DOI: 10.18605/2175-7275/cereus.v10n4p160-171

<<Recebido: 06/10/2018. Aceito: 30/11/2018>>

ARTIGO ORIGINAL

REVISTA CEREUS ISSN: 2175-7275

Utilização de Técnicas de Inteligência Computacional na Caracterização de Pacientes com Doenças Cardiovasculares

Use of Computational Intelligence Techniques in the Characterization of Patients with Cardiovascular Diseases

Juliana Baroni[1], Robson Mariano da Silva[2]

[1] M.sC Universidade Federal Rural do Rio de Janeiro

E-mail:

juliana.baroni@gee.inatel.br

[2] D.sC Universidade Federal Rural do Rio de Janeiro

E-mail:

rsmariano2010@gmail.com

RESUMO

Por estarem do topo da lista das doenças que mais matam no mundo, as doenças cardiovasculares estão assustando cada vez mais a classe médica, devido aos seus números alarmantes, sendo assim técnicas de Inteligência Computacional (IC) foram utilizadas para caracterizar pacientes da base de dados pública "*Heart Disease Database*", como cardiopatas ou não, a partir das variáveis fornecidas pela base. Máquina de Vetor de Suporte (SVM) e Regressão Linear Múltipla, foram escolhidas por terem desempenhos satisfatórios em aplicações similares na literatura. O modelo em que as variáveis foram diretamente introduzidas ao SVM, conseguiu em sua melhor simulação uma acurácia de 77%, sensibilidade de 91%, especificidade de 69% e falso negativo de 9%, enquanto na simulação em que as variáveis foram selecionadas por Regressão, os índices representaram 85%, 86%, 84% e 14%, respectivamente. O fator considerado como o de maior relevância foi o falso negativo, confirmando o melhor desempenho do modelo de SVM.

Palavras-chave: Inteligência Computacional, Máquina de Vetor de Suporte, SVM, Regressão Linear Múltipla, Doenças Cardiovasculares

ABSTRACT

Because they are at the top of the list of diseases that kill the most, such as cardiovascular diseases are always having a new medical class because of their alarming data, so that Computational Intelligence (CI) techniques were used to characterize patients from the public database. Diseases ", such as those with heart disease or not, from the variables predicted by the base. Support Vector Machine (SVM) and Multiple Linear Regression were chosen because they were considered as satisfactory results in the literature. The number was 91%, specificity 69% and false negative 9%, while the rate was filtered by Regression, the indices represented 85%, 86%, 84% and 14%, respectively. The most important factor as the most important was the false, confirming the best performance of the SVM model.

Keywords: Computational Intelligence, Vector Support Machine, SVM, Multiple Linear Regression, Cardiovascular Diseases

1. INTRODUÇÃO

Segundo a Organização Mundial da Saúde (OMS), as doenças cardiovasculares (DCV) são a maior causa de morte no mundo. São registrados cerca de 17,5 milhões de óbitos de pessoas anualmente vítimas de doenças do sistema cardiovascular. Estima-se que este valor chegue a 23,6 milhões até o ano de 2030 (OMS, 2017). Mansur *et al*. (2012) identificou que no Brasil, 20% das mortes registradas no período de um ano, por pessoas maiores de 30 anos, têm como causa as DCV, sendo que as regiões Sudeste e Sul, possuem um número ainda mais representativo. Em 2016, a Sociedade Brasileira de Cardiologia (SBC), estimou 350 mil mortes de brasileiros ocasionadas por essas doenças, um número 2,3 maior do que as mortes registradas nesse mesmo período ocasionadas por causas externas, como acidente e violência por exemplo (SBC, 2016).

O diagnóstico das doenças cardiovasculares é feito através de uma disposição de análises laboratoriais e estudos de exames de imagem lactente. O exame mais utilizado para auxiliar nesse diagnóstico é o Eletrocardiograma (ECG), pois esse exame permite um estudo das propriedades da musculatura cardíaca, como a formação e condução do estímulo cardíaco, podendo assim registrar os sinais elétricos emitidos durante a atividade do coração, em busca de anomalias no seu funcionamento.

