UFRRJ

INSTITUTO DE CIÊNCIAS EXATAS

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

DISSERTAÇÃO

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

Juliana Baroni Azzi

2018

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

UTILIZAÇÃO DE TÉCNICAS DE INTELIGÊNCIA COMPUTACIONAL NA CARACTERIZAÇÃO DE PACIENTES COM DOENÇAS CARDIOVASCULARES

JULIANA BARONI AZZI

Sob a Orientação do Professor
Robson Mariano da Silva

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre em Ciências**, no Curso de Pós-Graduação em Modelagem Matemática e Computacional, Área Concentração em Modelagem Matemática e Computacional

Seropédica, RJ
Maio de 2018

Universidade Federal Rural do Rio de Janeiro

Biblioteca Central / Seção de Processamento Técnico

Ficha catalográfica elaborada com os
dados fornecidos pelo(a) autor(a)

A999u Azzi, Juliana Baroni, 1990-
Utilização de Técnicas de Inteligência Computacional na Caracterização de Pacientes com Doenças Cardiovasculares / Juliana Baroni Azzi. - 2018.
66 f.

Orientador: Robson Mariano da Silva.
Dissertação(Mestrado). -- Universidade Federal Rural do Rio de Janeiro, Programa de Pós-Graduação em Modelagem Matemática e Computacional, 2018.

1. Inteligência Computacional. 2. Máquina de Vetor de Suporte. 3. SVM. 4. Regressão Linear Múltipla. 5. Doenças Cardiovasculares. I. Silva, Robson Mariano da, 1963-, orient. II Universidade Federal Rural do Rio de Janeiro. Programa de Pós-Graduação em Modelagem Matemática e Computacional III. Título.

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

JULIANA BARONI AZZI

Dissertação submetida como requisito parcial para obtenção do grau de Mestre em Ciências, no Curso de Pós-Graduação em Modelagem Matemática e Computacional, área de Concentração em Modelagem Matemática e Computacional.

DISSERTAÇÃO APROVADA EM 18/05/2018.

Prof. Robson Mariano da Silva (Dr.) – PPGMMC – UFRRJ
(Orientador)

Prof. Reinaldo Bellini (Dr.) – LNCC

Prof. Angel Ramon Sanchez Delgado (Dr.) – PPGMMC – UFRRJ

DEDICATÓRIA

Dedico este trabalho aos meus pais, Rogério e Silvana, e minhas irmãs, Thaís, Déborah e Lara.

AGRADECIMENTOS

Primeiramente à Deus, por me guiar a mais essa conquista, me concedendo saúde, coragem e sabedoria.

Aos meus pais e minhas irmãs, que são meu porto seguro e não medem esforços para me ver cada dia indo mais longe.

Aos familiares que sempre tinham uma palavra de incentivo.

Aos meus amigos, de perto e de longe, que estiveram ao meu lado.

A minha equipe de trabalho dos CGI, que me deram força e sempre cobriram minha ausência no período do curso.

Ao meu orientador, Robson Mariano, que não mediu esforços para me fazer crescer como profissional e como pessoa ao longo dessa orientação.

Ao professor Angel Delgado, que com tanta presteza me orientou no estágio à docência, e me ensinou a ter prazer em ser pesquisadora.

Aos laços que construí ao longo do curso, em especial aos professores Gizelle Kupac, Duílio Tadeu, José Weberszpil, Ronaldo Gregório, Wagner Tassinari, Rosane Ferreira, Carlos Andres, Marcelo Dib, Josiane Cordeiro e Moisés Monteiro, que me passaram um pouco do muito que sabem. À Janaina que sempre esteve pronta a me ajudar. Aos colegas que por aqui encontrei, e de forma ainda mais especial aos amigos, Soline, Fabiano e Thiago, que estiveram mais próximos e não permitiram que eu desistisse.

Aos professores Reinaldo Bellini (LNCC), Rafael Teixeira (UFRRJ) e Marcos Benac (IM/UFRRJ), pelos preciosos conselhos na qualificação, ajudando a melhorar a qualidade do trabalho final.

Enfim, a todos que direta ou indiretamente estiveram presentes nessa conquista.

Muito obrigada!

RESUMO

AZZI, Juliana Baroni. **Utilização de Técnicas de Inteligência Computacional na Caracterização de Pacientes com Doenças Cardiovasculares.** 2018. 66p. Dissertação (Mestre em Ciência em Modelagem Matemática e Computacional). Instituto de Ciências Exatas, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2018.

Este trabalho engloba técnicas de Inteligência Computacional (IC), como a Máquina de Vetor de Suporte (SVM) e a Regressão Linear Múltipla, a fim de classificar os 303 pacientes presentes na base de dados pública "*Heart Disease Database*", como cardiopatas ou não, a partir de uma série de informações concebidas em exames periódicos realizados nos mesmos. Em busca de reduzir ou antecipar o diagnóstico de cardiopatias, doenças que estão no topo da lista das que mais matam ao redor de todo o mundo, ambas as técnicas foram escolhidas para esta aplicação, baseando em experiências anteriores similares à realizada nesta dissertação, levando em consideração seus desempenhos satisfatórios. Buscando ser um método capaz de auxiliar médicos no diagnóstico de doenças cardiovasculares (DCV), esta comparação tornou-se necessária para a diminuição de diagnósticos errôneos. A partir das informações coletadas, obtivemos um valor de 77% de acurácia, 91% de sensibilidade, 69% de especificidade e 9% de Falso Negativo para a melhor simulação da técnica de Máquina de Vetor de Suporte, enquanto para as simulações feitas com seleção de variáveis por Regressão Linear Múltipla, foram obtidos 85%, 86%, 84% e 14% respectivamente, confirmando estudos anteriores que mostram que a Inteligência Computacional, pode sim ser um auxiliador de diagnóstico de doenças cardiovasculares, contando com a associação de simples informações como: idade, gênero, pressão arterial, colesterol, glicose no sangue, ritmo cardíaco máximo alcançado, angina induzida por exercício e depressão da onda ST, aplicados à Máquina de Vetor de Suporte, que apesar de ter uma acurácia um pouco mais baixa, apresentou um melhor desempenho com relação aos resultados Falsos Negativos, assim obtendo um resultado mais satisfatório.

Palavras-chave: doenças cardiovasculares, máquina de vetor de suporte, regressão linear múltipla.

ABSTRACT

AZZI, Juliana Baroni. **Use of Computational Intelligence Techniques in the Characterization of Patients with Cardiovascular Diseases**. 2018. 66p. Dissertation (Master in Science in Mathematical and Computational Modeling). Institute of Exact Sciences, Federal Rural University of Rio de Janeiro, Seropédica, RJ, 2018.

This work encompasses Computational Intelligence (CI) techniques, such as the Vector Support Machine (SVM) and Multiple Linear Regression, in order to classify the 303 patients present in the public database "Heart Disease Database", as cardiac patients or not, based on a series of information designed in periodical examinations carried out in them. In order to reduce or anticipate the diagnosis of cardiopathies, diseases that are at the top of the list of the ones that kill the most around the world, both techniques were chosen for this application, based on previous experiences similar to the one performed in this dissertation, satisfactory performance. Seeking to be a method capable of assisting physicians in the diagnosis of cardiovascular diseases (CVD), this comparison became necessary for the reduction of erroneous diagnoses. From the information collected, we obtained a value of 77% of accuracy, 91% of sensitivity, 69% of specificity and 9% of False Negative in the best simulation for the Support Vector Machine technique, while for the simulations made with selection of variables by Multiple Linear Regression, were obtained 85%, 86%, 84% and 14% respectively, confirming previous studies that show that Computational Intelligence can rather be a helper with the association of simple information such as: age, gender, blood pressure, cholesterol, blood glucose, maximum heart rate achieved, exercise induced angina and ST wave depression, applied to the Support Vector Machine , which in spite of having a slightly lower accuracy, presented a better performance in relation to the False Negative results, thus obtaining a more satisfactory result.

Keywords: cardiovascular diseases, supporting vector machine, multiple linear regression.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

As doenças cardiovasculares (DCV) são a maior causa de morte no mundo. Segundo a Organização Mundial da Saúde (2017), cerca de 17,5 milhões de pessoas vêm a óbito todos os anos vítimas dessas doenças do coração e dos vasos sanguíneos, sendo este um número extremamente expressivo, já que representa 31% do total das mortes registradas dentro do período analisado, e em sua grande maioria em países de média e baixa renda. Estima se que até o ano de 2030 este número chegue a 23,6 milhões de pessoas mortas por doenças cardíacas (OMS, 2016).

Em Mansur *et al*. (2016), concluiu-se que pelo menos 20% das mortes registradas no Brasil, no período de um ano, por pessoas maiores de 30 anos, têm como causa as DCV, sendo que nas regiões Sul e Sudeste, este número é ainda maior, e que apesar de ter apresentado uma diminuição no índice de mortalidade entre os anos de 1980 e 2012, ainda apresenta um valor bastante significativo, principalmente entre os homens. A Sociedade Brasileira de Cardiologia (SBC) estima que quase 350 mil mortes de brasileiros no ano de 2016 foram causadas por essas doenças, um número 2,3 maior do que todas as mortes por causas externas como por acidentes e violência, 3 vezes maior do que o de mortes por doenças respiratórias e ainda 6,5 maior do que as causadas por todas as infecções incluindo a AIDS.

Segundo Dariele *et al.* (2016), a Hipertensão Arterial Sistêmica (HAS), é um importante fator de risco para acidentes cerebrovasculares, sendo responsável por pelo menos 40% das mortes por acidente vascular cerebral e por 25% das mortes por doença arterial coronariana. Em análise feita com pacientes da rede pública de saúde, chegou-se à conclusão de que a prevalência dessa doença está entre os idosos, podendo-se afirmar que a baixa escolaridade é o principal fator que dificulta a prevenção da doença, visto que estes pacientes se recusam a mudar de hábitos diários e aderir ao tratamento.

As doenças cardiovasculares são diagnosticadas usando uma disposição de análises laboratoriais e de estudos da imagem lactente. O exame de diagnóstico mais utilizado tem sido o Eletrocardiograma (ECG), sendo este um exame que possibilita estudar diversas propriedades da musculatura do coração, através de um equipamento chamado Eletrocardiógrafo. Este, nada mais é que um Galvanômetro, termo físico utilizado para equipamentos que medem a diferença de potencial entre dois pontos, capaz de analisar a formação e condução do estímulo cardíaco, ou seja, registrando os sinais elétricos emitidos durante a atividade cardíaca, a fim de que permita então um estudo preciso da atividade deste músculo. Através do ECG é feito o diagnóstico de problemas no ritmo do coração (arritmias e bradicardias), problemas da condução cardíaca, sinais de insuficiência cardíaca, entre diversas outras doenças. Porém, este é apenas um exame complementar, não podendo ser utilizado sozinho no fechamento de diagnóstico de qualquer paciente.

Tomando como exemplo a diagnose de um Infarto Agudo do Miocárdio, segundo Ana Rita (2016), é necessária a associação de pelo menos dois dos três critérios seguintes, (com obrigatoriedade de elevação plasmática dos marcadores de necrose miocárdica [MNM]): dor torácica, alterações no eletrocardiograma (ECG; segmento ST e onda T) e/ou elevação dos MNM (creatinoquinase [CK], creatinoquinase MB [CK-MB], mioglobina, troponina). Em

estudo de caso recente, Moraes (2016), afirma a caracterização de pacientes cardiopatas ou não através da medição dos perímetros braquial, da cintura, do quadril, da coxa e da panturrilha como marcadores antropométricos de risco para doenças cardiovasculares, de acordo com o modelo geométrico desenvolvido *pelo New York Obesity Research Center* (NYORC), frisando também que este não deve ser utilizado como diagnóstico único, necessitando da associação de outros fatores, visto que os resultados obtidos apresentam uma diferença considerada pequena entre pacientes cardiopatas e pacientes saudáveis.

Na literatura pesquisada, é verificado grande interesse na aplicação de técnicas de Inteligência Computacional na categorização de doenças cardiovasculares, dentre as quais podemos citar, Rodrigues (2007), aplicando técnicas de redes neurais, obteve acurácia de 91% na caracterização de pessoas portadoras de doenças cardíacas. Perozin (2002), explorou a potencialidade das Redes Neurais estudando fatores de risco de doenças arteriais coronarianas. Aplicando a Teoria dos *Fuzzy Set*, aos princípios da lógica clássica e resultados encontrados através das Redes Neurais em cima dos dados de treinamento, conseguiu-se então definir um percentual do grau de risco das doenças coronárias em cada paciente, ponderando-se os dados clínicos e laboratoriais pré-existentes. Tavares (2013), utilizou Máquina de Vetor de Suporte, Redes Neurais MLP, Algoritmo Genéticos e Árvore de Decisão, associando essas técnicas de diferentes formas para classificar cardiopatias em crianças a partir de uma base de dados não normalizada, concluindo que SVM com pesos foi a técnica que apresentou os melhores resultados. Campos (2013), realizou uma análise multimodal de sinais cardiovasculares, estudando a amplitude, fase padrões de domínio do tempo e sensibilidade aos estímulos impostos, ou seja, drogas que bloqueiam o sistema autonômico. Tanto os ganhos, efeitos causais e relações dinâmicas foram estudados através de um modelo de Lógica *Fuzzy*, um modelo de tempo discreto e um de evento discreto. A fim de aumentar a precisão da modelagem e melhorar a estimativa da frequência cardíaca e da série de tempo da pressão arterial dos pacientes. Ubiratan (2014), apresentou a proposta de uma ferramenta a partir de técnicas de Algoritmos Genéticos (AG), Reconhecimento Baseado em Casos (RBC) e derivações da função de Distância Euclidiana para auxílio no diagnóstico de cardiopatia isquêmica, obtendo um índice de 97,01% de acertos nas etapas de treinamento com acurácia, especificidade e sensibilidades superiores a 92%. Já Garcia *et al*. (2004), identificou fatores associados a níveis elevados de pressão arterial em crianças. Analisando variáveis como: idade, sexo, cor da pele, índice de qualidade de vida urbana, estatura e índice de massa corpórea, realizou a comparação das médias através da análise de variância, e na comparação de proporções, o teste qui-quadrado. As variáveis associadas a níveis mais elevados de pressão arterial foram incluídas em análise de Regressão Linear Múltipla. Destacando que fatores como o sobrepeso e a obesidade estiveram associados com os mais elevados níveis de pressão arterial sistólica. 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, chegando a conclusão de que a associação entre as doenças cardiovasculares e fatores socioeconômicos é inversa, com destaque à escolaridade.

Neste trabalho foram usadas técnicas de Inteligência Computacional para caracterizar pacientes, a partir de algumas características, apresentadas em um banco de dados, se são cardiopatas ou não.

1.1 Objetivo Geral

Utilizar a técnica estruturada em Máquina de vetor de Suporte não linear na caracterização de pacientes portadores de doenças Cardiovasculares.

1.2 Objetivos Específicos

- Avaliar o impacto na performance do modelo proposto com a aplicação da regressão na seleção de atributos;
- Verificar o desempenho de cada modelo na caracterização de doenças Cardiovasculares;
- Verificar o desempenho do modelo na determinação de falso negativo.

2 REVISÃO DE LITERATURA

2.1 Sistema Cardiovascular

O Sistema Cardiovascular, também conhecido como Sistema Circulatório, é responsável pela circulação do sangue através do corpo humano. Sendo este seu componente principal, o sangue exerce as seguintes funções: recolher o oxigênio e demais alimentos necessários nos alvéolos pulmonares e nas vilosidades intestinais, e distribuí-los nas células; encaminhar as substâncias que necessitam ser expelidas pelas células, aos órgãos responsáveis por essa função (pulmão, rins, etc.); criar relação entre partes do organismo, a fim de distribuir por elas os produtos das glândulas de secreção interna; auxiliar no equilíbrio da temperatura, e do conteúdo em água, no organismo, contribuindo para sua defesa.

Os principais componentes deste sistema são o coração e os vasos sanguíneos, como pode ser visto na Figura 1.

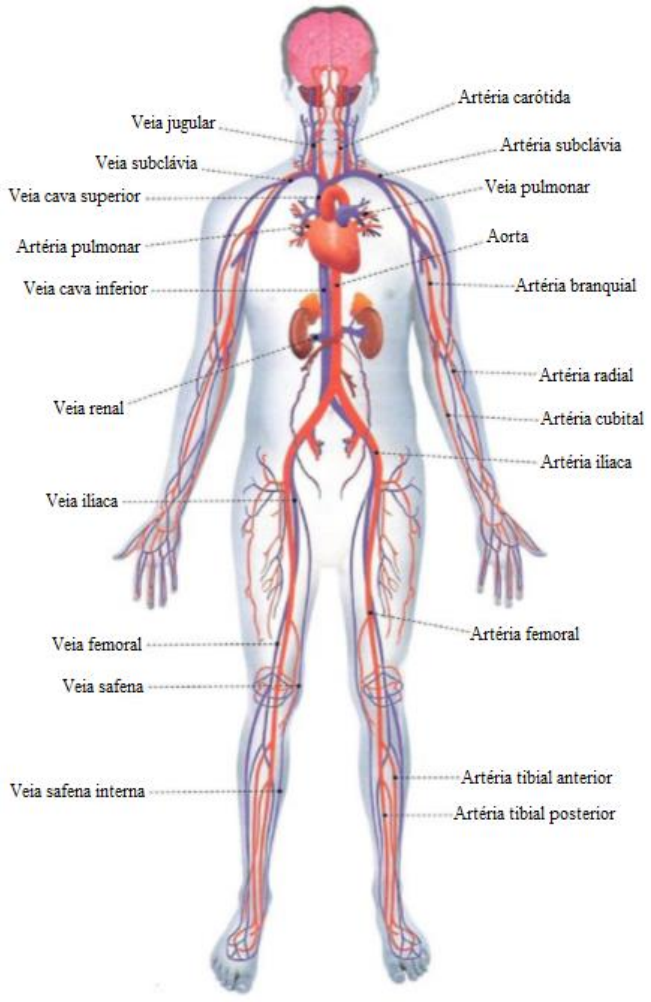

Figura 1: Sistema cardiovascular humano (Tortora, 2017)

O coração é um músculo oco, preparado para impulsionar o sangue que por suas cavidades passar para vasos denominados artérias, que se ramificam em vasos cada vez menores para que haja a condução do sangue para todo o corpo, inclusive suas extremidades. É o principal órgão do sistema circulatório. De formato aproximado a um cone, este se encontra na cavidade torácica, entre os pulmões (mediastino médio).

