XXI ENMC
ENCONTRO NACIONAL DE MODELAGEM COMPUTACIONAL
IX ECTM
ENCONTRO DE CIÊNCIA E TECNOLOGIA DE MATERIAIS

08 a 11 de Outubro de 2018
Instituto Federal Fluminense
Búzios - RJ

CARACTERIZAÇÃO DE PACIENTES COM CARDIOPATIAS UTILIZANDO TÉCNICAS DE INTELIGÊNCIA COMPUTACIONAL

Juliana Baroni[1] – juliana.baroni@gee.inatel.br
Fabiano Marcos de Lima[2] – fabianomarcos1@hotmail.com
Robson Mariano da Silva[3] – rsmariano2010@gmail.com
[1 2 3] Universidade Federal Rural do Rio de Janeiro, Instituto de Ciências Exatas, Programa de Pós-Graduação em Modelagem Matemática e Computacional – Seropédica, RJ, Brasil

Resumo. *Com o aumento progressivo no número de óbitos ocasionados por Doenças Cardiovasculares em torno do mundo, este assunto vem sendo cada vez mais abordado em estudos em diferentes áreas. A partir de treze variáveis e o resultado de diagnose presentes na Heart Diseases Database, foi possível caracterizar pacientes a partir de dois modelos. Para o modelo completo, no qual os pacientes foram classificados por Máquina de Vetor de Suporte, que foi considerado melhor principalmente por sua estabilidade, obteve-se em sua melhor simulação, dentre as 100 realizadas, uma acurácia de 92.1% e 6.8% de falso negativo. Enquanto para o modelo fit, ontem as variávei foram selecionadas por Regressão Linear e posteriormente classificadas por SVM, a acurácia foi de 89.8% e falso negativo de 11.1%.*

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

1. INTRODUÇÃO

As Doenças Cardiovasculares (DCV) se encontram no topo da lista de fatores de mortalidade no mundo, alcançando um percentual de 31% do total. A Organização Mundial de Saúde (OMS) registrou no ano de 2016, cerca de 17,5 milhões de mortes por doenças que afetam o sistema circulatório humano, e estatísticas apontam que esse número alcançará 23,9 milhões até 2030. No Brasil, a Sociedade Brasileira de Cardiologia (SBC) estimou que no ano de 2017, cerca de 381 mil pessoas foram a óbito por DCV, esta estimativa foi feita pelo Cardiômetro, um instrumento estatístico que calcula índices a partir de dados coletados da 10ª Classificação Internacional de Doenças (CID10), do sítio do DATASUS/MS, para os anos de 2006 a 2016.

Apesar desses dados apontarem um crescimento no número de casos fatais de doenças cardiovasculares, Soares *et al.* (2015), identificou uma redução da mortalidade nos grupos de

portadores de doenças do aparelho circulatório, doenças cerebrovasculares e doenças isquêmicas do coração, no Estado do Rio de Janeiro, em análise realizada no período de 1979 a 2010, e apontou que esta queda pode estar relacionada com a melhoria socioeconômica da população do estado, já que não pôde ser comprovada que se sucedeu a partir do desenvolvimento da tecnologia nos procedimentos necessários para o combate dessas anomalias, nem tampouco, por controle dos riscos dessas doenças.

A SBC, aponta como principais fatores para a incidência das doenças cardiovasculares a hipertensão arterial, colesterol, tabagismo, estresse, sedentarismo e diabetes. Dentre esses índices, alguns foram confirmados por Mansur *et al.* (2016), em estudo que concluiu que ao menos 20% das mortes registradas no Brasil no período de um ano, de adultos acima dos 30 anos, deram se por DCV.

O diagnóstico dessas doenças, dá-se pela associação de ao menos dois exames de diagnose. Ferreira *et al.* (2016), exemplificou que a identificação de um Infarto Agudo do Miocárdio exige que a elevação plasmática dos marcadores de necrose miocárdica [MNM] seja obrigatória, porém associada a dor torácica (analisada por exame clínico), ou alterações no eletrocardiograma (segmento ST e onda T).

