Dissertação para Mestrado

Proposta e Avaliação de um Modelo para Predição da Morbidade e Mortalidade em Pacientes Diagnosticados com Taquicardia Ventricular

Victor Gabriel Lima Holanda
vglh@ic.ufal.br

Orientadores:

Dr. Rafael de Amorim Silva
Dr. Bruno Almeida Pimentel

Maceió
29 de Agosto, 2022

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Taciana Sousa dos Santos – CRB-4 – 2062

H722p Holanda, Victor Gabriel Lima.
Proposta e avaliação de um modelo para predição da morbidade e mortalidade em pacientes diagnosticados com taquicardia ventricular / Victor Gabriel Lima Holanda. – 2022.
115 f. : il. color.

Orientador: Rafael de Amorim Silva.
Coorientador: Bruno Almeida Pimentel.
Dissertação (Mestrado em Informática) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2022.

Bibliografia: f. 95-104.
Apêndice: f. 105-115.

1. Morbidade e mortalidade – Prognóstico 2. Taquicardia ventricular. 3. Informática em saúde. I. Título.

CDU: 004: 616.12-008.311

Victor Gabriel Lima Holanda

Proposta e Avaliação de um Modelo para Predição da Morbidade e Mortalidade em Pacientes Diagnosticados com Taquicardia Ventricular

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Informática pelo Instituto de Computação da Universidade Federal de Alagoas.

Orientadores:

Dr. Rafael de Amorim Silva

Dr. Bruno Almeida Pimentel

Maceió
29 de Agosto, 2022

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Informática – PPGI
Instituto de Computação/UFAL
Campus A. C. Simões BR 104-Norte Km 14 BL 12 Tabuleiro do Martins
Maceió/AL - Brasil CEP: 57.072-970 | Telefone: (082) 3214-1401

Folha de Aprovação

VICTOR GABRIEL LIMA HOLANDA

PROPOSTA E AVALIAÇÃO DE UM MODELO PARA PREDIÇÃO DA MORBIDADE E MORTALIDADE EM PACIENTES DIAGNOSTICADOS COM TAQUICARDIA VENTRICULAR

Dissertação submetida ao corpo docente do Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas e aprovada em 29 de agosto de 2022.

Banca Examinadora:

gov.br Documento assinado digitalmente
RAFAEL DE AMORIM SILVA
Data: 27/10/2022 11:48:11-0300
Verifique em https://verificador.iti.br

Prof. Dr. RAFAEL DE AMORIM SILVA
UFAL – Instituto de Computação
Orientador

gov.br Documento assinado digitalmente
BRUNO ALMEIDA PIMENTEL
Data: 26/10/2022 19:10:18-0300
Verifique em https://verificador.iti.br

Prof. Dr. BRUNO ALMEIDA PIMENTEL
UFAL – Instituto de Computação
Coorientador

gov.br Documento assinado digitalmente
DIEGO DERMEVAL MEDEIROS DA CUNHA MA
Data: 26/10/2022 22:52:47-0300
Verifique em https://verificador.iti.br

Prof. Dr. DIEGO DERMEVAL MEDEIROS DA CUNHA MATOS
UFAL – Instituto de Computação
Examinador Interno

gov.br Documento assinado digitalmente
ALMIR PEREIRA GUIMARAES
Data: 26/10/2022 21:49:39-0300
Verifique em https://verificador.iti.br

Prof. Dr. ALMIR PEREIRA GUIMARÃES
UFAL – Instituto de Computação
Examinador Externo

Agradecimentos

Aos meus pais Mônica e Junior por todo incentivo e suporte em minha educação.

À todos meus familiares que de alguma forma demonstraram apoio, em especial minha avó Maria, por todo incentivo dado tanto na escolha pelo caminho do mestrado, quanto na superação dos desafios encontrados ao decorrer desta etapa.

À meu sobrinho e afilhado Miguel por ser mais uma grande motivação em minha vida.

Aos meus amigos de quatro patas: Nick (in memoriam), Mel, Ayra e Dom, por tornarem os dias mais leves.

Aos colegas do Centro de Pesquisa em Tecnologia Emergente, Maxwell Acioli, Arthur Moreno, Marcos Bento e Fabiano Conrado, que compartilharam de seu tempo e conhecimento durante as reuniões realizadas no decorrer deste mestrado.

Às médicas Adriana Fattori e Gabriella Duarte que se disporam a contribuir para os avanços desta pesquisa.

Ao meu orientador Prof. Dr. Rafael de Amorim Silva e co-orientador Prof. Dr. Bruno de Almeida Pimentel, pelo suporte, disponibilidade e orientações, sempre mostrando os melhores caminhos a serem percorridos, e os pontos que deveriam ser levados em consideração para cada tomada de decisão durante todo este período.

Aos professores Drs. Almir Pereira Guimarães e Diego Demerval Medeiros da Cunha Matos, pela presença na banca de defesa.

À todos os professores do Programa de Pós Graduação em Informática pelos ensinamentos ao longo destes 2 anos.

Resumo

Segundo a Organização Mundial da Saúde, as doenças cardiovasculares são a maior causa de morte no mundo. Milhares de pessoas vêm a óbito todos os anos em decorrência das complicações ocasionadas por este tipo de doença. Neste sentido, médicos cardiologistas buscam o diagnóstico precoce, evitando que a doença alcance a fase patológica (i.e. os estágios onde as limitações surgem e se tornam cada vez maiores para os seus portadores). No contexto das arritmias do tipo taquicardia, os pacientes precisam adotar o uso contínuo de medicamentos e a mudança de hábitos para reduzir o impacto causado por este tipo de doença, provendo assim uma vida mais saudável. Sendo assim, este trabalho propõe e avalia um modelo de prognóstico de morbidade e mortalidade para pacientes diagnósticados com taquicardia ventricular. O modelo proposto considera o cenário onde o paciente diagnósticado com taquicardia ventricular irá desenvolver o quadro de insuficiéncia cardíaca (IC) e fibrilaçao ventricular (FV). Neste cenário, utilizamos o score CHA_2DS_2-VASc para predição do risco de AVC, AIT e Embolia Sistêmica para o cenário de FV. Além de uma combinação do CHA_2DS_2-VASc com o score gerado a partir dos hábitos do paciente, para ser capaz de categorizar o paciente dentro das classes de NYHA, que por sua vez gera um nível de morbidade e mortalidade para cada categoria. Foram utilizadas técnicas computacionais como o uso de Máquinas de Vetores Suporte (SVM),K vizinhos mais próximos (KNN), Redes Neurais, Árvores Aleatórias e Naïve Bayes para validar o sistema. O modelo alcançou seus melhores resultados pela validação com Máquinas de Vetores Suporte (SVM), com acurácia de 0.997, F1, precision e recall de 0.976. A partir destas classificações e predições, é possível prover ao cardiologista um maior nível de informações, otimizando as tomadas de decisão, impactando no tratamento adotado, intervenções, e acompanhamento do paciente.

Palavras-chave: Informática na Saúde, Taquicardia, Predição, Morbidade, Mortalidade.

Abstract

According to the World Health Organization, cardiovascular diseases are the leading cause of death in the world. Thousands of people die every year due to complications caused by this type of disease. In this sense, cardiologists seek early diagnosis, preventing the disease from reaching the pathological phase (i.e. the stages where limitations arise and become increasingly greater for its bearers). In the context of tachycardia-type arrhythmias, patients need to adopt the continuous use of medication and change habits to reduce the impact caused by this type of disease, thus providing a healthier life. Therefore, this study proposes and evaluates a morbidity and mortality prognostic model for patients diagnosed with ventricular tachycardia. The proposed model considers the scenario where the patient diagnosed with ventricular tachycardia will develop heart failure (HF) and ventricular fibrillation (VF). In this scenario, we used the CHA_2DS_2-VASc score to predict the risk of stroke, TIA and Systemic Embolism for the VF scenario. In addition to a combination of the CHA_2DS_2-VASc with the score generated from the patient's habits, to be able to categorize the patient within NYHA classes, which in turn generates a morbidity and mortality level for each category. Computational techniques such as the use of SVM,KNN, Neural Networks, Random Trees and Naïve Bayes were used to validate the system. The model achieved its best results by validation with Support Vector Machines (SVM), with an accuracy of 0.997, F1, precision and recall of 0.976. From these classifications and predictions, it is possible to provide the cardiologist with a higher level of information, optimizing decision making, impacting on the adopted treatment, interventions, and patient follow-up.

Keywords: Health Informatics, Tachycardia, Prediction, Morbidity, Mortality.

Lista de Figuras

Lista de Tabelas

Sumário

1

Introdução

A tecnologia vem ganhando cada vez mais espaço na área da saúde, sendo responsável por importantes avanços na medicina, seus procedimentos e técnicas fornecem melhores resultados na identificação e prevenção de doenças e no desenvolvimento de tratamentos em casos mais complexos, se tornando ferramenta essencial para esta área. Dentre os responsáveis por isto, encontram-se grandes áreas da computação, como a Ciência dos Dados, a Inteligência Artificial (IA) e suas subáreas, como a própria Machile Learning (ML). Estas que se mantêm em constante processo de evolução entregando desempenhos cada vez mais satisfatórios em suas aplicações. Isto deve-se ao fato de suas características convergirem com o cenário que se desenvolve, não só na saúde, mas em todo ambiente que gera uma grande quantidade de dados, que devem ser processados e analisados para se obter informações suficientes para as melhores tomadas de decisões, trazendo melhores resultados em seus contextos.

