UNIVERSIDADE FEDERAL DO AMAPÁ
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO
DEPARTAMENTO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

METODOLOGIA BASEADA EM MÁQUINA DE VETORES DE SUPORTE PARA PREVISÃO DE ÓBITOS POR ATAQUE CARDÍACO

JOÃO PEDRO DOS SANTOS MONTEIRO

Orientador: Thiago Pinheiro do Nascimento

MACAPÁ
DEZEMBRO DE 2020

JOÃO PEDRO DOS SANTOS MONTEIRO

METODOLOGIA BASEADA EM MÁQUINA DE VETORES DE SUPORTE PARA PREVISÃO DE ÓBITOS POR ATAQUE CARDÍACO

Monografia de TCC apresentada à Universidade Federal do Amapá como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Thiago Pinheiro do Nascimento

MACAPÁ
DEZEMBRO DE 2020

Dados Internacionais de Catalogação na Publicação (CIP)
Biblioteca Central da Universidade Federal do Amapá
Elaborada por Cristina Fernandes– CRB-2/1569

Monteiro, João Pedro dos Santos.

Metodologia baseada em máquina de vetores de suporte para previsão de óbitos por ataque cardíaco. / João Pedro dos Santos Monteiro; orientador, Thiago Pinheiro do Nascimento. – Macapá, 2020.

44 f.

Trabalho de conclusão de curso (Graduação) – Fundação Universidade Federal do Amapá, Coordenação do Curso de Bacharelado em Ciência da Computação.

1. Máquina de Vetores de Suporte. 2. Ataques Cardíacos. 3. Predição. I. Nascimento, Thiago Pinheiro do, orientador. II. Fundação Universidade Federal do Amapá. III. Título.

005.1 M775m
CDD. 22 ed.

UNIVERSIDADE FEDERAL DO AMAPÁ
DEPARTAMENTO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
COORDENAÇÃO DO CURSO DE CIÊNCIA DA COMPUTAÇÃO

ATA DE DEFESA DE TCC

Realizou-se no dia 21 de dezembro de 2020, às 15:00, via videoconferência pelo Google Meet, a defesa de TCC intitulado “METODOLOGIA BASEADA EM MÁQUINA DE VETORES DE SUPORTE PARA PREVISÃO DE ÓBITOS POR ATAQUES CARDÍACOS”, do discente JOÃO PEDRO DOS SANTOS MONTEIRO. A Banca Examinadora foi composta pelo Prof. Me. THIAGO PINHEIRO DO NASCIMENTO, presidente da banca e orientador; Prof. Me. MARCO ANTÔNIO LEAL DA SILVA e Prof. Esp. ADEILDO TELLES DA SILVA, examinadores. Concluída da defesa, foram realizadas as arguições e comentários. Em seguida, procedeu-se o julgamento pelos membros da Banca Examinadora, tendo o trabalho sido APROVADO com NOTA 9,5.

E, para constar, eu, Prof. THIAGO PINHEIRO DO NASCIMENTO, orientador e presidente da Banca Examinadora, lavrei a presente ata que, após lida e achada conforme, foi assinada por mim e demais membros da Banca Examinadora.

Macapá, 15 de fevereiro de 2021

Thiago Pinheiro do Nascimento

THIAGO PINHEIRO DO NASCIMENTO

PROF. ME. MARCO ANTÔNIO LEAL DA SILVA

PROF. ESP. ADEILDO TELLES DA SILVA

JOÃO PEDRO DOS SANTOS MONTEIRO

METODOLOGIA BASEADA EM MÁQUINA DE VETORES DE SUPORTE PARA PREVISÃO DE ÓBITOS POR ATAQUE CARDÍACO

Monografia de TCC apresentada à Universidade Federal do Amapá como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Trabalho aprovado. Macapá, de de 2020.

Thiago Pinheiro do Nascimento
Orientador

Me. Marco Antônio Leal da Silva
Universidade Federal do Amapá

Esp. Adeildo Telles da Silva
Universidade Federal do Amapá

MACAPÁ
DEZEMBRO DE 2020

Aos meus pais que não medem esforços para me ajudar a tornar meus sonhos realidade, e, ao meu professor orientador pela paciência, dedicação e compreensão que tornaram possível a conclusão desta monografia.

Agradecimentos

Aos meus país que não medem esforços para garantir o meu sucesso.

Ao professor orientador, Thiago Pinheiro do Nascimento, pela orientação, apoio, compreensão e confiança que tornaram a realização dessa monografia possível.

A UNIFAP, seu corpo docente, direção e administração, pela oportunidade de fazer o curso e por todo o conhecimento que obtive que fortaleceram meu crescimento pessoal e profissional.

Aos meus amigos e colegas de turma que me acompanharam e me ajudaram a trilhar este jornada fazendo sempre com que eu desse o meu melhor.

Finalmente, agradeço àqueles que contribuíram de forma direta e indireta para a construção desse trabalho.

"Senhores, temer a morte é o mesmo que supor-se sábio quem não o é, por que é supor que se sabe o que não se sabe. Ninguém sabe o que é a morte; talvez seja ela o maior dos bens, mas todos a temem, como se fosse ela o maior dos males." – Sócrates

Resumo

Ataques cardíacos figuram como uma das principais causas de óbitos no mundo, sendo considerado a consequência de taxas de mortalidades elevadas em muitos países. Essa situação faz com que esforços sejam feitos para a diminuição do número de mortes por ataques cardíacos. Nesse contexto, este trabalho recomenda uma abordagem embasada em Máquina de Vetores de Suporte para predizer óbitos por ataques cardíacos, uma vez que essa predição pode ajustar tratamentos adequados a pacientes cardíacos.

Palavras-chave: Máquina de Vetores de Suporte. Ataques Cardíacos. Predição.

Sumário

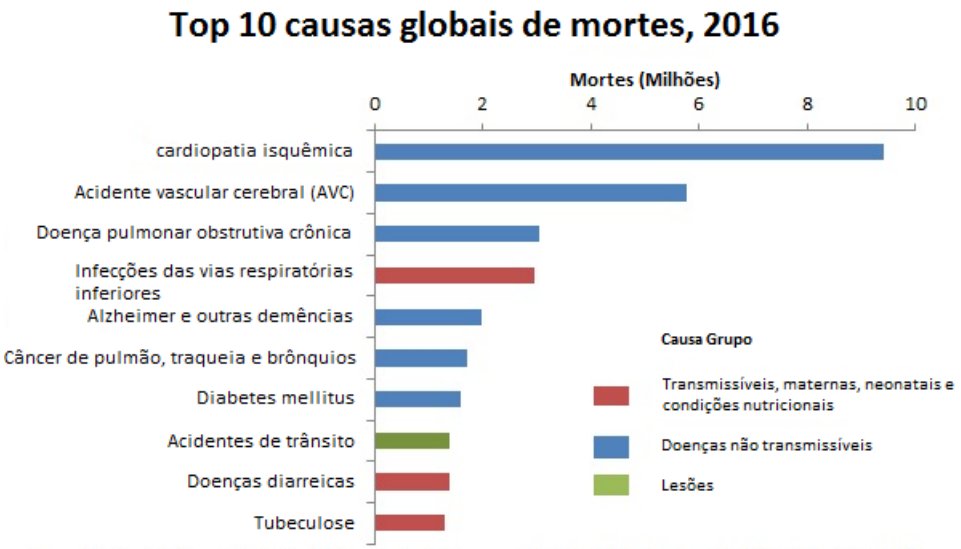
1 Introdução

Este trabalho propõe uma abordagem baseada em Máquina de Vetores de Suporte para predição de óbitos por ataques cardíacos. Nesse sentido, este capítulo introdutório visa a apresentação de aspectos essenciais sobre o referido trabalho, tais como a problemática e os objetivos pretendidos; a hipótese, a justificativa e relevância e as atividades a serem realizada. Finalmente, o cronograma e a organização do presente trabalho também são descritos.

1.1 Problemática

O ataque cardíaco é um dos principais motivos de morte no mundo [1], sendo a cardiopatia isquêmica, que motiva a ocorrência de ataques cardíacos, responsável pela morte de mais 9 milhões de pessoas nos últimos quatro anos. O acidente vascular cerebral é também considerado uma das principais justificativas relacionadas a morte de pessoas em todo o mundo, apesar de atingir uma taxa de mortalidade inferior ao ser comparada a cardiopatia isquêmica [2]. Para ilustrar essa situação, a Figura 1 apresenta as dez principais causas de mortes por doenças em todo mundo no ano de 2016, sendo esse o reflexo dos anos anteriores, uma vez que esses resultados foram muito parecidos.

Figura 1 – As dez maiores causas de morte por doença no ano de 2016.



Fonte: Organização Mundial de Saúde [2]

Os dados da pesquisa referida anteriormente sobre as principais causas de morte no mundo se refletem no Brasil. Em uma pesquisa feita entre os anos de 2008 a 2016

levando em consideração homens e mulheres na faixa etária de 30 a 59 anos concluiu-se que a taxa de mortalidade por ataques cardíacos ultrapassam 20.000 casos [3].

Durante muito tempo, os ataques do miocárdio foram considerados a principal causa de morte relacionada a problemas cardíacos, tendo, nos últimos anos, aumentado sua taxa de mortalidade em 48%. Possivelmente, no ano de 2020, caso essa situação não mude, o ataque cardíaco se tornará a principal causa de óbito por doença no Brasil [3].

1.2 Objetivo Geral

Desenvolver uma abordagem baseada em Máquina de Vetores de Suporte capaz de suportar a previsão de óbitos motivado por ataque cardíaco, onde essa previsão consiste na inferência se pacientes morrerão dentro de um ano ou não. Essa previsão permitirá a elaboração de cuidados mais precisos para pacientes com doenças cardíacas [4, 5].

1.3 Objetivos Específicos

- Entender possíveis causas que levam a ocasião de óbitos após ataques cardíacos;
- Coletar e analisar bases de dados reais e variáveis relacionadas a ataques cardíacos;
- Realizar o reconhecimento de padrões às bases de dados sobre ataques cardíacos;
- Aferir os resultados gerados do reconhecimento de padrões em ataques cardíacos.

1.4 Hipótese

O aprendizado de máquina foi utilizado atrativamente em muitos domínios de aplicações voltados ao diagnóstico e a previsão de doenças [6, 7, 8, 9], sintetizando uma situação semelhante ao contexto proposto no presente trabalho, uma vez que tratam problemas de classificação. Portanto, almeja-se que o aprendizado de máquina também seja atrativo para a previsão de óbitos por ataques cardíacos [10, 11].

1.5 Justificativa

O infarto agudo do miocárdio é um problema que se mostrou ter uma alta taxa de mortalidade. Por esse motivo, esforços são feitos para diminuir essa taxa de óbitos e para melhorar a qualidade de vida de pacientes cardíacos [2].

Entre os esforços feitos para a diminuição da taxa de mortalidade e a melhoria da qualidade de vida de pacientes cardíacos estão os sistemas computacionais, os quais podem ser empregados em diagnósticos e previsões de ataques cardíacos [12].

O diagnóstico de ataques cardíacos é considerado fundamental para o tratamento de pacientes cardíacos, no entanto esse diagnóstico é também considerado custoso e complexo; e é nesse contexto que os sistemas computacionais mostram relevância. Esses sistemas são capazes de automatizar o processo de diagnóstico médico, dando um apoio essencial para o processo de tomada de decisões médicas [1].