Ferreira (2016), utilizou a diagnose de um Infarto Agudo de Miocárdio para comprovar a necessidade de associação de pelo menos dois dos três critérios seguintes, (sendo obrigatória a elevação plasmática dos marcadores de necrose miocárdica [MNM]): alteração no Eletrocardiograma (segmento ST e onda T), dores torácicas e/ou elevação dos MNM (creatinoquinase [CK], creatinoquinase MB [CK-MB], mioglobina, troponina). De acordo com um modelo geométrico desenvolvido pelo *New York Obesity Research Center* (NYORC), Moraes (2016) afirma conseguir caracterizar pacientes como cardiopatas ou não a partir de medições feitas dos perímetros braquial, da cintura, do quadril, da coxa e da panturrilha, por serem marcadores antropométricos de risco para doenças cardiovasculares, mas destacou a necessidade da associação de outros fatores.

Na literatura, é possível identificar diversas aplicações de técnicas de Inteligência Computacional na categorização de doenças cardiovasculares, dentre as quais podemos citar Rodrigues (2008), que utilizando Redes Neurais, obteve acurácia de 91% na caracterização de pessoas portadoras de doenças cardíacas. Tavares (2013), utilizou uma base de dados desbalanceada em técnicas como Máquina de Vetor de Suporte, Redes Neurais MLP, Algoritmo Genéticos e Árvore de Decisão, para classificar cardiopatias em

crianças, e concluiu que SVM com pesos foi a técnica com melhor resultado. Ubiratan (2014), apresentou uma ferramenta auxiliadora do diagnóstico de cardiopatia isquêmica, baseada na associação de técnicas de Algoritmos Genéticos (AG), Reconhecimento Baseado em Casos (RBC) e derivações da função de Distância Euclidiana, obtendo um índice de 97,01% de acertos nas etapas de treinamento com acurácia, especificidade e sensibilidade superiores a 92%. Ishitani (2006), investigou a associação entre alguns indicadores de nível socioeconômico e mortalidade de adultos por DCV no Brasil, utilizando a Regressão Linear Simples e Múltipla, concluindo que a associação entre as doenças cardiovasculares e fatores socioeconômicos é inversa, com destaque à escolaridade.

Neste artigo será apresentado o resultado de simulações de dois modelos computacionais, sendo o primeiro um SVM implementado a todas as variáveis pré-selecionadas manualmente, e o segundo, essas variáveis são selecionadas por um modelo de Regressão Linear e posteriormente implantadas à Máquina de Vetor de Suporte.

2. MATERIAIS E MÉTODOS

O conjunto de dados utilizado, foi extraído da base de dados pública *Heart Disease Database*. Que possui informações de 303 pacientes, dos quais, 164 são saudáveis e 139 doentes. Com um total de 76 atributos de cada um dos pacientes, sendo selecionados 14, por possuírem um número irrisório de dados faltantes. As variáveis são: idade, gênero, tipo de dor no peito, pressão arterial em repouso, colesterol sérico, concentração de açúcar no sangue em jejum, resultados eletrocardiográficos em repouso, ritmo cardíaco máximo alcançado, angina induzida por exercício, depressão da onda ST induzida pelo exercício em relação ao repouso, inclinação do pico de segmento ST durante o exercício, número de grandes vasos coloridos por fluoroscopia, talassenia e o diagnóstico.

Para o modelo SVM foram utilizadas as variáveis: idade, gênero, pressão arterial em repouso, colesterol, açúcar no sangue em jejum, ritmo cardíaco máximo alcançado, angina induzida por exercício e depressão da onda ST induzida por exercício, sendo que as outras seis não apresentaram importância significativa na obtenção de resultado nas técnicas utilizadas.

O segundo modelo realizou uma nova seleção de atributos através da Regressão, e o Modelo Linear Generalizado (MLG) determinou as seguintes variáveis com correlação: gênero, colesterol, ritmo cardíaco, angina e depressão da onda ST. Em seguida o modelo SVM recalculou a performance para os novos atributos. Ou seja, o modelo de Regressão Linear foi utilizado na determinação das variáveis que seriam usadas na SVM.