Seu interior é dividido em quatro câmaras, sendo dois átrios e dois ventrículos, formando uma bomba aspirante (átrios) e propulsiva (ventrículos). Sua parede é formada por três camadas, endocárdio (revestimento interno que tem contato direto com o sangue), miocárdio (musculatura cardíaca) e epicárdio (parte do pericárdio que faz o revestimento externo). Como pode ser visto na Figura 2.

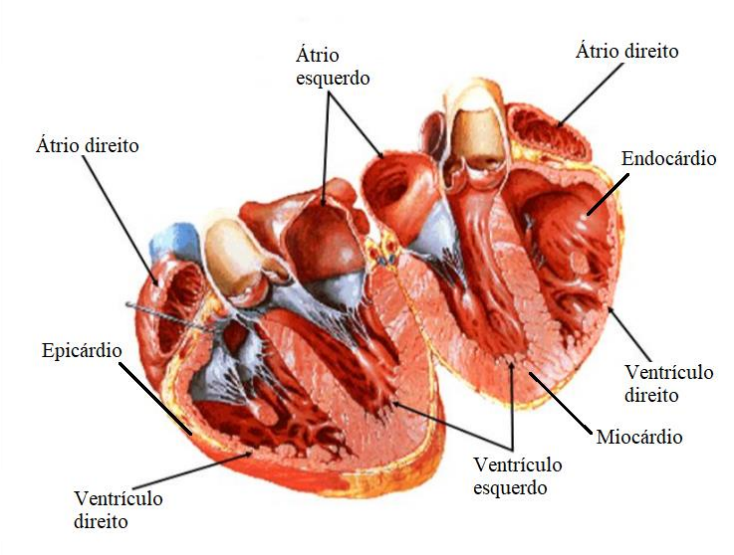

Figura 2: Coração humano (Netter, 2000)

O átrio direito apresenta quatro orifícios (três de entrada do sangue e um de saída), sendo os de entrada a veia cava superior (traz sangue da cabeça e membros superiores), veia cava inferior (traz sangue do tronco e membros inferiores) e veia coronária (recolhe o sangue que circula pelo próprio coração). O de saída, conhecido como válvula tricúspide, é a ligação com o ventrículo direito.

O ventrículo direito além do orifício da válvula tricúspide, apresenta outro de saída, a artéria pulmonar, que como o próprio nome diz, leva sangue aos pulmões. Este é munido de três válvulas sigmoides, que impedem que haja retorno do sangue ao ventrículo no momento da contração muscular do coração.

O átrio esquerdo, recebe o sangue vindo dos pulmões pelas quatro veias pulmonares nele presentes, e fornece sangue para o ventrículo esquerdo através da válvula mitral.

Já o ventrículo esquerdo possui o maior orifício de saída do coração, conhecido como artéria aorta, que também apresenta três válvulas sigmoides, impedindo o retrocesso do sangue.

A circulação sanguínea é dividida em quatro partes. Pequena circulação ou circulação pulmonar (coração/pulmão), grande circulação ou circulação sistêmica (coração/corpo), circulação colateral (artéria principal) e circulação portal (artéria secundária). Além de ter o auxílio da circulação linfática.

2.1.1 Doenças cardiovasculares

Doenças cardiovasculares são as principais responsáveis por óbitos em grande parte dos países, já que estas estão diretamente ligadas as condições de vida da população, sendo que no Brasil, nos últimos 5 anos atinge 38% dos óbitos na faixa etária considerada produtiva (18 a 65 anos).

Estas doenças não afetam apenas o coração, mas também os vasos sanguíneos, sendo assim oriundas de todo o sistema circulatório humano, formando o seguinte grupo de doenças: doença coronária, cerebrovascular, arterial periférica, cardíaca reumática, cardiopatia congênita, trombose venosa profunda e embolia pulmonar. Eventos agudos, como ataques cardíacos e acidentes vasculares, são bloqueios do fluxo sanguíneo para o coração e/ou cérebro, causados em sua maioria pelo acúmulo de gordura nas paredes dos vasos sanguíneos que irrigam esses órgãos, levando a conclusão de que os mais importantes fatores de risco comportamentais para estas doenças são o sedentarismo, dietas inadequadas, uso de tabaco e bebidas alcoólicas, gerando como efeito a elevação da pressão arterial, alta glicemia, hiperlipidemia, começando na fase do sobrepeso e se estendendo até a obesidade mórbida. Além desses fatores, pode-se considerar também como determinantes para esta enfermidade a pobreza (devido à falta de tratamento necessário aos fatores de risco já citados), estresse e hereditariedade.

Em revisão bibliográfica, realizada por Oliveira *et al.* (2015), chegou-se à conclusão de que outro fator preponderante para DCV são as modificações bi psíquicas decorrentes da condição hipoestrogênica, recorrente ao ciclo de menopausa em mulheres, considerando que nesta fase da vida ocorre um aumento de triglicerídeos e da lipoproteína de baixa intensidade no organismo, entre outros fatores coadjuvantes, sendo assim, essas doenças também são consideradas o fator que mais leva mulheres acima de 50 anos a óbito.

Conhecendo um pouco do sistema cardiovascular humano, pode-se entender melhor algumas doenças nesse sistema, onde ocorrem e os motivos pelos quais são originadas. A seguir, será apresentada uma breve descrição das doenças cardiovasculares mais comuns, sendo elas: doença coronária, cerebrovascular, arterial periférica, cardíaca reumática, cardiopatia congênita, trombose venosa profunda e embolia pulmonar. A seguir, será apresentada uma breve descrição das mais comuns.

2.1.1.1 Doença coronária

Também conhecida como Doença Arterial Coronária (DAC), é uma DCV causada pelo fornecimento inadequado de sangue ao coração. É válido lembrar que o bloqueio de uma artéria coronariana pode levar homens e mulheres a um ataque cardíaco, que é uma das principais causas de morte.

Causada principalmente pelo depósito de placas de gordura e ceráceos na parte interna das veias, podendo obstruí-las e enrijecê-las, tornando-as irregulares. Este bloqueio pode ser único ou múltiplo, dependendo da intensidade e localização. A diminuição do fluxo sanguíneo no músculo cardíaco pode causar dores no peito (angina), falta de ar, fadiga extrema com o esforço, inchaço nos pés, dores nos braços e ombros, entre outros sintomas. Caso o bloqueio da artéria se dê por completo, leva o indivíduo a ter um ataque cardíaco, que pode ser identificado através de fortes náuseas e dores nas costas ou mandíbula, podendo também ocorrer sem nenhum sinal aparente.

Os fatores de risco para as DAC’s são, a idade, histórico familiar de doença cardíaca, fumo, alta pressão arterial, alto colesterol LDL e baixo colesterol HDL, diabetes, sobrepeso ou obesidade, falta da prática de atividades físicas e estresse. Os homens apresentam um risco mais alto do que as mulheres (que sofrem um aumento do risco após a menopausa).

2.1.1.2 Doença cerebrovascular

As doenças cerebrovasculares são mais comumente chamadas de Acidente Vascular Cerebral (AVC) ou Acidente Vascular Encefálico (AVE). Segundo o Ministério da Saúde, esta é a principal causa de morte no Brasil, além de ser também a principal causa de incapacidade neurológica temporária ou permanente em pessoas acima dos 50 anos. Já em âmbito mundial, ela é a terceira maior causa de morte, ficando atrás apenas do infarto do miocárdio e câncer.

Os AVC’s podem ser classificados como isquêmicos ou hemorrágicos. Os isquêmicos ocorrem quando há uma diminuição ou parada de circulação sanguínea em um determinado vaso do sistema nervoso central, causando, em sua grande maioria, uma necrose irreversível, responsável pelo déficit neurológico focal apresentado pelo paciente. O AVC isquêmico pode ser dividido em trombótico ou embólico. O trombótico é representado pela oclusão progressiva de um vaso cerebral por um trombo de aterosclerose, sendo alguns dos principais fatores de risco para a aterosclerose o tabagismo, hipertensão arterial, diabetes, obesidade, hiperlipidemia, sedentarismo, estresse, consumo elevado de álcool, uso de drogas, uso de anticoncepcionais orais, entre outros. Já o AVC isquêmico embólico é caracterizado pelo desprendimento de um êmbolo das artérias carótidas, artérias vertebrais, arco da aorta ou das cavidades cardíacas a fim de impedir o fluxo sanguíneo em uma artéria cerebral, causando o infarto cerebral. Os fatores preponderantes para o seu acontecimento são: sopro de carótida, placas de ateroma no arco aórtico, fibrilação atrial, insuficiência cardíaca grave, presença de válvulas cardíacas metálicas devido substituição cirúrgica. Os AVC’s hemorrágicos acontecem quando ocorre o rompimento de um vaso e consequentemente o transbordamento de sangue para o parênquima cerebral subjacente, espaço subaracnóide e espaços subdural e extradural. Suas causas são a ruptura de aneurismas cerebrais, sangramento de malformações vasculares cerebrais, hipertensão arterial severa, sangramento de tumores cerebrais, traumatismo craniano, transformação hemorrágica de infarto cerebral, uso de anticoagulantes e distúrbios da coagulação.

As manifestações de um AVC isquêmico são dadas pela dificuldade de expressão ou compreensão da fala, fala enrolada, borramento visual unilateral, hemiparesia ou hemiplegia súbita, déficit de pares cranianos, perda súbita da coordenação motora e dificuldades para caminhar de início súbito. Os sintomas são variáveis, dependendo principalmente da intensidade do AVC e do local do cérebro que atingiu.

2.1.1.3 Cardiopatia congênita

O coração de um bebê se forma nas oito primeiras semanas de gestação, quando falamos de cardiopatia congênita, estamos falando de qualquer tipo de anormalidade que possa vir ocorrer nesse período de formação, proveniente de uma alteração no desenvolvimento embrionário da estrutura cardíaca. Pode ser descoberta ao nascimento ou anos mais tarde.

Os principais sintomas, nos bebês são: cianose (ponta dos dedos e/ou lábios roxos), irritação frequente, dificuldade para ganhar peso, transpiração e cansaço durante as mamadas e respiração ofegante até mesmo durante o sono. Já em crianças maiores, pode-se observar um cansaço excessivo em atividades físicas, o crescimento e o ganho de peso ocorrem de forma não adequada, constantes infecções pulmonares, cianose e pele pálida no decorrer de atividades físicas, além da taquicardia. Em casos mais graves ocorrem sinais de baixo débito cardíaco (incapacidade do coração de bombear o sangue de forma adequada) e crises de hipóxia (falta de oxigênio no sangue). Ainda é possível ocorrer síncope (desmaio precedido por tontura e vista turva) e dores torácicas.

A cardiopatia congênita não possui uma causa definida, porém a maioria dos casos ocorrem quando presente os seguintes fatores: mães com mais de 35 anos, parentes de primeiro grau com a doença, mães portadoras de diabetes, hipotireoidismo ou lúpus eritematoso sistêmico (LES), mães que adquirem toxoplasmose ou rubéola durante a gestação, mães que fazem o uso de medicamentos anticonvulsivantes, anti-inflamatórios, ácido retinóico, lítio, entre outros, no período da gestação, gestação múltipla ou por fertilização *in vitro*.

Dentre as cardiopatias mais frequentes, podemos citar a anomalia de Ebstein, atresia pulmonar, atresia tricúspide, coarctação da aorta (COA), comunicação interatrial (CIA), comunicação interventricular (CIV), defeito no septo atrioventricular, drenagem anômala das veias pulmonares, dupla via de saída de ventrículo direito (DSV), febre reumática, hipoplasia de ventrículo direito, miocardiopatias, persistência do canal arterial (PCA), síndrome de hipoplasia do coração esquerdo (SHCE) ou síndrome do coração esquerdo hipoplásico (SCEH), tetralogia de Fallot, tronco arterial comum ou *truncus*, transposição das grandes artérias (TGA), ventrículo único e válvula aórtica bivalvularizada.

2.1.1.4 Trombose venosa profunda

Também conhecida como flebite ou tromboflebite profunda. Tem como causa principal a coagulação do sangue no interior das veias, em local ou momento inoportunos, já que a coagulação é um mecanismo natural de defesa do organismo. Em 90% dos casos, as veias mais acometidas são as dos membros inferiores, causando dores e inchaço nos mesmos.

Ocorre mais comumente em pessoas que fazem o uso de anticoncepcionais ou tratamento hormonal, fumantes, obesos, que apresentem varizes, pacientes com insuficiência cardíaca, tumores malignos, ou que possuam história prévia de trombose venosa. A doença pode se desencadear após cirurgias de médio e grande porte, infecções graves, traumatismo, na fase final da gravidez ou no pós-parto, ou em qualquer outra situação que exija imobilização prolongada. É válido ressaltar que a idade avançada e anormalidades genéticas do sistema de coagulação, são predisposições para a trombose.

Em sua fase mais aguda, o paciente pode sofrer embolias pulmonares, em grande parte das vezes, fatais. Na fase crônica (de 2 a 4 anos após a fase aguda), pode ocorrer uma deficiência no funcionamento dos vasos sanguíneos, em decorrência de inflamações nas paredes das veias ao cicatrizarem. A síndrome pós-flebítica é caracterizada por um conjunto de lesões, como o

escurecimento da pele, a formação de grandes varizes, inchaços constantes, eczemas e úlceras nas pernas.

O diagnóstico clínico é difícil, portanto utiliza-se o exame *Eco Color Doppler* para o fechamento de um diagnóstico. E o tratamento é realizado com anticoagulantes ou fibrinolíticos.

2.1.1.5 Embolia pulmonar

O infarto pulmonar ou tromboembolismo pulmonar (TEP), é conhecido como uma complicação da trombose, dá-se início quando um coágulo ou êmbolo sanguíneo localizado em uma das veias das pernas ou pelve se solta, e bloqueia os vasos do pulmão. O bloqueio pode ocorrer por placas de gordura, pequenos fragmentos de ossos ou bolhas de ar. Dependendo do tamanho do trombo, pode causar morte súbita ao paciente.

A principal causa da embolia pulmonar são os êmbolos originários das tromboses, enquanto o principal fator de ocorrência é a imobilização prolongada do paciente, principalmente em repousos pós-cirúrgicos. Considera-se como causas também a obesidade, tabagismo, idade avançada, pressão alta, colesterol alto, insuficiência venosa dos membros inferiores, problemas de hipercoagulabilidade do sangue, problemas vasculares, insuficiência cardíaca, tratamentos hormonais, uso de tamoxifeno ou raloxifeno, trombofilia, anestesia, gravidez, inflamações, queimaduras, AVE e estrogênio suplementar.

Pode ser classificada como gasosa (causada por bolhas de gás formados na circulação), gordurosa (causada por fragmentos de tecido adiposo que entram na corrente sanguínea) ou amniótica (que ocorre no pós-parto, devido a passagem do líquido amniótico para a circulação sanguínea da mãe).

Em 50% dos casos, a embolia é assintomática, o que pode agravar ainda mais a doença. Já quando os sintomas se fazem presentes, eles podem ser bastante parecidos aos da trombose venosa profunda, sendo eles: aumento do tamanho do fígado e baço, ansiedade, inchaço nas pernas, falta de ar, taquicardia, cianose, sibilancia, dor no peito, tosse, falta de fôlego, respiração curta e ofegante, pulso fraco e rápido, tontura e sudorese intensa.

O diagnóstico precoce permite que seja feito um tratamento seguido à risca. Este é indicado para evitar a formação de coágulos, ou dissolver os já existentes, podendo ser medicamentosa ou cirúrgica.

2.1.2 Exames de diagnóstico

Existem vários exames para o diagnóstico das doenças descritas, sendo que para cada caso há uma indicação mais precisa feita pelo médico, a partir de uma avaliação prévia. De uma forma geral, a investigação se inicia pelos exames mais simples, evoluindo até os mais complexos.

Os exames podem ser classificados como invasivos (ecografia transesofágica, cintilografia, cateterismo cardíaco), ou não invasivos (eletrocardiograma (ECG), radiografia de tórax, monitorização do ECG por *Holter*, ecocardiograma, teste de esforço, tomografia do coração e vasos, ressonância magnética do coração e vasos (RM), angiografia digital). Os mais utilizados para diagnosticar cardiopatias serão listados a seguir.

2.1.2.1 Cateterismo cardíaco

O cateterismo cardíaco se trata da introdução de um cateter (tubo flexível extremamente fino) na artéria do braço ou da perna do paciente, que será conduzido até o coração, a fim de diagnosticar ou tratar diversas condições cardíacas, como por exemplo avaliar o entupimento de artérias coronárias, desobstruir artérias e válvulas com acúmulo de tecido adiposo, verificar possíveis lesões ou alterações no músculo cardíaco ou nas válvulas ou mostrar em detalhes malformações congênitas em recém-nascidos e crianças.

Quando associado a outras técnicas de angioplastia coronariana, é possível desobstruir um vaso coronarianos, além de poder ser utilizada como o implante de um *stend*. Quando associado à valvuloplastia percutânea com balão, é utilizado no tratamento de doenças das válvulas cardíacas, como estenose pulmonar, estenose aórtica e estenose mitral.