Dentre os exames mais utilizados para controle do funcionamento do sistema circulatório, podemos destacar o Eletrocardiograma (ECG). Este exame permite estudar diversas propriedades da musculatura do coração, através de um Galvanômetro, que mede a diferença de potencial (ddp) entre dois pontos, permitindo uma análise da formação e condução de estímulo cardíaco, e assim o diagnóstico de problemas no ritmo cardíaco, problemas de condução cardíaca, sinais de insuficiência cardíaca dentre outras doenças. Como já foi visto, apesar da importância desse exame, o mesmo não pode ser usado de forma isolada para diagnóstico de qualquer DCV.

De acordo com um modelo geométrico desenvolvido pelo *New York Obesity Research Center* (NYORC), Moraes (2016), afirmou ser possível caracterizar pacientes com cardiopatias através da medição dos perímetros braquial, da cintura, do quadril, da coxa e da panturrilha, porém, esta também é uma técnica que exige associação de outros exames para diagnose de doenças, principalmente por seus resultados apresentarem uma diferença considerada pequena entre pacientes cardiopatas e saudáveis.

Técnicas de Inteligência Computacional são facilmente encontradas na literatura, sendo utilizadas para fins de categorização de doenças cardiovasculares. A associação das técnicas Máquina de Vetor de Suporte (SVM), Redes Neurais MLP, Algoritmos Genéticos e Árvore de Decisão, foi feita por Tavares (2013), para classificar cardiopatias em crianças, utilizando uma base de dados não normalizada. O destaque deu-se a técnica de SVM com pesos, que obteve melhores resultados. Ubiratan (2014), obteve acurácia, especificidade e sensibilidade superiores a 92% no auxílio do diagnóstico de cardiopatia isquêmica, utilizando técnicas de Algoritmo Genético, Reconhecimento Baseado no Casos (RBC) e derivações da função de Distância Euclidiana. A Regressão Linear Simples e Múltipla, foi utilizada por Ishitani (2006), para investigar a associação entre indicadores de nível socioeconômico e mortalidade por DCV de adultos no Brasil, concluindo que existe uma relação inversa entre esses fatores, dando destaque a educação.

Para este trabalho serão apresentados dois modelos, um chamado de completo, onde todas as variáveis do banco serão aplicadas à técnica de Máquina de Vetor de Suporte, e o segundo, chamado de modelo *fit*, em que teve as variáveis aplicadas ao modelo SVM selecionadas por Regressão Linear Múltipla. Após a caracterização dos pacientes será realizada uma comparação entre as técnicas, a fim de selecionar os melhores resultados.

2. REVISÃO BIBLIOGRÁFICA

- Doenças Cardiovasculares

As doenças cardiovasculares são provenientes de todo o sistema circulatório humano, podendo afetar então o músculo cardíaco ou os vasos sanguíneos. Doenças coronárias, cerebrovasculares, arterial periférica, cardíaca reumática, cardiopatia congênita, trombose venosa profunda e embolia pulmonar, formam o grupo de DCV.

Oliveira *et al.* (2015), concluiu que um dos fatores preponderantes para DCV são as modificações bi psíquicas geradas pela condição hipoestrogênica, oriundas do ciclo de menopausa em mulheres. Nesta fase da vida a mulher apresenta um aumento no triglicerídeos e na lipoproteína de baixa intensidade no organismo, podendo então considerar que as doenças cardiovasculares também são consideradas o fator que mais leva mulheres acima de 50 anos a óbito.

Pode-se destacar a doença arterial coronariana, acidente vascular cerebral (AVC), acidente vascular encefálico (AVE), carditite reumática, cardiopatias congênitas em suas diversas formas, flebite, trombose e embolia pulmonar, como as DCV as com maior incidência na população. Sendo que estas, podem ser diagnosticadas através de exames como ecografia transesofágica, cintilografia, cateterismo, ECG, radiografia de tórax, ecocardiograma, ressonância magnética, entre outros.

- Regressão Linear

A relação existente entre uma variável dependente e uma ou várias variável (is) independente (s), pode ser encontrada de forma estatística, através de uma análise de regressão. Esta, é representada por um modelo matemático representado por uma equação que associa as variáveis, em um gráfico, chamado de diagrama de dispersão. Quando a relação existente é de uma variável dependente pra uma independente, o modelo é de Regressão Linear Simples, caso exista mais de uma variável independente, é um modelo de Regressão Linear Múltiplo.