1.1 Problemática

Segundo a Organização Mundial da Saúde (OMS), as doenças cardiovasculares são a principal causa de morte no mundo com cerca de 17,9 milhões de óbitos por ano, representando 32% de todas as mortes no mundo, como podemos observar na Figura 1.1 (Ritchie (2019),BBC). Mais de 80% destas mortes são em decorrência de ataques cardíacos e derrames, e $\frac{1}{3}$ delas ocorre prematuramente em indivíduos com menos de 70 anos de idade. As doenças cardiovasculares são um grupo de distúrbios do coração e dos vasos sanguíneos que incluem doenças coronárias, doenças cerebrovasculares, doenças cardíacas reumáticas, arritmias, dentre outras. O exame mais utilizado para o diagnóstico deste tipo de doença é o Eletrocardiograma (ECG), um exame básico e inicial para avaliar a saúde cardiovascular e apontar as anormalidades cardíacas. O ECG tem como finalidade avaliar a atividade elétrica do coração por meio de eletrodos fixados na pele, sendo possível detectar o número de batimentos por minuto e ritmo do coração

em tempo real. Este pode ser realizado em todas as fases da vida, fetal (realizado quando o bebê está no útero), pediátrica (a partir do período recém-nascido até a adolescência) e adulta.

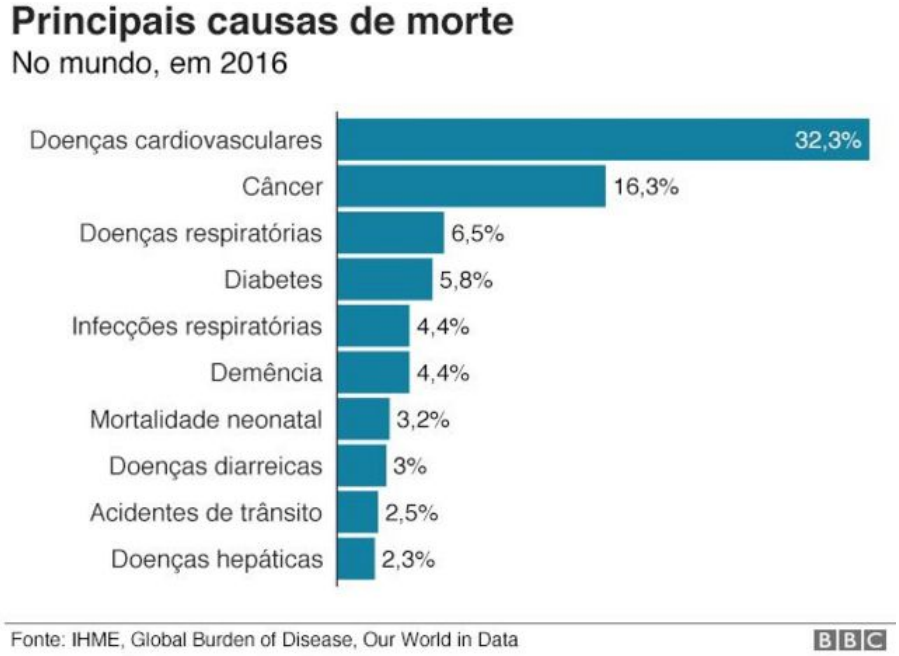


Figura 1.1: Maiores causas de morte no mundo
BBC

Uma vez que os sinais e sintomas avaliados forneçam indícios de alguma doença cardiovascular após realização do ECG, faz-se necessário a realizaçao de exames mais completos. Podemos citar como alguns destes:

- Holter: Tipo de ECG coletado pelo período de 24, 48 ou 72 horas, com o auxílio de um equipamento e dos registros das atividades feitas pelo paciente durante este intervalo;
- Teste ergométrico: Avaliação ampla do funcionamento cardiovascular, quando submetido a esforço físico gradualmente crescente, em esteira rolante;
- Ecocardiograma: Diagnóstico da estrutura e do funcionamento do coração baseados no uso de ultrassom, ou seja, as ondas acústicas com frequência de mais de 20 mil Hz, geralmente em torno de 2 a 4 MegaHertz;
- Monitorização ambulatorial de pressão arterial - MAPA: Através do uso de uma braçadeira ligado a um aparelho eletrônico durante 24h, registra a pressão arterial ao longo do dia.
- Cintilografia do Miocárdio: Feito de forma não invasiva através de modernas câmaras gama, coleta imagens dos órgãos sem necessidade de grandes quantidades de radiação.
- Cateterismo: Técnica mais invasiva que pode ser realizado de forma eletiva, para confirmar a presença de obstruções das artérias coronárias ou avaliar o funcionamento das valvas e do músculo cardíaco.

- Outras informações ou exames relevantes: Histórico familiar, hemogramas, hábitos de vida e alimentação saudável.

A análise da maioria dos exames supracitados tem alto grau de complexidade, o que por sua vez traz a necessidade de mais ferramentas tecnológicas para aumentar a celeridade e acurácia neste processo de análise e diagnóstico das doenças cardiovasculares. O mesmo ocorre com as arritmias, uma vez que a análise do eletrocardiograma deve ser minuciosa, identificando em cada ponto e segmentos os padrões que apontem que o paciente possa ter bradicardia, taquicardia ou fibrilação atrial. A tecnologia já tem amplo campo de estudo no que diz respeito à diagnósticos, não só das doenças cardiovasculares, mas de muitas outras frentes da medicina. Entretanto, o volume de estudos no campo do prognóstico é recente, sendo este menor quando comparado ao diagnóstico. Em nosso contexto, o das arritmias, o prognóstico tem grande impacto em como o paciente irá conviver com a doença diagnosticada.

O Prognóstico em nosso contexto, taquicardia ventricular, é de suma importância para que os pacientes avaliados, sejam eles diagnosticados ou não, recebam os tratamentos ideais a partir de uma tomada de decisão médica auxiliada pelas informações fornecidas pelos sistemas computacionais. Provendo desta maneira uma melhor qualidade de vida e longevidade por parte dos pacientes, em nosso contexto, aqueles que sejam previamente diagnosticados com taquicardia ventricular.

1.2 Proposta

Este trabalho tem como objetivo propor e avaliar um modelo para o prognóstico de pacientes diagnosticados com taquicardia. Prognóstico este que visa estimar a morbidade (como os sintomas se apresentam no paciente, em outras palavras, as limitações ocasionadas pela taquicardia) e a mortalidade destes pacientes, afim de prover ao cardiologista o maior volume de informações possível para auxiliar as tomadas de decisão acerca de que medidas adotar sobre aquele paciente. As medidas incluem, por exemplo, quais medicamentos serão receitados, em quais dosagens, o nível de rigidez com os hábitos, dentre outras medidas. Outro objetivo é a partir das informações adquiridas, expor ao paciente a importância de seguir o tratamento corretamente, em contraste com os impactos decorrentes de uma negligência com aquele tipo de enfermidade. Trazendo aos cardiologistas uma ferramenta que possa auxiliar no contexto de tomadas de decisão dado o prognóstico obrido.

Para predição dos valores de morbidade e mortalidade o sistema leva em consideração o score CHA_2DS_2-VASc(Joundi et al. (2016)), que se baseia em uma série de fatores relacionados a fibrilação, que é uma das consequências da taquicardia maligna sustentada (Taquicardia Ventricular (TV), por exemplo), que será mais detalhado na Seção 3, bem como um maior detalhamento de como a taquicardia evolui para uma fibrilação. Além das variáveis cobertas pelo CHA_2DS_2-VASc, foram considerados outros fatores de risco (Feitosa et al. (2002)) associados

os hábitos do paciente e outros indicadores. Por fim, com o score gerado é feita uma classificação com a técnica New York Heart Association (NYHA)(Bredy et al. (2018), Bennett et al. (2002)), onde cada uma das 4 categorias indica como os sintomas se apresentam, e a taxa de mortalidade associada a ela.

1.3 Objetivo Geral

O objetivo geral deste trabalho é propor e avaliar um modelo de predição da morbidade e mortalidade em pacientes diagnosticados com taquicardia. Sendo capaz de determinar em que nível os sintomas da taquicardia iram atingir o paciente, e estimar a taxa de mortalidade, levando em consideração scores definidos na literatura, como o caso do CHA_2DS_2-VASc(Joundi et al. (2016)), fatores de risco associados aos hábitos (Feitosa et al. (2002)), e o classificador NYHA (Bredy et al. (2018), Bennett et al. (2002)).

1.4 Objetivos Específicos

Com o objetivo de alcançar o objetivo geral supracitado, faz-se necessário distribuir os objetivos específicos que irão sustentar o modelo proposto.