O procedimento é realizado com anestesia local, e para a introdução do cateter a uma das artérias (radial, femoral ou braquial), é feita uma pequena abertura na pele da virilha ou antebraço. O cateter será conduzido até o coração, localizando as entradas das artérias coronárias direita e esquerda. É injetado uma substância a base de iodo para permitir a visualização das artérias e possíveis entupimentos através do Raio X.

É indicado para pacientes em estudo para realização de revascularização ou procedimento de dilatação das coronárias, pacientes com possibilidade de intervenções por apresentarem angina instável ou angina pós infarto, também é realizado em pacientes que possuem estenose da válvula aórtica e que apresentam indícios de doença isquêmica, pacientes submetidos a cirurgia de revascularização, pacientes com insuficiência cardíaca, pacientes com arritmias graves ou dores no peito desconhecidas.

2.1.2.2 Eletrocardiograma (ECG)

O ECG está dentre os exames mais utilizados para o fechamento de diagnóstico de diversas doenças cardiovasculares. Esta é uma técnica não invasiva, capaz de avaliar a atividade elétrica do coração, a partir de eletrodos fixados na pele do paciente. Seu resultado é dado em forma de gráficos, que permitem a comparação da atividade cardíaca do paciente à atividade cardíaca normal, a fim de verificar a existência de distúrbios.

Este é considerado um dos métodos mais eficazes, sendo assim, precisão e eficiência são características de suma importância para a obtenção de um bom resultado. A velocidade e a amplitude em que o ECG é passado para o papel são de suma importância, para uma boa leitura do resultado, o ideal é 25 mm/s e 1 mV por 10 mm de amplitude. Ele deve ser avaliado tanto de forma individual (das 12 derivações), quanto o conjunto de derivações, sempre relacionando com os sintomas do paciente. Cumprindo esses requisitos básicos para a interpretação, parte-se para as análises.

O cálculo da frequência cardíaca e a análise do ritmo cardíaco são primordiais, partindo em seguida para o cálculo do intervalo PR (fim da linha P até o início do QRS) e do intervalo QT (sístole elétrica ventricular). Depois, faz-se necessário o estabelecimento do eixo elétrico, para buscar alterações no seguimento ST (tempo de não atividade entre a despolarização e a repolarização ventricular). Determinados esses resultados, pode haver a busca por alterações eletrocardiográficas. Na Figura 3, podemos visualizar um exame ECG normal, identificando os segmentos.

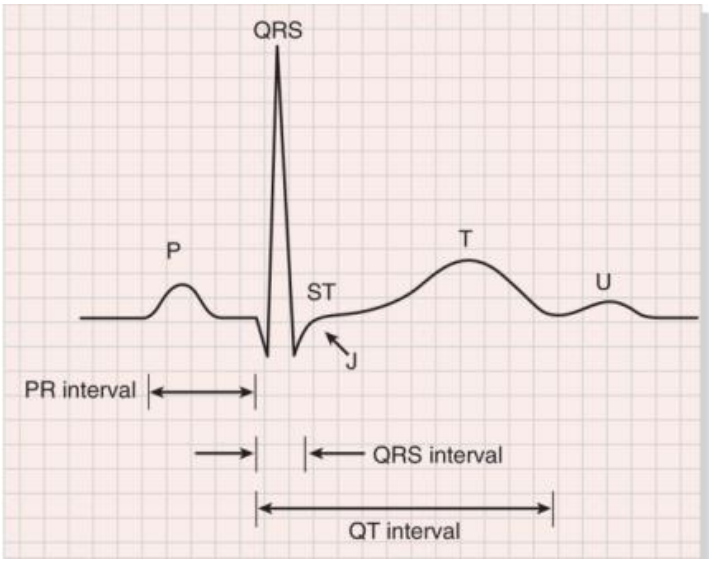

Figura 3: Representação ECG normal (Dias, 2012)

As ondas P, T e U, além do complexo QRS, formam um conjunto de eventos fundamentais do exame. A linha horizontal é a chamada de linha base. E por análise indireta é possível definir os intervalos PR, ST e QT. Sendo cada um dos eventos descritos a seguir:

- Onda P: com duração típica de 80 ms, é originada ao longo da despolarização auricular normal, quando o vetor do campo elétrico cardíaco está direcionado do nódulo sino-auricular para o nódulo auriculo-ventricular, sendo distribuído da aurícula direita para a esquerda.

- Onda T: sua duração típica é de 160 ms, e caracteriza a repolarização ventricular. Seu pico a divide em duas regiões distintas (período refratário absoluto e relativo).

- Onda U: geralmente é a última onda visível, e representa a repolarização do septo interventricular. Na grande maioria das vezes apresenta amplitude baixa ou até mesmo quase nula, porém sua ausência pode representar patologias como hipertiroidismo ou hipercalcemia.

- Complexo QRS: de duração tipicamente variável entre 80 a 120 ms, indica as variações rápidas da despolarização ventricular. A amplitude será muito maior do que a da onda P, devido a quantidade a mais de massa presente nos ventrículos quando comparados às aurículas.

- Linha base: é o ponto de partida das atividades elétricas cardíacas referentes a despolarizações e repolarizações.

- Intervalo PR: duração típica entre 120 a 200 ms, indica o tempo de travessia do impulso elétrico do nó sinusal até o nó aurículo-ventricular.

- Segmento ST: com duração típica de 80 a 120 ms, representa o intervalo entre o final da onda S até o início da T, que caracteriza a excitação ventricular.

- Intervalo QR: Representa o tempo da sístole elétrica, vai do início do complexo QRS até o final da onda T, e alterações em sua estrutura pode dar indícios para a prevenção de taquicardias e morte súbita.

Existem três diferentes tipos de eletrocardiograma, porém todos apresentam finalidades iguais, e são capazes de atingir o mesmo resultado, o que diferencia um do outro é a forma com que são realizados. Estes são conhecidos como ECG padrão, que possui uma duração em torno de 5 minutos e é realizado no paciente em repouso, o ECG de esforço, também conhecido como teste ergométrico, e como o próprio nome diz, é realizado enquanto o paciente pratica alguma atividade física (na maioria das vezes esteira ou bicicleta) e o *Holter*, conhecido como

monitorização de ECG ambulatorial, qual ocorre o registro das atividades cardíacas do paciente no período de 24 horas, e seu resultado é relacionado às atividades que o paciente realizou ao longo do dia.

Os resultados anormais desse exame podem caracterizar problemas como angina, isquemia, arritmia, lesões nas válvulas cardíacas, bloqueio no sistema de condução de impulsos, cardiopatias congênitas, modificação de eletrólitos no organismo ou danos a partes do coração.

2.1.2.3 Ecocardiograma

Também conhecido como ultrassom, o ecocardiograma utiliza ondas sonoras de alta frequência para a criação de imagens que permitem avaliar o funcionamento e a estrutura do coração. É utilizado no diagnóstico e acompanhamento de patologias como insuficiência cardíaca, sopro, sequelas de um infarto, cardiopatia congênita, para indicar uma cirurgia cardíaca e acompanhar a evolução do pós-operatório. Pode avaliar sintomas como palpitações, falta de ar, cansaço, elevação da pressão arterial e inchaço.

Os principais tipos de ecocardiograma são o transtorácico com doppler colorido, que além de ser o mais comum, pode ser realizado inclusive em fetos na barriga da mãe, ecocardiograma sob estresse e o ecocardiograma transesofágico, já falado anteriormente.

2.1.2.4 Tomografia computadorizada (TC)

O uso da TC na cardiologia vem se expandindo rapidamente, por permitir muito mais imagens, de maneira bem mais rápida e com melhor qualidade, sendo assim, é capaz de fornecer informações valiosas, como por exemplo, o diagnóstico de uma aterosclerose ainda em fase subclínica. Dividido em duas etapas (escore de cálcio e angiotomografia coronariana), que podem ser realizadas de forma simultânea ou separadas, à critério do médico responsável.

O Escore de Cálcio é um exame preditor de risco, que realiza cortes axiais sobre o coração, para a investigação da presença de placas de cálcio nas coronárias, analisando a quantidade de placas nas artérias, e não o grau de obstrução das mesmas. Essa etapa permite a medição do risco de infarto do miocárdio.

A Angiotomografia Coronariana, diferente do escore de cálcio, necessita do uso do contraste para visualizar se existem placas obstruindo as coronárias. Quando comparada essa técnica ao cateterismo, ela apresenta uma precisão em torno de 93%, sensibilidade de 95% a 96%, e valor preditivo negativo próximo a 100%, portanto, dificilmente o problema não será detectado por esse exame.

Outras indicações da tomografia computadorizada são para doenças da aorta, como coágulos ou tumores, avaliação de aneurismas, doenças do pericárdio e da válvula aórtica.

2.2 Aprendizado de Máquina

Partindo dos princípios de Aprendizado de Máquina (AM), parte de IC na qual SVM faz parte, pode-se destacar algumas informações importantes. Essa técnica estuda o desenvolvimento de métodos computacionais capazes de extrair conhecimento a partir de amostra de dados, bem como construir sistemas capazes de adquirir conhecimento de forma automática. Sendo assim, seus algoritmos têm a capacidade de gerar classificadores, de forma indutiva, para um conjunto de exemplos, designando à amostra um rótulo definindo a qual classe do conjunto determinada amostra pertence, a partir de um conjunto inicial de treino, que seja capaz de definir o mais próximo do real de tal rótulo, ou seja, este é um programa computacional capaz de tomar decisões, baseadas em experiências acumuladas de soluções anteriores bem sucedidas, sendo então uma poderosa ferramenta na aquisição de conhecimento automaticamente.

Dois principais padrões de AM são utilizados para definir a maneira com que o algoritmo se relaciona com o seu meio ambiente, ou seja, este padrão é a maneira com que ocorre o aprendizado por meio de uma base de dados (Haykin, 1999).

O aprendizado não-supervisionado, também conhecido como clusterização, e objetiva agrupar as entradas segundo uma medida de qualidade, não existindo então uma classe pré-definida para nenhum dos atributos, então é dado um conjunto de observações com o intuito de estabelecer a existência dos clusters (classes). É utilizado quando existe a intenção de encontrar tendências ou padrões que ajudem no entendimento dos dados.

No aprendizado supervisionado, ou classificador, o algoritmo recebe um conjunto de dados, proveniente do treinamento, que definem o que o algoritmo deve buscar. Após ter treinado o algoritmo é utilizado um conjunto de dados, já mapeados para cada categoria desejada, e então este algoritmo deve ter a capacidade de verificar os resultados previstos e esperados, encontrando a diferença dos resultados, ajustando seus parâmetros internos, a fim de obter determinado resultado. Este é o caso mais utilizado no treinamento de redes neurais e árvores de decisão, por ter como objetivo principal, produzir padrões de saída corretos para novas entradas não apresentadas previamente.

A obtenção de classificadores através do Aprendizado de Máquina baseado em uma amostra de dados, considera-se um processo de busca, buscando acima de tudo e em todas as hipóteses, que o algoritmo tenha a capacidade de gerar a partir dos dados, a busca com melhor capacidade de descrever o ambiente em que está acontecendo o aprendizado.

2.2.1 Máquina de vetor de suporte

A Máquina de Vetor de Suporte é uma técnica de Inteligência Computacional, referenciada do inglês *Support Vector Machines* (SVM) embasada na Teoria de Aprendizado Estatístico, que visa a proposição de técnicas de aprendizado de máquina que buscam a maximização da capacidade de generalização e minimização do risco estrutural (Haykin, 2001). A maximização da capacidade de generalização em técnicas de aprendizado de máquina é a capacidade da máquina na classificação eficiente perante o conjunto de treinamento, e a minimização do risco estrutural é a probabilidade de classificação errônea de padrões ainda não apresentados a máquina.

Na literatura é encontrado o termo máquinas de vetor de suporte ligado a problemas de classificação e regressão (Hearst, 1998; Lima, 2004; Stitson *et al*., 1996), e o termo vetores-suporte ou ainda, vetores de suporte utilizado para encontrar um hiperplano ótimo de separação, responsável pela separação de classes, ou uma função de separação com margem máxima entre classes distintas. A teoria que define rigorosamente os conceitos e demonstrações matemáticas da função do hiperplano ótimo é a teoria de aprendizado estatístico, tratado por Vapnik como dimensão Vapnik-Chervonenkis, ou simplesmente dimensão VC (Haykin 2001, Lorena & Carvalho 2003, Semolini 2002). Essa dimensão é de fundamental importância, pois, sua estimativa correta garante o aprendizado de maneira confiável, em outras palavras, a dimensão VC engloba o princípio de minimização de risco estrutural, que envolve a minimização de um limite superior sobre o erro de generalização, tornando a máquina com uma habilidade alta para generalizar padrões ainda não apresentados.

Vem sendo bastante utilizada nos últimos anos em diversas atividades de reconhecimento de padrão e tem se destacado devido aos seus resultados superiores (principalmente em reconhecimento de faces em imagens, na categorização de textos e em aplicações de bioinformática) aos de outras técnicas similares, como as Redes Neurais Artificiais (RNA).

Algumas características de um classificador são de suma importância para esse processo, sendo elas, a precisão com que se classifica os dados, a velocidade, que pode consistir tanto no tempo para construir o modelo quanto no tempo para executar o mesmo, a robustez do sistema que nada mais é do que a capacidade do mesmo lidar com ruídos e valores faltantes (*missing*), a escalabilidade, que é a eficiência em banco de dados existente em disco, a interpretabilidade ou a clareza com que o modelo fornece as informações e a relevância na seleção de regras como o tamanho da árvore de decisão, e as regras de classificação compactas.

As características que fazer com que o SVM se destaque com relação as outras técnicas de IC são: boa capacidade de generalização, robustez em grandes dimensões, convexidade da função objetivo e teoria bem definida.

De um modo geral, os classificadores gerados por uma máquina de vetor de suporte, apresentam bons resultados, que são medidos pela sua eficiência na classificação dos dados que não são pertencentes ao conjunto de treino. Porém, para essa técnica é evitado o *overfitting* (preditor muito especializado no conjunto de treino, não atingindo o desempenho esperado quando aplicado ao conjunto de teste). Quando falamos da convexidade da função objetivo, admitimos que há a aplicação de otimização de uma função quadrática na aplicação de SVM, que possui somente um mínimo global.

Dentre as características de destaque, podemos citar a capacidade de generalização, a qual teve resultados apresentados por Vapnik e Chernovenkis, se baseando na Teoria de Aprendizado Estatístico, propostas pelos mesmos autores nas décadas de 60 e 70, fazendo com que os resultados diretos dessa técnica gerassem o SVM (Lorena & Carvalho, 2003).

A seguir será descrita uma formulação básica da SVM, apresentando a SVM para classificação do caso linearmente separável e do caso não linearmente separável.

2.2.1.1 Máquina de vetor de suporte linearmente separável

O problema de classificação binária, problema de classificação inicial tratado pela SVM, trata da classificação de duas classes, sem perda de generalidade, através de um hiperplano ótimo a partir de um conjunto de treinamento linearmente separável. Um conjunto de treinamento é dito linearmente separável se for possível separar os padrões de classes diferentes contidos no mesmo por pelo menos um hiperplano (Haykin, 2001; Semolini, 2002).

Considere o conjunto de treinamento $\{(\mathbf{x}_i, d_i)\}_{i=1}^{N}$, onde $\mathbf{x}_i$ é o padrão de entrada para o *i*-ésimo exemplo e d_i é a resposta desejada, $d_i = \{+1,-1\}$, que representa as classes linearmente separáveis.

A equação que separa os padrões através de hiperplanos pode ser definida por:

$$\mathbf{w}^T.\,\mathbf{x} + b = 0 \tag{2.1}$$

onde, $\mathbf{w}^T.\,\mathbf{x}$ é o produto escalar entre os vetores **w** e **x**, em que **x** é um vetor de entrada que representa os padrões de entrada do conjunto de treinamento, **w** é o vetor de pesos ajustáveis e *b* é um limiar também conhecido como bias.

A Figura 4 mostra o hiperplano de separação (**w**, *b*) em um espaço bidimensional para um conjunto de treinamento linearmente separável.

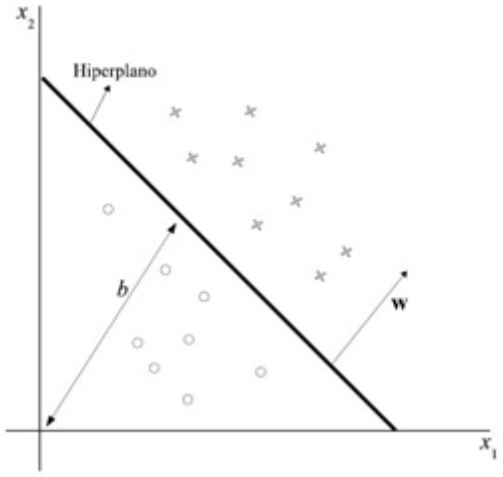

Figura 4: Hiperplano de separação (**w**,*b*) para um conjunto de treinamento bidimensional.

A equação (2.1) pode ser reescrita por:

$$\begin{cases} \boldsymbol{w}^T \,.\, \boldsymbol{x}_i + b \geq 0, \; se\; d_i = +1 \\ \boldsymbol{w}^T \,.\, \boldsymbol{x}_i + b < 0, \; se\; d_i = -1 \end{cases} \tag{2.2}$$

A margem de separação, distância entre o hiperplano definido na equação (2.1) e o ponto mais próximo de ambas as classes, é representado por ρ. O objetivo de uma SVM é encontrar um hiperplano que separe o conjunto de treinamento sem erro e maximize a margem de separação, sobre essa condição, o hiperplano é referido como hiperplano ótimo. A Figura 5 ilustra o hiperplano ótimo para um espaço de entrada bidimensional.