Esta técnica, portanto, define a influência de uma variável X (investigativa), sobre um valor esperado de uma variável Y (resposta), objetivando analisar e identificar alterações em $E[Y]$, ou seja, no valor esperado de Y, a fim de verificar se o mesmo sofre alterações devido as condições de interação com a variável Y, enquanto a variável X informa sobre o comportamento de Y.

A variável resposta desse modelo, em geral se comporta de forma linear, quadrática, cúbica, exponencial ou logarítmica. Essa identificação é de suma importância para o modelo, para que o mesmo seja explicado. A aproximação dos pontos no diagrama de dispersão é feita pelo Método de Mínimos Quadrados (MMQ). Esse método realiza a soma dos quadrados da distância entre os pontos do modelo e os pontos do diagrama, gerando uma relação entre X e Y, sempre buscando o menor erro. A utilização dessa técnica é fundamental, já que nem todos os pontos do modelo se ajustarão perfeitamente aos pontos do diagrama.

- Máquina de Vetor de Suporte

Com base na teoria do Aprendizado de Máquina (AM), do inglês *Machine Learning*, pode-se dizer que esta técnica estuda o desenvolvimento de algoritmos capazes de aprender com os próprios erros, para então realizar previsões sobre dados, ou seja, suas decisões serão

tomadas a partir de experiências anteriores acumuladas, adquirindo automaticamente novos conhecimentos.

O AM é dividido em dois padrões, na maioria das vezes. A forma com que o algoritmo se relaciona com o meio pode ser supervisionada ou não supervisionada. No aprendizado supervisionado, o algoritmo recebe um conjunto de treino, responsável por definir o que o algoritmo irá buscar, em seguida, o conjunto de teste é mapeado dentro de cada categoria desejada, produzindo padrões de saída corretos para cada nova entrada. Já o aprendizado não supervisionado, agrupa as entradas de acordo com medidas de qualidade, não existindo classes pré definidas para os atributos, portanto, este padrão visa o estabelecimento de classes.

Características como a precisão e velocidade de classificação dos dados, robustez e escalabilidade do sistema e a clareza de resposta do modelo são fundamentais para que esse processo se cumpra obtendo os melhores resultados possíveis. O SVM se destaca devido a características como essas, e por ser uma teoria bem definida.

O *Support Vector Machine* (SVM) foi criado em cima da Teoria do Aprendizado Estatístico, que propõe uma maximização da capacidade de generalização, buscando uma classificação eficiente do conjunto de treino, e minimização do risco estrutural do sistema, que representa a probabilidade de classificação errônea de padrões ainda não conhecidos pela máquina.

Na literatura, a Máquina de Vetor de Suporte, encontra se correlacionada a problemas de classificação e regressão. Os vetores de suporte são necessários para a definição de um hiperplano ótimo, ou seja, uma função capaz de separar as classes. As funções desse hiperplano ótimo são definidas com base na teoria do aprendizado estatístico, desenvolvido por Vipnik, e chamado de dimensão Vipnik-Chervonenkis ou dimensão VC. A importância dessa dimensão se encontra no fato de que se a mesma for definida corretamente, o aprendizado se torna confiável.

Essa técnica se SVM vem se destacando dentre outras técnicas de Inteligência Computacional, quando se trata de reconhecimento de padrão, devido aos seus resultados superiores, quando comparado com as Redes Neurais, por exemplo.

3. MATERIAIS E MÉTODOS

Para aplicação das técnicas escolhidas para este trabalho, foram utilizados os dados extraídos da *Heart Disease Database*, esta que é uma base de dados de domínio público, da qual foi subdividida em quatro subconjuntos e utilizaremos o de Claverland, obtidos por Robert Detrano no *Cleverland Clinic Foundation*. Trezentos e três pacientes foram incluídos nessa base, dos quais, 164 foram classificados como saudáveis e 139 doentes. Para cada um desses, foram expostos 76 atributos, porém, apenas 13 foram utilizados, devido ao número irrisório de dados faltantes, permitindo uma melhor análise. As variáveis aplicadas ao modelos foram: 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 e talassenia, além da décima quarta coluna que apresenta o diagnóstico de cada paciente.