- Revisão sistemática da literatura para identificar quais são as técnicas computacionais, doenças cardiovasculares, e tipos de dados mais utilizadas e consequentemente mais relevantes no contexto das arritmias, com foco nas taquicardias;
- Identificar das variáveis mais relevantes para predição da Morbidade e Mortalidade em arritmias;
- Identificar uma base de dados a mais próxima possível do ideal dentro do nosso contexto, isto é, que possua as informações que alimentem os scores adotados e demais variáveis consideradas;
- Aplicar os conceitos da mineração de dados sobre a base (seleção, pré-processamento, transformação, mineração dos dados, interpretação das informações obtidas)
- A partir das informações adquiridas, ser capaz de predizer a morbidade do paciente avaliado.
- A partir da morbidade do paciente, ser capaz de predizer a mortalidade do mesmo.
- Avaliar do modelo proposto;

1.5 Estrutura do Trabalho

Este trabalho está estruturado da seguinte maneira: A primeira e atual Seção 1, trouxe uma visão geral do uso da tecnologia na área médica, bem como elencou os objetivos gerais e específicos do modelo proposto, por fim trouxe um breve resumo da metodologia adotada. Na Seção 2 é exposto os resultados obtidos pela revisão sistemática feita anteriormente. Revisão esta com foco na compreensão do uso da aprendizagem de máquina no contexto das doenças cardiovasculares. Por fim, o capítulo encerra trazendo artigos relacionados as especificidades do estudo, o que inclui a predição da morbidade e mortalidade de arritmias, baseado em scores da literatura, como o CHA_2DS_2-VASc (Joundi et al. (2016)), e classificações como o NYHA (Bredy et al. (2018)). Na Seção 3 trazemos todos os conceitos necessários para a compreensão do modelo proposto, incluindo cardiologia, arritmias, taquicardias, fatores de risco, score CHA_2DS_2-VASc, classificação NYHA, conceitos de ciência de dados, algoritmos para seleção de características, aprendizagem supervisionada e seus tipos (supervisionada, não supervisionada e por reforço), regressão, KNN, Redes Neurais e Árvore de Decisão. Na Seção 4 trazemos as características do modelo proposto. Na Seção 5 abordamos sobre as estratégias adotadas e ferramentas utilizadas. Na Seção 6 trazemos o conjunto dedados obtido, bem como os resultados expostos com base na estatística. Na Seção 7 é feito uma avaliação sobre os resultados obtidos, ameaças a validade do modelo proposto, resumo, benefícios obtidos, limitações do trabalho, além de trabalhos futuros. Por fim, encerramos com a conclusão e considerações finais que estão presentes Na Seção 8.

2

Trabalhos Relacionados

Nesta seção iremos descrever a revisão sistemática feita sobre o uso da aprendizagem de máquina no diagnóstico e prognóstico das doenças cardiovasculares, e como isso permitiu identificarmos os metadados associados a esta área. Por fim, iremos citar alguns artigos que utilizam a abordagem de uso de scores da literatura para a predição de morbidade e mortalidade em pacientes diagnosticados com arritmias.

2.1 Protocolo de Revisão Sistemática

No processo de busca pelos trabalhos relacionados às arritmias, foi possível notar um grande volume de estudos presentes na literatura, principalmente devido à sua grande área, a cardiologia. Neste cenário, foi realizada uma revisão sistemática da literatura, com a finalidade de identificar trabalhos que fossem específicos ao contexto das arritmias, precisamente as taquicardias. Define-se revisão sistemática como a revisão da literatura a partir de uma pergunta de pesquisa definida, através da qual se busca identificar, avaliar, selecionar e sintetizar evidências de estudos empíricos que atendam os critérios de elegibilidade definidos. Uma vez que a área convergiu para o prognóstico, foi realizada uma pesquisa mais específica, para os artigos que adotam os scores e classificadores definidos nesta proposta.

2.1.1 Aprendizagem de Máquina para o Diagnóstico e Prognóstico na Cardiologia

A metodologia adotada nesta revisão sistemática assume as convenções recomendadas pelo PRISMA (Galvao et al. (2015)), as técnicas, e protocolos fundamentados por Kitchenham (Kitchenham and Charters (2007)). O seguinte conjunto de ferramentas foi adotado: (i) gerenciadores de citações de artigos Mendeley; (ii) Zotero, responsável por transformar arquivos do

Tabela 2.1: Questões de Pesquisa - Revisão Sistemática

Questão Principal
QP: Como a aprendizagem de máquina vem sendo aplicada no auxílio ao diagnóstico de doenças cardiovasculares?
Questões Secundárias
QS1: Quais as doenças cardiológicas mais abordadas com o uso de aprendizagem de máquina?
QS2: Quais algoritmos ou técnicas computacionais mais aplicados no uso de aprendizagem
QS3: Quais são os tipos de dados mais utilizados no uso de aprendizagem de máquina para diagnóstico na cardiologia?
QS4: Qual a acurácia alcançada pelos sistemas que aplicam aprendizagem de máquina nos diagnósticos da cardiologia?

formato .csv em arquivos .bib; (iii) Mendeley, sendo a ferramenta principal de gerenciamento dos artigos tratados; e (iv) Google, utilizado para organizar os dados específicos que eram buscados nos artigos, bem como registrar as informações relacionadas as filtragens aplicadas.

O protocolo adotado consiste em cinco etapas: (i) definição das questões primárias e secundárias de pesquisa; (ii) definição das palavras-chave; (iii) seleção das fontes de pesquisa; (iv) definição dos critérios de inclusão e exclusão que serão adotados nas filtragens dos artigos; (v) estratégia adotada na pesquisa dos artigos. O objetivo final deste protocolo de filtragem é encontrar os artigos mais relevantes da área de conhecimento considerada.

Uma das etapas do mapeamento sistemático consiste em definir o norte da sua pesquisa baseado em uma série de questionamentos. Define-se a questão principal responsável pelo foco do estudo, seguida das questões secundárias, compostas por perguntas mais específicas acerca do objeto tratado na questão principal. Dentro do nosso contexto, elaboram-se as questões visando atacar a área de doenças cardiológicas sendo diagnosticadas a partir de técnicas de aprendizagem de máquina, como pode ser observado na Tabela 2.1.

O foco desta revisão foi identificar quais as doenças cardiológicas são mais abordadas, quais as técnicas ou algoritmos mais utilizados e que tipos de bases de dados são mais analisados no diagnóstico de doenças cardiovasculares. Além disto, busca-se compreender o desempenho do uso dessas tecnologias, principalmente em relação às soluções propostas anteriormente, além de quanto cada estudo tem sido relevante para a cardiologia. Dentro deste contexto, é possível definir as palavras chaves: (i) Cardiologia, (ii) Aprendizado de Máquina, (iii) Diagnóstico e (iv) Prognóstico.

Foram adotados três critérios de elegibilidade durante a investigação dos artigos: (i) critérios de inclusão; (ii) critérios de exclusão e (iii) critérios de qualidade. Os critérios de inclusão consideram aptos os artigos primários publicados na língua inglesa em periódicos nos últimos cinco anos (2017 - 2021), de forma eletrônica ou impressa. Assim consegue-se agregar a esta pesquisa, características como atualidade dos conceitos identificados, além da inovação e impacto

que envolvem um artigo primário

Já como critérios de exclusão, foram descartados artigos secundários como surveys e revisões sistemáticas, artigos terciários, e artigos feitos em outras línguas que não a inglesa. De forma mais específica, excluem-se artigos duplicados e artigos que não estão relacionados a diagnósticos, como é o caso dos prognósticos, por exemplo, que também possuem grande representação neste meio. Tais características podem ser identificadas no primeiro momento através de uma análise de título, posteriormente na análise do abstract, seguida por uma análise mais a fundo, na leitura do artigo em si, onde foi adotada a técnica PICOS (Amir-Behghadami and Janati (2020)).

Os critérios de qualidade seguiram a pontuação estratégica indicada pelo PICOS, estratégia esta que consiste em analisar 5 critérios presentes no artigo: população/paciente (P), intervenção/indicação/interesse (I), comparação (C), *Outcome*/desfecho/resultado (O), *Study/design* do mesmo (S). Cada um dos cinco critérios são avaliados no artigo em uma escala de 0 - 1, onde cada um deles podem ser pontuados em 0 quando não atendidos, 0,5 quando parcialmente atendidos, e 1 quando atendidos completamente, de forma que o PICOS atribui uma pontuação de 0 a 5 para cada artigo. Como pontuação de corte para este estudo, foi adotada a avaliação em 3, ou seja, aqueles artigos que não atingiram nota igual ou superior a 3 foram excluídos

As 5 fontes de pesquisa selecionadas foram: (i) ACM; (ii) ScienceDirect (pertence a Elsevier); (iii) PubMed; (iv) Xplore; (v) Springer. Algumas destas fontes são mais gerais em sua área de conhecimento, já outras possuem focos em determinadas áreas, como no caso da ACM que é voltada para a computação, PubMED para medicina, e IEEE para engenharia. O protocolo foi executado no primeiro trimestre de 2021.