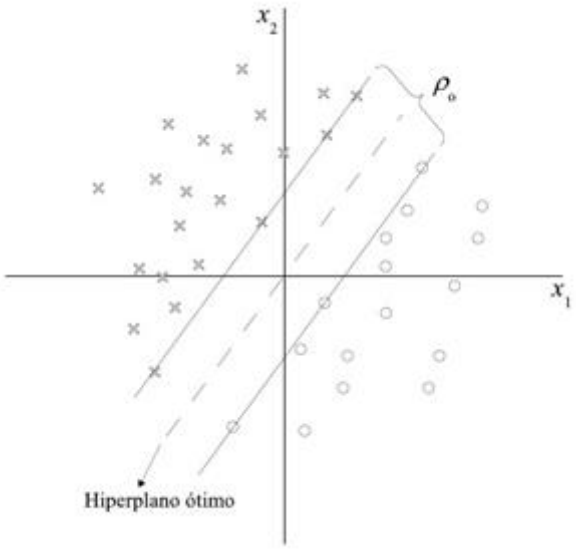

Figura 5: Hiperplano ótimo com máxima margem ρ_o de separação dos padrões linearmente separáveis

Considerando que $\boldsymbol{w}_o$ e b_o representam os valores ótimos do vetor peso e do bias, respectivamente, a equação (2.1) do hiperplano pode ser reescrita para o hiperplano como:

$$\boldsymbol{w}_o^T \,.\, \boldsymbol{x}_o + \, b_o = 0 \tag{2.3}$$

A função discriminante

$$\mathrm{g}(\mathbf{x}) = \boldsymbol{w}_o^T \,.\, \boldsymbol{x}_o + \, b_o \tag{2.4}$$

fornece uma medida algébrica de distância r entre $\mathbf{x}$ e o hiperplano ($\mathbf{w}_o$, b_o) que pode ser representado por:

$$\mathbf{x} = \mathbf{x}_p + \mathbf{x}_n \tag{2.5}$$

onde, $\mathbf{x}_p$ é a projeção normal de $\mathbf{x}$ sobre o hiperplano ótimo e $\mathbf{x}_n$ é o vetor normal com distância r, onde

$$\mathbf{x}_n = r \,.\frac{\mathbf{w}_o}{||\mathbf{w}_o||} \tag{2.6}$$

A Figura 6 ilustra a distância r entre $\mathbf{x}$ e o hiperplano ($\mathbf{w}_o$, b_o), onde, r é positivo se $\mathbf{x}$ estiver no lado positivo do hiperplano ótimo, caso contrário será negativo.

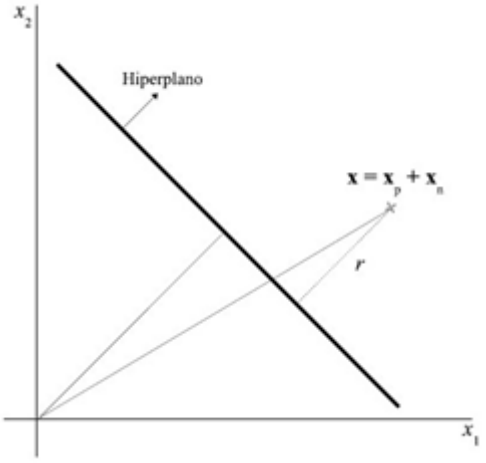

Figura 6: Interpretação gráfica da distância **x** até o hiperplano para o caso bidimensional

Considerando g ($\mathbf{x}_p$) = 0

$$\text{g}(\mathbf{x}) = \boldsymbol{w}_o^T . \boldsymbol{x}_o + b_o = r . ||\boldsymbol{w}_o|| \tag{2.7}$$

onde, através da equação (2.7) é obtido a distância *r*

$$r = \frac{|\, g(\mathbf{x})|}{||\boldsymbol{w}_o||} \tag{2.8}$$

O conjunto de treinamento é linearmente separável se $\boldsymbol{w}_o$ e b_o satisfazem a restrição

$$\begin{cases} \boldsymbol{w}_0^T . \ \boldsymbol{x}_i + \ b_0 \geq \ +1, \ se \ d_i = +1 \\ \boldsymbol{w}_0^T . \ \boldsymbol{x}_i + \ b_0 \leq \ -1, \ se \ d_i = -1 \end{cases} \tag{2.9}$$

onde os parâmetros $\boldsymbol{w}_0$ e b_0 são obtidos somente através do conjunto de treinamento.

A equação (2.9) pode ser reescrita por:

$$d_i(\boldsymbol{w}_0^T \cdot \boldsymbol{x}_i + b_0) \geq 1 \tag{2.10}$$

Os pontos (***x***,*d*), onde a equação (2.10) é satisfeita para o sinal de igualdade são denominados de vetores-suporte, e são esses pontos que influenciam diretamente na localização do hiperplano ótimo de máxima margem, pois, esses pontos estão mais próximos da superfície de decisão.

Considerando um ponto $\boldsymbol{x}^{(s)}$ vetor-suporte de classe positiva $d^{(s)}$ = +1, então por definição:

$$g\left(\boldsymbol{x}^{(s)}\right) = \ \boldsymbol{w}_0^T \cdot \boldsymbol{x}^{(s)} + b_0 - 1 \ \ para \ \ d_i = +1 \tag{2.11}$$

Da equação (2.8), a distância do vetor de suporte $\boldsymbol{x}^{(s)}$ até o hiperplano ótimo é dado por:

$$r = \frac{\boldsymbol{w}_0^T \cdot \boldsymbol{x}^{(s)} + b_0}{\|\boldsymbol{w}_0\|} = \begin{cases} +\dfrac{1}{\|\boldsymbol{w}_0\|} & se\ d^{(s)} = +1 \\ -\dfrac{1}{\|\boldsymbol{w}_0\|} & se\ d^{(s)} = -1 \end{cases} \tag{2.12}$$

onde, o sinal positivo indica que $\boldsymbol{x}^{(s)}$ pertence ao lado positivo do hiperplano ótimo e o sinal negativo o contrário. Considerando ρ a margem de separação máxima entre duas classes de um conjunto de treinamento, então:

$$\rho = 2r = \frac{2}{\|\boldsymbol{w}_0\|} \tag{2.13}$$

Logo, a equação (2.13) mede a distância entre os hiperplanos da equação (2.10), da mesma forma que a distância entre os hiperplanos $\boldsymbol{w}^T \cdot \boldsymbol{x} + b = 0$ e $\boldsymbol{w}^T \cdot \boldsymbol{x} + b = 1$ ou $\boldsymbol{w}^T \cdot \boldsymbol{x} + b = -1$ é dado por $\frac{1}{\|\boldsymbol{w}_0\|}$. Como é suposto que a margem de separação é sempre maior que esta última distância, a minimização de $\|\boldsymbol{w}\|$ leva a uma maximização da margem.

O hiperplano ótimo para classes linearmente separáveis definido para os parâmetros $\boldsymbol{w}$ e b que satisfaçam as desigualdades da equação (2.10), pode ser reescrito como:

$$d_i(\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b) \geq 1 \tag{2.14}$$

O objetivo da SVM é encontrar um procedimento computacional que, utilizando o conjunto de treinamento $\{(\boldsymbol{x}_i, d_i)\}_{i=1}^N$ encontra o hiperplano ótimo sujeito às restrições da equação (2.14). Este problema pode ser resolvido através do problema de otimização com restrições, minimizando a função custo Φ em relação ao vetor de peso $\boldsymbol{w}$ e satisfazendo as restrições da equação (2.14)

$$\Phi(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^T \cdot \boldsymbol{w} \tag{2.15}$$

A partir da função custo Φ da equação (2.15) pode ser formulado o problema de otimização com restrições, denominado de problema primal:

Minimizar: $\frac{1}{2}\boldsymbol{w}^T \cdot \boldsymbol{w}$ (2.16)

Sujeito as restrições: $d_i(\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b) \geq 1,\ para\ i = 1, \dots, n$

Este é um problema clássico em otimização de programação quadrática (Hearst, 1998) sob o aspecto de aprendizado de máquina. O problema de otimização analisado sob o ponto de vista de otimização de função quadrática pode ser resolvido introduzindo uma função lagrangiana, definida em termos de $\boldsymbol{w}$ e b:

$$J(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{N} \alpha_i (d_i(\boldsymbol{w}^T \cdot \boldsymbol{x} + b) - 1) \tag{2.17}$$

onde, os α_i são denominados de multiplicadores de Lagrange não-negativos.

O problema passa a ser então a minimização da equação (2.17) em relação a $\boldsymbol{w}$ e b e maximização de α_i, com $\alpha_i \geq 0$. Os pontos ótimos desta equação são obtidos diferenciando a equação (2.17) em relação a $\boldsymbol{w}$ e b igualando os resultados a zero, obtendo as condições de otimização:

$$\begin{aligned} &\text{Condição 1:} && \frac{\partial J(\boldsymbol{w},b,\alpha)}{\partial \boldsymbol{w}} = 0 \\ &\text{Condição 2:} && \frac{\partial J(\boldsymbol{w},b,\alpha)}{\partial \boldsymbol{b}} = 0 \end{aligned} \tag{2.18}$$

A aplicação das condições de (2.18) à função lagrangiana da equação (2.17) levam ao resultado:

$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i d_i \boldsymbol{x}_i \tag{2.19}$$

$$\sum_{i=1}^{N} \alpha_i d_i = 0$$

Substituindo a equação (2.19) em (2.17), obtém-se o problema dual de otimização:

$$\begin{aligned} &\text{Maximizar:} && \textstyle\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \boldsymbol{x}_i^T \cdot \boldsymbol{x}_j \\ &\text{Sujeito as restrições:} && \begin{cases} (1)\alpha_i \geq 0, \ i = 1, \ldots, N \\ (2)\sum_{i=1}^{N} \alpha_i d_i = 0 \end{cases} \end{aligned} \tag{2.20}$$

Tendo encontrado os multiplicadores de Lagrange ótimos, representados por α_{oi}, pode-se calcular o vetor de peso ótimo $\boldsymbol{w}_0$ através da equação (2.19):

$$\boldsymbol{w}_0 = \sum_{i=1}^{N} \alpha_{0i} d_i \boldsymbol{x}_i \tag{2.21}$$

O valor do bias ótimo b_0 é encontrado utilizando os pesos ótimos $\boldsymbol{w}_0$ encontrados na equação (2.21) e descrito como:

$$b_0 = 1 - \boldsymbol{w}_0^T \cdot \boldsymbol{x}^{(s)} \ para \ d^{(s)} = 1 \tag{2.22}$$

O problema dual (2.20) é formulado totalmente em termos dos padrões de treinamento, além disso, a equação a ser maximizada da equação (2.20) depende somente dos padrões de entrada. O hiperplano ótimo depende somente dos vetores de suporte, considerados os padrões mais significativos do conjunto de treinamento. Os multiplicadores de Lagrange $\alpha_0 > 0$

(diferente de zero) são justamente os padrões de entrada com margem igual a 1, chamados de vetores de suporte.

O hiperplano ótimo é expresso em termos do conjunto de vetores de suporte descrito pela função sinal como:

$$f(\boldsymbol{x}) = sgn(\sum_{i=1}^{N_{SV}} d_i \alpha_{0i} \boldsymbol{x}^T \cdot \boldsymbol{x} + b_0) \tag{2.23}$$

Os padrões de entrada que não são vetores de suporte não influenciam na função de decisão da escolha do hiperplano ótimo pela da SVM.

Um ponto importante para este problema de otimização é que fornece solução única encontrado eficientemente, diferente das outras técnicas.

2.2.1.2 Máquina de vetor de suporte não linear

O problema de classificação binária, onde, classes distintas não são linearmente separáveis no espaço original, mas, com um mapeamento não linear através de um produto interno *kernel* transforma o espaço original em um espaço de características de dimensão maior, e, o problema que era não linearmente separável no espaço original passa-se a ser linearmente separável no espaço de características é representado pela SVM não linearmente separável ou SVM para classes linearmente separáveis no espaço de características.

O espaço de características, mencionado acima, corresponde a uma representação do conjunto de treinamento, um mapeamento do espaço de entrada original em um novo espaço utilizando funções reais φ_i, ..., φ_M. A Figura 7 ilustra este conceito.

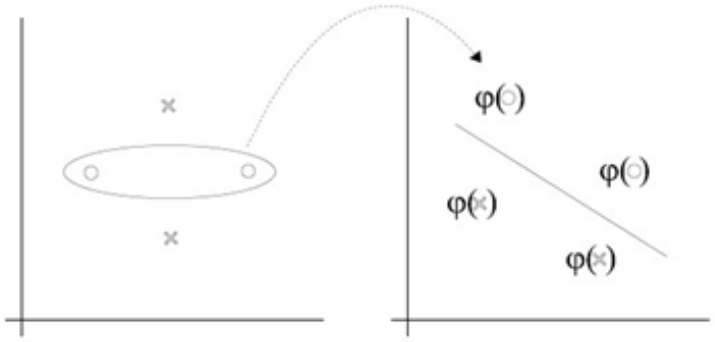

Figura 7: Mapeamento de características

Para a construção da SVM no caso não linear, a ideia depende de duas operações matemáticas. Primeiro: o mapeamento não linear do vetor de entrada para um espaço de características de alta dimensionalidade. O teorema que trata dessa operação é o teorema de Cover (Haykin, 2001), onde as funções φ_i são não lineares e a dimensionalidade do espaço de características M é suficientemente alta. Segundo: a construção de um hiperplano ótimo para separação das características descobertas no primeiro, uma vez que o teorema de Cover não procura o hiperplano ótimo de separação. A fundamentação desta última operação está na teoria da dimensão VC que busca o princípio de minimização do risco estrutural (Haykin, 2001; Lorena & Carvalho, 2003; Semolini, 2002).

Considerando, em uma visão geral, o problema de classificação, onde, as classes são não linearmente separáveis, a construção do hiperplano de separação, dado os padrões de treinamento, possivelmente gerará erros de classificação. O objetivo da SVM neste caso é encontrar um hiperplano que minimiza a probabilidade de erro de classificação junto com o conjunto de treinamento.

Existem alguns casos onde, não é necessário fazer um mapeamento de características no conjunto de treinamento. Esses casos são tratados pela SVM linear com margens de separação entre classes suaves ou flexíveis (*soft*), pois, poderão existir pontos ($\boldsymbol{x}_i$,d_i) que violarão a equação (2.14)

Esta violação pode ocorrer em três diferentes situações descritas a seguir:

1. O ponto ($\boldsymbol{x}_i$,d_i) se encontra dentro da região de separação e no lado correto da superfície de decisão, ilustrado na Figura 8(a). Neste caso, houve uma escolha incorreta do hiperplano.
2. O ponto ($\boldsymbol{x}_i$,d_i) se encontra dentro da região de separação e no lado incorreto da superfície de decisão, ilustrado na Figura 8(b). Neste caso, houve uma escolha incorreta do hiperplano de margem maior.
3. O ponto ($\boldsymbol{x}_i$,d_i) se encontra fora da região de separação e no lado incorreto da superfície de decisão, ilustrado na Figura 8(c).

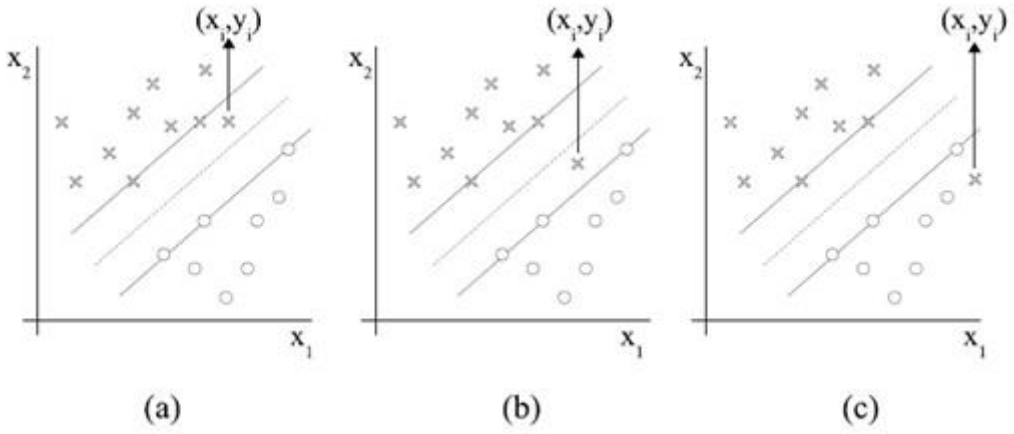

Figura 8: (a) O ponto ($\mathbf{x}_i$,d_i) se encontra na região de separação, mas do lado correto. (b) O ponto ($\mathbf{x}_i$,d_i) se encontra na região de separação, mas do lado incorreto. (c) O ponto ($\mathbf{x}_i$,d_i) se encontra fora da região de separação, mas do lado incorreto.

Para tratar desses problemas introduz-se uma variável não negativa $\{\xi_i\}_{1 \le i \le N}$ na definição do hiperplano de separação:

$$d_i(\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i \qquad (2.24)$$

As variáveis ξ_i são denominadas de variáveis soltas, e medem os desvios dos pontos ($\boldsymbol{x}_i$, d_i) para a condição ideal de separação de classes. Quando ξ_i satisfazer $0 \leq \xi_i \leq 1$ o ponto encontra-se dentro da região de separação, mas do lado correto da superfície de decisão. Quando $\xi_i > 1$ o ponto encontra-se do lado incorreto do hiperplano de separação.

Os vetores-suportes são os pontos que o resultado da equação (2.24) é igual a 1 - ξ_i mesmo que $\xi_i > 0$. Ao retirar um padrão do conjunto de treinamento em que $\xi_i > 0$ a superfície

de decisão tem possibilidade de mudança, porém, ao retirar um padrão em que $\xi_i = 0$ e o resultado da equação (2.24) for maior que 1 a superfície de decisão permanecerá inalterada.