Foram realizadas cem simulações para cada um dos dois modelos simulados, e para ambos os modelos foi utilizada a função *kernel* C-svc, com um valor de C unitário . Dos 303 paciente, foram considerados 297 para as simulações, levando em consideração que os outros 6 possuiam dados faltantes, impossibilitando um melhor desempenho dos modelos, desse 297, 70% formava o conjunto de treino, e 30% o conjunto de teste.

O primeiro modelo simulado foi o chamado de completo, onde todas as 13 variáveis do banco foram consideradas em uma rotina de Máquina de Vetor de Suporte. O segundo modelo, chamado de *fit*, as 13 variáveis foram analisadas por Regressão Linear Múltipla, onde um Modelo Linear Generalizado (MGL) selecionou apenas 6 variaveis que apresentaram correlação, sendo elas: gênero, pressão arterial em repouso, tipo de dor no peito, angina induzida por exercício, número de grandes vasos coloridos por fluoroscopia e talassenia, em seguida essas foram classificadas pelo SVM.

Ambos os modelos foram implemendado no *software* livre chamado de R. Este que vem sendo vastamente utilizado para aplicações de modelos lineares e não lineares, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento e técnicas gráficas altamente expansíveis. O próprio *software* gerou em meio a simulação um sumário dos dados para cada um dos conjuntos (treino e teste), nas Tabelas 1 e 2 serão apresentados esses valores estatísticos das variáveis contínuas.

Tabela 1: 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	210.8	131.8	0
Mediana	55.5	130	239	154.5	0.75
Média	54.58	132.3	244.6	149.2	1.058
3° Quartil	61.25	140	271.8	167.2	1.65
Valor máximo	76	192	564	202	6.2

Tabela 2: Sumário de valores do conjunto de teste

	Idade	Pressão arterial	Colesterol	Ritmo cardíaco	Depressão onda ST
Valor mínimo	35	100	149	97	0
1° Quartil	47	120	214	138	0
Mediana	56	130	254	151	0.8
Média	54.46	130.3	253.8	150.5	1.051
3° Quartil	60	140	288	163	1.6
Valor máximo	77	200	409	194	4.4

Os resuldados das 100 simulações de cada um dos modelos, forneceram valores de erro de treino, erro de validação cruzada, número do vetor suporte, acurácia, sensibilidade, especificidade e valor falso negativo. A partir desses resultados, é possível construir a Matriz Confusão do modelo, como pode ser visto através da formação de cada um desses índices nas fórmulas seguintes.

A acurácia pode ser calculada conforme a Eq. 1, e esta garante o grau de confiabilidade do modelo.

$$\text{Acurácia} = \frac{VP+VN}{N} = \frac{(Verdadeiro\ positivo + Verdadeiro\ negativo)}{Total\ lote} \quad (1)$$

A sensibilidade apresenta a capacidade do sistema de reconhecimento dos pacientes doentes, enquanto a especificidade a capacidade de reconhecimento dos saudáveis. Estas são calculadas pelas Eq. 2 e 3, respectivamente.

$$\text{Sensibilidade} = \frac{VP}{VP+FN} = \frac{Número\ de\ resultados\ de\ testes\ verdadeiros\ positivos}{Todos\ os\ doentes\ afetados} \quad (2)$$

onde VP = Verdadeiro Positivo e FN = Falso Negativo.

$$\text{Especificidade} = \frac{VN}{VN+FP} = \frac{\text{Número de resultado de teste verdadeiros negativos}}{\text{Todos os doentes não afetados}} \quad (3)$$

onde VN = Verdadeiro Negativo e FP = Falso Positivo.

4. RESULTADOS E DISCUSSÕES

Em busca de melhores resultados uma mesma base foi aplicada a dois distintos modelos (completo e *fit*), que nos permitiram analisar cada um de forma individual e por fim compará-los, a fim de escolher o melhor. Foram realizadas 100 simulações para cada um dos modelos, ambas com um conjunto de treino formado por 70% dos dados, e um conjunto de teste com os 30% restantes. Com o *software* R foram aplicados SVM e Regressão Linear para selecionar as variáveis do modelo *fit* (gênero, pressão arterial em repouso, tipo de dor no peito, angina induzida por exercício, número de grandes vasos coloridos por fluoroscopia e talassenia), e então, foi obtido como resposta do sistema o erro de treino, erro de validação cruzada, número de vetores de suporte, acurácia, sensibilidade, especificidade e falso negativo.