As buscas desta revisão sistemática foram feitas apenas utilizando bases eletrônicas em motores de busca da World Wide Web. A estruturação da string de busca considerou as palavras chaves da pesquisa, ficando da seguinte forma:

(("Cardiology") OR ("Heart diseases")) AND (("Machine learning") OR ("Artificial Intelligence")) AND (("Diagnosis") OR ("Prognosis"))

Ainda dentro de algumas fontes era possível inserir filtros, como publicações dos últimos 5 anos, publicações em periódicos, aprendizagem de máquina e inteligência artificial. Porém, como estes filtros não estão disponíveis em todas as 5 fontes, todos os artigos foram verificados posteriormente e filtrados no Mendeley de acordo com os critérios de exclusão adotados.

A seleção dos dados consistiu de 7 etapas. Na string de busca, a soma dos artigos localizados pelos motores de busca das 5 fontes totalizou 530 artigos. A primeira etapa consistiu na remoção das duplicatas. A segunda etapa consiste na filtragem a partir do título, onde foram categorizados como diagnóstico, prognóstico ou outros. Os títulos de artigos classificados como prognóstico foram selecionados para uma análise mais detalhada, enquanto que os demais títulos foram descartados. A quarta etapa consistiu na remoção de alguns artigos de conferência, dando destaque então naqueles que foram publicados em periódicos. A quinta etapa efetua uma filtragem a partir da leitura dos abstracts, onde por vezes, o título indicava classificação/di-

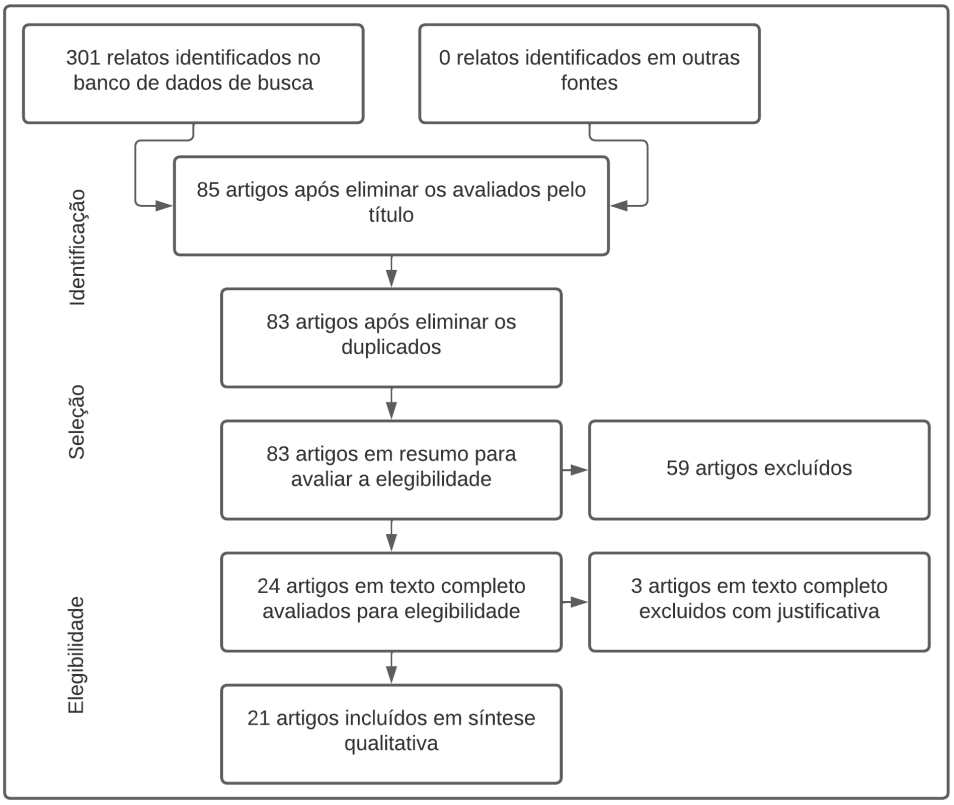


Figura 2.1: Revisão Sistemática

agnóstico, mas o estudo do artigo destinava-se a predição/prognóstico. A sexta etapa consiste na avaliação pelo critério PICOS, onde os avaliados abaixo de 3 foram eliminados. A sétima etapa consiste na leitura do artigo em si, para categorizar estes de acordo com as doenças abordadas, técnicas ou algoritmos utilizados e tipos de dados analisados.

Todos os artigos desta pesquisa foram analisados e organizados pela ferramenta Mendeley em conjunto com as planilhas disponibilizadas pela ferramenta Google Sheets, onde buscou-se observar quais doenças, técnicas de aprendizagem de máquina e bases de dados eram mais utilizados nas abordagens das doenças cardíacas. Tais dados foram organizados em gráficos no intuito de identificar o que vem sendo mais abordado nesta área.

Na primeira etapa, as strings de buscas retornaram um total de 530 artigos advindos das 5 fontes selecionadas. O primeiro passo filtrou artigos duplicados através de uma ferramenta do Mendeley, onde o total de artigos resultantes foi de 519. Os 519 artigos foram analisados pelo seu título, onde apenas 398 estavam de fato relacionados à cardiologia com uso de aprendizagem de máquina. Destes, apenas 71 estavam relacionados ao diagnóstico, e a filtragem indicou que 16 deles eram *conference* paper ou short paper, e prontamente removidos. Restaram então 55 artigos que foram avaliados pelo conteúdo presente e seus abstracts. Destes, restaram apenas 25 que atendiam ao critério. Estes 25 artigos foram avaliados pelo critério PICOS (corte para avaliação < 3), restando apenas 20 artigos que podem ser vistos na Tabela 2.1.

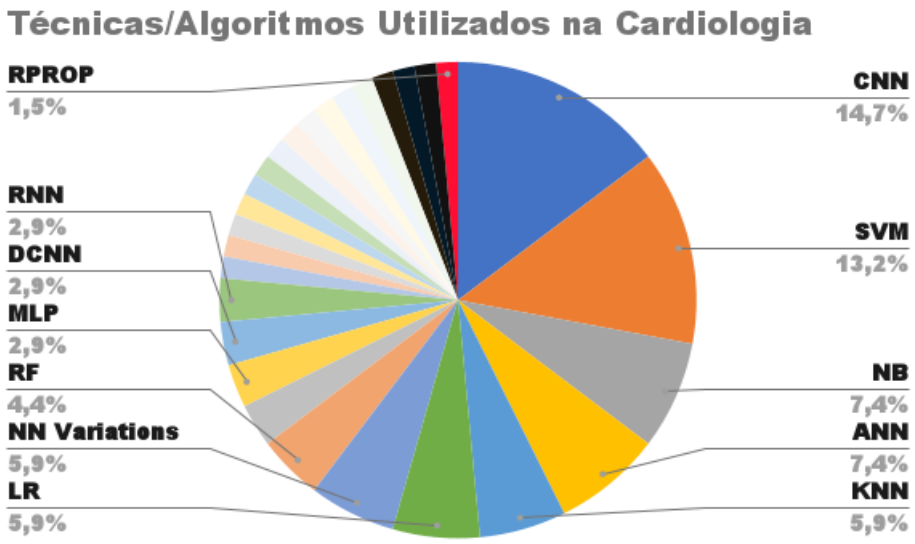


Figura 2.2: Técnicas mais Recorrentes no Diagnóstico e Prognóstico da Cardiologia

Dos 519 artigos, após a aplicação dos critérios de exclusão e qualidade, chegou-se aos 20 artigos apresentados na tabela de metadados, distribuídos da seguinte forma de acordo com cada uma das fontes: Dentro da **ACM**, apenas 1 artigo foi selecionado, representando 5% dos 20 selecionados e 1,25% dos 80 extraídos da mesma. Na **IEEE** 8 artigos foram selecionados, sendo o maior representante com 40% dos selecionados, e 12,7% dos extraídos na mesma. Na **PubMed** 5 artigos foram selecionados representando 25% dos selecionados e 4% dos 208 extraídos do mesmo. Na **ScienceDirect** 2 artigos foram selecionados, representando 10% dos selecionados e 2% dos 46 extraídos. Por fim, a **Springer** teve 4 artigos, 20% dos selecionados e 3% dos artigos exportados do mesmo.

A partir da análise dos artigos selecionados, ficou evidente que Redes Neurais possuem um amplo uso no contexto doenças cardiovasculares. Logo em seguida estão as SVM. Outras técnicas também estão presentes, como Naïve Bayes(NB), Redes Neurais (RN), KNN, Regressão Logística, Randon Forest (RF), Multicamada Perceptron, e outras variações de RN, como podemos observar na Figura 2.2. Quanto as doenças mais presentes neste conjunto dos 20 artigos, temos a Doença Arterial Coronariana e a Classificação dos Sons de Batimentos Cardíacos como as principais, seguidas pelas Arritmia e Insuficiência Cardíaca, como podemos observar na Figura 2.3. Vale salientar, que estas estão correlacionadas.

É possível observar na Figura 2.4 o uso disseminado da base de dados da MIT, a MIT-BIH Arrhythmia Database (Moody and Mark (1992)), uma base de dados de acesso livre, que ainda possui outras duas bases relacionadas, MIT-BIH Noise Stress Test Database, e MIT-BIH P-wave Annotations. Ambos utilizam dados gerados por exame de eletrocardiograma (ECG), evidenciando que este é de fato o principal exame no contexto das doenças cardiovasculares.