O objetivo é encontrar um hiperplano de separação onde o erro de classificação incorreta seja mínimo perante o conjunto de treinamento, podendo ser feito minimizando a equação:

$$\Phi(\xi) = \sum_{i=1}^{N} I(\xi_i - 1) \tag{2.25}$$

em relação ao vetor peso $\boldsymbol{w}$, sujeito à restrição da equação do hiperplano de separação da equação (2.24) e a restrição sobre $\boldsymbol{w}^T\boldsymbol{w}$. A função $I(\xi - 1)$ é uma função indicadora, definida por:

$$I(\xi - 1) = \begin{cases} 0 & se\ \xi \leq 0 \\ 1 & se\ \xi > 0 \end{cases} \tag{2.26}$$

A minimização de $\Phi(\xi)$ é um problema de otimização não convexo de classe NP-completo não determinístico em tempo polinomial. Para fazer este problema de otimização matematicamente tratável, aproxima-se a função $\Phi(\xi)$ por:

$$\Phi(\xi) = \sum_{i=1}^{N} \xi_i \tag{2.27}$$

Para a simplificação de cálculos computacionais a função a ser minimizada em relação ao vetor peso $\boldsymbol{w}$ segue:

$$\Phi(\boldsymbol{w}, \xi) = \frac{1}{2}\boldsymbol{w}^T \cdot \boldsymbol{w} + C \sum_{i=1}^{N} \xi_i \tag{2.28}$$

onde, o parâmetro C controla a relação entre a complexidade do algoritmo e o número de amostras do conjunto de treinamento classificados incorretamente, sendo denominado de parâmetros de penalização.

A minimização do primeiro termo da equação (2.28) está relacionada à minimização da dimensão VC da SVM. O segundo termo pode ser visto como um limitante superior para o número de erros no padrão de treinamento apresentados à máquina. Logo, a equação (2.28) satisfaz os princípios de minimização do risco estrutural.

O problema de otimização em sua representação primal para encontrar o hiperplano ótimo de separação para classes não linearmente separáveis pode ser escrito como:

Minimizar: $\frac{1}{2}\boldsymbol{w}^T \cdot \boldsymbol{w} + C \sum_{i=1}^{N} \xi_i$

Sujeito a: $\begin{cases} (1) d_i(\boldsymbol{w}^T \cdot \boldsymbol{w} + b) \geq 1 - \xi_i, para\ i = 1, \dots, N \\ (2) \xi_i \geq 0,\ \forall i = 1, \dots, N \end{cases}$ (2.29)

Utilizando o método dos multiplicadores de Lagrange, pode-se formular o problema de otimização primal em seu correspondente problema dual de maneira similar à descrita na seção 2.2.1.1.

$$\begin{aligned} &\text{Maximizar:} && \textstyle\sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \boldsymbol{x}_i^T \cdot \boldsymbol{x}_j \\ &\text{Sujeito a:} && \begin{cases} (1)\sum_{i=1}^{N} d_i \alpha_i \\ (2) 0 \le \alpha_i \le C, para\ i = 1, \dots, N \end{cases} \end{aligned} \tag{2.30}$$

onde, $C > 0$ é especificado pelo usuário.

A principal diferença entre o caso de classes linearmente separáveis, e o caso de classes não linearmente separáveis é que a restrição $\alpha_i \geq 0$ é substituída por uma mais forte $0 \leq \alpha_i \leq C$.

O vetor de pesos ótimos $\boldsymbol{w}_0$ é calculado da mesma maneira do caso de classe linearmente separáveis, equação (2.21). O bias ótimo *b* também segue um procedimento similar ao descrito anteriormente, equação (2.22).

Existem casos também onde, é necessário mapear o espaço de entrada não linear para um espaço de características. Para realizar esse mapeamento, as funções *kernel* ou produto do núcleo interno são utilizadas e que serão apresentados a seguir.

Existem muitos casos onde não é possível dividir satisfatoriamente os padrões do conjunto de treinamento através de um hiperplano, mesmo observando as variáveis soltas. Para a realização desta tarefa é feito um mapeamento no domínio do espaço de entrada do conjunto de treinamento para um novo espaço, o espaço de características, usando uma função *kernel* apropriada.

Um *kernel k* é uma função que recebe dois pontos $\boldsymbol{x}_i$ e $\boldsymbol{x}_j$ do espaço de entrada e computa o produto escalar $\varphi^T(\boldsymbol{x}_i) \cdot \varphi(\boldsymbol{x}_j)$ no espaço de características.

O termo $\varphi^T(\boldsymbol{x}_i) \cdot \varphi(\boldsymbol{x}_j)$ representa o produto interno dos vetores $\boldsymbol{x}_i$ e $\boldsymbol{x}_j$, sendo o *kernel* representado por:

$$k(\boldsymbol{x}_i \boldsymbol{x}_j) = \varphi^T(\boldsymbol{x}_i) \cdot \varphi(\boldsymbol{x}_j) \tag{2.31}$$

Adaptando a equação (2.21) envolvendo um espaço de características, pode ser reescrito como:

$$\boldsymbol{w} = \sum_{i,j=1}^{N} \alpha_i d_i \varphi^T(\boldsymbol{x}_i) \cdot \varphi(\boldsymbol{x}_j) \tag{2.32}$$

onde, o vetor de característica $\varphi(\boldsymbol{x}_i)$ corresponde ao padrão de entrada $\boldsymbol{x}_i$ no i-ésimo exemplo.

Dessa forma, pode ser usado o produto interno $k(\boldsymbol{x}_i,\boldsymbol{x}_j)$ para construir um hiperplano ótimo no espaço de características sem ter que considerar o próprio espaço de características de forma explícita, observe a equação (2.32) em (5.1):

$$\sum_{\mathrm{i,j=1}}^{\mathrm{N}} \alpha_{\mathrm{i}} \mathrm{d}_{\mathrm{i}} \mathrm{k}(\mathbf{x}_{\mathrm{i}}, \mathbf{x}_{\mathrm{j}}) \tag{2.33}$$

A utilização de *kernel* está na simplicidade de cálculos e na capacidade de representar espaços muito abstratos.

As funções φ devem pertencer a um domínio em que seja possível o cálculo de produtos internos. No geral, utiliza-se o teorema de Mercer para satisfazê-las. Segundo o teorema, os *kernels* devem ser matrizes positivamente definidas, isto é, $k_{ij} = k(\boldsymbol{x}_i,\boldsymbol{x}_j)$, para todo $i,j = 1, ..., N$, deve ter auto-vetores maiores que 0.

Alguns *kernels* mais utilizados são: os polinomiais, os gaussianos ou RBF (*Radial Basis Function*) e o sigmoidais.

Tabela 1: Principais *kernels* utilizados nas SVMs

Kernel	Função $k(\boldsymbol{x}_i.\boldsymbol{x}_j)$	Comentários
Polinomial	$(\boldsymbol{x}_i^T \cdot \boldsymbol{x}_j + 1)^p$	p é especificado *a priori* pelo usuário
RBF	$e^{(-\frac{1}{2\sigma^2}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2)}$	a largura de σ^2 é especificada *a priori* pelo usuário
Sigmoidal	$\tanh(\beta_0 \boldsymbol{x}_i^T \cdot \boldsymbol{x}_j + \beta_1)$	Teorema de Mercer satisfeito somente para β_0 e β_1

A obtenção de um classificador por meio do uso de SVMs envolve a escolha de uma função *kernel* apropriada, além de parâmetros desta função e do algoritmo de determinação do hiperplano ótimo. A escolha do *kernel* e de seus parâmetros afetam o desempenho do classificador através da superfície de decisão.

2.3 Regressão Linear

A análise de regressão estuda a relação entre uma única variável dependente e uma ou várias variável (is) independente (s), a partir de uma análise estatística. É representada por um modelo matemático (equação) que faz a associação entre essas variáveis, através de um gráfico, chamado diagrama de dispersão, afim de encontrar essa relação. Este modelo é conhecido como Modelo de Regressão Linear Simples, se representa uma relação linear entre uma variável dependente e uma variável independente, ou é chamado de Modelo de Regressão Linear Múltiplo, caso se trate da incorporação de mais de uma variável independente. Existem outros modelos de regressão, que estão ligados ao número e distribuição das variáveis explicativas (independentes), cujos os efeitos na variável resposta (dependente), se deseja estudar, mas abordaremos apenas os dois já citados por serem os mais comuns.

De uma forma geral, pode-se dizer que esta técnica é utilizada para definir a influência de uma variável investigativa (denominada *X*), sobre um valor esperado de uma variável resposta (denominada *Y*), ou seja, o objetivo é identificar e analisar as alterações do valor esperado de *Y* ($E[Y]$), se é afetado pela alteração das condições que interagem com a variável *Y*. Enquanto a variável *X* deve fornecer informações a respeito do comportamento de *Y*.

O comportamento da variável resposta com relação a variável investigativa, pode ser representada de forma linear, quadrática, cúbica, exponencial, logarítmica, entre outras. Identificar o tipo de curva e uma equação de um modelo matemático mais próximo aos pontos encontrados no diagrama de dispersão, é primordial para o estabelecimento de um modelo que explique este fenômeno. Porém, alguns pontos do diagrama de dispersão não se ajustarão perfeitamente à curva do modelo proposto. Este fato se dá pelo fenômeno não ser propriamente matemático, e sim um fenômeno que se sujeita a influências acontecidas ao acaso, explicando então o objetivo principal da regressão linear.

Alguns fatores devem ser levados em consideração no momento de decisão do modelo que mais se adeque a situação, sendo elas:

1. Para que o modelo represente em termos práticos o fenômeno em estudo, ele deve ser condizente tanto no grau, quanto em seu aspecto.
2. Apenas as variáveis relevantes para explicação do fenômeno devem estar contidas no modelo

O Método dos Mínimos Quadrados (MMQ) é utilizado nessa situação para realizar a aproximação dos pontos do diagrama de dispersão aos pontos do modelo matemático escolhido. De forma resumida, este método realiza a soma dos quadrados da distância entre os pontos do modelo e os pontos do diagrama, obtendo assim uma relação entre X e Y, para o modelo que for escolhido, visando o menor erro possível.

2.3.1 Regressão linear simples

Como já foi dito anteriormente, na Regressão Linear Simples, apenas uma variável explicativa é utilizada para determinar o comportamento esperado de *Y*. Essa relação linear pode ser vista na equação a seguir.

$$E[Y_i] = \alpha + \beta X_i + \varepsilon_i \tag{2.34}$$

onde α = coeficiente linear da reta, também conhecido como termo constante da equação, β = coeficiente de regressão ou angular, o índice *i* é a referência de cada observação ($i = 1, 2, ..., n$),

X_i são as n observações da variável explicativa e ε_i = desvio entre a observação real e o valor estimado por $E[Y_i]$. Os erros dessa última variável apresentam o valor médio igual a zero e a variância σ^2, com distribuição Normal.

Os parâmetros α e β serão obtidos a partir de uma amostra de n pares de valores das variáveis (X_i,Y_i), $i = 1, 2, ..., n$, observados. O método dos mínimos quadrados é utilizado a fim de minimizar a soma dos quadrados dos desvios, ε_i, $i = 1, ..., n$ (Silva, 2016). Vale ressaltar que as estimativas dos parâmetros de β, obtidos pela técnica dos mínimos quadrados são imparciais ou não tendenciosos, ou seja,

$$E(a) = \alpha \qquad \text{e} \qquad E(b) = \beta$$

onde a e b são as estimativas de mínimos quadrados de α e β, respectivamente, e são definidos pelas fórmulas:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} X^2 \sum_{i=1}^{n} Y - \sum_{i=1}^{n} (XY) \sum_{i=1}^{n} X}{n \sum_{i=1}^{n} X^2 - \left(\sum_{i=1}^{n} X\right)^2} \tag{2.35}$$

$$\hat{\beta} = \frac{n \sum_{i=1}^{n} (XY) - \sum_{i=1}^{n} X \sum_{i=1}^{n} Y}{n \sum_{i=1}^{n} X^2 - \left(\sum_{i=1}^{n} X\right)^2} \tag{2.36}$$

Seu desenvolvimento é a partir da definição do método de mínimos quadrados, visando determinar α e β, e minimizando

$$\sum_{i=1}^{n} (Y_i - \beta X_i - \alpha)^2 \tag{2.37}$$

desenvolvendo o quadrado e descartando os termos constantes, temos

$$\beta^2 \sum_{i=1}^{n} X^2 + n\alpha^2 - 2\beta \sum_{i=1}^{n} (XY) - 2\alpha \sum_{i=1}^{n} Y + 2\alpha\beta \sum_{i=1}^{n} X \tag{2.38}$$

usando propriedades de cálculo e transformação de coordenadas

$$\alpha = \alpha_1 - \frac{\sum_{i=1}^{n} X}{n} \beta = \alpha_1 - \beta \bar{X} \tag{2.39}$$

substituindo teremos

$$\beta^2 \sum_{i=1}^{n} X^2 + n\alpha_1^2 - \frac{\left(\sum_{i=1}^{n} X\right)^2}{n} \beta^2 - 2\beta \sum_{i=1}^{n} (XY) - 2\alpha_1 \sum_{i=1}^{n} Y + 2\bar{X} \sum_{i=1}^{n} Y \beta \tag{2.40}$$

e separando na soma de duas expressões quadráticas independentes, que possibilitam a minimização usando matemática elementar

$$n\alpha_1^2 - 2\alpha_1 \sum_{i=1}^{n} Y \tag{2.41}$$

$$\beta^2 \sum_{i=1}^{n} X^2 - \frac{\left(\sum_{i=1}^{n} X\right)^2}{n}\beta^2 - 2\beta \sum_{i=1}^{n} (XY) + 2\frac{\sum_{i=1}^{n} X \sum_{i=1}^{n} Y}{n}\beta \tag{2.42}$$

admitindo como valores minimizadores:

$$\alpha_1 = \frac{\sum_{i=1}^{n} Y}{n} \tag{2.43}$$

$$\alpha = \bar{Y} - \bar{X}\beta \tag{2.44}$$

$$\beta = \frac{n\sum_{i=1}^{n}(XY) - \sum_{i=1}^{n} X \sum_{i=1}^{n} Y}{n\sum_{i=1}^{n} X^2 - \left(\sum_{i=1}^{n} X\right)^2} \tag{2.45}$$

Definindo que a relação entre $\hat{\alpha}$ e $\hat{\beta}$ é dada por

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \tag{2.46}$$

A partir do momento em que definimos um modelo como de Regressão Linear Simples, podemos pressupor que a relação entre *X* e *Y* é linear como já foi dito. Que os valores de *X* são fixos, portanto, esta não é uma variável aleatória. A média do erro é representada por $E(u_i) = 0$, ou seja, a média é nula. Para um dado valor de *X*, a variância do erro *u* é sempre σ^2, e é chamada de variância residual, que pode ser representada pelas equações

$$E(u_i^2) = \sigma^2 \tag{2.47}$$

ou

$$E[Y_i - E(Y_i|X_i)]^2 = \sigma^2 \tag{2.48}$$

Esse erro então é dito homocedástico ou dizemos que temos homocedasticia. Podemos definir também que os erros não são correlacionados entre as observações, ou seja, $E(u_i u_j) = 0$, para $i \neq j$. Além de que os erros apresentam distribuição Normal (Hoffmann, 2016).

Para o ajuste de uma regressão desse tipo, precisamos de no mínimo, três observações. Tendo em vista que quando houver apenas duas, a determinação da reta passa a ser um problema de geometria analítica, não sendo possível a realização de uma análise estatística.

Na figura abaixo está representado um modelo de regressão linear simples, em que $E(Y_i) = \alpha + \beta X_i$, ou seja, as médias das distribuições *X* | *Y* estão sobre a reta.

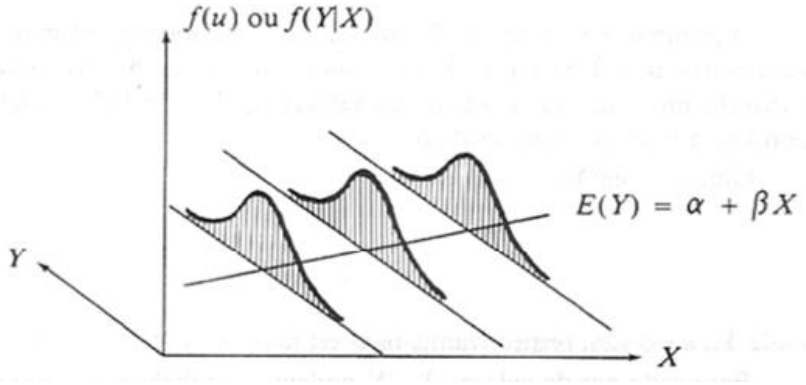

Figura 9: Representação de um modelo estatístico de uma regressão linear simples (Hoffmann, 2014)

2.3.2 Regressão linear múltipla

Regressão Linear Múltipla é um método estatístico de estimação e previsão de valores, com uma variável de resposta a partir de um conjunto de variáveis regressoras. O objetivo principal dessa técnica é realizar uma avaliação da relação de uma única variável de interesse *Y* (dependente ou resposta) com as *n* variáveis X_j (independentes), tal que *j* = 1, 2, 3, ..., *n*.

O modelo estatístico de uma regressão linear múltipla com *k* variáveis explanatórias é dado por:

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_k X_{kj} + u_j, \quad j = 1, \ldots, n \tag{2.49}$$

ou

$$Y_j = \alpha + \sum \beta_i X_{ij} + u_j \tag{2.50}$$

Este modelo também pode ser representado de forma matricial, como mostrado a seguir, a partir da fórmula

$$Y = X\beta + \varepsilon \tag{2.51}$$

com

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

onde $\boldsymbol{Y}$ é um vetor $n \times 1$ cujos os componentes correspondem às n respostas, $\boldsymbol{X}$ é uma matriz dimensão $n \times (p+1)$ chamada de matriz do modelo, ε é um vetor de dimensão $n \times 1$ composto pelos erros e β é um vetor $(p+1) \times 1$ onde os elementos são coeficientes de regressão.