De forma a classificar as 98 mulheres e 205 homens, dentre os quais, 270 se encontram em idade útil (18 a 65 anos). Pôde-se distribuí-los dentro de cada uma das características fornecidas, de forma estatística, apresentadas na Tabela 1, facilitando a interpretação das informações. Todos os valores utilizados como parâmetro para classificação dos pacientes são dados da Sociedade Brasileira de Cardiologia (SBC) e da Organização Mundial de Saúde (OMS). Em decorrência da antiguidade da base, alguns parâmetros já foram alterados e/ou atualizados pelos órgãos competentes.

Tabela 1. Estatística de pacientes para cada variável no conjunto total

Característica	Parâmetro	Pacientes no parâmetro	Pacientes fora do parâmetro
Pressão arterial	≤ 130mmHg	171 (56,4%)	132 (44,6%)
Colesterol sérico	< 240 mg/dl	147 (48,5%)	156 (51,5%)
Açúcar no sangue	≤ 120 mg/dl	258 (85,1%)	45 (14,9%)
Ritmo cardíaco máximo	≤ 220 - idade	238 (78,5%)	65 (21,5%)
Angina no exercício	-	99 apresentaram (32,7%)	204 não apresentaram (67,3%)
Depressão da onda ST	≤ 1mm	180 (59,4%)	123 (40,6%)

Os modelos foram implementados no R, um *software* livre de linguagem própria, que possibilita o trabalho com modelos lineares e não lineares, testes estatísticos, análise de séries temporais, classificação, agrupamento e técnicas gráficas altamente expansíveis.

Para a simulação, foi utilizado o conjunto de treino com 80% dos dados e de teste com 20%, baseado nessa divisão foi feito um sumário dos dados, onde foram apresentados os valores mínimos, os 1° e 3° quartis, a mediana, a média dos valores e o valor máximo. Essas informações serão apresentadas nas Tabelas 2 e 3, apenas para as variáveis de valores contínuos, separados para cada um dos conjuntos.

Tabela 2. Sumário de valores do conjunto de treino

	Idade	Pressão arterial	Colesterol	Ritmo cardíaco	Depressão onda ST
Valor mínimo	29	94	126	71	0
1° Quartil	48	120	208	133	0
Mediana	55	130	239	152,5	0,8
Média	54,17	131	245,2	149,7	1,074
3° Quartil	60	140	273,8	165	1,800
Valor máximo	77	200	564	202	6,2

Tabela 3. Sumário de valores do conjunto de teste

	Idade	Pressão arterial	Colesterol	Ritmo cardíaco	Depressão onda ST
Valor mínimo	37	94	157	103	0
1° Quartil	45	120	209	142	0
Mediana	56	130	243	158	0,6
Média	54,15	129,6	250,9	155,6	0,9164
3° Quartil	62	140	290	172	1,4
Valor máximo	77	170	417	187	4,4

Foram realizadas 100 simulações para cada um dos modelos e obtivemos como resultado dessas simulações valores de erro de treino, erro de validação cruzada (*loocv*), número do vetor suporte, acurácia, sensibilidade, especificidade e valor falso negativo.

Os resultados serão dados a partir da comparação dos índices estatísticos, formadores da matriz de confusão. Um desses índices é a acurácia, que é definida como o grau de confiança do modelo. Quanto mais próximo ao resultado esperado, maior será o valor da acurácia do modelo.

$$\text{Acurácia} = \frac{VP+VN}{N} = \frac{(Verdadeiro\ positivo+Verdadeiro\ negativo)}{Total\ lote}$$

A sensibilidade é a capacidade do sistema de reconhecimento dos pacientes doentes, enquanto a especificidade é a capacidade de reconhecimento dos saudáveis, podendo ser calculas a partir das seguintes fórmulas:

$$\text{Sensibilidade} = \frac{VP}{VP+FN} = \frac{Número\ de\ resultados\ de\ testes\ verdadeiros\ positivos}{Todos\ os\ doentes\ afetados}$$

onde VP = Verdadeiro Positivo e FN = Falso Negativo.

$$\text{Especificidade} = \frac{VN}{VN+FP} = \frac{Número\ de\ resultado\ de\ teste\ verdadeiros\ negativos}{Todos\ os\ doentes\ não\ afetados}$$

onde VN = Verdadeiro Negativo e FP = Falso Positivo.