A aplicação da técnica de SVM foi escolhida dentre as de Aprendizado de Máquina devido aos bons resultados já encontrados anteriormente na literatura em aplicações semelhantes a realizada neste trabalho. Além de querer confrontá-la com resultados já obtidos em modelos bastante parecidos aos aqui desenvolvidos.

O erro de treino analisa sempre os mesmos valores presentes no conjunto de treino da base, portanto é considerado mais simples, e apresentará um valor inferior ao do erro de validação cruzada que é calculado toda vez que uma nova informação é introduzida no modelo, sendo assim, este garante uma maior robustez ao sistema.

Com base nos resultados do sistema foi construída a Tabela 3.

Tabela 3: Estatística dos resultados do modelo completo

	Erro de treino	Erro de *loocv*	N° de vetores suporte	Acurácia	Sensibilidade	Especificidade	Falso Negativo
Valor mín.	7.2%	12.1%	101	75.2%	59%	71.4%	2.5%
Mediana	10.5%	17.3%	114	83.1%	78.3%	87.2%	21.6%
Média	10.5%	17.3%	114.4	82.8%	77.8%	87.2%	22.1%
Valor máx.	13.9%	22.1%	128	92.1%	97.5%	97.9%	40.9%

Diante da diferença, já vista, entre os erros de treino e erro de validação crussada, e observando os valores mínimos e máximos do modelo completo, 7.2% e 13.9% para erro de treino e 12.1% e 22.1% para o de *loocv*, podemos perceber que até mesmo suas variações são diferentes, sendo a de treino inferior a de validação cruzada. A complexidade do sistema também não apresentou uma alteração brusca entre os valores de mínimo e máximo (101 e 128), confirmando isso pela comparação entre a média e mediana desse índice.

Este modelo completo, em sua melhor simulação, obteve uma acurácia de 92.1%, garantindo que em sua simulação mais acertiva, existe pouco mais de 92% de chance do modelo acertar no diagnóstico de um paciente. E em sua pior simulação, esse valor foi de 75.2%, o que não é um valor consideravelmente baixo.

Com valores elevados para sensibilidade e especificidade, em suas melhores simulações destes pontos de vista, 97.5% dos diagnósticos positivos foram para pacientes realmente doentes, enquanto 97.9% foram de resultados negativos para pacientes saudáveis.

Falso negativo é um fator de suma importância quando se trata de resultados na área da biomedicina, principalmente em diagnósticos médicos. Este índice representa a porcentagem de diagnósticos assertivos, ou seja, menores valores representam menos erros de resultados de exames. Para este caso obteve-se um valor bastante baixo, 2.5%, o que garante uma baixa probabilidade de erro.

Para escolha de melhor e pior simulações, foi escolhido o valor de acurácia, por esse ser o real valor de avalição de confiabilidade do modelo, portanto, na Tabela 4 serão apresentadas essas simulações com seus valores.

Tabela 4: Pior e melhor simulações do modelo completo

Simulações	Erro de treino	Erro de *loovc*	N° de vetores suporte	Acurácia	Sensibilidade	Especificidade	Falso Negativo
Pior	7.6%	13.6%	101	75.2%	64.8%	82.6%	35.1%
Melhor	12.9%	21%	123	92.1%	93.1%	91.1%	6.8%

Para o valor mais alto de acurácia, obteve-se um valor de 6.8% de falsos negativos, e valores de sensibilidade e especificidade acima de 91%, mostrando que este poderia ser um modelo confiável, para auxiliar em diagnósticos de doenças cardiovasculares.

Para o modelo *fit*, em que as variáveis foram escolhidas por Regressão Linear, foi feita a análise estatística apresentada na Tabela 5.