A predição possibilitada pelo sistema computacional proposto nesse trabalho será capaz de inferir óbitos por ataques cardíacos, no sentido de classificar se pacientes cardíacos morrerão dentro de um ano ou mais. Esse diagnóstico é importante, uma vez que permitirá a elaboração de cuidados diferenciados, bem como será responsável pela adequação de tratamentos para redução de riscos [4, 5].

1.6 Contribuição Científica

O trabalho proposto pretende gerar duas contribuições científicas, a se saber:

- Um auxilio ao tratamento de pacientes baseado na predição de óbito;
- Um conjunto de dados capazes de auxiliar o desenvolvimento de outros estudos;

1.7 Cronograma e Atividades

O presente trabalho é pautado na realização das atividades listadas, como segue:

1. Revisão bibliográfica;
2. Investigação dos objetivos;
3. Identificação de recursos técnicos;
4. Implementação e avaliação de um estudo de caso;
5. Submisão de artigos relacionados à pesquisa desenvolvida.
6. Redação do projeto de pesquisa final para a conclusao de curso.
7. Apresentação final do trabalho de pesquisa para a conclusão do curso.

As atividades acima são efetuadas conforme o cronograma descrito na Tabela 1.

Tabela 1 – Cronograma

Etapa	jan.	fev.	mar	abr.	mai.	jun.	jul.	ago.	set.	out.	nov.	dez.
1	x	x	x	x	x	x	x	x	x	x	x	x
2	x	x	x									
3	x	x	x	x	x							
4					x	x	x	x	x	x		
5						x	x	x	x	x	x	x
6				x	x	x	x	x	x	x	x	x
7												x

1.8 Organização

Além desse capítulo introdutório, este trabalho terá outros quatro capítulos. O segundo capítulo objetivará discutir sobre o referencial teórico relacionado ao trabalho proposto; o terceiro capítulo detalhará a metodologia a ser desenvolvida; o quarto visará a apresentação dos resultados; e, por fim, o quinto capítulo descreverá as conclusões.

2 Referencial Teórico

O presente capitulo se encontra dividido em duas partes. Primeiramente se fala sobre o Infarto agudo do miocárdio, que consiste no problema abordado por esta pesquisa, explanando sobre o que de fato é essa condição, suas causas, complicações, diagnóstico e tratamento. Logo em seguida fala-se sobre Máquina de Vetores de Suporte (SVM), algoritmo de aprendizado de máquina utilizado para a tarefa de previsão proposta na presente pesquisa.

2.1 Infarto Agudo do Miocárdio (IAM)

O Infarto Agudo do Miocárdio (IAM) é a necrose de parte do tecido muscular do coração provocada pela interrupção da irrigação sanguínea para essa região [13, 14, 15]. Em 90% dos casos, o IAM é uma consequência da doença aterosclerótica. Possibilidades bem menos frequentes envolvem Arterites, Goma sifilítica, origem anômala da artéria coronária, dissecção aórtica, Anemia e Causas Hematológicas [16]. A aterosclerose é o nome que descreve o processo de formação de uma placa de gordura e outros elementos (placa aterosclerótica) na parede de uma artéria, a estreitando e enrijecendo, dificultando a passagem de sangue nesta região [17, 18]. Ela é resultado de uma série de fatores de risco [13, 17]. Os principais fatores, de acordo com [13], são:

- História familiar
- Faixa etária (Homem > 45 anos e mulher > 55 anos)
- Tabagismo
- Hipercolesterolemia
- Hipertensão arterial sistêmica.
- Diabete melito.
- Obesidade.
- Gordura abdominal.
- Sedentarismo.
- Dieta pobre em frutas e vegetais.
- Estresse psicossocial.

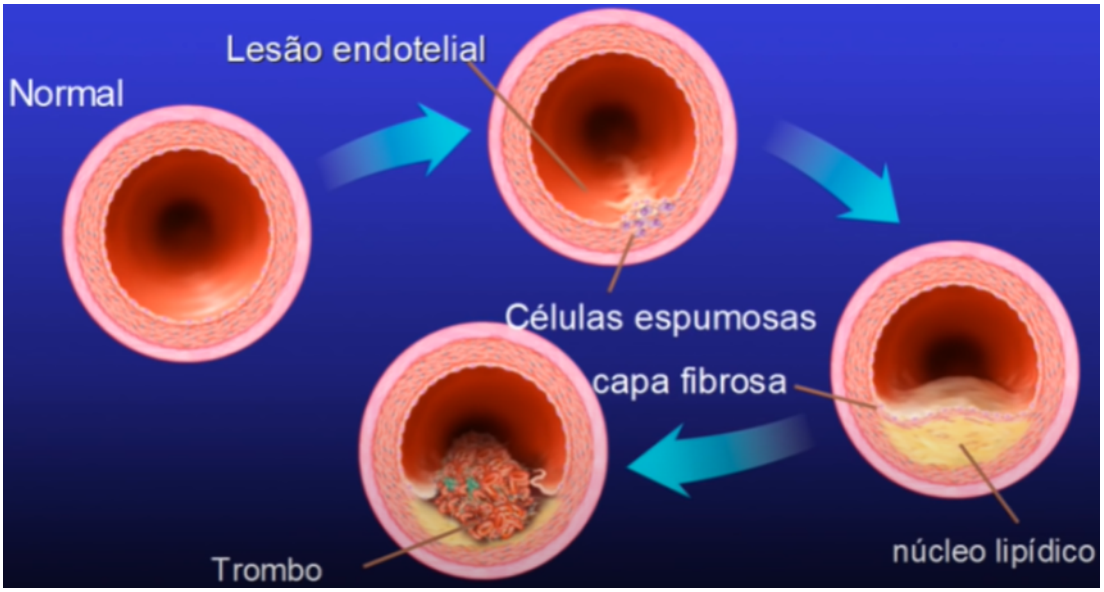
Os dois primeiros fatores de risco listados, História familiar e faixa etária, pertencem ao grupo dos não modificáveis que, como o próprio nome já indica, não há possibilidades de mudança, enquanto que todos os restantes pertencem ao grupo dos modificáveis e podem ser controlados ou prevenidos de alguma forma, evitando, o desenvolvimento da aterosclerose [13, 17]. A aterosclerose é uma doença que pode acontecer em algumas localidades do corpo, como o coração (mais especificamente nas artérias responsáveis pela irrigação da musculatura cardíaca) onde pode levar a complicações do infarto do miocárdio [17].

Quando os fatores risco mencionados por [13] não são devidamente controlados por um indivíduo, esses fatores então influenciam a aterosclerose. Em contato com o sangue, existe uma camada da artéria chamada de endotélio. A aterosclerose começa quando essa camada é danificada por algum tipo de fator, como tabagismo, hipertensão arterial, etc [19, 20]. Uma parte do colesterol que circula livremente pelo vaso através do sangue, começa adentrar na área danificada, se acumulando na região abaixo do endotélio onde começa a se oxidar. A oxidação desse colesterol induz a atração de monócitos. Esses monócitos se diferenciam em macrofagos, processo também promovido pela oxidação do colesterol, e começam a fagocitar esse colesterol oxidado. Os macrofagos não conseguem eliminar o colesterol fagocitado, se enchendo dessa substância, transformando-se em células espumosas que contribuem para o acúmulo de conteúdo na parede da artéria junto ao colesterol. Quando os macrofagos morrem neste processo, eles liberam substâncias que induzem a atração de mais monócitos. As células musculares lisas da parede da artéria cobrem o conteúdo que se forma com o acúmulo de colesterol e células espumosas, para que esse conteúdo não entre em contato com a circulação sanguínea formando uma capa em sua volta denominada de capa fibrosa. Ao longo de anos esse conteúdo vai aumentando de volume fazendo com que parte do fluxo de sangue do vaso seja obstruído [19, 20, 21, 22].

Muitas das vezes, a capa fibrosa da placa aterosclerótica rompe-se expondo todo conteúdo lipídico ao sangue. Neste momento, o sangue reage com o conteúdo formando um coágulo (trombo) que impede que ele vaze. Esse processo pode ser benigno não causando mal algum. No entanto, esse coágulo pode terminar de obstruir o vaso, cortando todo o fluxo sanguíneo para todo o tecido que estiver adiante à artéria. O resultado dessa oclusão podem ser o infarto, a angina instável ou até mesma a morte súbita [23]. Na figura 2, ilustra-se uma artéria saudável que após o processo aterosclerótico tem parte do seu fluxo sanguíneo obstruído.

O IAM se engloba no conjunto das Síndromes Coronarianas Agudas (SCA). Entre as condições que fazem parte desse conjunto, junto com angina instável, são as que possuem o pior prognóstico com risco de óbito e possibilidades de sequelas elevados [24]. Após a oclusão de uma artéria coronária e a consequente interrupção de

Figura 2 – Processo de Formação de um trombo.



Fonte: Canal Médico, 2015 [18]

fluxo sanguíneo para parte do músculo cardíaco, essa parte do músculo começa entrar em processo de necrose. Esse processo não é imediato, ocorrendo de maneira gradual, traduzindo a importância da ação rápida de profissionais qualificados para o diagnóstico assim como reverter o quadro do paciente antes que o danos ao tecido muscular do coração sejam irreversíveis. Fator esse que justifica a máxima na cardiologia que diz: "Tempo é músculo" [13, 25]. O IAM pode levar à complicações como insuficiência cardíaca (O coração não consegue bombear sangue para o corpo adequadamente), Tromboembolismo sistêmico (Tendência à formação de trombos) e Complicações mecânicas. [26].

Para realizar o diagnóstico de infarto agudo do miocárdio profissionais da área podem levar em conta o quadro clínico do paciente (conjunto de sintomas que este apresenta), elevação de marcadores bioquímicos de necrose (Elementos presentes no sangue que indicam a ocorrência de um infarto), o eletrocardiograma, assim também como a história clínica [13, 14, 15]. Quando o diagnóstico de infarto do miocárdio é realizado, o paciente é classificado de acordo com uma tabela que mostra a evolução desse infarto. Um tipo de classificação comumente utilizada é a de Killip-kimball mostrada na figura 3 [27]. Os pacientes na faixa "Killip I" têm um baixa taxa de mortalidade devido a não apresentarem sinais de insuficiência cardíaca e possivelmente apresentarem um infarto sucinto. Pacientes em "Killip II" já possuem uma insuficiência cardíaca moderada e seu risco de óbito já aumentam entre 8% a 10%. Pacientes em "Killip III" já apresentam edema agudo de pulmão (quando há acúmulo de líquido nos pulmões) e chances de óbito entre 20% a 25%. Por fim, pacientes em "Killip IV" apresentam choque cardiogênico (O coração não bombeia sangue adequadamente para o corpo). [13, 28, 27].

Figura 3 – Classificação de Killip.

Parâmetros	Classe	Risco de óbito (%)
Sem sinais de insuficiência cardíaca	I	2-3
Insuficiência cardíaca discreta (estertores nas bases e presença de terceira bulha)	II	8-10
Edema agudo de pulmão	III	20-25
Choque cardiogênico	IV	45-70

Fonte: V Diretriz da Sociedade Brasileira de Cardiologia sobre Tratamento do Infarto Agudo do Miocárdio com Supradesnível do Segmento ST, 2015 [28]

Antes do advento das unidades coronárias, não era feito muito para se reverter o quadro de infarto. Eram apenas oferecidas morfina para dor e oxigênio para falta de ar junto ao repouso, o que justifica a mortalidade ser por volta de 30% durante esse período. Com o advento das Unidades Coronárias nos anos 60, a mortalidade caiu de 30% para 15%. Considerada um grande avanço no tratamento do infarto do miocárdio, nas unidades coronárias, os pacientes recebem uma atenção especial com: profissionais mais bem treinados para conduzir a situação, monitorização cardíaca constante, melhor manejo das arritmias, entre outros cuidados e particularidades. Da época do surgimento das unidades coronarianas até os anos mais recentes, a taxa de mortalidade por ataque cardíaco caiu para 5% devido a diversos avanços nos tratamentos [29]. As terapias de reperfusão são exemplos desses avanços realizados. Essas terapias têm como função reverter a obstrução que é a causa primária do infarto. Os trombolíticos e a angioplastia são duas dessas estratégias. A finalidade dos trombolíticos é dissolver o trombo que está ocluindo a artéria e causando o infarto, devolvendo a circulação no vaso [13]. A angioplastia se trata de um método mais eficaz do que os trombolíticos, no qual é inserido uma prótese na artéria obstruída, que expande o vaso e libera o fluxo sanguíneo [30, 31].