Assim como no modelo de regressão linear simples, no múltiplo podemos apresentar algumas pressuposições.

1. A variável dependente (Y_j) é a função linear das variáveis X_{ij}, $i = 1, ..., k$.
2. Os valores das variáveis X são fixos.
3. A média dos erros também é zero ($E(u_j) = 0$), e pode ser representada em forma de vetor de zeros ($E(\boldsymbol{u}) = 0$).
4. Os erros são homocedásticos, ou seja, $E(u_j^2) = \sigma^2$.
5. Não há correlação entre os erros ($E\ (u_j\ u_h) = 0, j \neq h$).
6. Os erros apresentam distribuição Normal (Silva, 2016).

A estimativa de coeficientes de regressão também é feita pelo método de Mínimos Quadrados, assim como na simples. O Coeficiente de Determinação (R^2), é o parâmetro utilizado no julgamento de adequação de um modelo de regressão, através da fórmula (2.52)

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2 \Big/ (n-1)}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \Big/ (n-1)} \qquad (2.52)$$

em que $e_i = Y_i - \hat{Y}_i$ é o resíduo da i-ésima observação.

Seu intervalo de compreensão é $0 \leq R^2 \leq 1$. Quanto mais próximo R^2 está de 1, significa que as variáveis independentes do modelo possuem uma relação mais próxima à linear com a variável dependente.

3 MATERIAIS E MÉTODOS

O conjunto de dados utilizado para a aplicação dessas técnicas de Inteligência Computacional, foi extraído da base de dados pública "*Heart Disease Database*". Esta base em sua estrutura original apresenta quatro subconjuntos de dados, retirados de diferentes localidades (Hungria, Suíça e Estados Unidos), porém, devido aos dados faltantes, neste trabalho foi usado apenas o subconjunto de Cleveland, que foram obtidos por Robert Detrano no *Cleveland Clinic Foundation*. Este apresenta informações de 303 pacientes, dos quais, 164 foram considerados saudáveis e 139 doentes. A lista original de dados possui setenta e seis diferentes atributos a respeito de cada um dos pacientes, mas foram pré-selecionadas apenas treze desses setenta e seis, que foram considerados de maior relevância pelos médicos responsáveis, além de possuírem um número irrisório de dados faltantes. A seguir serão listados esses treze diferentes atributos que foram selecionados a respeito dos pacientes:

- Idade – variando de 29 a 77 anos;
- Gênero – feminino ou masculino, representados respectivamente por 0 e 1;
- Tipo de dor no peito – quatro distintos tipos de dores no peito:
  - 1: angina típica
  - 2: angina atípica
  - 3: sem dor anginal
  - 4: assintomático
- Pressão arterial em repouso – representada em mm Hg;
- Colesterol sérico – representado em mg/dl;
- Concentração de açúcar no sangue em jejum – >120 mg/dl
  - 1: verdadeiro
  - 0: falso
- Resultados eletrocardiográficos em repouso
  - 0: Normal
  - 1: Com onda ST-T anormal
  - 2: Mostrando provável hipertrofia do ventrículo esquerdo
- Ritmo cardíaco máximo alcançado;
- Angina induzida por exercício
  - 1: Presença
  - 0: Ausência
- Depressão da onda ST induzida pelo exercício em relação ao repouso;
- Inclinação do pico de segmento ST durante o exercício
  - 1: Inclinação ascendente
  - 2: Plano
  - 3: Inclinação descendente
- Número de grandes vasos coloridos por fluoroscopia - valor de 0 a 3;
- Talassenia
  - 3: Normal
  - 6: Defeito irreparável
  - 7: Defeito reversível.

Desses, oito foram utilizados inicialmente no modelo SVM apresentado neste trabalho: 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 os outros quatro, não apresentaram uma importância significativa para a obtenção de qualquer resultado nas técnicas computacionais utilizadas. Vale ressaltar que dos trezentos e três pacientes, seis ainda apresentavam dados faltantes, porém esses dados estavam dentre os quatro que foram descartados.

O modelo em que foi incluída a regressão, houve uma nova seleção de atributos ao ser implementado, 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 calculou novamente a performance com as variáveis que apresentaram correlação. Ou seja, o modelo de Regressão Linear foi utilizado apenas para determinar as variáveis que seriam novamente implementadas em um novo modelo SVM.

O fluxograma a seguir (Figura 10) representa de forma simplificada o processo que será realizado ao longo deste trabalho, a fim de demonstrar os modelos.

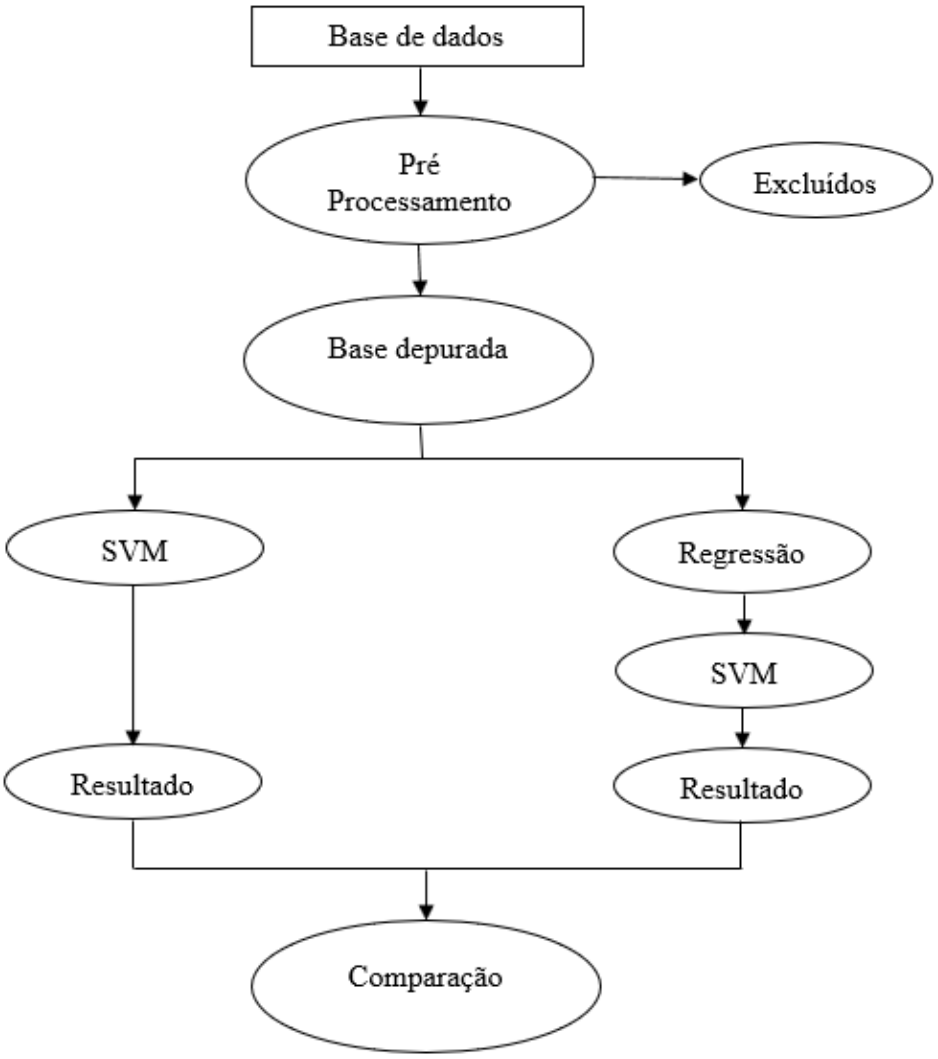

Figura 10: Fluxograma da construção dos modelos

De forma a classificar as 98 mulheres e 205 homens presentes no banco, dentre esses 270 considerados em idade útil (18 a 65 anos), pôde-se dividi-los dentro de cada uma das características fornecidas, de forma estatística, apresentadas na tabela a seguir, a fim de facilitar a interpretação de algumas informações. É válido lembrar que todos os valores utilizados como parâmetro para classificação dos pacientes são fornecidos pela Sociedade Brasileira de Cardiologia (SBC), juntamente com a Organização Mundial de Saúde (OMS), e que como a base de dados é antiga, alguns parâmetros que foram deixados como limiares para variáveis binárias, já foram alterados e/ou atualizados pelos órgãos competentes.

Tabela 2: Estatística de pacientes para cada variável no conjunto total

Característica	Parâmetro	Pacientes dentro do parâmetro	Pacientes fora do parâmetro
Pressão arterial em repouso	≤ 130mmHg	171 (56.4%)	132 (44.6%)
Colesterol sérico	< 240 mg/dl	147 (48.5%)	156 (51.5%)
Açúcar no sangue em jejum	≤ 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%)

Toda a parte estatística do trabalho foi implementada em R, sendo elas, os dois modelos e a comparação estatística dos resultados, para análise de diferença significativa entre ambos. O R é um software *Open Source* de linguagem própria, também chamada de R. Este possibilita o trabalho com modelos lineares e não lineares, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento e técnicas gráficas altamente expansíveis.

Para a simulação, utilizamos os conjuntos de treino e teste, com 80% e 20% dos dados respectivamente, e a partir dessa divisão foi feito um sumário dos dados, pelo próprio R, onde foram apresentados os valores mínimos, o 1°, 2° e 3° quartis (onde o segundo quartil vem sendo chamado de mediana), a média dos valores e o valor máximo. Essas informações serão apresentadas nas tabelas 3 e 4, a seguir, apenas para as variáveis de valores contínuos, separados para cada um dos conjuntos.

É válido lembrar o que significa cada um dos índices, já citados, apresentados no sumário para cada um dos conjuntos.

Valor mínimo: é o elemento de menor valor do conjunto.

Quartis: são considerados 3 valores, que dividem a amostra em 4 partes de mesmo tamanho. Esses valores são dados a partir da ordenação de forma crescente de todos os valores pertencentes ao conjunto de amostras. Sendo assim, o 1° quartil, é o número que deixa 25% das observações acima de si e 75% abaixo. O 2° quartil, também chamado de mediana, deixa 50% acima e 50% abaixo. O 3° quartil, é o inverso do primeiro, deixa 75% das observações acima e 25% abaixo. Em situações onde o conjunto possuir um número par de amostras, é feita uma média dos dois valores que se encontram na posição de encontro de cada um dos quartis.

Média: é representada pela soma de todos os elementos do conjunto, dividida pelo número de amostras.

Valor máximo: é o elemento de maior valor do conjunto.

Tabela 3: 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	1.07
3° Quartil	60	140	273.8	165	1.8
Valor máximo	77	200	564	202	6.2

Tabela 4: 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	0.92
3° Quartil	62	140	290	172	1.4
Valor máximo	77	170	417	187	4.4

Para o conjunto de treino, foram selecionados 242 pacientes (80% do total de pacientes do banco de dados), com idade variando entre 29 e 77 anos. E para o conjunto de teste, com 61 pacientes (20% do total de pacientes do banco de dados), com a idade variando entre 37 e 77 anos. Apesar da idade do grupo de teste ser maior, todos os outros dados (pressão arterial, colesterol, ritmo cardíaco e depressão da onda ST), tiveram como valor máximo, valores mais baixos dos que os apontados no conjunto de treino, sendo assim, as médias para essas variáveis também ficaram mais baixas para o conjunto de teste. Já se tratando dos valores mínimos, esse em sua maioria, foi mais baixo para o conjunto de treino, levando ao fato de que variáveis como colesterol e ritmo cardíaco apresentassem um intervalo maior entre valor mínimo e valor máximo. Os valores para depressão da onda ST foram baixos, pois o valor de normalidade para esta característica encontra se dentro da variação de 0 e 1 mm, portanto, considera-se pela média, que a maioria dos pacientes se encontram dentro desse padrão. O ritmo cardíaco é a variável mais complexa de ser calculada, pois ela apresenta um valor limite diferente para cada paciente, por ter um valor resposta dependente da idade de cada um, sendo assim, torna-se mais complicado uma avaliação a respeito dos números dessa variável. A pressão arterial e o colesterol apresentaram dados bastante alarmantes, desde suas medianas, os valores encontrados já são superiores aos definidos pela OMS como valores limites para esses índices, garantindo assim, que quase metade dos pacientes possuem alterações.

Foi utilizado o pacote *kernlab* (*Kernel-Based Machine Learning Lab*) que define a função *kernel* a ser utilizada no modelo. Para esta aplicação foi utilizada a função RBF (*Radial Basis Function*), esta pode ser considerada uma das funções de *kernel* mais comuns, principalmente quando sua aplicação é na caracterização de informações por Máquina de Vetor de Suporte.

O parâmetro C, é essencialmente de regularização, capaz de realizar a troca entre uma classificação diferente de exemplos do conjunto de treinamento, contra uma simples superfície de decisão. Este valor pode variar entre 0 e infinito, e caso seja considerado baixo, torna a superfície de decisão mais suave, se for alto, seu objetivo será de classificar corretamente os exemplos do treino. O valor unitário (C = 1) foi definido para estes modelos, a partir de testes com diferentes valores, chegando a esse valor, que foi satisfatório ao problema.

Os resultados finais serão avaliados a partir da comparação desses índices estatísticos, formadores da matriz de confusão, que serão encontrados como resultado parcial de cada modelo, como por exemplo, a acurácia, que é definida como o grau de confiança do modelo. Quanto mais próximo ao resultado esperado, maior será a acurácia do sistema, esta pode ser definida pela equação (3.1).

$$\text{Acurácia} = \frac{VP+VN}{N} = \frac{(Verdadeiro\ positivo+Verdadeiro\ negativo)}{Total\ lote} \quad (3.1)$$

A sensibilidade (**s**), é a probabilidade de um indivíduo acometido de uma cardiopatia ter o resultado de seu teste alterado, ou seja, receber como resultado do teste que ele se encontra saudável. Este índice pode ser calculado a partir da equação (3.2).

$$\text{Sensibilidade} = \frac{VP}{VP+FN} = \frac{Número\ de\ resultados\ de\ testes\ verdadeiros\ positivos}{Todos\ os\ doentes\ afetados} \quad (3.2)$$

onde VP = Verdadeiro Positivo e FN = Falso Negativo.

Já a especificidade (**e**) é o número de resultados negativos em pacientes que não possuem a doença, ou seja, os resultados positivos negativos. Calculada a partir da equação (3.3).

$$\text{Especificidade} = \frac{VN}{VN+FP} = \frac{Número\ de\ resultado\ de\ teste\ verdadeiros\ negativos}{Todos\ os\ doentes\ não\ afetados} \quad (3.3)$$

onde VN = Verdadeiro Negativo e FP = Falso Positivo.

Prevalência (**p**), que é a fração de indivíduos doentes na população avaliada, dada pela equação (3.4).

$$\text{Prevalência} = \frac{Número\ de\ indivíduos\ doentes}{Número\ total\ da\ população} \quad (3.4)$$

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

$$\text{VPP} = \frac{VP}{VP+FP} = \frac{Número\ de\ doentes\ positivos}{Todos\ os\ resultados\ positivos} \quad (3.5)$$

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

$$\text{VPN} = \frac{VN}{VN+FN} = \frac{Número\ de\ pacientes\ saudáveis}{Todos\ os\ resultados\ negativos} \quad (3.6)$$

onde VN = Verdadeiro Negativo e FN = Falso negativo.

A partir dessas equações é possível montar a matriz de confusão completa, que pode ser vista na Figura 11.

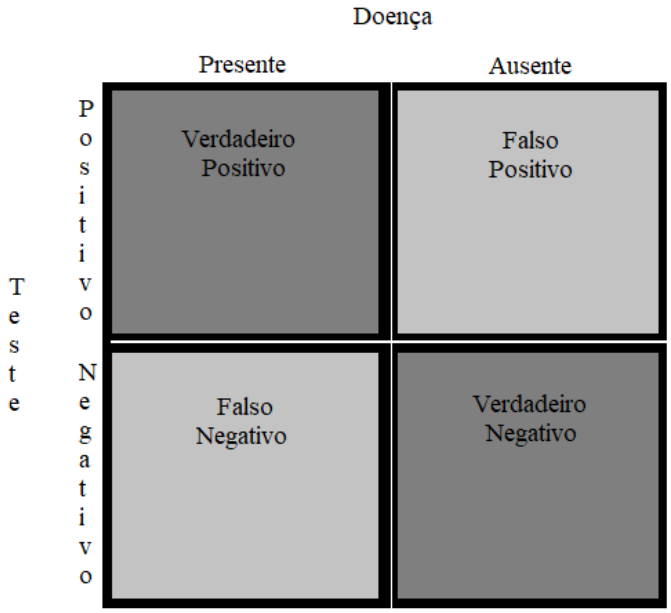

Figura 11: Matriz de confusão

Através da matriz de confusão gerada para um determinado ponto de operação, torna-se possível o cálculo dos acertos e erros dos classificadores, desde que a decisão de classificação tenha sido tomada neste ponto, tendo em vista que a matriz de confusão é formada por valores reais e por valores preditos por seus classificadores, estes já vistos anteriormente. Estes acertos e erros são chamados de:

Verdadeiro Positivo (VP): testes positivos para portadores da doença;
Falso Positivo (FP): testes positivos para pacientes saudáveis;
Falso Negativo (FN): testes negativos em pacientes doentes;
Verdadeiro Negativo (VN): testes negativos para pacientes saudáveis.

4 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 foram feitas por um modelo de Regressão Linear, e foram escolhidas aquelas que apresentaram índice de correlação a partir de um modelo GLM. Para ambos os casos foram realizadas cem simulações no software R, utilizando 80% dos dados como conjunto de treino e 20% no conjunto teste, dessa forma foram fornecidos dados de erro de treino, erro de validação cruzada, número do vetor suporte, acurácia, sensibilidade, especificidade e valor de falso negativo para estas simulações.

