26	0.22	0.03
Disfunção Diastólica	0.01	-0.79	0.15	0.01
Déficit Segmentar	0.0	0.05	0.34	0.0
Holter - Distúrbio na Condução Atrioventricular	0.0	-0.04	0.0	0.01
Disfunção Nódulo Sinusal	0.0	-0.04	-0.23	0.0
Holter - Fibrilação/Flutter Atrial	0.0	-0.31	-0.11	Não selecionada
Frequência Cardíaca Média	0.05	0.59	0.04	Não selecionada
Taquicardia Ventricular Sustentada	0.0	-2.0	-0.05	0.0
Holter - Taquicardia Ventricular Não Sustentada	0.49	2.59	0.86	0.25
Holter - Extrassístole Ventricular	0.01	-0.79	-0.12	0.0
Extrassístole Ventricular Total	0.08	2.97	0.76	0.3

4. Conclusão

A principal contribuição deste artigo é fornecer uma série de modelos de diagnóstico por computador treinados usando uma amostra de dados significativa, bem como um sistema de balanceamento e métricas adequadas e originais. Além disso, estes modelos contêm diferentes níveis de interpretação sobre as variáveis clínicas, servindo no auxílio ao diagnóstico médico precoce de comorbidades mais sérias causadas pela doença de Chagas. Todos os algoritmos foram avaliados em termos de acurácia, precisão, sensibilidade e *F1*

score. Os algoritmos KNN, BRF, LR e SVM obtiveram os melhores resultados em termos de sensibilidade sem redução significativa nas demais métricas. Outro importante achado foi que, por meio dos coeficientes de peso de cada atributo fornecidos pelos modelos, notou-se que as variáveis Taquicardia Ventricular Não Sustentada e Extrassístole Ventricular Total tiveram um grande impacto na decisão da morte súbita. Tal constatação sinaliza que estas variáveis podem servir como futuros objetos de estudo.

Além disso, foram apresentadas 3 metodologias para avaliação de eficácia de técnicas de seleção de atributos e balanceamento dos dados para este estudo. Foi constatado que a combinação do SMOTE com subamostragem aleatória garantem um resultado melhor em termos de sensibilidade. Enquanto que a seleção de atributos afetou levemente a sensibilidade e precisão negativamente, porém contribuiu para a diminuição drástica da variação dos resultados.

Para trabalhos futuros, a fim de diminuir ainda mais a variância dos resultados, métodos comitê poderão ser propostos. Outros métodos de seleção de atributos podem ser utilizados, e diversos atributos podem ser extraídos dos sinais de eletrocardiograma de cada paciente, podendo aumentar o poder preditivo do modelo atual.

Agradecimentos

Os autores agradecem o apoio da Universidade Federal do Ceará (PIBIC-UFC), da Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (BICT-FUNCAP Processo IC7-0170-00004.01.93/20) e do CNPq (processos 426002/2016-4, 440092/2020-5 e 310317/2019-3).

Referências

Alberto, A. C., Limeira, G. A., Pedrosa, R. C., Zarzoso, V., and Nadal, J. (2017). Ecg-based predictors of sudden cardiac death in chagas' disease. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE.

Alberto, A. C., Pedrosa, R. C., Zarzoso, V., and Nadal, J. (2020). Association between circadian holter ecg changes and sudden cardiac death in patients with chagas heart disease. *Physiological Measurement*, 41(2):025006.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Coura, J. R. and Viñas, P. A. (2010). Chagas disease: a new worldwide challenge. *Nature*, 465(7301):S6–S7.

de Souza, A. C. J., Salles, G., Hasslocher-Moreno, A. M., de Sousa, A. S., do Brasil, P. E. A. A., Saraiva, R. M., and Xavier, S. S. (2015). Development of a risk score to predict sudden death in patients with chaga's heart disease. *International journal of cardiology*, 187:700–704.

Guedes, P. M. M., Gutierrez, F. R. S., Silva, G. K., Dellalibera-Joviliano, R., Rodrigues, G. J., Bendhack, L. M., Rassi Jr, A., Rassi, A., Schmidt, A., Maciel, B. C., et al. (2012). Deficient regulatory t cell activity and low frequency of il-17-producing t cells correlate with the extent of cardiomyopathy in human chagas' disease. *PLoS Negl Trop Dis*, 6(4):e1630.

Hernandez, J., Carrasco-Ochoa, J. A., and Martínez-Trinidad, J. F. (2013). An empirical study of oversampling and undersampling for instance selection methods on imbalance datasets. In *Iberoamerican Congress on Pattern Recognition*, pages 262–269. Springer.

Jović, A., Brkić, K., and Bogunović, N. (2015). A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 1200–1205. Ieee.

Li, P.-J., Jin, T., Luo, D.-H., Shen, T., Mai, D.-M., Hu, W.-H., and Mo, H.-Y. (2015). Effect of prolonged radiotherapy treatment time on survival outcomes after intensity-modulated radiation therapy in nasopharyngeal carcinoma. *PloS one*, 10(10):e0141332.

Rassi Jr, A., Rassi, A., Little, W. C., Xavier, S. S., Rassi, S. G., Rassi, A. G., Rassi, G. G., Hasslocher-Moreno, A., Sousa, A. S., and Scanavacca, M. I. (2006). Development and validation of a risk score for predicting death in chagas' heart disease. *New England Journal of Medicine*, 355(8):799–808.

Silva, L. E. V., Moreira, H. T., Bernardo, M. M. M., Schmidt, A., Romano, M. M. D., Salgado, H. C., Fazan Jr, R., Tinós, R., and Marin-Neto, J. A. (2021). Prediction of echocardiographic parameters in chagas disease using heart rate variability and machine learning. *Biomedical Signal Processing and Control*, 67:102513.

WHO (2021). Chagas disease (american trypanosomiasis). https://www.who.int/health-topics/chagas-disease.