O Valor Preditivo Positivo (VPP), ou taxa de precisão, indica a proporção de pacientes doentes com resultado de testes positivo. Diretamente relacionado à sensibilidade e a especificidade, pode ser calculado da seguinte forma:

$$\text{VPP} = \frac{VP}{VP+FP} = \frac{Número\ de\ doentes\ positivos}{Todos\ os\ resultados\ positivos}$$

onde VP = Verdadeiro Positivo e FP = Falso Positivo.

O Valor Preditivo Negativo (VPN) representa a proporção de pacientes presentes no controle com resultados negativos e corretamente diagnosticados. Calculados a partir de:

$$\text{VPN} = \frac{VN}{VN+FN} = \frac{Número\ de\ pacientes\ saudáveis}{Todos\ os\ resultados\ negativos}$$

onde VN = Verdadeiro Negativo e FN = Falso negativo.

3. RESULTADOS E DISCUSSÕES

Com o intuito de realizar uma comparação entre dois modelos distintos, ambos com aplicação final de SVM, sendo que no primeiro as variáveis aplicadas foram escolhidas por fatores externos, como por exemplo a ausência de dados faltantes, e o segundo modelo, a seleção das variáveis se deu em função do índice de correlação, obtida por meio da Regressão Linear.

Baseado nos cem resultados obtidos para cada um dos modelos, foi realizada uma análise estatística das métricas usadas (Tabela 4).

Tabela 4. Estatística dos resultados das simulações SVM

	Erro de treino	Erro de *loocv*	N° de vetores suporte	Acurácia	Sensibilidade	Especificidade	Falso Negativo
Valor mín.	0,13	0,20	141	0,62	0,47	0,59	0,09
1° Quartil	0,15	0,23	148	0,72	0,63	0,78	0,23
Mediana	0,16	0,25	152	0,77	0,70	0,83	0,29
Média	0,17	0,24	152	0,76	0,70	0,82	0,30
3° Quartil	0,18	0,26	156	0,80	0,77	0,87	0,36
Valor máx.	0,21	0,30	163	0,87	0,91	0,96	0,53
Desvio padrão	0,01	0,02	5,08	0,05	0,10	0,07	0,10

Vale destacar a diferença teórica entre os erros. O erro de validação cruzada é sempre superior ao erro de treino, devido a sua formação. O erro de treino é um erro considerado simples, ele é calculado a partir das diferenças que acontecem dentro desse mesmo conjunto de treino, sendo assim, um valor relativamente baixo por sempre apresentar os mesmos valores de entrada. Já o erro de validação cruzada é calculado a partir da introdução de uma nova informação ao modelo, sendo este dado externo ao conjunto treino, tornando o erro de validação cruzada superior ao erro de treino, mas sendo assim, garantindo robustez ao sistema. Observando apenas os valores mínimo e máximo dos erros de treino e validação cruzada, temos 0,13 e 0,21, e 0,20 e 0,30, respectivamente.

No que tange ao número de vetores suporte, podemos concluir que quanto maior o número de vetores de suporte, maior será a complexidade do sistema. Apesar do desvio padrão com um valor mais alto do que o encontrado para as outras respostas do sistema (5,08), é possível identificar que a variação total de *N* foi de apenas 22 vetores, sendo o valor mínimo 141 e o máximo 163, com uma média e mediana iguais, com 152 vetores de suporte para que se tomasse a decisão naquela simulação correspondente.