Na Tabela 2.2, é possível observar em uma amostragem dos artigos extraídos da revisão sistemática, como os estudos atuais performam em relação à classificação (diagnóstico) das

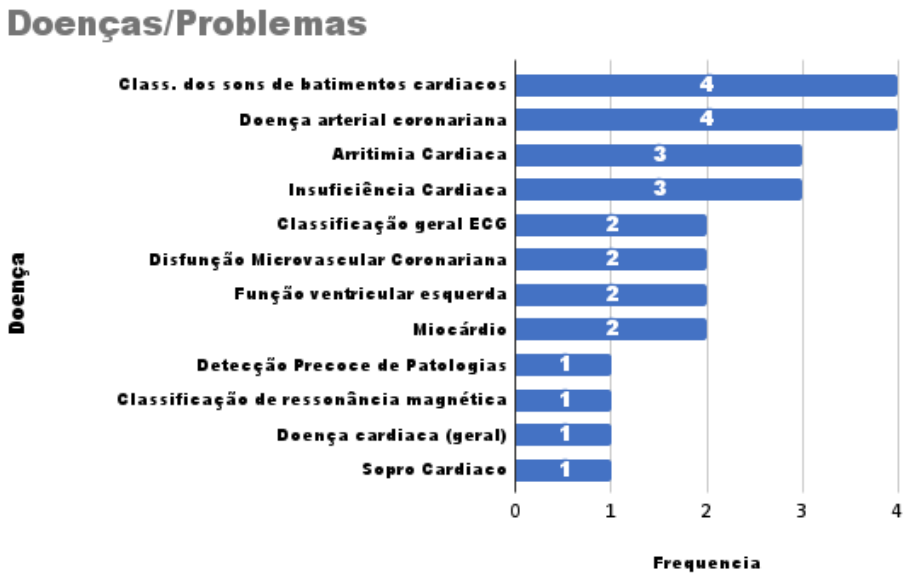


Figura 2.3: Doenças mais Recorrentes no Diagnóstico e Prognóstico da Cardiologia

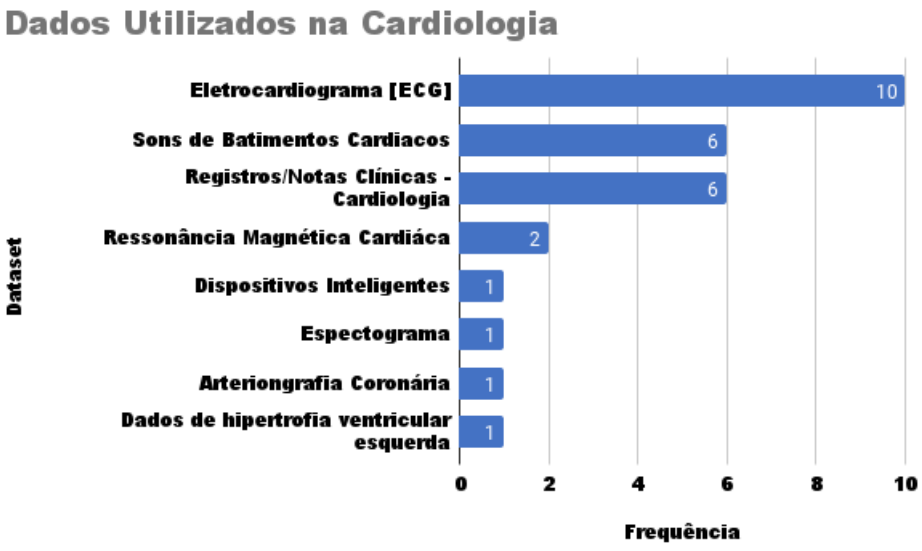


Figura 2.4: Dados mais Recorrentes no Diagnóstico e Prognóstico da Cardiologia

ARTIGO	TÉCNICAS / ALG.	ACC (%)	SEN (%)	SPE (%)	CLASS	DATASET
i. Arrhythmia Detection using deep convolutional neural network with long duration ECG signals	Deep Convolutional Neural Networks (DCNN), Convolutional Neural Networks (CNN)	86,67	83,91	99	17	MIT-BIH (ECG)
ii. Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats	Convolutional Neural Network (CNN), Long Short Term Memory (LSTM)	98,10	97,50	98,70	4	MIT-BIH (ECG)
iii. ADDHard: Arrhythmia Detection with Digital Hardware by Learning ECG Signal	Pre processing, Feature Extraction, Update, Simulation Matlab	97,28	78,70	98,25	2	Simulation Matlab
iv. Mixed convolutional and long short-term memory network for the detection of lethal ventricular arrhythmia	Convolutional Neural Network (CNN), Log Short Term Memory (LSTM)	99,1	99,70	98,9	1	MIT-BIH (ECG)
v. Arrhythmia Classification of ECG Signals Using Hybrid Features	Feature Extraction, Discrete Wavelet Transform (DWT), Neural Network (NN)	99,75	98,70	99,9	18	MIT-BIH (ECG)

Tabela 2.2: Desempenho dos estudos atuais no Diagnóstico das Arritmias

arritmias. Performace está exposta em termos da acurácia (ACC), sensibilidade (SEN) e especificidade (SPE). Além disto, são determinados quais técnicas e algoritmos (TÉCNICAS / ALG) foram utilizados, quantos tipos de arritmias (CLASS) estão sendo classificadas em cada artigo, e por fim o tipo de dado (DATASET) utilizado. As informações expostas na Tabela 2.2 (i. Özal Yıldırım et al. (2018), ii. Oh et al. (2018), iii. Pudukotai Dinakarrao and Jantsch (2018), iv. Picon et al. (2019), v. Anwar et al. (2018)), que trazem informações que vão de encontro com o que foi dito anteriormente, em relação aos algoritmos mais utilizados, com destaque para as Redes Neurais e suas variações. E em relaçao aos dados utilizados, os ECGs, em específico uma base antiga da MIT, a MIT-BIH (Moody and Mark (1992)).

2.2 Estudos Relacionados

Como dito anteriormente, a pesquisa evoluiu ao longo dos últimos meses e convergiu para o prognóstico de morbidade e mortalidade de taquicardias. Entretanto, seu início teve foco na compreensão da aplicação da Aprendizagem de Máquina no cenário das doenças cardiovasculares, como visto no tópico anterior. Uma vez identificado as principais características deste contexto, torna-se necessário especificar o objeto da pesquisa dentro deste espectro da cardiologia, visto que se trata de uma gigantesca variedade de doenças. Foi definido para este estudo, que a doença cardiovascular abordada seriam as arritmias, ou seja, as alterações nos batimentos que implicam no funcionamento anormal do sistema cardiovascular, podendo levar à morte. Seus principais diagnósticos quanto a alteração do batimento são a taquicardia, bradicardia ou fibrilação, ou ainda quanto ao foco, podem ser ventriculares, atriais ou sinusal.

A literatura apresenta uma série de artigos relacionados ao objetivo do modelo aqui proposto. As linhas dos artigos envolvem o uso do score CHA_2DS_2-VASc, os fatores de risco associados as arritmias em geral, bem como o uso do NYHA para classificar os pacientes com Insuficiência Cardíaca (IC).

Em relação ao CHA_2DS_2-VASc, temos desde sua revisão sistemática (Joundi et al. (2016)), até artigos mais específicos, como predizer a mortalidade em pacientes diagnosticados com a COVID-19 (Cetinkal et al. (2020)), avaliar o risco da ocorrência de um Ataque Vascular Cerebral (AVC) (Lip) ou AVC Isquemico (Siddiqi et al. (2022)); Boriani et al. (2011)) em pacientes com fibrilação. Enquanto que outros artigos já são mais pautados em detalhar os impactos dos valores resultantes (Friberg et al. (2012b); Camm (2012)) do score CHA_2DS_2-VASc. De todo modo, este score será mais detalhado na Seção 3.

Em se tratando de arritmias, foi necessário identificar artigos que abordassem esta doença de uma forma geral, incluindo seus fatores de risco (Feitosa et al. (2002)), etapas até se chegar ao diagnóstico, tipos de exames, dentre outras especificidades que também iremos detalhar na Seção 3.

Por fim, artigos relacionados ao classificador NYHA, desde sua origem (Bredy et al. (2018)), como aprimoramentos em seu questionário (Kubo et al. (2004)), e suas validações em outros contextos, como mensurar reabilitação em pacientes portadores de doenças cardíacas (Bennett et al. (2002)).

A associação das informações extraídas na revisão sistemática em conjunto com as especificidades identificadas neste conjunto de artigos relacionados aos scores, fatores de risco, classificadores e características referentes às arritmias, foram a base para fundamentar o modelo proposto. Na seção seguinte iremos falar sobre estes e outros conceitos que fomentam o conhecimento necessário para o modelo de predição proposto.