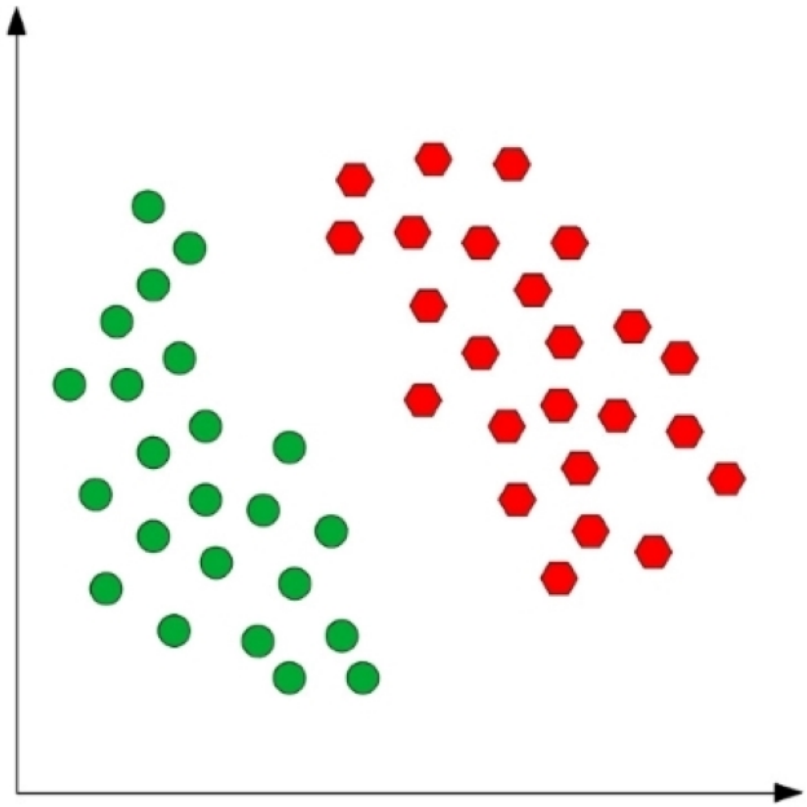
2.2 Máquina de Vetores de Suporte (SVM)

2.2.1 Visão geral das SVMs

Máquinas de vetores de suporte (SVM) foram introduzidas por Vladimir Vapnik em 1979 [32]. Se tratam de um método de aprendizagem supervisionada de máquina que podem abordar tarefas de classificação e regressão [33]. Enquanto aos problemas de classificação, originalmente, o aprendizado com SVM teve foco em problemas binários, onde os dados de entrada possuíam apenas duas classes possíveis de saída. Porém, algumas técnicas possibilitam o seu uso para problemas de classificação multiclasses, e uma delas é a decomposição desses problemas em uma série de subproblemas de classificação binária [34].

As SVMs realizam a tarefa de classificação separando os dados, primeiramente organizados em um espaço n-dimensional, através de um hiperplano. O hiperplano pode ser descrito como uma divisa que separa um espaço de n dimensões em duas partes, possuindo n - 1 dimensões [33, 34, 35]. Em um espaço de duas dimensões, por exemplo, essa fronteira seria uma reta [33]. A forma mais simples da SVM, chamada SVM com margens rígidas, permite tratar problemas de classificação onde os dados são linearmente separáveis [33, 34]. Na figura 4, considera-se um problema de classificação ilustrativo onde se quer classificar corretamente todos os dados que são círculos ou hexágonos de um conjunto de instâncias.

Figura 4 – Base de dados linearmente separável.



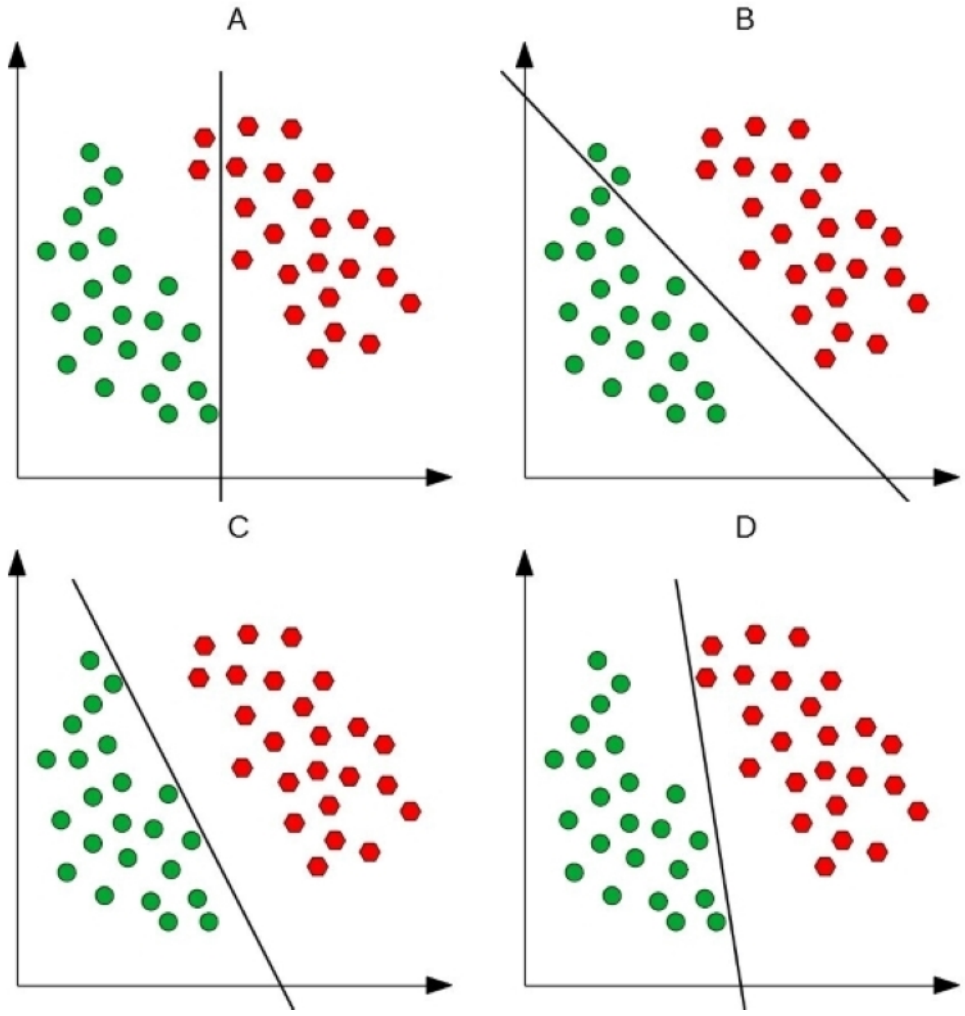
Fonte: Adaptado de Ales (2008) [33]

O problema abordado acima é linearmente separável, pois os dados estão distribuídos de tal forma que é possível encontrar uma fronteira linear que discrimina muito bem todas as instâncias das duas classes consideradas. No entanto, a SVM de margens rígidas, assim como os outros tipos de SVM, não só tentará encontrar um hiperplano que segrega todos as instâncias de classes distintas, mas também irá tentar encontrar um hiperplano que tenha máxima margem entre os pontos de dados mais próximos dele, levando o classificador gerado a ter maior poder de generalização [33, 34, 36].

A figura 5 mostra alguns exemplos de hiperplanos dentro do amplo espectro a serem considerados. Nos exemplos A e B, os hiperplanos calculados não realizam

uma separação adequada dos dados, pois alguns deles estão presentes no subespaço pertencente a outra classe. Logo, fazem a classificação errônea de alguns dos pontos, o que caracteriza que o classificador encontrado classifica alguns hexágonos como sendo círculos no exemplo A e o contrário no exemplo B. Os hiperplanos C e D conseguem segregar às classes distintas sem que nenhum ponto de dados de uma classe esteja no subespaço de outra. No entanto, a margem considerada em ambos os hiperplanos não é adequada, o que pode levar pouca generalização dos modelos encontrados. Por fim, o hiperplano da figura 6 é o que melhor separa os dados em questão, com uma margem máxima entre os pontos de dados mais próximos. Esses pontos de dados são chamados de vetores de suporte e são aqueles que influenciam na complexidade do classificador encontrado [33, 34].

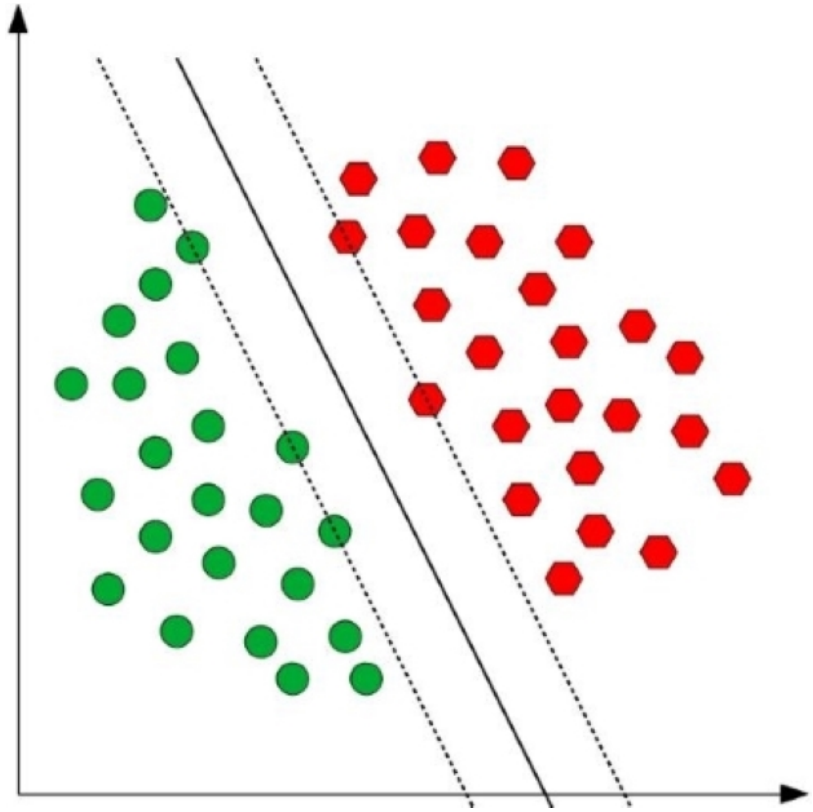
Figura 5 – Possíveis Hiperplanos Separadores.



Fonte: Adaptado de Ales (2008), Carvalho (2005) e Josh Readhead (2014) [33, 36, 35]

São raros os conjuntos de dados linearmente separáveis. Muitas das vezes, pela própria natureza do problema, algumas das instâncias de dados de uma classe invadem o espaço da outra, assim como também muitas bases de dados sofrem com a presença

Figura 6 – Hiperplano Separador Ótimo.