Pode-se perceber que para a utilização das 8 variáveis, foi possível chegar a uma acurácia máxima de 87%, ou seja, este método em sua simulação mais assertiva tem 87% de chance de fornecer uma informação correta a quem ele o utilizar. Este índice tem ainda mais valor quando analisado com os outros valores desta mesma simulação, quando temos 85% e 89% de sensibilidade e especificidade, respectivamente. Estes valores garantem que o modelo apresenta uma ótima capacidade de reconhecer pacientes doentes e de mesma forma, pacientes saudáveis, além de um valor consideravelmente baixo de falsos negativos (15%).

Em sua pior simulação, do ponto de vista da acurácia, conseguiu-se um valor de 62%, seguido por 47% de sensibilidade e 79% de especificidade. Esses valores bastante discrepantes, acompanhados por um valor de falso negativo de 53%, fazem com que esta simulação não tenha tanta credibilidade.

A sensibilidade é um dos fatores com grande significância na avalição do resultado de um teste, por ela representar a chance de um teste com resultado positivo, realmente estar correto, ou seja, ela fornece a probabilidade de uma pessoa doente receber seu exame com uma resposta positiva. O valor mínimo fornecido pelo teste é bem diferente daquilo que foi esperado, 47%, ou seja, o teste não tem nem 50% de chance de dar um resultado correto para o paciente. Mas por um outro lado, em sua melhor simulação, foi alcançado um valor de 91% de chance de fornecer um resultado positivo a uma pessoa portadora de uma doença cardiovascular. Este já é um valor considerado excelente.

Dentre as cem simulações realizadas, apesar da distância existente entre os valores mínimo e máximo, foi conseguida uma média de 70% de sensibilidade para os testes, um valor considerado bom para este dado. Em uma análise feita a todos os resultados obtidos nas simulações, em apenas 17 delas, foi obtido como resultado para sensibilidade um valor abaixo de 60%, em contrapartida, 55 resultados para sensibilidade foram acima dos 70%.

A especificidade nos informa a respeito da probabilidade de um teste ter resultado negativo quando o paciente não está doente, sendo então outro fator muito importante a ser observado, por estar diretamente ligado aos valores de Verdadeiro Negativo e Falso Positivo, valores presentes na matriz de confusão. Com uma média dos resultados obtidos consideravelmente elevada (82%), podemos afirmar que seus valores para estas simulações foram relativamente bons, apesar de seu valor mínimo ter sido de 59%. O valor encontrado como 1° quartil já está bem acima do mínimo (78%), reafirmando a baixa probabilidade desse resultado não ser satisfatório ao problema. Já o valor mais elevado, chegou a 96%, valor bem próximo à perfeição (100%), ou seja, a chance de obter um teste negativo para uma pessoa saudável é perto do ideal.

O valor de Falso Negativo pode ser dito um dos mais importante para aplicações de Inteligência Computacional na área médica, tendo em vista que este valor representa a ausência de uma anormalidade no exame de um paciente que apresenta alguma doença. Sendo assim, quanto menor esse valor, menor será a chance de fornecer um diagnóstico errôneo ao paciente. Seu valor mínimo encontrado foi consideravelmente baixo, 9% é a chance desse erro acontecer para a melhor simulação da aplicação de SVM. Sua pior

simulação obteve um valor de 53% de Falsos Negativos, sendo este um valor que já não é mais tão proveitoso.

O desvio padrão calculado para estes valores foi de 0,10, valor também considerado elevado quando comparado aos outros elementos resposta obtidos nas simulações deste modelo, e semelhante ao encontrado no cálculo de desvio padrão da sensibilidade, isso pode ter sido caracterizado pela relação presente entre esses classificadores. Por ter obtido em sua melhor simulação um valor considerado ótimo, a média de todos os resultados não pode ser avaliada como um valor bom, já que este indica 30% de chance de erro de diagnóstico.

O valor de Falso Negativo foi escolhido dentre todos os índices apresentados, como o fator de maior importância para os testes, sendo que a partir dele foram escolhidos a melhor e a pior simulação. Estas estão descritas na Tabela 5.