Baseando nos cem resultados obtidos para cada um dos modelos, foi realizada uma análise estatística, indicativa dos valores mais significativos e menos significativos dentre o conjunto de informações resposta do sistema. Para o primeiro modelo foram encontrados os valores de erros (treino e *cross*) presentes na Tabela 5. É válido ressaltar que os valores fornecidos pelo software apresentavam até 9 casas decimais depois da vírgula para cada resultado, porém foi adotado o arredondamento, para que sejam usadas apenas 2 casas decimais após a vírgula, assim como já vem acontecendo ao longo do trabalho.

Tabela 5: Estatística dos erros nos resultados das simulações SVM

	Erro de treino	Erro de *cross*
Valor mínimo	0.13	0.20
1° Quartil	0.15	0.23
Mediana	0.16	0.25
Média	0.17	0.24
3° Quartil	0.18	0.26
Valor máximo	0.21	0.30
Desvio padrão	0.01	0.02

Para os erros a primeira coisa que deve ser observada é a diferença entre eles. 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 *cross* superior ao erro de treino. Porém, esta operação de cálculo do erro de *cross* é totalmente necessária, por esta garantir uma robustez maior do sistema e a capacidade de generalização do modelo, com base em um determinado conjunto de dados, ou seja, este erro é o erro encontrado no método de aprendizado das observações que não foram utilizadas no treino, analisando assim o comportamento do sistema com a introdução de novos dados, desde que seja mantida a mesma probabilidade conjunta das variáveis explicativas e da resposta do treino. 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. É possível identificar que até a variação deles é diferente, sendo novamente a de erro de treino inferior, comprovada também pelo valor do desvio padrão encontrado.

Na Tabela 6 serão apresentados os valores de número de erro de vetor suporte para as simulações.

Tabela 6: Estatística para o número de vetor suporte encontrados nas simulações SVM

	N° de vetores suporte
Valor mínimo	141
1°Quartil	148
Mediana	152
Média	152
3°Quartil	156
Valor máximo	163
Desvio padrão	5,08

Este valor de número de vetores suporte está diretamente ligado às decisões, como pode ser visto na fórmula de decisão a seguir

$$f(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i) + b \tag{4.1}$$

onde *N* é o número de vetores de suporte, *K* é a função *kernel*, α_i e *b* são parâmetros encontrados durante o treinamento e *x* é o vetor de características.

A partir desta fórmula podemos concluir que quanto maior o número de vetores de suporte, maior será a complexidade do sistema, já que este apresentará um somatório cada vez maior. 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.

Ao analisarmos a acurácia, mostrada na Tabela 7, podemos concluir que obtivemos resultados bastante satisfatórios para este modelo.

Tabela 7: Estatística da acurácia das simulações SVM

	Acurácia
Valor mínimo	0.62
1° Quartil	0.72
Mediana	0.77
Média	0.76
3° Quartil	0.80
Valor máximo	0.87
Desvio padrão	0.05

Com base nas informações presentes nesta tabela, podemos 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.

Em apenas 11 simulações foram encontrados valores inferiores a 70% para a acurácia, esta, que em sua mediana já apresenta um valor considerado bom (77%), principalmente quando é analisado com seus valores consequentes, como sensibilidade, especificidade e falso negativo, que são apresentados a seguir, nas análises desses valores, nas tabelas a seguir, a começar pela sensibilidade (Tabela 8).

Tabela 8: Estatística da sensibilidade das simulações SVM

	Sensibilidade
Valor mínimo	0.47
1° Quartil	0.63
Mediana	0.70
Média	0.70
3° Quartil	0.77
Valor máximo	0.91
Desvio padrão	0.10

O desvio padrão encontrado para este índice é consideravelmente alto, quando comparado ao encontrado para outros classificadores, mostrando assim a grande variação que existe entre valores mínimo e máximo e a média, já que o desvio padrão é a medida que indica o grau de variação do conjunto de elementos, em outras palavras, o desvio padrão diz o quanto os valores dos quais se extraiu a média estão distantes dela.

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. Esta média ainda nos mostra, seguida do valor de mediana e 1° quartil, 70% e 63%, respectivamente, que os casos em que foram registrados uma sensibilidade muito baixa, foram raros, casos isolados. 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%.

Tabela 9: Estatística da especificidade das simulações SVM

	Especificidade
Valor mínimo	0.59
1° Quartil	0.78
Mediana	0.83
Média	0.82
3° Quartil	0.87
Valor máximo	0.96
Desvio padrão	0.07

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 confusão já apresentada. 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 elevadíssimo, perto do ideal.

Tabela 10: Estatística do falso negativo das simulações SVM

	Falso Negativo
Valor mínimo	0.09
1° Quartil	0.23
Mediana	0.29
Média	0.30
3° Quartil	0.36
Valor máximo	0.53
Desvio padrão	0.10

O valor de Falso Negativo pode ser dito um dos mais importantes para aplicações de Inteligência Computacional na área médica, tendo em vista o 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. Ao olhar por um lado ético e humano para o diagnóstico, podemos considerar esse não um dos índices mais importantes, mas sim o mais importante de todos os citados e descritos anteriormente, principalmente pelo valor consideravelmente baixo, encontrado como valor mínimo, 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.

Ao analisarmos o conjunto total de dados resposta do sistema, podemos adotar como melhor e pior simulações as descritas na Tabela 11.

Tabela 11: Pior e melhor simulação do modelo SVM

	Erro de treino	Erro de *cross*	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

Foi escolhido, após análise, o critério valor de Falso Negativo como o critério mais importante do modelo, já que há uma diferença consideravelmente alta entre os valores encontrados para esse índice, garantindo então que para a melhor simulação teríamos um bom valor de Falso Negativo, e consequentemente de sensibilidade, além de já ter sido visto que a chance de ter um bom resultado de acurácia para esta escolha seria também consideravelmente alta. Partindo deste ponto, quando o valor de Falso Negativo se torna o mais importante para a definição de pior e melhor simulação, foi buscada a simulação em que este classificador obteve seus valores extremos, sendo este então os valores de máximo e mínimo de falso negativo, respectivamente, das 100 simulações realizadas.

Para a melhor simulação escolhida, o valor de acurácia encontrado foi de 77%, sendo este o valor encontrado como mediano dentre as simulações. Já na pior simulação, foi encontrado para este classificador, o valor de 62%, que foi considerado o mais baixo dentre todas as simulações. 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.

Quanto a complexidade presente na decisão dessas duas simulações, que pode ser analisada a partir do número de vetores de suporte utilizados pelo sistema, não há uma diferença considerável deste critério, já que tivemos 147 para a pior simulação e 149 para a melhor.

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%.

De forma a realizar uma comparação entre os dois modelos implementados, a partir de agora, será feita uma análise estatística de forma individual do segundo modelo, para que logo em seguida este seja comparado ao primeiro. Lembrando que este segundo modelo se deu a partir da associação das técnicas de Regressão Linear Múltipla, para a seleção de variáveis correlacionadas que serão aplicadas à Máquina de Vetor de Suporte.

Novamente começaremos a análise de resultados pelos valores dos erros de treino e de validação cruzada, dados na Tabela 12.

Tabela 12: Estatística dos erros nos resultados das simulações Regressão + SVM

	Erro de treino	Erro de *cross*
Valor mínimo	0.15	0.18
1° Quartil	0.18	0.21
Mediana	0.19	0.22
Média	0.19	0.22
3° Quartil	0.19	0.23
Valor máximo	0.21	0.26
Desvio padrão	0.01	0.02

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 apresentou um valor 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 por exemplo, para as simulações apenas com SVM os valores mínimos dos erros foram 0,13 e 0,20, apresentando 0,07 de diferença, enquanto para os valores mínimos do modelo que foi aplicado a regressão, a diferença foi de apenas 0,03, como pode ser visto na tabela acima. Apesar dessa diferença, o desvio padrão permaneceu o mesmo.

A diferença entre os valores encontrados para os dois modelos foi bem pequena, mostrando que o fato de ter acrescentando a regressão ao SVM só reduziu a diferença entre os treinos dentro do próprio modelo. Os erros de treino do modelo com aplicação da regressão foram um pouco maiores aos do SVM, e os erros de *cross* foram um pouco mais baixos.

Continuando a análise estatística dos resultados, a seguir veremos a complexidade na decisão a ser tomada pelo modelo.

Tabela 13: Estatística do número de vetor suporte encontrados das simulações Regressão + SVM

	N° de vetores suporte
Valor mínimo	130
1° Quartil	138
Mediana	143
Média	143
3° Quartil	148
Valor máximo	156
Desvio padrão	6.61

A partir da Tabela 13, que mostra o número de vetores de suporte necessários na decisão do problema, podemos ver que para o modelo com presença da regressão, a complexidade de decisão diminuiu, pois para o modelo anterior, 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. A média apresentou quase 9 vetores a menos para este modelo, sendo 152 no primeiro e 143 no segundo.

A acurácia pode ser analisada a partir dos valores presentes na Tabela 14.

Tabela 14: Estatística da acurácia encontrada das simulações Regressão + SVM

	Acurácia
Valor mínimo	0.67
1° Quartil	0.74
Mediana	0.79
Média	0.78
3° Quartil	0.82
Valor máximo	0.87
Desvio padrão	0.05

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 X 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, por exemplo, 76% para o primeiro e 78% para o segundo. Podemos observar em uma análise geral das simulações do modelo com presença da regressão, que em 75% delas, o valor encontrado para acurácia, estiveram acima dos 74%, permitindo avaliar o modelo, em um contexto geral e do ponto de vista da acurácia, como um modelo bastante preciso, já que em grande maioria das simulações, os resultados dos testes realizados, forneceriam valores reais como resultado, garantindo uma precisão de resultados. Um valor similar foi encontrado para o 1° quartil do modelo de SVM (72%), estes valores demonstram que ambos os modelos apresentam bons valores de acurácia, ou seja, os valores fornecidos pela rede estão bem próximos do valor real dos dados, garantindo a autenticidade de ambos os resultados.

Tabela 15: Estatística da sensibilidade encontrada das simulações Regressão + SVM

	Sensibilidade
Valor mínimo	0.5
1° Quartil	0.63
Mediana	0.70
Média	0.69
3° Quartil	0.75
Valor máximo	0.86
Desvio padrão	0.08

A sensibilidade para este modelo apresentou uma variação menor de valores, 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.

Este modelo, conseguiu em sua pior simulação, um valor de 50% de sensibilidade, e o modelo anterior conseguiu também para sensibilidade, 46%, portanto para este modelo com a associação de técnicas, conseguiu-se uma porcentagem maior de chance do teste diagnosticar positivo a um indivíduo que realmente esteja doente. Apesar deste modelo apresentar um valor mínimo maior, seu valor máximo foi inferior ao já visto anteriormente, sendo este 86% e o anterior 91%. Já os valores de média e mediana variaram pouco, apenas 1%, caindo de 70% e 71%, para 69% e 70%, mostrando assim, que para este quesito nos resultados das simulações não ocorre uma variação comprometedora para nenhum dos modelos, tornando-os válidos e com resultados bastante proveitosos.

Tabela 16: Estatística da especificidade encontrada das simulações Regressão + SVM

	Especificidade
Valor mínimo	0.72
1° Quartil	0.83
Mediana	0.86
Média	0.86
3° Quartil	0.89
Valor máximo	1
Desvio padrão	0.05

Já a especificidade para este modelo, é indiscutivelmente melhor do que a encontrada pelas simulações do modelo anterior. Em uma das simulações do modelo em que ocorre a associação entre Regressão Linear e SVM, 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%, um valor que já é considerado bom para este classificador. Estes valores fizeram com que a média de todas as simulações fosse elevada com relação a média das simulações do modelo anterior, passando de 82% para 86%.

Tabela 17: Estatística do falso negativo nos resultados das simulações Regressão + SVM

	Falso Negativo
Valor mínimo	0.14
1° Quartil	0.25
Mediana	0.30
Média	0.31
3° Quartil	0.38
Valor máximo	0.5
Desvio padrão	0.08

Já 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. Apesar de 14% ser um valor mais alto quando comparado ao anterior, ainda assim é um valor bastante satisfatório quando se trata de falso negativo, pois ele garante que em apenas 14% das vezes que o modelo for aplicado, que um paciente receberá um exame com resultado negativo, sendo que ele é portador de qualquer doença cardiovascular.

Ao separarmos a pior e a melhor simulação dentre as 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 18: Pior e melhor simulação do modelo Regressão + SVM

	Erro de treino	Erro de *cross*	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

Seguindo a mesma ideia de escolher como melhor simulação aquela que apresentou o menor valor de falso negativo, e como pior simulação a que apresentou o maior valor, garantindo assim na melhor situação que a chance de erro de diagnóstico seja reduzida, foram escolhidas as simulações transcritas na tabela X.

A diferença entre os valores máximos e mínimos de falso negativo para as simulações deste modelo obtiveram um intervalo menor do que o obtido pelo primeiro modelo, apresentando uma variação de 36% entre a melhor (14%) e a pior (50%) simulação. Ao analisarmos de forma individual o valor de falso negativo para a pior simulação, concluímos que 50% é um valor consideravelmente baixo quando falamos de precisão de resultados de exame. Como consequência deste resultado, temos uma sensibilidade baixa (50%), principalmente quando comparada a da melhor simulação que apresentou um valor bem elevado (86%), relembrando a importância deste valor, ele representa a precisão de um teste ter resultado positivo, para uma pessoa cardiopata.

Quanto à diferença entre os valores de erro de validação cruzada da melhor e da pior simulação, ela pode ser considerada irrisória, 0,01, mostrando que para este modelo o erro de *cross* foi bem parecido para todas as simulações. Já a complexidade de decisão de ambas as simulações foi bem diferente, enquanto na pior simulação foram usados 132 vetores de suporte, na melhor simulação foram usados 152.

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%). Já na pior simulação, a acurácia foi de 70%, acompanhada de uma sensibilidade de 50% e especificidade de 88%.

5 CONCLUSÕES

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 bastante satisfatórios, visto que não foi necessária a utilização das 13 variáveis fornecidas pelo banco de dados 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 identificada 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 possam ser grandes auxiliadores 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 outro exame de custo mais elevado, desde que isso não interfira no diagnóstico final da doença.

6 TRABALHOS FUTUROS

Tendo em vista os resultados obtidos neste trabalho, fica como sugestão para trabalhos futuros a implementação de outras técnicas de Inteligência Computacional para este mesmo banco de dados, como por exemplo, Redes Neurais Artificiais, Lógica *Fuzzy* e Computação Evolutiva. Além destas implementações é possível que sejam realizados novos testes com variáveis distintas, do mesmo banco de dados, ou até mesmo de um novo banco de dados.

7 REFERÊNCIAS BIBLIOGRÁFICAS

ALBUQUERQUE, G.R. Estudo Epidemiológico dos Pacientes Submetidos ao Cateterismo no Centro de Diagnóstico de Imagem e Cardiológico de Estado de Roraima. UFR, Boa Vista, RR. 2016.

ARAÚJO, J.D. Polarização Epidemiológica no Brasil. Informe Epidemiológico do SUS. Brasília, 2012.

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.

CAMPOS, L.A., *et al*. Mathematical biomarkers for the autonomic regulation of cardiovascular system. Frontiers in Physiology, Integrative Physiology. v.4. n279. p.1-9. Outubro de 2013.

D.W. Aha. *Heart Disease Databases*. Disponível em: www.ics.uci.edu/pub/machine-learning-databases/heart-disease.names. Acessado em 12/2017

DIAS, R. Correspondência Entre Alterações da Voz e do Eletrocardiograma. Disponível em: https://paginas.fe.up.pt/~ee07135/ecg.html Acessado em 09/2017

DUTRA, D.D., DUARTE, M.C.S., ALBUQUERQUE *et al.* Doenças Cardiovasculares e Fatores Associados em Adultos e Idosos Cadastrados em uma Unidade Básica de Saúde. Revista de pesquisa Cuidado é fundamental online. Universidade Federal do Estado do Rio de Janeiro. Rio de Janeiro, 2016.

FELDMAN, J. & GOLDWASSER, G.P. Eletrocardiograma: Recomendações para sua interpretação, Universidade Federal do Rio de Janeiro, Faculdade de Medicina Souza Marques, Universidade Gama Filho. Rio de Janeiro, RJ. 2004

FERREIRA, A.R.P.A., SILVA, M.V., MACIEL, J. Eletrocardiograma no infarto agudo do miocárdio: O que esperar?, *International Journal of Cardiovascular Science*. v.29. n3. p.198 – 209. Coimbra, Portugal. 2016.

GOLDSMIDT, R.R. Uma Introdução à Inteligência Computacional: Fundamentos, Ferramentas e Aplicações. Primeira Edição. Rio de Janeiro, 2010. p.142.

HAYKIN, S. Neural Networks: A Comprehensive Foundation. *Second Edition*. Canada: Pearson Education, 1999. p.823.

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*, v17, p.683-695. Taiwan. 2001

HOFFMANN, R. Análise de Regressão: Uma introdução à Econometria. 4. Ed. Piracicaba: Hucitec, 2014. 393p. v.1.

ISHITANI, L.H., *et al*. Desigualdade social e mortalidade precoce por doenças cardiovasculares no Brasil. Rev Saúde Pública. v.40. n4. p.1-8. Março de 2006.

JUNIOR, R.R.C. Redes Neurais Artificiais no Auxílio do Diagnostico de Cardiopatias. 2011. p.37. Monografia. UFV. Viçosa, 2011.

KAWAMURA, T. Interpretação de um este sob a visão epidemiológica. Eficiência de um teste. Arquivo Brasileiro de Cardiologia. v.79. n4. p.437-441. Araçatuba, SP. 2002.