3

Background

Nesta seção será apresentada toda a fundamentação teorica, contendo todos os conceitos necessários para compreensão e execução do modelo proposto. Desde a grande área médica que estamos abordando, a cardiologia, o subconjunto explorado, das arrimitas, e a doença específica, a taquicardia. Todavia, os demais tipos de arritmias serão detalhados, casos da bradicardia e fibrilação, visto que existe forte relação entre elas, principalmente entre a fibrilação e as demais, que, em geral é uma consequência das outras. Iremos detalhar a relação da taquicardia ventricular com a insuficiência cardíaca e a fibrilação ventricular. Será detalhado o escore que foi utilizado para identificar os riscos no paciente avaliado, o escore CHA_2DS_2-VASc. Os fatores de risco associados a taquicardia, descritos na literatura médica. O classificador NYHA na predição de morbidade e mortalidade em pacientes com Insuficiência Cardíaca. Conceitos de Ciência dos Dados, aplicados em nosso estudo, como compreensão de um cenário que lida com Big Data, e as etapas da Mineração dos Dados, com o objetivo de se obter as informações desejadas. Conceitos de Aprendizagem de Máquina, bem como seus tipos de aprendizado: supervisionado, não supervisionado e por reforço. Por fim, serão apresentados os conceitos dos algoritmos utilizados no modelo proposto, que incluem Floresta Aleatória, Naïve Bayes), Máquinas de Vetores Suporte, Redes Neurais e K-Vizinhos mais próximos (KNN) .

3.1 Cardiologia

A cardiologia pode ser definida como ramo da medicina responsável por estudar, cuidar e tratar o coração e vasos sanguíneos. Sendo uma das especialidades da medicina que cuida do coração e do sistema circulatório (Kadi et al. (2017)). Cardiologistas podem atuar em diversas áreas, sendo as mais comuns: i. Interpretação dos exames como eletrocardiograma (ECG), monitorização ambulatorial da pressão, holter, ecocardiograma, dentre outros. ii. Análise clínica de casos cardiológicos, buscando tratar da maneira menos invasiva possível em casos de problemas simples, bem como na prevenção de problemas futuros, utilizando de fármacos para casos

de hipertensão arterial e hipotensão arterial, assim como algumas anormalidades de desempenho cardíaco. iii. Microcirurgia cardiológica, i.e., procedimentos ambulatoriais de pequena complexidade como cauterização de pequenos vasos. iv. Cirurgião cardiológico, atuando em intervenções invasivas a partir de diagnósticos anteriores de especialistas.

A cardiologia (Kadi et al., 2017) se divide em duas grandes áreas, como toda a medicina, a preventiva e a curativa. A cardiologia preventiva foca a atenção na prevenção de doenças cardíacas ou do sistema circulatório, e sua importância é justamente pelo fato das doenças cardiovasculares se tornarem problemas maiores e mais graves no futuro. Por outro lado, a cardiologia curativa é voltada para a cura e tratamento das doenças cardíacas, sejam elas desenvolvidas com o tempo, ou congênitas. Esta fundamenta-se nos estudos dos sintomas das doenças para diagnosticá-las. A cardiologia curativa ocorre sempre que o paciente apresenta sintomas relacionados à doenças cardiovasculares. Para a partir disto, diagnosticar e tratar o problema identificado. Nesta área, dentre os exames mais comuns para acompanhamento dos pacientes, está o ECG, este é utilizado para registrar e monitorar a atividade elétrica do coração. Estes dados coletados são utilizados para controlar a regularidade dos batimentos do coração, provendo aos cardiologistas a possibilidade de identificar doenças cardiovasculares como a arritmia, isquemia, infarto do miocárdio, dentre outras. É justamente nesta etapa que a Inteligência Artificial e a Aprendizagem de Máquina, entram para auxiliar os médicos cardiologistas e tornar mais efetivos os procedimentos de diagnósticas e prognóstico deste grupo de enfermidades.

O coração é um órgão localizado na parte inferior do mediastino médio. Encontra-se na cavidade torácica, mais precisamente na região posterior ao osso esterno e acima do músculo diafragma. Esse órgão tem o tamanho aproximado de uma mão fechada e apresenta maior parte da sua massa à esquerda da linha mediana. O coração possui um formato de cone invertido, com seu ápice voltado para baixo, e seu peso é de cerca de 300 g. O coração é um órgão formado, principalmente, por tecido muscular estriado cardíaco, o qual se caracteriza por apresentar contração involuntária. Esse órgão realiza contrações e relaxamentos que seguem um ritmo cíclico. Sua contração garante o bombeamento de sangue, enquanto seu relaxamento permite que as câmaras encham-se dele. A fase de contração recebe a denominação de sístole, enquanto a fase de relaxamento é denominada diástole.

O batimento cardíaco é conseguido graças a uma massa de células especializadas, encontradas nesse órgão, chamadas de nó sinoatrial. Essa massa de células está na parede do átrio direito e é responsável por gerar impulso elétrico. Esse impulso propaga-se pelo tecido muscular até chegar ao nó atrioventricular, o qual está localizado na parede, entre os átrios esquerdo e direito. Esse nó constitui um ponto de transmissão de impulsos nervosos, os quais partem dessa região e seguem em direção aos ventrículos e ao ápice do coração através de estruturas chamadas de ramos do feixe e fibras de Purkinje.

O coração apresenta paredes constituídas por três camadas (Figura 3.1):

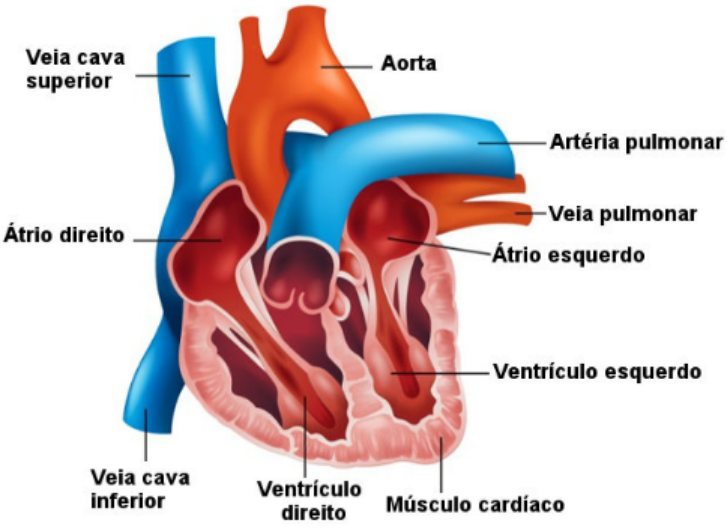


Figura 3.1: Camadas do Coração
SANTOS (2020)

- **Endocárdio**: é a camada mais interna e é formado por endotélio que está sobre uma camada subendotelial delgada de tecido conjuntivo. Essa última camada conecta-se ao miocárdio por uma camada de tecido conjuntivo que apresenta nervos, veias e alguns ramos do sistema responsável pela condução do impulso do coração. O endocárdio reveste as cavidades do coração, as válvulas e os músculos associados com as válvulas.

- **Miocárdio**: é a camada média do coração e a mais espessa. O miocárdio é rico em células musculares cardíacas, sendo a camada responsável pela capacidade de contração desse órgão. O miocárdio do ventrículo esquerdo é mais espesso que o do ventrículo direito. Isso se deve ao fato de que a contração nessa região deve ser mais vigorosa, de modo a garantir que o sangue siga para o corpo.

- **Pericárdio**: é uma espécie de saco invaginado constituído de uma camada mais externa, chamada de pericárdio parietal, e de uma camada mais interna, chamada de pericárdio visceral. É este último que adere ao coração e forma a camada mais externa do órgão (epicárdio).

O coração humano é formado por quatro cavidades, assim como o de todos os mamíferos. Nesse órgão é possível observar dois átrios e dois ventrículos (Figura 3.2).

- **Átrios**: apresentam paredes relativamente delgadas e funcionam como câmaras receptoras de sangue. O átrio esquerdo recebe o sangue proveniente do pulmão, enquanto o direito recebe o que vem de outras partes do corpo, com exceção do pulmão. Dessas cavidades o sangue segue para os ventrículos.

- **Ventrículos**: apresentam paredes mais grossas — uma característica importante, uma vez que sua contração garante que o sangue siga para os pulmões e outras partes do

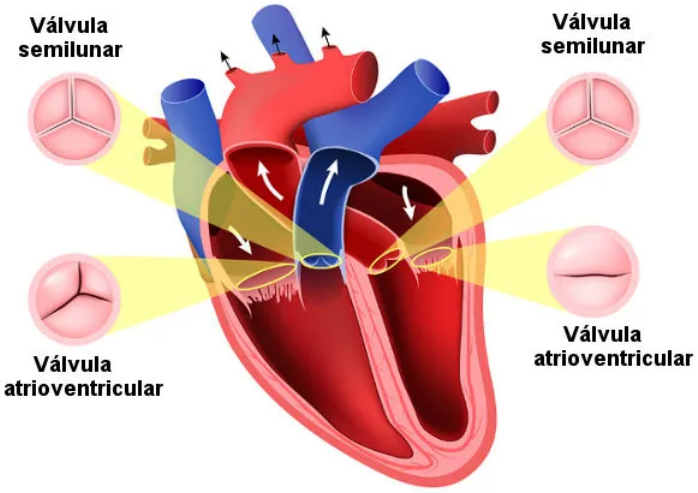


Figura 3.2: Válvulas do Coração
SANTOS (2020)

corpo. A contração do ventrículo direito impulsiona-o em direção aos pulmões, enquanto a do esquerdo garante o seu impulso ao restante do corpo.