Tabela 5. Pior e melhor simulação do modelo SVM

	Erro de treino	Erro de *loocv*	N° vetores de suporte	Acurácia	Sensibilidade	Especificidade	Falso Negativo
Pior simulação	0,14	0,21	147	0,62	0,47	0,79	0,53
Melhor simulação	0,16	0,25	149	0,77	0,91	0,69	0,09

Para a melhor simulação, o valor de acurácia encontrado foi de 77%, sendo este o valor encontrado como mediano dentre as simulações. Já na pior simulação, obtida pelo classificador foi de 62%. A acurácia poderia ter sido o fator determinante para melhor e pior simulação, como acontece em grande parte dos trabalhos na área, mas devido a variação de valores considerada pequena, tornou-se mais valioso buscar bons resultados de classificadores que não foram tão satisfatórios no contexto geral das simulações realizadas.

A sensibilidade, como já foi dito, está diretamente ligada ao valor de falso negativo, portanto, a diferença encontrada entre os valores para este critério entre a pior e melhor simulação foi bem elevada, sendo que na pior tivemos apenas 47% de chance de um teste positivo ser de um paciente acometido pela doença, enquanto para a melhor simulação essa probabilidade sobe para 91%. Já a especificidade não apresentou uma diferença tão considerável para as simulações, sendo que para a pior tivemos 79% e para a melhor 69%.

Na Tabela 6 é apresentada a mesma análise estatística para o segundo modelo.

Tabela 6. Estatística dos resultados das simulações Regressão+SVM

	Erro de treino	Erro de *loocv*	N° de vetores suporte	Acurácia	Sensibilidade	Especificidade	Falso Negativo
Valor mín.	0,15	0,18	130	0,67	0,5	0,72	0,14
1° Quartil	0,18	0,21	138	0,74	0,63	0,83	0,25
Mediana	0,19	0,22	143	0,79	0,70	0,86	0,30
Média	0,19	0,22	143	0,78	0,69	0,86	0,31
3° Quartil	0,19	0,23	148	0,82	0,75	0,89	0,38
Valor máx.	0,21	0,26	156	0,87	0,86	1	0,5
Desvio padrão	0,01	0,02	6,61	0,05	0,08	0,05	0,08

Assim como nas simulações realizadas para o modelo de SVM, e por motivos já explicados entre os dois tipos de erro presentes na tabela, verificamos que novamente o erro de validação cruzada foi superior ao erro de treino, porém, por eliminar as variáveis que não apresentaram correlação, essa diferença se tornou menor, quando comparados valores mínimos e valores máximos.

O número de vetores de suporte necessários na decisão do problema, demonstrou uma menor complexidade do modelo quando comparado ao anterior devido a existência de correlação entre as variáveis, encontrada pela Regressão Linear. Em sua simulação mais simples, foram necessários 141 vetores de suporte, enquanto para este modelo, foram necessários 130 vetores. Já para a simulação mais complexa, no primeiro modelo foram necessários 163 vetores, enquanto para a segunda 156.

Apesar do valor máximo encontrado para acurácia nessa segunda aplicação ter sido igual ao encontrado na aplicação anterior (87%), o valor mínimo representado na Tabela 6 foi maior do que o obtido no primeiro modelo, sendo eles 67% e 62%, respectivamente. Esta informação pode explicar também uma média mais elevada para este modelo.

A sensibilidade para este modelo apresentou uma variação menor de valores em torno na média, quando comparada ao modelo anterior, e podemos confirmar tal fato, a partir do valor de desvio padrão, enquanto para este ficou em 0,08, para o modelo anterior foi encontrado 0,10.

Já a especificidade para este modelo, é indiscutivelmente melhor do que a encontrada pelas simulações do modelo anterior. Em uma das simulações do modelo completo, o valor de especificidade atingiu 100%, ou seja, para esta simulação não há chance de não identificar corretamente os pacientes que não apresentam cardiopatia. Em sua pior simulação, chegou-se a um valor de 72%, considerado bom para este classificador.

Os valores de Falso Negativo, assim como os de sensibilidade, para este modelo possuem um intervalo de variação de valores inferior ao intervalo encontrado no modelo de

SVM puro. Apesar do valor máximo para este modelo ter sido menor do que o da pior simulação do SVM, 50% e 53%, respectivamente, o valor mínimo para este modelo representa um número consideravelmente mais alto quando comparado ao modelo anterior, 14% e 9%, respectivamente.