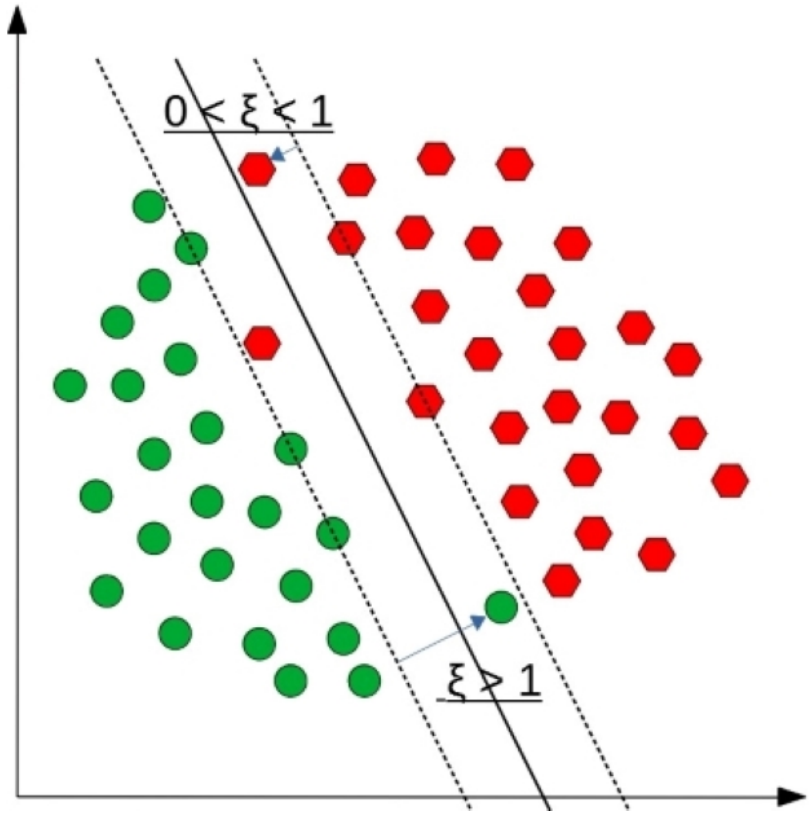


Fonte: Adaptado de Ales (2008) e Carvalho (2005) [33, 36]

de ruídos que interferem na linearidade do problema [33, 34]. Levando em conta essas características comuns da maioria das bases de dados, surge então a SVM com margens suaves ou flexíveis, onde determinados pontos podem ser separados incorretamente implicando em um maior poder de generalização [33, 32, 34]. Os pontos que violarem a margem definida para sua área sofrem um tipo de penalização através do uso das variáveis de folga (ξ). Um ponto que estiver em seu subespaço fora da margem tem o valor dessa variável como sendo igual a 0, enquanto que pontos dentro da margem possuem valor entre 0 e 1 e os que se situam além do hiperplano possuem valores maiores que 1 [33]. O exemplo de separação realizada por uma SVM de margens suaves é ilustrada na figura 7.

Apesar da SVM com margens suaves poderem tratar dados que não são linearmente separáveis, esses dados ainda tem que ter um certo grau de linearidade para poderem ser tratados por esse tipo de SVM. Mesmo com a SVM de margens suaves, existem dados organizados de tal forma que a separação não pode ser feita através de um hiperplano, ou seja, uma fronteira linear não fornece uma separação adequada para a maneira como os dados estão distribuídos [33, 34]. O exemplo ilustrado na figura 8, demonstra uns desses tipos de dados, em que uma fronteira circular seria mais adequada do que um hiperplano para realizar essa separação [34].

Figura 7 – Hiperplano de Margens Suaves.



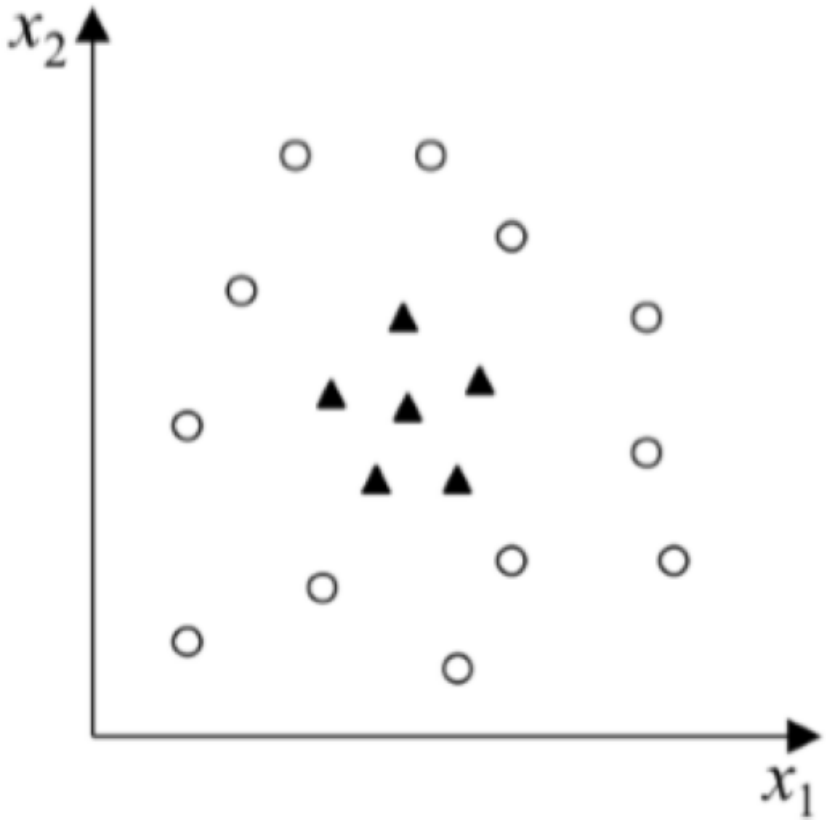
Fonte: Adaptado de Carvalho (2005) [36]

As SVMs com margens rígidas ou suaves são ditas SVMs lineares, pois definem fronteiras lineares para realizar a classificação dos dados [34]. Para tratar dos tipos de dados descritos acima, um novo tipo de SVM é criado, chamada de SVM não linear. Para poder separar um conjunto de dados não linearmente separável, a SVM não linear utiliza uma função Kernel. O objetivo dessas funções é mapear os dados de treinamento para um espaço de maior dimensão, onde esses dados são linearmente separáveis [33, 34]. O novo espaço, para onde os dados foram mapeados, é nomeado como espaço de características, enquanto o espaço original, onde os dados se encontravam, é nomeado de espaço de entradas [34]. A figura 9 ilustra o processo realizado pelos SVMs não lineares.

2.2.2 Visão geral matemática das SVMs

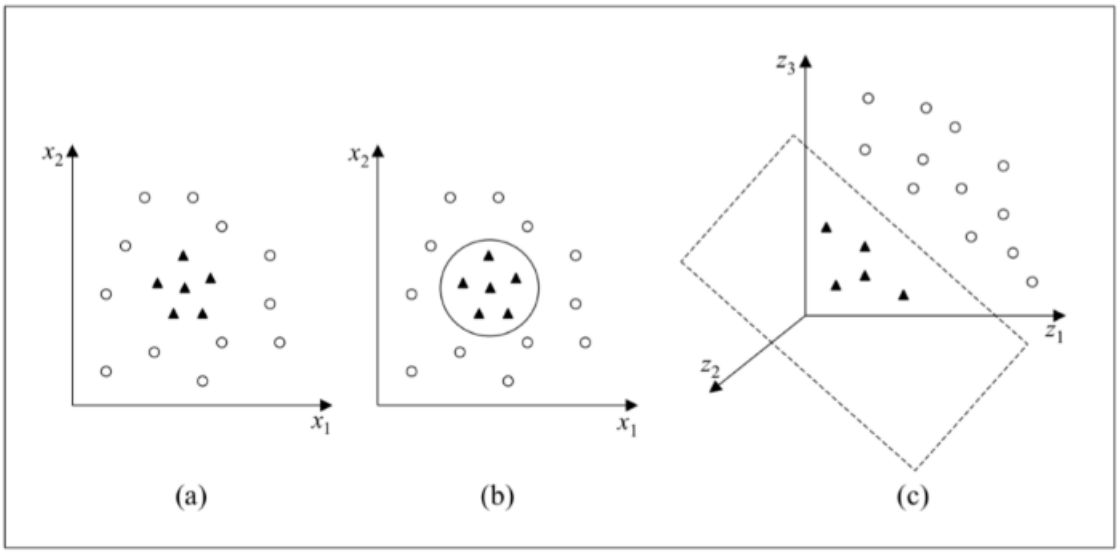
A fórmula que define o hiperplano é mostrada na equação 1, onde $f(\vec{x}) = 0$ se trata do hiperplano que realiza a separação do conjunto de dados, separando eles entre os espaços $f(\vec{x}) > 0$ e $f(\vec{x}) < 0$; $\vec{w} \cdot \vec{x}$ um produto escalar entre o vetor $\vec{w}$, vetor normal

Figura 8 – Dados não linearmente separáveis.



Fonte: Retirado de Lorena e Carvalho (2007) [34]

Figura 9 – Função Kernel.



Fonte: Retirado de Lorena e Carvalho (2007) [34]

ao hiperplano, e o vetor $\vec{x}$, vetor de entrada [34, 32].

$$f(\vec{x}) = \vec{w} \cdot \vec{x} + b \tag{1}$$

Os planos $f(\vec{x}) = +1$ e $f(\vec{x}) = -1$, são paralelos ao hiperplano separador, onde residem os pontos mais próximos à ele [32, 34]. A distância entre os planos é dada pela equação $\frac{2}{\|\vec{w}\|}$ [34]. Levando em conta que o valor de w e o valor de b foram selecionados para que nenhum dado fique entre os planos $f(x) = +1$ e $f(x) = -1$ a margem mínima então é definida através da equação $m = \frac{1}{\|\vec{w}\|}$ [34, 32].