Tabela 5: Estatística dos resultados do modelo *fit*

	Erro de treino	Erro de *loocv*	N° de vetores suporte	Acurácia	Sensibilidade	Especificidade	Falso Negativo
Valor mín.	12%	13.8%	92	69.6%	59.5%	69.3%	10.8%
Mediana	15.8%	19.8%	114	80.8%	76.8%	84.6%	23.1%
Média	15.6%	19.9%	112.99	80.2%	76.5%	83.6%	23.4%
Valor máx.	19.7%	26.6%	130	89.8%	89.1%	95.4%	40.4%

Tanto o erro de treino quanto o de validação cruzada, em seus menores valores, apresentaram um percentual acima do encontrado para o modelo completo, afirmando então, que o modelo *fit* apresentou mais erros em seus resultados quando comparado ao modelo completo. Já quando falamos do número de vetores de suporte, este modelo (*fit*), mostrou-se menos complexo em sua melhor simulação, apresentando a necessidade de 92 vetores de suporte, enquanto o modelo completo precisou de 101.

Com uma acurácia de 89.8%, o modelo *fit* também pode ser considerado confiável, já que este é um valor alto, porém, sua pior simulação ficou abaixo de 70%, o que pode mostrar uma certa instabilidade do modelo. Essa instabilidade é confirmada pela variação entre os valores mínimos e máximos de sensibilidade e especificidade. Apesar de apresentarem valores máximos elevados, os valores mínimos baixo, não transferem a confiança desejada por quem busque bons resultados. Da mesma forma, podemos avaliar o valor de falso negativo, que obteve 10.8% de diagnósticos errados em sua melhor simulação, contudo apresentou 40.4% de exames negativos para doença, de pacientes saudáveis.

Na Tabela 6, serão mostradas as simulações consideradas pior e melhor, também tomando como parâmetro, como no modelo anterior, os valores da acurácia.

Tabela 6: Pior e melhor simulações do modelo *fit*

Simulações	Erro de treino	Erro de *loocv*	N° de vetores suporte	Acurácia	Sensibilidade	Especificidade	Falso Negativo
Pior	14.9%	19.3%	107	69.6%	68.1%	71.1%	31.8%
Melhor	19.7%	23.5%	119	89.8%	88.8%	90.9%	11,1%

Para a melhor simulação do modelo chamado de *fit*, em que a acurácia ficou próxima de 90%, o sistema mostrou-se pouco menos complexo por ter a necessidade de 119 vetores de suporte, porém, apresentou valores de erros um pouco acima do que foi visto para o modelo completo. Com sensibilidade e especificidade consideradas boa para esta melhor simulação, o valor de falso negativo, também representou uma quantidade de erros considerada baixa, mas não tanto quanto no modelo anterior.

A pior simulação confirma o que foi dito anteriormente, com relação a confiabilidade do modelo, devido a grande diferença de valor para os índices relacionados a matriz confusão, quando comparados a melhor simulação.

5. CONCLUSÕES

Diante dos resultados obtidos no modelo completo que utilizou todas as variáveis do banco de dados (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 e talassenia), e nos obtidos no modelo *fit*, que selecionou as variáveis (gênero, pressão arterial em repouso, tipo de dor no peito, angina induzida por exercício, número de grandes vasos coloridos por fluoroscopia e talassenia) por Regressão Linear, pode-se dizer que além de satisfatórios, foram conclusivos para que fosse escolhido um modelo que se destacasse perante o outro apresentado.

Em aplicações de Inteligência Computacional, em geral, duas variáveis de resposta do sistema são fundamentais para que se diga que um modelo é confiável, e em ambas o modelo completo, obteve destaque diante do modelo *fit*, com valores de acurácia de 92.1% e falso negativo de 6.8% em uma mesma simulação, comparados a 89.8% de acurácia e 11.1% de falso negativo no modelo *fit*. A importância desses valores, associados aos valores de sensibilidade (93.1% para a melhor simulação do modelo completo e 88.8% para o *fit*), que representam a capacidade do modelo em acertar diagnósticos de pacientes doentes, e em geral é muito utilizado em modelos médicos, dá-nos a segurança de se ter um resultado de exame com diagnóstico correto, sendo assim, podemos considerar que os valores obtidos neste trabalho foram satisfatórios e superiores aos de Bhatia *et al*.(2008), que aplicou a mesma base de dados a um modelo SVM, e obteve uma acurácia de 72,55% em sua melhor simulação, e ao de Ho & Chou (2001), que apresentou um percentual de erro de 81% em suas respostas para diagnósticos de tais doenças, utilizando também um modelo de Máquina de Vetor de Suporte.