A Tabela 7 representa a pior e a melhor simulação, no conjunto das cem simulações realizadas para o modelo em que as variáveis foram escolhidas pela Regressão Linear, e posteriormente aplicada à Máquina de Vetor de Suporte, conseguimos os seguintes valores.

Tabela 7. Pior e melhor simulação do modelo Regressão + SVM

	Erro de treino	Erro de *loocv*	N° vetores de suporte	Acurácia	Sensibilidade	Especificidade	Falso Negativo
Pior simulação	0,17	0,21	132	0,70	0,50	0,88	0,50
Melhor simulação	0,19	0,22	152	0,85	0,86	0,84	0,14

Igualmente ao primeiro modelo, o valor de Falso Negativo foi utilizado na escolha da pior e melhor simulação. A diferença entre os valores extremos deste modelo apresentou um intervalo menor do que o obtido pelo primeiro modelo. Como consequência do menor valor de Falso Negativo, temos uma sensibilidade baixa (50%), principalmente quando comparada a da melhor simulação que apresentou um valor bem elevado (86%).

A acurácia encontrada para a simulação, que consideramos melhor, apresentou um valor significativo (85%), que quando comparado ao modelo anterior, este se destaca nesse classificador, principalmente por vir acompanhada de valores igualmente significativos de sensibilidade (86%) e especificidade (84%), enquanto na pior simulação, a acurácia foi de 70%, acompanhada de uma sensibilidade de 50% e especificidade de 88%.

4. CONSIDERAÇÕES FINAIS

Os resultados obtidos com as simulações realizadas no desenvolvimento deste trabalho indicaram que a utilização das variáveis: idade, gênero, pressão arterial em repouso, colesterol sérico, açúcar no sangue em jejum, ritmo cardíaco máximo alcançado, angina induzida por exercício e depressão da onda ST, no modelo de Máquina de Vetor de Suporte, e das variáveis: gênero, colesterol sérico, ritmo cardíaco máximo alcançado, angina induzida por exercício e depressão da onda ST, no modelo em que estas foram selecionadas por possuírem um índice de correlação encontrado por um processo estatístico de Regressão Linear, e posteriormente classificadas pelo SVM, para estudar a

classificação de pacientes com doenças cardiovasculares ou saudáveis, foram suficientes.

Em ambos os modelos, os resultados obtidos foram satisfatórios, visto que não foi necessária a utilização das 14 variáveis fornecidas pelo banco para se conseguir valores resposta significativos para os modelos. Com a utilização de apenas 8 dessas variáveis foi possível estimar o diagnóstico de uma DCV com percentuais de acerto elevados, quando comparados a trabalhos similares encontrados na literatura. Se destacando diante do modelo de SVM apresentado por Bhatia *et al.* (2008), aplicando a mesma base de dados, tornando em um contexto generalizado o trabalho bastante similar a este, e ainda assim foi obtida uma acurácia de 72,55% em sua melhor simulação. Ou ainda quando comparado ao modelo de SVM apresentado por Ho & Chou (2001), que apresentou um percentual de erro de 81% em suas respostas para diagnósticos de tais doenças.

Foram utilizados dois modelos distintos já conhecidos na literatura, de forma que seus resultados fossem comparados estatisticamente, para que fosse identificado dentre os dois modelos qual obteria uma maior precisão de resultado quando feita a comparação das respostas dos sistemas para os classificadores. A partir de uma análise mais detalhada, comparando essas respostas, classificador a classificador, concluímos que ambos os modelos são precisos. Com relação a acurácia os dois modelos atingiram um valor máximo para este classificador de 87%, porém ao analisarmos estatisticamente as 100 simulações de cada modelo, foi possível identificar uma diferença significativa no teste das médias desse classificador, identificando o modelo que associou a Regressão Linear com o SVM, pouco superior neste quesito. Porém, o classificador Falso Negativo, pertencente a matriz confusão do modelo, foi o escolhido para definirmos o melhor modelo implementado, por ser extremamente importante para o diagnóstico, a inexistência de erros dos testes. Para esta variável aplicamos o teste *t-Student*, para comparar as médias com *p*-valor inferior a 5%, e não foi verificada diferença significativa após este teste de normalidade, sendo assim, o modelo completo, em que foi usado apenas SVM, foi escolhido como melhor, já que conseguiu em sua melhor simulação, 9% de diagnósticos errôneos.