Aborda-se em seguida o problema de otimização na equação 2, para maximizar as margens em questão pela minimização de $\|\vec{w}\|$ [32, 34]. Nas restrições y_i se trata da saída correta da instância de número i, ou seja, $+1$ ou -1, relacionada ao vetor de entrada em questão definido por x_i. N é a quantidade de exemplos de treinamento na base de dados [32].

$$\min_{\vec{w},b} \frac{1}{2}\|\vec{w}\|^2 \tag{2}$$

$$\text{Sujeito as restrições: } y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1, \forall_i, 1, ..., N \tag{3}$$

Pode-se obter uma solução para o problema de otimização descrito pela equação anterior usando funções Lagrangianas, restringindo a função objetivo ψ através de limitações relacionadas a um conjunto de multiplicadores de Lagrange α_i [34, 32]. Através do uso de uma função lagrangeana ao problema em questão, obtém-se então a equação 4, onde é chamada de problema dual, sujeito às restrições descritas em (5) [32]:

$$\min_{\vec{\alpha}} \psi(\vec{\alpha}) = \min_{\vec{\alpha}} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j (\vec{x}_i \cdot \vec{x}_j) \alpha_i \alpha_j - \sum_{i=1}^{N} \alpha_i \tag{4}$$

$$\text{Sujeito as restrições: } \begin{cases} \alpha_i \geq 0, & \forall_i, 1, ..., N \\ \sum_{i=1}^{N} y_i \alpha_i = 0 \end{cases} \tag{5}$$

Tanto o vetor $\vec{w}$ quanto o termo de polarização b podem ser definidos a partir dos multiplicadores de lagrange já definidos [32].

$$\vec{w} = \sum_{i=1}^{N} y_i \alpha_i \vec{x}_i, b = \vec{w} \cdot \vec{x}_k - y_k \text{ para cada } \alpha_k > 0 \tag{6}$$

Para tratar problemas não linearmente separáveis, a solução encontrada acima não é adequada e pode resultar em um processo infinito. Para esses problemas utilizam-se as variáveis de folga ξ que permitem que alguns dados ultrapassem as margens

ou sejam classificados erroneamente [32]. O problema de otimização descrito em (2) é então modificada para incluir as variáveis de folga dando forma a o novo problema de otimização descrito abaixo em (7) [32].

$$\min_{\vec{w},b,\vec{\xi}} \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{N}\xi_i \tag{7}$$

$$\text{Sujeito as restrições: } y_i(\vec{w}\cdot\vec{x}_i - b) \geq 1 - \xi_i, \forall_i, 1, ..., N \tag{8}$$

A constante C segundo [34]: "é um termo de regularização que impõe um peso à minimização dos erros no conjunto de treinamento em relação à minimização da complexidade do modelo". Com a introdução de uma função lagrangeana como feita anteriormente na versão de margens rígidas a equação obtida é semelhante à descrita em 4, no entanto, agora com as restrições descritas em 9 [32].

$$\text{Sujeito as restrições: } \begin{cases} 0 \leq \alpha_i \leq C, & \forall_i, 1, ..., N \\ \sum_{i=1}^{N} y_i\alpha_i = 0 \end{cases} \tag{9}$$

Para os conjuntos de dados em que um hiperplano não é satisfatório na separação das classes, é necessário a utilização de SVM não linear, capaz de mapear os dados de treinamento para um espaço de dimensão maior através de uma função kernel que mais se adequa aquele problema, tornando os dados, nesse novo espaço, linearmente separáveis. A base dessa SVM é descrita na equação abaixo, onde K se trata da função kernel [32, 34].

$$u = \sum_{j=1}^{N} y_j\alpha_j K(\vec{x}_j, \vec{x}) - b \tag{10}$$

O problema em questão é então transformado no seguinte problema de otimização descrito em 11 com restrições em 12 [32].

$$\min_{\vec{\alpha}} \psi(\vec{\alpha}) = \min_{\vec{\alpha}} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j K(\vec{x}_i\cdot\vec{x}_j)\alpha_i\alpha_j - \sum_{i=1}^{N}\alpha_i \tag{11}$$

$$\text{Sujeito as restrições: } \begin{cases} 0 \leq \alpha_i \leq C, & \forall_i, 1, ..., N \\ \sum_{i=1}^{N} y_i\alpha_i = 0 \end{cases} \tag{12}$$

Por fim, a função Kernel presente no problema deve acatar condições do teorema de mercer com objetivo de que o problema de otimização seja convexo e a possibilidade que produtos escalares sejam realizados no mapeando do Kernel [34].

2.3 Considerações Finais

Este capítulo teve como objetivo apresentar o Infarto do miocárdio e as Máquinas de Vetores de suporte. Complementando o conteúdo da problemática, a condição do Infarto do miocárdio foi explicada para que se tenha o entendimento do que é essa doença. Seguindo o mesmo veio, as máquinas de vetores de suporte foram explicadas primeiro levando em conta sua teoria, e depois sua formulação matemática para se ter o entendimento sobre essa tecnologia. No próximo capítulo, metodologia, o trabalho feito para se alcançar um bom modelo de previsão de óbitos motivados por ataque cardíaco foi descrito passo a passo.

3 Metodologia

Para realizar o processo de descoberta de conhecimento do trabalho proposto, foi utilizada a ferramenta Weka. O Weka é um software de código aberto que disponibiliza algoritmos de aprendizado de máquina para utilização em bases de dados. O Weka também dispõe ferramentas para dar suporte ao processo de descoberta de conhecimento. Muitas funcionalidades dessa ferramenta foram utilizadas, desde a etapa de pré-processamento; onde os dados foram limpos, até a parte de mineração de dados; onde o Weka não só disponibilizou o algoritmo de aprendizado de máquina escolhido, mas também forneceu mecanismos para escolher o melhor algoritmo.

A metodologia utilizada baseou-se no processo KDD. Este processo define 5 etapas que permitem retirar conhecimento de volumes de dados. O cerne da pesquisa está em prever óbitos motivados por ataque cardíaco. A figura 12 explica o processo utilizado. Portanto, os dados são transformados ao longo do processo, saindo de sua forma mais bruta até chegar nos padrões que antes eram implícitos aos dados, revelados pelo algoritmo de aprendizado de máquina.

O KDD não é uma linha continua, pois quase sempre é necessário voltar etapas anteriores para corrigir problemas afim de se obter melhores resultados. Por exemplo, na etapa de mineração de dados, se após exaustivas tentativas os algoritmos não estiverem apresentando um desempenho ao menos regular, pode ser que o problema esteja nos próprios dados e as etapas anteriores deverão ser revistas para otimizar o resultado final. Na presente pesquisa, precisou-se retornar as etapas de pré-processamentos para avaliar a importância de variáveis e como essas variáveis afetavam o classificador gerado. A etapa de mineração de dados também foi executada repetidas vezes à procura pelo algoritmo que melhor se saía bem na tarefa de classificação. A metodologia proposta nesse trabalho é apresentada na figura 12. Essa figura sintetiza 5 atividades:

- **Seleção:** Se concentra em coletar os dados necessários para a descoberta de padrões.
- **Pré-processamento:** Se concentra em adequar os dados para a etapa de mineração. Os dados são limpos para que ruídos não interfiram na qualidade do conhecimento que se deseja obter no final.
- **Transformação:** Se concentra na organização dos dados para a etapa de mineração.
- **Mineração de dados:** É onde finalmente a extração de conhecimento ocorre. Utilizando um algoritmo de aprendizado de máquina, os dados são interpretados para gerar um modelo capaz de prever mortes por ataque cardíaco. Essa etapa

Figura 10 – Processo KDD para descoberta de conhecimento

Fonte: The KDD process for extracting useful knowledge from volumes of data, 1996. [37]

também inclui a procura pelo algoritmo que apresentou o melhor modelo para a previsão proposta.

- **Interpretação e avaliação:** É onde as informações geradas são interpretadas e avaliadas para se obter a previsão, levando em consideração acurácia do modelo, taxa de erro, entre outros fatores.

3.1 Coleta da Base de dados

A base de dados escolhida para a previsão de óbitos por ataque cardíaco foi obtida a partir do estudo em [38]. Cada instância contida na base contém uma serie de medidas tiradas de um paciente que sofreu um Infarto do miocárdio. A base de dados contém medidas utilizando o ecocardiograma (ultrassom do coração). Profissionais utilizam essas medidas para poder prever se uma pessoa que sofreu um ataque do

coração possui alguma chance de sobrevivência. Apesar de ser construída por um profissional da área, a base não está livre de ruído. O ruído está presente nas variáveis que a medição está relacionado com métodos não precisos. O exemplo disso é a variável "Wall-motion-score"que é uma medida subjetiva feita por um médico observando um ecocardiograma [39].

Figura 11 – Seleção

Fonte: Modificado de The KDD process for extracting useful knowledge from volumes of data, 1996. [37]

3.2 Pré-Processamento

A base de dados, ao todo, possui 132 instâncias com 13 atributos. No entanto, como recomendado em [38], nem todas as instâncias são utilizadas para classificação. Algumas das instâncias são de pacientes que não foram acompanhados por tempo suficiente para dizer que sobreviveram ou não ao ataque cardíaco durante o período de um ano. Essas instâncias precisaram ser retiradas da base de dados para que não interferissem no classificador. Não se sabe se esses pacientes de fato sobreviveram ou se morreram ao período estipulado de doze meses ao ataque cardíaco, então suas características não puderam ser ligadas com certeza ao óbito ou a sobrevivência.

Para identificar as instâncias de pacientes que não foram acompanhados dentro do tempo adequado, precisou-se olhar para os valores das variáveis ‘survival’ e ‘still-alive’ em conjunto. A variável ‘survival’ pode indicar o período em meses que o paciente veio a óbito depois de ter sofrido um ataque cardíaco, ou pode indicar o período de sobrevivência desse paciente caso ele ainda esteja vivo. Isso dependerá do valor da variável ‘still-alive’, que indica o status do paciente e pode assumir apenas dois valores: 0, caso o paciente esteja morto, então ‘survival’ indicará o período de óbito, ou 1 caso o paciente esteja vivo, então ‘survival’ indicará o período de sobrevivência. As instâncias que possuem o valor de ‘still-alive’ como 1, e apresentam um período menor que doze meses indicado pela variável ‘survival’ são de pacientes que não foram acompanhados

pelo período mínimo de um ano necessário para contribuir com seu status para a tarefa de classificação.

Para o pré-processamento de dados, o weka fornece diversos filtros para manipulação da base de dados bruta, permitindo sua transformação. O filtro para retirar as instâncias dos pacientes que não foram acompanhados por tempo suficiente foi o SubsetByExpression. Esse filtro permitiu a retirada de um conjunto de instâncias da base de dados, através de expressões regulares. Das 132 instâncias que a base continha sobraram 96 após a aplicação do filtro.

As instâncias da base de dados escolhida apresentaram alguns atributos irrelevantes para a tarefa de classificação. Dos 13 atributos, 2 dos atributos, 'mult' e 'group', foram apagados como recomendado dentro da descrição da própria base de dados que manda ignorá-los. Outra recomendação da própria descrição da base de dados é substituir a variável 'wall-motion-score' pela variável 'wall-motion-index'. A variável 'name' também foi desconsiderada.

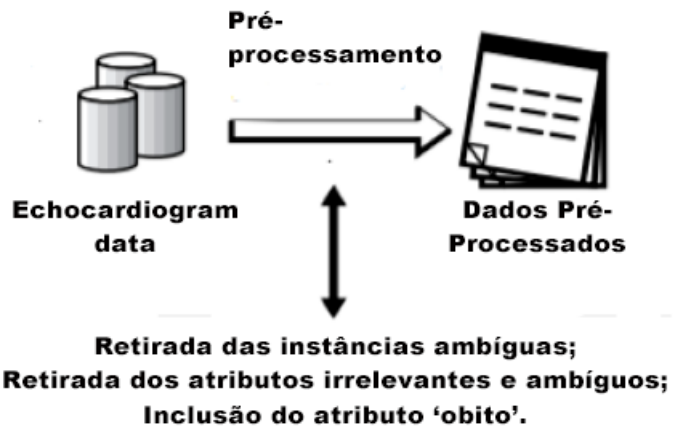
A variável 'alive-at-one' é uma variável cujo o valor é derivado dos valores das variáveis 'survival' e 'still-alive'. A 'alive-at-one' assumiria o valor 0 se o paciente estivesse morto após um ano ou se ele foi acompanhado por menos de um ano, e assumiria o valor 1 se ele estivesse vivo aos um ano. Essa variável é o mais próximo que sem tem de uma classe na base de dados para o problema abordado. No entanto, em detrimento da utilização dessa variável como classe para a pesquisa presente, foi criada uma nova variável que assumirá esse papel, 'óbito', que pode conter apenas dois valores, 0 e 1, sendo 1 para o caso de pacientes que não sobreviveram ao período de um ano, e 0 caso contrário. O motivo da substituição da variável 'alive-at-one' pela variável criada 'óbito', e consequentemente a sua não utilização, é o fato que essa variável não representa muito bem a classificação proposta aqui, porque além ela não levar em consideração os pacientes que morreram antes do período de um ano isoladamente, sua distribuição de valores na base de dados é confusa por não bater com a descrição que se tem dela.