Agradecimentos

Agradecemos ao apoio do Programa de Pós-Graduação em Modelagem Matemática e Computacional e do Instituto de Ciências Exatas da UFRRJ pelo incentivo ao desenvolvimento da pesquisa.

REFERÊNCIAS

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.

Cardiômetro: Mortes por doenças cardiovasculares no Brasil. Sociedade Brasileira de Cardiologia; 2016. Disponível em: http://www.cardiometro.com.br/ - Acessado em Agosto/2018

Cardiovascular diseases. WHO: *World Health Organization*; 2017. Disponível em: http://www.who.int/news-room/cardiovascular-diseases-(cvds) – Acessado em Junho/2018

Doenças Cardiovasculares. Organização Pan-Americana de Saude & Organização Mundial de Saúde Brasil, 2017. Disponível em: www.paho.org/bra/index.php?option=com_content&view=article&id=5253%3 Adoencas-cardiovasculares&catid=845%3Anoticias&Itemid=839 – Acessado em Maio/2017

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.

Heart Disease Databases. D.W. Aha. Disponível em: www.ics.uci.edu/pub/machine-learning-databases/heart-disease.names - Acessado em Dezembro/2017

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.

Hoffmann, R. Análise de Regressão: Uma introdução à Econometria. 4 ed. Piracicaba: Hucitec, 2014.

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.

Interpretação dos resultados dos testes. Thermo Scientific; 2012. Disponível em: www.phadia.com/pt-BR/Diagnostico-de-auto-imunidade/Saber-mais/Avaliacao-dos-Resultados-dos-Testes/#Sens Spec – Acessado em Novembro/2017

Lorena, A.C., Carvalho, A.C.P.L.F. Introdução às máquinas de vetor suporte. Relatório técnico. São Paulo: Universidade de São Paulo, 2003.

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.

Oliveira, A.S. Fatores de Risco Cardiovascular em Mulheres Pós-Menopausa. Canoas: UNILASALLE, 2015.

Passos, U.R.C. Computação evolutiva e aprendizado de máquina aplicados ao apoio do diagnóstico da cardiopatia isquêmica. Dissertação. Campos dos Goytacazes: Universidade Cândido Mendes, 2014.

Soares, G.P., Klein, C.H., Silva, N.A.S., Oliveira, G.M.M. Evolução da Mortalidade por Doenças do Aparelho Circulatório nos Municípios do Estado do Rio de Janeiro, de 1979 a 2010. Arq. Bras. Cardiol., v.104, n.5, p.356-365, 2015.

Stitson, M.O., Weston, J.A.E., Gammerman, A., Vovk, V. & Vapnik, V. *Theory of support vector machines*. Relatório técnico. Londres: *University of London*, 1996.

Tavares, T.R. Utilização de técnicas de inteligência artificial para classificação de crianças cardiopatas em base de dados desbalanceada. Dissertação. Recife: Universidade Federal de Pernambuco, 2013.

The R *Project for Statistical Computing*. *The* R *Foundation*; 2017. Disponível em: https://www.r-project.org/ - Acessado em Outubro/2017

CHARACTERIZATION OF PATIENTS WITH HEART DISEASE USING COMPUTATIONAL INTELLIGENCE TECHNIQUES

Abstract*. With the progressive increase in the number of deaths caused by Cardiovascular Diseases around the world, this subject has been increasingly addressed in studies in different areas. From thirteen variables and the diagnostic result present in the Heart Diseases Database, it was possible to characterize patients from two models. For the complete model, in which the patients were classified by the Support Vector Machine, which was best considered mainly for their stability, the best simulation, among the 100 performed, was an accuracy of 92.1% and 6.8% of false negative. While for the fit model, the variables were selected by Linear Regression and later classified by SVM, the accuracy was 89.8% and the false negative was 11.1%.*

***Key words**: Computational Intelligence, Cardiovascular Diseases, Multiple Linear Regression, Support Vector Machine*