Acredita-se que os resultados sejam significativos, e que as técnicas podem ser auxiliadoras a condutas médicas no diagnóstico de doenças cardiovasculares, sendo utilizado como um parâmetro significativo na investigação dessas doenças, assim como na prevenção das mesmas, tendo em vista, inclusive o baixo custo de aplicação dessa técnica, podendo então, substituir exames de custo superior, desde que isso não interfira no diagnóstico da doença.

REFERÊNCIAS

AHA, D. W. ***Heart Disease Databases***. Disponível em: <www.ics.uci.edu/pub/machine-learning-databases/heart-disease.names> Acesso em: 02 dez. 2017

BHATIA, S., *et al*. **SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features**. *World Congress on Engineering and Computer Science*. San Francisco, USA, 2008.

FERREIRA, A. R. P. A.; SILVA, M. V.; MACIEL, J. **Eletrocardiograma no infarto agudo do miocárdio: O que esperar?** *International Journal of Cardiovascular Sciences*, v. 3, n. 29, p. 198-209, 2016.

HO, C. S.; CHOU, J. S. **Fuzzy ARTRON:** A general-purpose classifier empowered by fuzzy ART and error back-propagation learning. *Journal of Information Science and Engineering*, v. 13, n. 17, p. 683-695, 2001.

ISHITANI, L. H., *et al*. **Desigualdade social e mortalidade precoce por doenças cardiovasculares no Brasil**. Rev Saúde Pública, v. 40, n. 4, p. 1-8, 2006.

MANSUR, A. D. P.; FAVARATO, D. **Mortalidade por doenças cardiovasculares no Brasil e na região metropolitana de São Paulo**. São Paulo: Instituto do Coração (InCor) – HCFMUSP, 2012.

MORAES, V. C. S., *et al*. **Identificação do risco de cardiopatia através do estudo combinado de circunferências corporais**. Acta Biomédica Brasiliensia, v. 7, n. 1, p. 31-39, 2016.

PASSOS, U. R. C. **Computação evolutiva e aprendizado de máquina aplicados ao apoio do diagnóstico da cardiopatia isquêmica**. Dissertação (Mestrado). Campos dos Goytacazes, Universidade Cândido Mendes, 2014.

RODRIGUES, T. B.; MACRINI, J. L. R.; MONTEIRO, E.C. **Seleção de variáveis e classificação de padrões por redes neurais como auxílio do diagnóstico de cardiopatia isquêmica**. Pesquisa Operacional, v. 28, n. 2, p. 285-302, 2008.

Sociedade Brasileira de Cardiologia. **Cardiômetro:** Mortes por doenças cardiovasculares no Brasil, 2016. Disponível em: <http://www.cardiometro.com.br/default.asp> Acesso em: 06 jun. 2017

TAVARES, T.R. **Utilização de técnicas de inteligência artificial para classificação de crianças cardiopatas em base de dados desbalanceada**. Dissertação (Mestrado). Recife, Universidade Federal de Pernambuco, 2013.

The R *Foundation*. ***The* R *Project for Statistical Computing***. Disponível em: <https://www.r-project.org> Acesso em: 10 out. 2017

Thermo Scientific. **Interpretação dos resultados dos testes**. 2012. Disponível em: <www.phadia.com/pt-BR/Diagnostico-de-auto-imunidade/Saber-mais/Avaliacao-dos-Resultados-dos-Testes/#Sens Spec> Acesso em: 12 nov. 2017

WHO: *World Health Organization*. **Cardiovascular diseases**, 2017. Disponível em: <http://www.who.int/news-room/cardiovascular-diseases-(cvds)> Acesso em: 20 jun. 2018