As variáveis 'survival' e 'still-alive' são importantes para a derivação dos valores da classe 'óbito' e o pré-processamento da base, porém, essas variáveis não serão utilizadas para o aprendizado de máquina. Incluir a variável 'survival' no aprendizado não é adequado para geração de uma boa hipótese, já que a hipótese gerada pelo algoritmo de aprendizado de máquina estaria ligada ao número de meses de sobrevivência ou de óbito de um paciente para prever um resultado. Para utilização dessa previsão como meio de apoio a decisão de tratamento de pacientes, não haverá número de meses, seja de sobrevivência ou de óbito, para se ter como base, o que acontecerá, é que as características fisiológicas dos pacientes serão retiradas o quanto antes possível dele para a indicação de seu status. A variável 'still-alive' também não pode ser utilizada,

porque para a tarefa de previsão proposta aqui o fato de o paciente ter morrido em qualquer momento após o período de um ano não é importante, o que se deve levar em consideração é se ele veio a óbito entre um ano ou não, característica essa já é indicada pela classe recém criada e nomeada como 'óbito'.

Algumas instâncias da base de dados possuem informações pendentes. No entanto, como isso pode ser uma característica comum para essa base de dados devido a forma e complexidade como os dados são coletados para compor cada instância, no presente trabalho, levou-se em consideração a escolha de um algoritmo que melhor se adequasse ou tratasse de maneira automática tal falta de dados. A figura 14 resume o processo feito nessa etapa de pré-processamento em tópicos dentro do contexto do processo utilizado. Uma visão geral das variáveis retiradas (variáveis que se encontram na cor vermelha), das variáveis mantidas e suas descrições é encontrada na tabela 2.

Figura 12 – Pré-processamento



Fonte: Modificado de The KDD process for extracting useful knowledge from volumes of data, 1996. [37]

3.3 Transformação

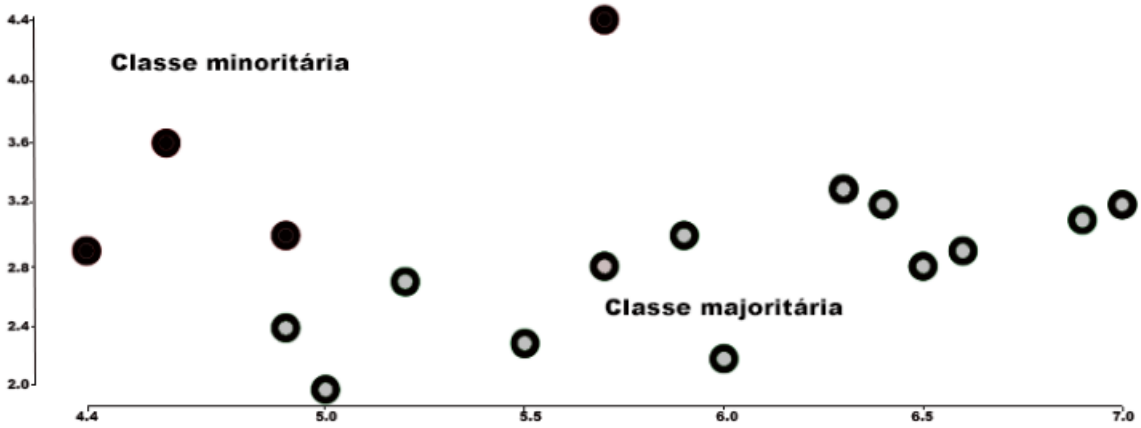
Após os valores da variável 'óbito' serem distribuídos a todas as instâncias, verificou-se que a base de dados utilizada estava desbalanceada. Para tratar o desbalanceamento a ferramenta SMOTE (Synthetic Minority Over-sampling Technique) foi utilizada. O SMOTE realiza o balanceamento de uma base de dados levando em contas os vizinhos mais próximos da mesma classe de uma instância. Uma instância nova é gerada por esse algoritmo em algum ponto ao longo da distância de uma instância considerada e seu vizinho ou vizinhos mais próximos. Na figura 15, tem-se um exemplo ilustrativo de uma base de dados desbalanceada.

Tabela 2 – Variáveis da base de dados

Nome	**Descrição**
Survival	Tempo de sobrevivência do paciente dado em meses esteja ele morto ou não.
Still-alive	Variável que indica se o paciente está morto ou não. 0 = morto ao fim do período indicado pela variável 'survival'; 1 = vivo.
Age-at-heart-attack	Idade do paciente no momento do ataque.
Pericardial-effusion	Fluído em volta do coração. 0 = sem fluído; 1 = com fluído.
Fractional-shortening	Mensuração da contratilidade ao redor do coração. Valores baixos indicam anormalidade.
Epps	Outra mensuração de contratilidade do coração. Números altos indicam anormalidade.
Lvdd	Mensura o tamanho do coração na diástole final. Há a tendencia de doença em corações grandes.
Wall-motion-score	Mensura a movimentação dos segmentos do ventrículo esquerdo.
Wall-motion-index	Semelhante a variável Wall-motion-score com a divisão do número de segmentos vistos.
Mult	Pode ser ignorada.
name	Nome dos pacientes cuja a instância representa.
group	Pode ser ignorada.
Alive-at-1	Se deriva de 'survival' e 'still-alive'. 0 = morto após um ano ou não acompanhado por esse período. 1 = vivo no período de um ano.

Os pontos totalmente preenchidos são as instancia da classe minoritária. Essa classe, em relação a outra, classe majoritária, possui bem menos dados, oque pode prejudicar o treinamento do algoritmo de aprendizado de máquina e por consequência o modelo gerado. Como o número de pontos não preenchidos e predominante maior, o

Figura 13 – Base de dados desbalanceada

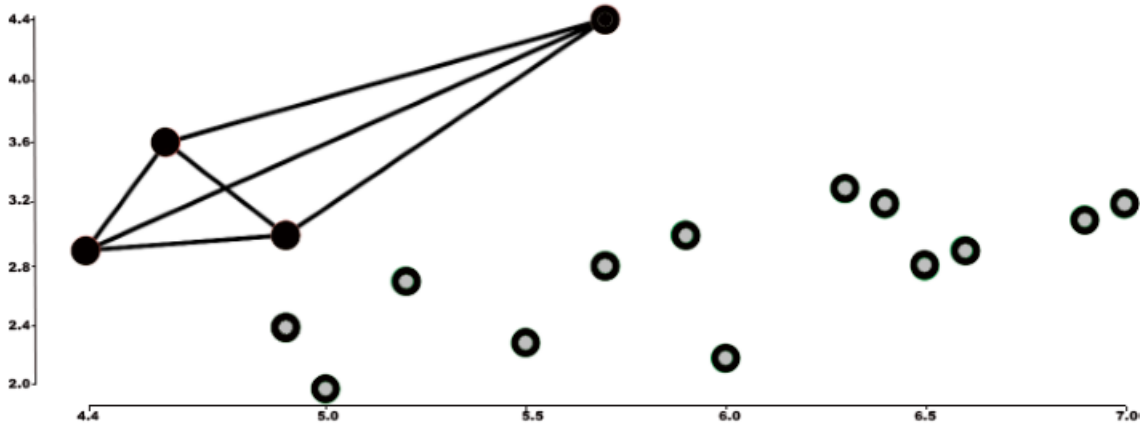


Fonte: Adaptado de SMOTE explained for noobs - Synthetic Minority Over-sampling TEchnique line by line (acessado em 23/10/2020). [40]

algoritmo tende a classificar todos dessa maneira. Na presente pesquisa o número de instâncias negativas é predominantemente maior, para evitar a tendência de classificação erronia das amostras positivas, mais dessas amostras foram geradas artificialmente através do SMOTE.

No campo de aprendizado de máquina esse método para reverter desbalanceamento é chamado de over-sample, que constitui o nome da ferramenta. A vantagem de se utilizar o SMOTE é que ele não duplica os dados, e sim, através de seu modelo matemático gera novos dados válidos. A figura 16 e 17 ilustram, em dois passos simplificados, como o SMOTE gera as amostras sintéticas, primeiramente calculando a distância entre elas, depois gerando os dados entre essas distâncias.

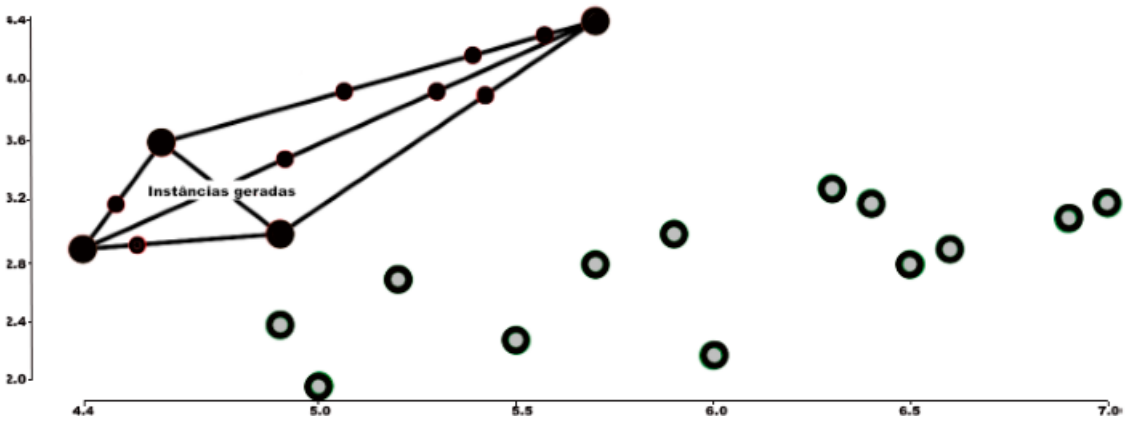
Figura 14 – Calculando as distâncias



Fonte: Adaptado de SMOTE explained for noobs - Synthetic Minority Over-sampling TEchnique line by line (acessado em 23/10/2020). [40]

Após a utilização do SMOTE passou-se a ter 48 instâncias da classe positiva em

Figura 15 – Pontos entre as distancias



Fonte: Adaptado de SMOTE explained for noobs - Synthetic Minority Over-sampling TEchnique line by line (acessado em 23/10/2020). [40]

relação a 92 instâncias da classe negativa, reduzindo o desbalanceamento consideravelmente. Outros valores também foram testados, mas essa proporção foi a que teve os melhores resultados. O capitulo 4 entra em mais detalhes sobre isso. A figura 18 resume o processo feito nessa etapa de transformação em tópicos dentro do contexto do processo utilizado.

Figura 16 – Transformação

Fonte: Modificado de The KDD process for extracting useful knowledge from volumes of data, 1996. [37]

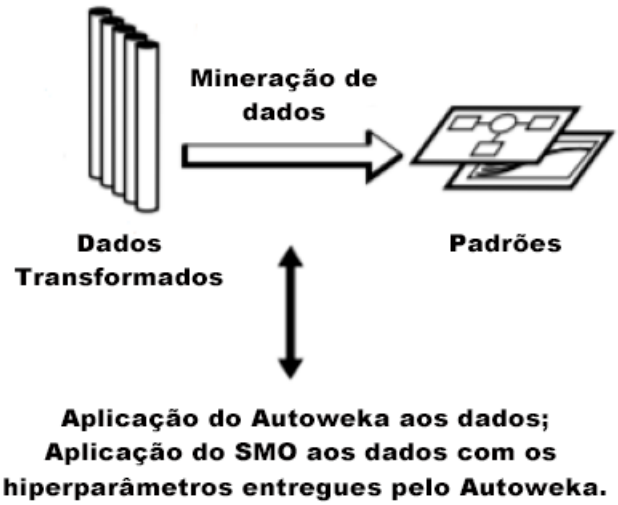
3.4 Extração de padrões

Para a descoberta de um melhor algoritmo que levaria a um melhor modelo preditivo sobre os dados considerados, foi utilizada a ferramenta Autoweka cujo o objetivo é, através do amplo espectro de algoritmos de aprendizado de máquina dis-

ponibilizados pela ferramenta weka, achar o que melhor gera um classificador para os dados em questão, já que dependendo das características dos dados alguns algoritmos podem se sair melhor do que outros nessa tarefa. Além do algoritmo em si, o autoweka também faz uma busca pelos melhores hiperparâmetros, o que também é essencial para gerar um classificador ótimo.