A presença de válvulas no coração faz com que o refluxo do sangue não ocorra e ele siga sempre uma mesma direção. No total existem quatro válvulas no coração: duas válvulas atrioventriculares e duas válvulas semilunares. As primeiras situam-se entre cada átrio e cada ventrículo; já as segundas estão localizadas nos dois locais por onde o sangue sai do coração. Uma válvula semilunar está localizada na região onde a aorta sai do ventrículo esquerdo, e a outra, no local onde a artéria pulmonar sai do ventrículo direito (Figura 3.2).

O coração funciona como uma espécie de bomba que garante o impulsionamento do sangue para todas as partes do corpo. O sangue, proveniente das várias partes do corpo, com exceção do pulmão, chega ao coração pelo átrio direito. As veias cavas superior e inferior trazem sangue do corpo e desembocam-no nessa cavidade, assim como o seio coronário, o qual é responsável por drenar o sangue presente no coração. Após chegar ao átrio direito, o sangue segue em direção ao ventrículo direito. Do ventrículo, ele parte em direção aos pulmões, sendo levado pela artéria pulmonar. Nos pulmões, o sangue, que até então estava rico em gás carbônico, recebe o oxigênio proveniente da respiração, tornando-se oxigenado. O sangue oxigenado retorna então para o coração por meio das veias pulmonares. Essas veias desembocam no átrio esquerdo. O sangue, daí, segue para o ventrículo esquerdo, local de onde será impulsionado para todo o corpo, com exceção do pulmão, sendo levado pela artéria aorta.

3.1.1 Doenças Cardíacas

A sociedade atualmente possuí um modo de vida, e boa parte disso se deve ao isolamento da pandemia, que nos levou ao amento da fatia da sociedade que chegou ao sedentarismo, à má

alimentação e a hábitos nada saudáveis, como o consumo de álcool e fumo em excesso. Como isso, cresce a incidência de doenças do coração e por consequência, o número de mortes, com um aumento de 132% durante a pandemia (UFMG). Entre as enfermidades, o infarto e o Acidente Vascular Cerebral (AVC) foram aquelas que mais se destacaram. Somente o AVC levou a mais de 230 mil óbitos no país em 2021. A maior parte foi na faixa-etária entre 70 e 79 anos.

As doenças cardiovasculares podem ser congênitas (quando o indivíduo já nasce com a patologia), ou desenvolvidas com o tempo. Aquelas que são desenvolvidas com o tempo podem ser tanto uma consequência de hábitos pouco saudáveis, quanto pela idade ou até mesmo por conta de infecções provocadas por vírus, fungos ou bactérias. As doenças congênitas tem seu desenvolvimento ainda no desenvolvimento do bebê, podendo a mesma ser identificada ainda no útero através de exames com ultrassom e ecocardiograma.

A cardiopatia congênita pode variar de casos leves a graves, e seu tratamento depende da avaliação do cardiologista. Já as doenças cardíacas desenvolvidas ao longo do tempo em decorrência dos hábitos pouco saudáveis, geram uma lista enorme de problemas. Segue abaixo uma relação de doenças cardíacas, como são adquiridas, seus sintomas e fatores de risco associados.

- Hipertensão, também conhecida popularmente como “pressão alta”, trata-se do aumento da pressão arterial ocasionando o mau funcionamento do coração, em geral, está ligada a falta de prática de exercícios físicos, alimentação baseada em gorduras e consumo excessivo de sal.

- Insuficiência cardíaca (IC) - podendo ser uma consequência da hipertensão, é uma das doenças do coração mais comuns e mais graves, uma vez que o coração não consegue bombear o sangue rico em oxigênio de maneira suficiente para o organismo. Com o tempo, surgem sintomas como dificuldade para respirar, fadiga, fraqueza, inchaços, palpitações, entre outros. A causa mais comum para insuficiência cardíaca no Brasil é a Doença Arterial Coronariana (DAC), pois provoca o estreitamento dos vasos sanguíneos. Dentre os outros fatores de risco associados estão maus hábitos, infecções, e outros problemas como diabetes, infarto e apneia do sono. Os casos mais extremos da IC exigem um transplante de coração.

- Infarto, popularmente chamado de ataque cardíaco, é uma das doenças do coração mais comuns e fatais. Ocorre quando há falha na passagem de sangue para o coração, no qual o fluxo de sangue no músculo cardíaco (miocárdio) é interrompido por um período prolongado, danificando-o ou causando a morte do tecido. O infarto ocorre quando a artéria coronária, responsável por levar sangue ao coração, é bloqueada por uma placa de gordura ou coágulo, obstruindo-a. Este caso pode ser uma consequência da arteriosclerose, ou outros fatores como tabagismo, obesidade, alcoolismo, colesterol alto ou

predisposição genética.

- Endocardite, uma inflamação que atinge o endocárdio, isto é, a túnica que reveste o interior do miocárdio e limita as cavidades cardíacas. Geralmente causado por fungos ou bactérias que chegam ao coração através da corrente sanguínea, porém não sendo a única causa, uma vez que pode acontecer em decorrência de outras doenças como câncer e doenças autoimunes. Porém, o tecido só é afetado se já estiver debilitado por outra doença, como uma insuficiência cardíaca ou imunidade baixa.

- Arteriosclerose, caracteriza-se pelo endurecimento e estreitamento das artérias que levam oxigênio e nutrientes do coração para o resto do corpo. Isto é consequência do acúmulo colesterol, gordura saturada, entre outras substâncias. Prejudica-se o fluxo de sangue, trazendo dores no peito, fraqueza muscular, fadiga, dentre outros sintomas. Dentre as doenças do coração, esta é a mais associada ao envelhecimento, já que a gordura acumula de forma progressiva, quanto maior a idade, maior a gordura acumulada. Dentre outros fatores de risco associados a arteriosclerose, podemos citar a hipertensão, diabetes, colesterol alto, e o histórico familiar.

- Angina, é caracterizada pela forte dor no peito em consequência de esforços físicos que acabem exigindo mais do coração. Fortes emoções ou até mesmo baixas temperaturas podem também ocasiona-la. Quando ocorre em intervalos curtos, é classificada como estável, enquanto que mesmo que em menor intensidade, perdure por mais tempo, é chamada de instável. Tanto estável quanto instável podem ser consequência da arteriosclerose, anemias, colesterol alto, diabetes, sedentarismo e a predisposição genética. A angina é uma das enfermidades do coração que geralmente indica a presença de outros problemas cardíacos mais graves, e por este motivo deve ser acompanhada de perto pelo médico cardiologista.

- **Arritmia**, são as alterações nos ritmos dos batimentos, sejam eles acelerados, lentos ou fora do padrão. O portador sente mal-estar, fadiga, desconforto, palpitações ou até mesmo dificuldades para respirar. Seus casos extremos levam a perda da consciência. A arritmia por se só não é um problema grave, mas a ausência de seu diagnóstico e respectivo tratamento pode ocasionar outros problemas que levam o paciente ao óbito. É adquirida principalmente por condições genéticas, estresse, ansiedade ou outras doenças cardíacas. Assim como a angina, também pode representar um sintoma de algo mais perigoso.

- Cardiomiopatia, é uma condição onde o músculo cardíaco fica inflamado e inchado, perdendo força. Desta maneira, o órgão não consegue bombear sangue suficiente para o corpo, levando a uma IC. Justamente por esta razão, os sintomas da cardiomiopatia confundem-se com os da insuficiência cardíaca, com a presença de fraqueza, inchaço,

fadiga. Em casos extremos, faz-se necessário um transplante do coração. As causas específicas da cardiomiopatia seguem desconhecidas, apesar de existir uma linha de pesquisadores que associem fatores genéticos, maus hábitos e outras doenças cardíacas como influência. Seu tratamento é através de medicamentos que controlem a doença, associados a mudanças de hábitos de seu portador.

- Doença Arterial Periférica, é uma condição que afeta artérias que são responsáveis por bombear o sangue para os membros inferiores, i.e., pernas e pés. Ao estreitar e endurecer estas artérias, dificulta-se a chegada de sengue nessas regiões periféricas, e por consequência surgem as dores, feridas que não cicatrizam, inchaços ou mesmo outros sintomas. Nos homens isso por ocasionar disfunção erétil. As principais causas da doença arterial periférica são a aterosclerose, o acúmulo de placas de gordura, cálcio ou outras substâncias nas artérias. Além de outros fatores como colesterol alto e o próprio tabagismo pode contribuir. O tratamento da doença arterial periférica adota o tratamento da aterosclerose, com o objetivo de evitar que ocorram novos episódios, uma vez que a progressão destes dois problemas pode levar o paciente ao infarto, AVC, formação de edemas, dentre outras complicações.

3.1.2 Eletrocardiograma

Como dito anteriormente, o Eletrocardiograma, também chamado de ECG, é o exame mais comum para detecção das doenças cardiovasculares. Em geral, realizado nas primeiras consulta, este será o responsável por dar os indícios de determinadas doenças, que posteriormente serão diagnosticadas com auxílio de outros exames, como ecocardiograma, holter, teste ergonômico, dentre outros.