LIMA, C.A.M. Comitê de Máquinas: Uma Abordagem Unificada Empregando Máquinas de Vetores-Suporte, Tese de Doutorado, Universidade Estadual de Campinas. Campinas/SP, 2004.

LORENA, A.C. & CARVALHO A.C.P.L.F. Introdução às máquinas de Vetor Suporte, Relatório técnico, Universidade de São Paulo. São Paulo/SP, 2003.

MANSUR, A. D. P.; FAVARATO, D. Mortalidade por doenças cardiovasculares no Brasil e na região metropolitana de São Paulo. Arquivos Brasileiros de Cardiologia. p.755-761. Setembro de 2012.

MORAES, V.C.S., *et al*. Identificação do risco de cardiopatia através do estudo combinado de circunferências corporais. Acta Biomédica Brasiliensia. v.7. n1. p.31-39. Julho de 2016.

NETTER, Frank H. Atlas de Anatomia Humana. 2ed. Porto Alegre: Artmed, 2000.

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

Organização Pan-Americana de Saude & Organização Mundial de Saúde Brasil, Doenças Cardiovasculares. Disponível em: www.paho.org/bra/index.php?option=com_content&view=article&id=5253%3Adoencas-cardiovasculares&catid=845%3Anoticias&Itemid=839 Acessado em 05/2017

PASSOS, U.R.C. Computação Evolutiva e Aprendizado de Máquina Aplicados ao Apoio do Diagnóstico da Cardiopatia Isquêmica. 2014. p.83. Dissertação. Universidade Cândido Mendes. Campos dos Goytacazes, 2014.

PEROZIN, A.R. Inteligência Computacional Avaliando o Risco Coronariano. 2002. p.160. Dissertação. UFSC. Florianópolis, 2002.

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, n2, p.285-302. Maio a agosto de 2008.

RUSSELL, S., NORVIG, P. Artificial Intelligence: A Modern Approch. *Third Edition*. New Jersey: Pearson Education, 2010. p.1132.

SEMOLINI, R. Support vector machines, inferência transdutiva e o problema de classificação. Dissertação. Universidade Estadual de Campinas. Campinas/SP, 2002.

SILVA, J.P.B.C. Modelos de Regressão Linear e Logística Utilizando o Software R. 2016. p.146. Dissertação. UAB. Portugal, 2016.

STITSON, M.O., WESTON, J.A.E, GAMMERMAN, A., VOVK, V. & VAPNIK, V. *Theory oh support vector machines*. Relatório Técnico. *University of London*. Londres, UK, 1996.

Sociedade Brasileira de Cardiologia. Cardiômetro: Mortes por doenças cardiovasculares no Brasil. Disponível em: http://www.cardiometro.com.br/default.asp Acessado em 06/2017

TAVARES, T.R. Utilização de Técnicas de Inteligência Artificial para Classificação de Crianças Cardiopatas em Base de Dados Desbalanceada. 2013. p.109. Dissertação. UFPE. Recife, 2013.

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

Thermo Scientific. Interpretação dos resultados dos testes. Disponível em: www.phadia.com/pt-BR/Diagnostico-de-auto-imunidade/Saber-mais/Avaliacao-dos-Resultados-dos-Testes/#Sens Spec. Acessado em 11/2017

TORTORA, G.J., DERRICKSON, B. Corpo humano: Fundamentos da anatomia e fisiologia. 10. ed. Porto Alegre: Artmed, 2017.

World Health Organization. Disponível em www.who.int/en/. Acessado em 09/2016

ANEXOS

A - Resultados das 100 simulações do modelo SVM

B - Resultados das 100 simulações do modelo de associação da Regressão ao SVM

Anexo A – Resultados das 100 simulações do modelo SVM

Tabela 19: Resultados das 100 simulações do modelo SVM (continua)

	Erro_treino	Erro_*cross*	N_Vetor_suporte	Acurácia	Sensibilidade	Especificidade	Falso Negativo
1	0,16	0,24	150	0,74	0,54	0,91	0,46
2	0,18	0,24	149	0,77	0,70	0,82	0,30
3	0,16	0,22	156	0,74	0,85	0,65	0,15
4	0,17	0,27	151	0,79	0,67	0,96	0,33
5	0,15	0,22	145	0,74	0,67	0,81	0,33
6	0,15	0,25	145	0,75	0,63	0,87	0,37
7	0,17	0,28	160	0,84	0,74	0,91	0,26
8	0,17	0,24	159	0,82	0,70	0,94	0,30
9	0,17	0,26	152	0,77	0,64	0,86	0,36
10	0,16	0,23	147	0,72	0,78	068	0,22
11	0,17	0,25	161	0,82	0,77	0,87	0,23
12	0,15	0,27	155	0,82	0,80	0,83	0,20
13	0,19	0,25	148	0,74	0,86	0,67	0,14
14	0,18	0,27	155	0,84	0,76	0,88	0,24
15	0,17	0,28	153	0,79	0,64	0,91	0,37
16	0,18	0,26	157	0,84	0,74	0,91	0,26
17	0,16	0,27	149	0,77	0,61	0,87	0,39
18	0,17	0,24	151	0,72	0,71	0,73	0,29
19	0,14	0,20	150	0,67	0,56	0,76	0,44
20	0,15	0,22	145	0,70	0,64	0,80	0,36
21	0,16	0,22	148	0,69	0,48	0,90	0,52
22	0,19	0,25	158	0,84	0,79	0,86	0,21
23	0,18	0,25	155	0,79	0,69	0,90	0,31
24	0,15	0,25	148	0,69	0,56	0,83	0,44
25	0,17	0,25	146	0,77	0,69	0,84	0,31
26	0,18	0,24	162	0,80	0,68	0,91	0,32
27	0,15	0,21	143	0,66	0,56	0,74	0,44
28	0,14	0,24	155	0,70	0,85	0,59	0,15
29	0,16	0,24	146	0,67	0,52	0,78	0,48
30	0,16	0,27	158	0,79	0,68	0,88	0,32
31	0,15	0,22	146	0,70	0,63	0,76	0,37

Tabela 19. Continuação

32	0,16	0,26	154	0,79	0,79	0,78	0,21
33	0,19	0,25	156	0,80	0,76	0,84	0,24
34	0,18	0,26	146	0,77	0,68	0,87	0,32
35	0,16	0,25	149	0,77	0,91	0,69	0,09
36	0,15	0,25	148	0,69	0,63	0,73	0,38
37	0,18	0,25	159	0,80	0,77	0,84	0,23
38	0,16	0,21	151	0,72	0,52	0,91	0,48
39	0,15	0,25	151	0,74	0,67	0,81	0,33
40	0,19	0,25	163	0,87	0,85	0,89	0,15
41	0,17	0,24	153	0,77	0,77	0,77	0,23
42	0,15	0,21	145	0,70	0,1	0,70	0,29
43	0,18	0,25	153	0,79	0,79	0,78	0,21
44	0,14	0,25	152	0,77	0,75	0,79	0,25
45	0,20	0,24	153	0,75	0,76	0,75	0,24
46	0,18	0,28	156	0,84	0,75	0,89	0,25
47	0,17	0,21	147	0,69	0,68	0,70	0,32
48	0,18	0,24	159	0,80	0,76	0,84	0,24
49	0,13	0,22	149	0,64	0,56	0,69	0,44
50	0,14	0,23	151	0,69	0,55	0,86	0,45
51	0,17	0,24	151	0,77	0,75	0,78	0,25
52	0,16	0,26	160	0,82	0,77	0,87	0,23
53	0,21	0,26	156	0,85	0,86	0,84	0,14
54	0,15	0,21	144	0,70	0,55	0,84	0,45
55	0,17	0,23	152	0,77	0,72	0,81	0,28
56	0,18	0,27	159	0,82	0,77	0,84	0,23
57	0,17	0,24	154	0,77	0,80	0,74	0,20
58	0,17	0,23	153	0,75	0,59	0,88	0,41
59	0,17	0,23	145	0,72	0,57	0,85	0,43
60	0,15	0,22	151	0,75	0,73	0,77	0,27
61	0,14	0,22	145	0,70	0,54	0,83	0,46
62	0,15	0,25	148	0,72	0,63	0,79	0,37
63	0,17	0,24	149	0,74	0,69	0,79	0,31
64	0,14	0,22	145	0,75	0,61	0,88	0,39
65	0,18	0,25	157	0,80	0,72	0,88	0,28
66	0,17	0,25	156	0,79	0,79	0,79	0,21

Tabela 19. Continuação

67	0,15	0,20	141	0,70	0,82	0,82	0,43
68	0,15	0,22	145	0,72	0,63	0,78	0,38
69	0,15	0,24	148	0,72	0,60	0,81	0,40
70	0,17	0,30	152	0,80	0,67	0,91	0,33
71	0,14	0,23	145	0,69	0,52	0,89	0,48
72	0,17	0,21	146	0,77	0,65	0,85	0,35
73	0,19	0,24	150	0,79	0,66	0,93	0,34
74	0,14	0,25	153	0,74	0,74	0,74	0,26
75	0,19	0,27	157	0,77	0,82	0,74	0,18
76	0,15	0,25	154	0,72	0,79	0,68	0,21
77	0,17	0,26	150	0,77	0,71	0,81	0,29
78	0,18	0,27	161	0,84	0,77	0,90	0,23
79	0,17	0,24	159	0,75	0,63	0,87	0,37
80	0,14	0,21	147	0,62	0,47	0,79	0,53
81	0,18	0,28	154	0,80	0,78	0,82	0,22
82	0,17	0,24	155	0,79	0,69	0,86	0,31
83	0,19	0,26	156	0,85	0,75	0,94	0,25
84	0,15	0,24	151	0,80	0,73	0,85	0,27
85	0,16	0,25	155	0,75	0,75	0,76	0,25
86	0,17	0,28	154	0,82	0,73	0,89	0,27
87	0,18	0,26	147	0,82	0,0	0,83	0,20
88	0,16	0,23	152	0,75	0,70	0,81	0,30
89	0,18	0,25	155	0,77	0,83	0,72	0,17
90	0,17	0,27	157	0,79	0,70	0,87	0,30
91	0,15	0,23	147	0,70	0,48	0,93	0,52
92	0,17	0,26	159	0,82	0,81	0,83	0,19
93	0,19	0,28	153	0,80	0,74	0,85	0,26
94	0,19	0,26	157	0,84	0,76	0,89	0,24
95	0,17	0,27	152	0,79	0,80	0,78	0,20
96	0,16	0,23	156	0,75	0,65	0,82	0,35
97	0,18	0,25	156	0,80	0,79	0,81	0,21
98	0,16	0,24	160	0,79	0,74	0,82	0,26
99	0,18	0,24	161	0,80	0,69	0,90	0,31
100	0,16	0,25	148	0,75	0,71	0,81	0,29

Anexo B – Resultados das 100 simulações do modelo de associação da Regressão ao SVM

Tabela 20: Resultados das 100 simulações do modelo Regressão + SVM (continua)

	Erro_treino	Erro_*cross*	N_Vetor_suporte	Acurácia	Sensibilidade	Especificidade	Falso Negativo
1	0,17	0,25	152	0,79	0,64	0,91	0,36
2	0,19	0,23	145	0,80	0,70	0,88	0,30
3	0,20	0,25	138	0,79	0,85	0,74	0,15
4	0,20	0,24	152	0,84	0,72	1	0,28
5	0,18	0,21	145	0,75	0,70	0,81	0,30
6	0,18	0,23	137	0,75	0,60	0,90	0,40
7	0,19	0,24	149	0,84	0,70	0,94	0,30
8	0,17	0,23	154	0,80	0,73	0,87	0,27
9	0,19	0,22	132	0,75	0,60	0,86	0,40
10	0,17	0,20	134	0,74	0,70	0,76	0,30
11	0,19	0,22	146	0,84	0,81	0,87	0,19
12	0,20	0,24	148	0,85	0,85	0,85	0,15
13	0,18	0,22	142	0,75	0,82	0,72	0,18
14	0,19	0,21	146	0,79	0,62	0,88	0,38
15	0,20	0,25	138	0,80	0,68	0,91	0,32
16	0,19	0,25	156	0,87	0,78	0,94	0,22
17	0,19	0,24	141	0,79	0,70	0,84	0,30
18	0,20	0,23	147	0,80	0,75	0,85	0,25
19	0,17	0,22	141	0,73	0,56	0,88	0,44
20	0,16	0,18	134	0,70	0,64	0,80	0,36
21	0,16	0,19	131	0,69	0,52	0,87	0,48
22	0,18	0,24	155	0,84	0,71	0,92	0,29
23	0,19	0,23	137	0,79	0,75	0,83	0,25
24	0,19	0,23	138	0,74	0,63	0,86	0,38
25	0,17	0,21	147	0,75	0,66	0,84	0,34
26	0,19	0,23	141	0,77	0,64	0,88	0,36
27	0,18	0,23	132	0,77	0,70	0,82	0,30
28	0,19	0,24	153	0,84	0,81	0,85	0,19
29	0,16	0,20	147	0,72	0,56	0,83	0,44
30	0,18	0,21	155	0,77	0,71	0,82	0,29
31	0,17	0,19	138	0,69	0,52	0,82	0,48

Tabela 20. Continua

32	0,20	0,24	141	0,84	0,75	0,89	0,25
33	0,19	0,22	150	0,82	0,72	0,91	0,28
34	0,20	0,23	139	0,82	0,71	0,93	0,29
35	0,17	0,21	154	0,74	0,68	0,77	0,32
36	0,20	0,23	143	0,82	0,75	0,86	0,25
37	0,20	0,25	141	0,84	0,77	0,90	0,23
38	0,18	0,21	138	0,74	0,59	0,88	0,41
39	0,17	0,21	131	0,69	0,60	0,77	0,40
40	0,20	0,23	145	0,84	0,70	0,94	0,31
41	0,18	0,20	145	0,75	0,77	0,74	0,23
42	0,16	0,20	149	0,70	0,63	0,76	0,38
43	0,21	0,24	142	0,82	0,63	0,95	0,38
44	0,18	0,23	143	0,79	0,68	0,88	0,32
45	0,18	0,20	143	0,75	0,72	0,78	0,28
46	0,20	0,25	150	0,85	0,75	0,92	0,25
47	0,17	0,21	135	0,74	0,61	0,85	0,39
48	0,19	0,22	142	0,80	0,69	0,91	0,31
49	0,16	0,19	148	0,70	0,52	0,83	0,48
50	0,16	0,21	137	0,72	0,58	0,89	0,42
51	0,20	0,24	142	0,77	0,67	0,84	0,33
52	0,21	0,23	145	0,82	0,77	0,87	0,23
53	0,19	0,22	152	0,85	0,86	0,84	0,14
54	0,17	0,22	135	0,74	0,58	0,88	0,41
55	0,19	0,26	141	0,84	0,79	0,88	0,21
56	0,19	0,23	150	0,84	0,82	0,85	0,18
57	0,21	0,24	143	0,82	0,80	0,84	0,20
58	0,19	0,23	137	0,75	0,63	0,85	0,37
59	0,17	0,23	140	0,79	0,68	0,88	0,32
60	0,17	0,20	152	0,72	0,65	0,77	0,35
61	0,18	0,21	139	0,77	0,62	0,89	0,38
62	0,17	0,21	150	0,79	0,70	0,85	0,30
63	0,18	0,20	141	0,72	0,66	0,79	0,34
64	0,17	0,21	132	0,70	0,50	0,88	0,50
65	0,19	0,21	147	0,82	0,72	0,91	0,28
66	0,18	0,22	139	0,79	0,76	0,82	0,24
67	0,15	0,19	130	0,67	0,54	0,79	0,46

Tabela 20. Continuação

68	0,17	0,20	141	0,74	0,63	0,81	0,38
69	0,19	0,22	132	0,74	0,56	0,86	0,44
70	0,21	0,23	143	0,79	0,63	0,91	0,37
71	0,19	0,22	130	0,69	0,52	0,89	0,48
72	0,18	0,21	138	0,79	0,61	0,89	0,39
73	0,18	0,23	138	0,79	0,69	0,90	0,31
74	0,18	0,19	134	0,74	0,68	0,81	0,32
75	0,21	0,25	148	0,84	0,82	0,85	0,18
76	0,20	0,23	151	0,80	0,79	0,81	0,21
77	0,19	0,22	143	0,84	0,83	0,84	0,17
78	0,19	0,25	145	0,82	0,70	0,94	0,30
79	0,20	0,23	142	0,80	0,70	0,90	0,30
80	0,17	0,22	148	0,74	0,63	0,86	0,38
81	0,19	0,22	150	0,80	0,74	0,85	0,26
82	0,19	0,23	152	0,80	0,73	0,86	0,27
83	0,19	0,23	145	0,82	0,71	0,91	0,29
84	0,20	0,22	137	0,80	0,68	0,87	0,32
85	0,19	0,22	150	0,77	0,75	0,79	0,25
86	0,19	0,21	146	0,80	0,69	0,89	0,31
87	0,19	0,22	142	0,80	0,76	0,83	0,24
88	0,20	0,24	138	0,77	0,70	0,84	0,30
89	0,18	0,22	150	0,82	0,79	0,84	0,21
90	0,21	0,22	145	0,80	0,73	0,87	0,27
91	0,19	0,23	146	0,75	0,55	0,97	0,45
92	0,18	0,23	146	0,77	0,61	0,89	0,38
93	0,17	0,22	148	0,79	0,70	0,85	0,30
94	0,19	0,24	148	0,82	0,72	0,89	0,28
95	0,19	0,24	152	0,85	0,84	0,86	0,16
96	0,20	0,22	147	0,87	0,70	0,97	0,30
97	0,18	0,20	145	0,80	0,75	0,84	0,25
98	0,19	0,23	131	0,77	0,67	0,85	0,33
99	0,20	0,23	150	0,84	0,76	0,91	0,24
100	0,19	0,25	131	0,74	0,65	0,85	0,35