Para utilizar o autoweka em uma determinada base de dados estipula-se um tempo máximo que essa ferramenta deverá realizar o processo de busca pelo algoritmo e hiperparâmetros ótimos. No presente trabalho, foi estipulado arbitrariamente um tempo de 5 horas para essa busca o que gerou dados bem satisfatórios como discutidos mais à frente na sessão 4. Como resultado o Autoweka apontou o SMO como sendo o algoritmo que melhor geraria um classificador para base de dados tratada no trabalho proposto. O SMO nada mais é do que um algoritmo de treinamento de SVM (Algoritmo de aprendizado de máquina discutido anteriormente) mais rápido é otimizado que outros com o mesmo objetivo. O SMO foi testado novamente utilizando a ferramenta weka com hiperâmetros apontados como melhores para a base de dados Echocardiogram pelo autoweka e o resultado foi compatível com o gerado pela ferramenta, no entanto, a busca de atributos ótimos não mostrou resultados significativos que pudessem levar em conta sua utilização. Esses resultados serão apresentados melhor na sessão 4. A figura 17 resume em tópicos o processo feito nesta etapa de extração de padrões dentro do contexto do processo utilizado.

Figura 17 – Mineração de dados.

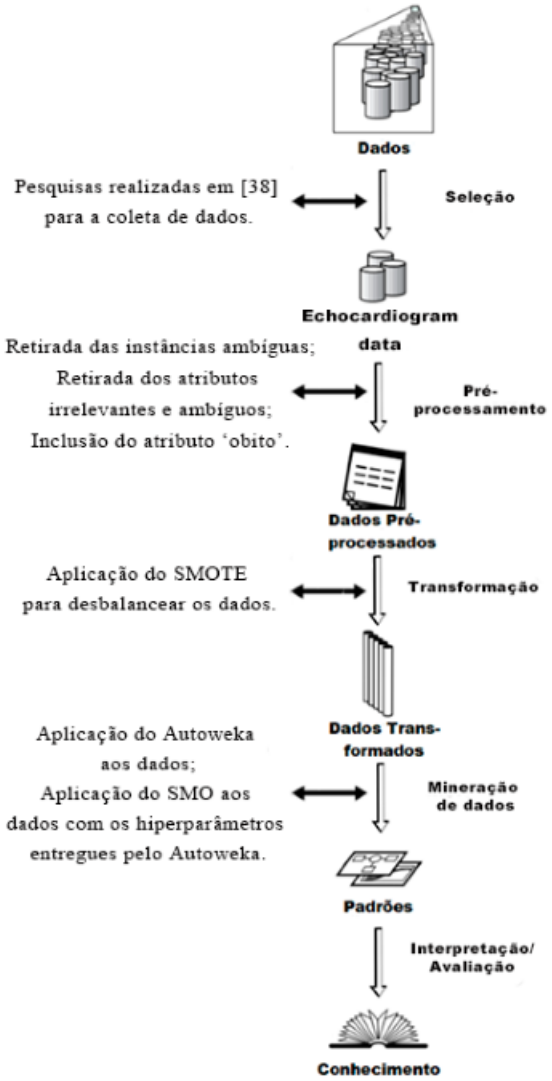


Fonte: Modificado de The KDD process for extracting useful knowledge from volumes of data, 1996. [37]

Como uma observação extra, o Autoweka também apontou em outras tentativas durante a pesquisa realizada o algoritmo de aprendizado de máquina AdaBoostM1 também como uma possível solução. Na presente pesquisa optou-se pelo SMO pelo fato

de ele trazer melhores resultados, mas isso não excluí a possibilidade de pesquisas futuras com o mesmo objetivo utilizando o AdaBoostM1 para se ter melhores ponderações sobre as diferenças da utilização das duas técnicas. A figura 18 fornece um panorama geral e resumido do que foi feito durantes as etapas para transformar os dados alvo em padrões.

Figura 18 – Panorama geral do processo



Fonte: Modificado de The KDD process for extracting useful knowledge from volumes of data, 1996. [37]

3.5 Considerações Finais

Neste capitulo foi descrito todo o processo para a obtenção de um classificador que desse suporte a previsão de óbitos motivados por ataques cardíacos. No próximo capítulo, resultados, foi discutido sobre o classificar gerado e seus resultados levando

em conta fatores como a taxa de acerto e recall. Considerações sobre a utilização desse modelo como meio de suporte a diagnóstico médico também compõe o capítulo de resultados.

4 Resultados

Como descrito na sessão anterior, o Autoweka retornou o SMO como o algoritmo que melhor gerou um classificador para o problema de previsão que se propõe nessa monografia, em um tempo estipulado de 5 horas. As figuras 22 e 23 abaixo mostram os principais resultados retornados pela ferramenta. Das 140 instancias presentes, 138 foram classificadas corretamente, valendo destacar que todas as 48 instâncias da classe 'obito' = 1 estão inclusas nesse conjunto. Como será discutido posteriormente este número demostra uma certa importância para avaliação inicial devido a natureza do problema que se trata neste pesquisa devido ao baixo grau de falsos negativos que o modelo gerou.

Figura 19 – Matriz de confusão - Autoweka

```
=== Confusion Matrix ===

  a  b   <-- classified as
 48  0 |  a = 1
  2 90 |  b = 0
```

Figura 20 – Taxa de classificações corretas e incorretas - Autoweka

```
Correctly Classified Instances         138               98.5714 %
Incorrectly Classified Instances         2                1.4286 %
```

Após o resultado do Autoweka, a próxima tarefa a ser realizada se trata da verificação dos resultados gerados pelo SMO isoladamente com os hiperparâmetros apontados pelo proprio Autoweka e aplicando a Validação cruzada com 10 subconjuntos para se ter uma melhor mensuração da acurácia mostrada pelo modelo. As figuras 24 e 25 mostram os novos resultados gerados pelo modelo. A taxa de instancias classificadas corretamente caíram em um valor não significativo de aproximadamente 3% gerando uma alta de mesmo valor na taxa de instancias classificadas incorretamente. Mostrou-se então uma taxa de acerto de 95% indicando um bom desempenho mostrado pelo algoritmo SMO em prever óbitos motivados por ataque cardíacos.

Figura 21 – Matriz de confusão - Validação Cruzada

```
=== Confusion Matrix ===

  a  b   <-- classified as
 47  1 |  a = 1
  6 86 |  b = 0
```

Fonte: O autor.

Figura 22 – Taxa de classificações corretas e incorretas - Validação Cruzada

Correctly Classified Instances	133	95 %
Incorrectly Classified Instances	7	5 %

Fonte: O autor.

Sabe-se que o problema abordado possui uma extrema delicadeza e que o resultado da previsão deve ser visto com os maiores cuidados. Apesar da boa acurácia mostrada pelo modelo, é essencial no problema abordado na presente pesquisa, dar igual importância ao grau de falsos negativos, que são o números de instâncias que possuíam como saída a classe positiva, mas foram classificadas como negativas. Um falso negativo poderia relaxar as decisões de tratamento mais rígidas para um paciente em risco de vida. O contrário (falso positivo) pode ser menos ou igualmente danoso, pois pode dar suporte a uma decisão de submeter desnecessariamente um paciente a tratamentos invasivos e inadequados para sua situação. No classificador obtido, das 48 instâncias positivas apenas uma foi classificada erroneamente como sendo negativa, ficando 47 verdadeiros positivos e 1 falso negativo. Essa proporção pode ser traduzida para o valor do Recall (uma medida que indica quantas instâncias positivas foram classificadas como positivas; valor esse que varia entre zero e um) sendo em torno de 0,979. Das 92 instâncias de classe negativa, 86 delas foram classificadas corretamente, ficando 86 verdadeiros negativos e 6 falsos positivos. Na tabela abaixo é possível conferir o valor do recall gerado diretamente pelo modelo.

Figura 23 – Recall e outras medidas

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,979	0,065	0,887	0,979	0,931	0,894	0,957	0,875	1
	0,935	0,021	0,989	0,935	0,961	0,894	0,957	0,967	0
Weighted Avg.	0,950	0,036	0,954	0,950	0,951	0,894	0,957	0,936	

Fonte: O autor.

O essencial seria que o classificador gerado não errasse uma previsão se quer, mas isso é extremamente improvável de acontecer devido a diversos fatores, um deles seria a própria natureza dos dados. No caso da pesquisa proposta o bom a ser feito e reduzir ao máximo o número de falsos negativos, já que se sabe que eles podem trazer grande complicações além de trazer mais desvantagens do que vantagens na utilização das previsões como suporte a decisão de profissionais para tratamento dos pacientes. Levando em consideração o resultado do classificador gerado para a presente pesquisa, é possível que o classificador gerado para essa tarefa de previsão tenho o número de falsos negativos próximo a zero.

5 Conclusões

Para a tarefa de previsão proposta na presente pesquisa o algoritmo SVM foi o que mais se mostrou eficiente, gerando um classificador com uma acurácia acima de 90%. Além disso o classificador gerado possui um baixo índice de falsos negativos que podem gerar impactos prejudiciais no tratamento de pacientes que precisam de uma atenção elevada.

O desbalanceamento da base de dados, apesar de minimizado pelo uso da ferramenta SMOTE, foi um grande limitador mostrado por ela para a descoberta de um classificador ótimo que realiza a tarefa de previsão proposta. Apesar de os dados sintéticos gerados pelo SMOTE serem válidos para utilização, desbalancear a base, se possível, com mais dados reais permitiria tratar o problema mais próximo da realidade. Aumentar o número de instâncias na base também contribuiria para o treinamento do classificador proposto.

Algumas variáveis a mais poderiam contribuir para a previsão do status do paciente além das presentes na própria base, por exemplo, uma variável que indicasse o tempo de demora para a reperfusão poderia ser incluída pois se sabe que quanto mais tempo se demora para realizar esse procedimento mais a necrose avança e maiores são os danos ao coração do paciente. Alguma informações sobre histórico clínico também poderia ser utilizada, informações essas que indicariam por exemplo predisposição ao ataque cardíaco, ou a presença de algum complicador para sua recuperação.

Deve-se levar em consideração em estudos mais aprofundados o impacto de uma previsão negativa para casos positivos (falso negativo) e sua minimização, para poder utilizar tais previsões com confiança como apoio a decisão de profissionais para tratamento dos pacientes. Deve-se também levar em consideração estudos sobre o impacto que uma previsão positiva em casos negativos trariam ao paciente e ao contexto em geral.

Referências

[1] M. Raihan, S. Mondal, A. More, M. O. F. Sagor, G. Sikder, M. A. Majumder, M. A. A. Manjur, and K. Ghosh, "Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design," in *2016 19th International Conference on Computer and Information Technology (ICCIT)*, pp. 299–303, Dec 2016.

[2] "Opas/oms brasil - opas/oms brasil." `https://www.paho.org/bra/index.php?option=com_content&view=article&id=5638:10-principais-causas-de-morte-no-mundo&Itemid=0`. (Accessed on 02/02/2019).

[3] R. M. B. D. e. N. M. G. d. S. Tatiana Laís Fonsêca de Medeiros, Paloma Cibelle Nascimento Silva de Andrade, "Mortalidade por infarto agudo do miocÁrdio," *Revista de Enfermagem UFPE On Line*, pp. 565 – 573, 2018.

[4] D. K. Ravish, K. J. Shanthi, N. R. Shenoy, and S. Nisargh, "Heart function monitoring, prediction and prevention of heart attacks: Using artificial neural networks," in *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, pp. 1–6, Nov 2014.

[5] F. Bulut, "Heart attack risk detection using bagging classifier," in *2016 24th Signal Processing and Communication Application Conference (SIU)*, pp. 2013–2016, May 2016.

[6] A. Rabeh, B. Faouzi, and H. Amiri, "Diagnosis of alzheimer diseases in early step using svm (support vector machine)," pp. 364–367, 03 2016.

[7] S. Hongzong, W. Tao, Y. Xiaojun, L. Huanxiang, H. Zhide, L. Mancang, and F. BoTao, "Support vector machines classification for discriminating coronary heart disease patients from non-coronary heart disease," *West Indian Medical Journal*, vol. 56, pp. 451 – 457, 10 2007.

[8] M. Ahmad, V. Tundjungsari, D. Widianti, P. Amalia, and U. A. Rachmawati, "Diagnostic decision support system of chronic kidney disease using support vector machine," in *2017 Second International Conference on Informatics and Computing (ICIC)*, pp. 1–4, Nov 2017.

[9] R. M. SILVA, M. LEAL, and F. LIMA, "Predico do Câncer de Mama com Aplicação de Modelos de Inteligência Computacional," *TEMA (São Carlos)*, vol. 20, pp. 229 – 240, 08 2019.

[10] A. Bhattacharya, M. Mishra, A. Singh, and M. K. Dutta, "Machine learning based portable device for detection of cardiac abnormality," in *2017 International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT)*, pp. 1–4, Nov 2017.

[11] J. Nahar, T. Imam, K. S. Tickle, and D. G. Alonso, "Medical knowledge based data mining for cardiac stress test diagnostics," in *2015 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, pp. 1–7, Dec 2015.

[12] C. Sowmiya and P. Sumitra, "Analytical study of heart disease diagnosis using classification techniques," in *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, pp. 1–5, March 2017.

[13] A. R. Langowiski, Paraná. Secretaria de estado da saúde, and Cardiologista da Divisão de Risco Cardiovascular, "Linha guia de infarto agudo do miocárdio 2017 | secretaria da saúde." `http://www.saude.pr.gov.br/arquivos/File/LinhaGuiaInfartoMiocardio_2017.pdf`. (Accessed on 06/09/2020).

[14] A. E. P. Pesaro, C. V. Serrano Jr., and J. C. Nicolau, "Infarto agudo do miocárdio: sindrome coronariana aguda com supradesnível do segmento ST," *Rev. Assoc. Med. Bras.*, vol. 50, pp. 214 – 220, 04 2004.

[15] M. H. V. Bruna, "Infarto do miocárdio (ataque cardíaco) | drauzio varella - drauzio varella." `https://drauziovarella.uol.com.br/doencas-e-sintomas/infarto-do-miocardio-ataque-cardiaco/`. (Accessed on 06/09/2020).

[16] Canal Médico, "Infarto agudo no miocárdio - etiologia 1/10 - youtube." `https://www.youtube.com/watch?v=iSa5wVnwShU`. (Accessed on 06/09/2020).

[17] Canal Médico, "Etiopatogenia - doença arterial coronariana 5/10 - youtube." `https://www.youtube.com/watch?v=z8zAgMAiTe0&list=PLuQOoDFb07pGdR2u_s9HqVhRBI3_yqehx&index=5`. (Accessed on 06/09/2020).

[18] Canal Médico, "Etiopatogenia 2 - doença arterial coronariana 6/10 - youtube." `https://www.youtube.com/watch?v=TRqUwR-N1XM&list=PLuQOoDFb07pGdR2u_s9HqVhRBI3_yqehx&index=6`. (Accessed on 06/09/2020).

[19] A. A. Faludi, M. C. d. O. Izar, J. F. K. Saraiva, A. P. M. Chacra, H. T. Bianco, A. Afiune Neto, A. Bertolami, A. C. Pereira, A. M. Lottenberg, A. C. Sposito, A. C. P. Chagas, A. Casella Filho, A. F. Simão, A. C. d. Alencar Filho, B. Caramelli, C. C. Magalhães, C. E. Negrão, C. E. d. S. Ferreira, C. Scherr, C. M. A. Feio, C. Kovacs, D. B. d. Araújo, D. Magnoni, D. Calderaro, D. M. Gualandro, E. P. d. Mello Junior,

E. R. G. Alexandre, E. I. Sato, E. H. Moriguchi, F. H. Rached, F. C. d. Santos, F. H. Y. Cesena, F. A. H. Fonseca, H. A. R. d. Fonseca, H. T. Xavier, I. C. P. Mota, I. d. C. B. Giuliano, J. S. Issa, J. Diament, J. B. Pesquero, J. E. d. Santos, J. R. Faria Neto, J. X. d. Melo Filho, J. T. Kato, K. P. Torres, M. C. Bertolami, M. H. V. Assad, M. A. H. Miname, M. Scartezini, N. A. Forti, O. R. Coelho, R. C. Maranhão, R. D. d. Santos Filho, R. J. Alves, R. L. Cassani, R. T. B. Betti, T. d. Carvalho, T. L. d. R. Martinez, V. Z. R. Giraldez, and W. Salgado Filho, "Atualização da Diretriz Brasileira de Dislipidemias e Prevenção da Aterosclerose – 2017," *Arquivos Brasileiros de Cardiologia*, vol. 109, pp. 1–76, 08 2017.

[20] K. Academy, "Aterosclerose (vídeo) | khan academy." `https://pt.khanacademy.org/science/health-and-medicine/circulatory-system-diseases/coronary-artery-disease/v/atherosclerosis`. (Accessed on 06/09/2020).

[21] P. R. T. Gonçalves, G. Y. B. de Moraes, B. de Almeida Pereira, and A. Gritti, "Aterosclerose e sua relaÇÃo com as doeÇas cardiovasculares atherosclerosis and its relationship with cardiovascular diseases," *Revista Saúde em Foco* , p. 711, 2018.

[22] P. J. Barter, S. Nicholls, K.-A. Rye, G. Anantharamaiah, M. Navab, and A. M. Fogelman, "Antiinflammatory properties of hdl," *Circulation Research*, vol. 95, no. 8, pp. 764–772, 2004.

[23] Canal Médico, "Infarto agudo no miocárdio - fisiopatologia i 2/10 - youtube." `https://www.youtube.com/watch?v=CNnp4YK5_Hs`. (Accessed on 06/09/2020).

[24] L. V. da Rosa, F. Medeiros, M. A. Deway, and R. K. Filho, "Infarto com Supradesnivelamento do ST," 06 2010.

[25] "Terapia de reperfusão no infarto agudo do miocárdio - trombólise e terapia trombolítica 1/3 - youtube." `https://www.youtube.com/watch?v=14BxhwOun9M`. (Accessed on 10/30/2020).

[26] "Infarto agudo no miocárdio - fisiopatologia ii 3/10 - youtube." `https://www.youtube.com/watch?v=nLnyl3lUnX4`. (Accessed on 10/30/2020).

[27] Canal médico, "Infarto agudo no miocárdio - fisiopatologia iii 4/10 - youtube." `https://www.youtube.com/watch?v=7LJE1_hxVWI`. (Accessed on 06/10/2020).

[28] L. Piegas, A. Timerman, G. Feitosa, J. Nicolau, L. Mattos, M. Andrade, A. Avezum, A. Feldman, A. De Carvalho, A. Sousa, A. Mansur, A. Bozza, B. Falcão, B. Markman Filho, C. Polanczyk, C. Gun, C. Serrano Junior, C. Oliveira, D. Moreira,

D. Précoma, D. Magnoni, D. Albuquerque, E. Romano, E. Stefanini, E. Santos, E. God, E. Ribeiro, F. Brito Júnior, G. Feitosa-Filho, G. Arruda, G. Oliveira, G. Oliveira, G. Lima, H. Dohmann, I. Liguori, J. Costa, J. Saraiva, L. Maia, L. Moreira, M. Arrais, M. Canesin, M. Coutinho, M. Moretti, N. Ghorayeb, N. Vieira, O. Dutra, O. Coelho, P. Leães, P. Rossi, P. Andrade, P. Lemos, R. Pavanello, R. Vivacqua Costa, R. Bassan, R. Esporcatte, R. Miranda, R. Giraldez, R. Ramos, S. Martins, V. Esteves, and W. Mathias Junior, "V Diretriz da Sociedade Brasileira de Cardiologia sobre Tratamento do Infarto Agudo do Miocárdio com Supradesnível do Segmento ST," *Arquivos Brasileiros de Cardiologia*, vol. 105, pp. 1 – 121, 08 2015.

[29] Canal médico, "Infarto agudo no miocárdio - tratamento i 7/10 - youtube." `https://www.youtube.com/watch?v=bzMGUgTNiB0`. (Accessed on 06/10/2020).

[30] Canal médico, "Terapia de reperfusão no infarto agudo do miocárdio - angioplastia primária 3/3 - youtube." `https://www.youtube.com/watch?v=op4NejCqkIo`. (Accessed on 06/10/2020).

[31] Drauzio Varella, "Tempo é músculo | episódio 2 - youtube." `https://www.youtube.com/watch?v=P7oV-M8br14&feature=emb_title`. (Accessed on 06/10/2020).

[32] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Tech. Rep. MSR-TR-98-14, April 1998.

[33] V. T. Ales, "O algoritmo sequential minimal optimisation para resolução do problema de support vector machine: uma técnica para reconhecimento de padrões," Master's thesis, Dissertação (Mestrado em Ciências)–Universidade Federal do Paraná, Curitiba, 2008.

[34] A. C. Lorena and A. C. P. L. F. de Carvalho, "Uma introdução às support vector machines," *Revista de Informática Teórica e Aplicada; Vol. 14, No 2 (2007); 43-67*, vol. 14, 12 2007.

[35] "Aprendizagem de máquina: como as máquinas de vetores de suporte podem ser utilizadas nas negociações - artigos mql5." `https://www.mql5.com/pt/articles/584`. (Accessed on 10/31/2020).

[36] B. P. R. de Carvalho, "Sucesu 2005–tecnologias–inteligência artificial o estado da arte em métodos para reconhecimento de padrões: Support vector machine," 2005.

[37] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The kdd process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, pp. 27–34, Nov. 1996.

[38] "Uci machine learning repository: Echocardiogram data set." `https://archive.ics.uci.edu/ml/datasets/echocardiogram`. (Accessed on 06/15/2020).

[39] S. Salzberg, *Exemplar-based learning: Theory and implementation (Technical Report TR-10-88)*. (33 Oxford Street; Cambridge, MA 02138): Harvard University, Center for Research in Computing Technology, Aiken Computation Laboratory, 1988.

[40] "Smote explained for noobs - synthetic minority over-sampling technique line by line · rich data." `https://rikunert.com/SMOTE_explained`. (Accessed on 10/23/2020).