Como uma ferramenta de diagnóstico não invasiva, como podemos observar na Figura 3.3, o ECG de 12 derivações registra a atividade elétrica do coração como formas de onda. Desde que feita com precisão, sua interpretação pode detectar e monitorar uma série de condições cardíacas, de arritmias a doenças cardíacas coronárias e desequilíbrio eletrolítico. Desde o primeiro telecardiograma registrado em 1903, grandes avanços foram feitos na gravação e interpretação do ECG. Hoje, o ECG de 12 derivações continua sendo uma ferramenta de diagnóstico padrão entre paramédicos, paramédicos e funcionários do hospital(Cables and Sensors)

Durante o exame de ECG são posicionadas 10 eletrodos, também chamados de leads, em diferentes pontos do corpo do paciente, como podemos observar na Figura 3.3, desta forma o aparelho fará a impressão de 12 visões distintas do órgão. Os resultados deste procedimento consistem em uma sequência de sinais que são subdivididos em intervalos com pontos específicos (PQRST), como observado na Figura 3.4, que representa um batimento. A relevância de cada um destes segmentos e intervalos de sinais dependem da doença abordada. Onde a partir daí serão identificados os padrões que caracterizam os indícios de que o paciente pode

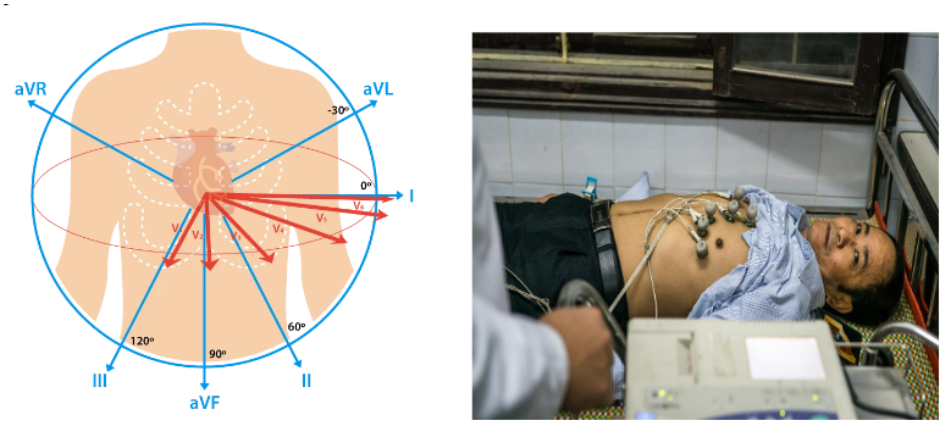


Figura 3.3: Plano vertical (Condutores Frontais) do ECG
(Cables and Sensors, OMS)

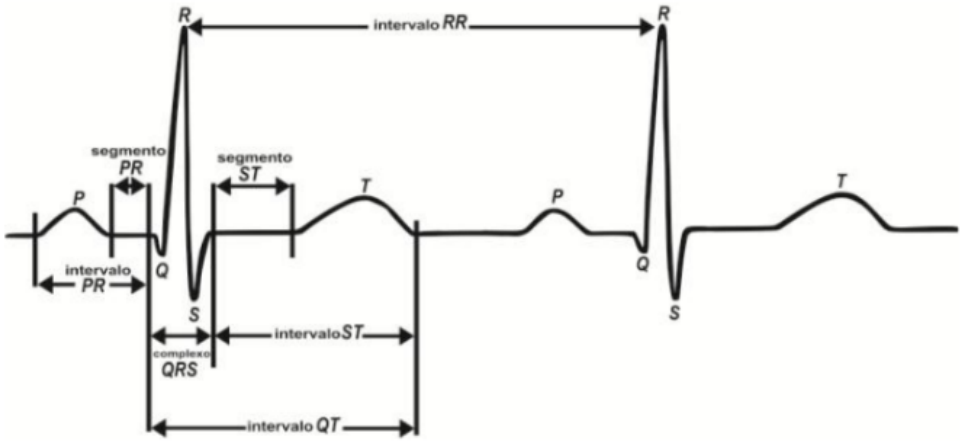


Figura 3.4: Segmentos e Intervalos do ECG (PQRST)
Pereira et al. (2019)

ou não ser portador daquele determinado tipo de arritmia. Na Figura 3.5, podemos observar a saída de um exame de ECG, que por sua vez será analisado pelo médico cardiologista que acompanhe aquele paciente.

Um ECG de 12 derivações mostra uma imagem completa da atividade elétrica do coração registrando informações através de 12 perspectivas diferentes. Imagine 12 pontos de vista diferentes de um objeto entrelaçado para criar uma história coesa: a interpretação do ECG. Essas 12 visualizações são coletadas colocando eletrodos ou pequenos adesivos pegajosos no peito (precordial), pulsos e tornozelos. Esses eletrodos são conectados a uma máquina que registra a atividade elétrica do coração. Para medir a atividade elétrica do coração com precisão, a colocação adequada dos eletrodos é crucial. Em um ECG de 12 derivações, existem 12 derivações calculadas usando 10 eletrodos. Podemos observar os posicionamentos dos eletrodos com colação do peito (Precordial) na Figura 3.6, e os eletrodos com colação do membro (extremidades) 3.7 (Cables and Sensors), com suas respectivas legendas.

Os eletrodos dos membros também podem ser colocados na parte superior dos braços e

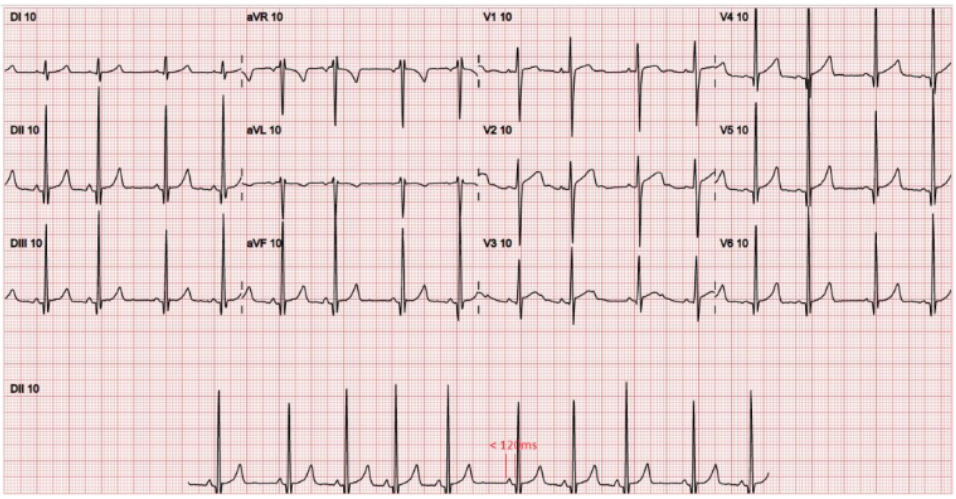
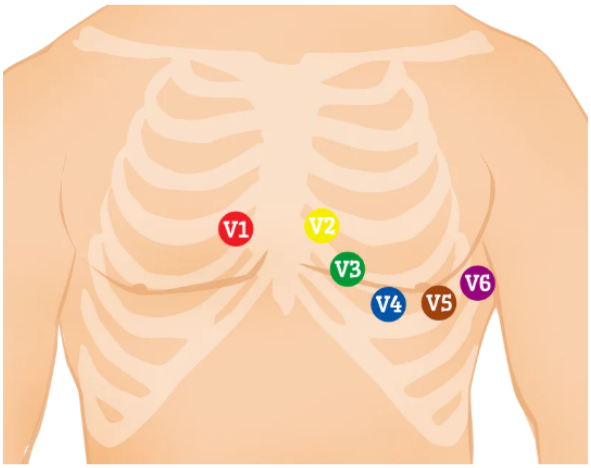
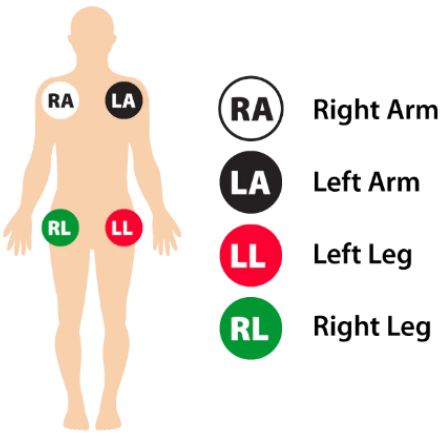


Figura 3.5: Resultado do Eletrocardiograma (ECG)
Lima (2018)



» V1 - Quarto espaço intercostal no esterno direito
» V2 - Quarto espaço intercostal no esterno esquerdo
» V3 - Meio caminho entre a colocação de V2 e V4
» V4 - Quinto espaço intercostal na linha hemiclavicular
» V5 - Linha axilar anterior na mesma horizontal nível como V4
» V6 - Linha axilar média no mesmo nível horizontal de V4 e V5

Figura 3.6: Derivações Precordiais ou do Plano Horizontal
Cables and Sensors



» RA (braço direito) - em qualquer lugar entre o ombro direito e o cotovelo direito
» RL (perna direita) - em qualquer lugar abaixo do tronco direito e acima do tornozelo direito
» LA (braço esquerdo) - em qualquer lugar entre o ombro esquerdo e o cotovelo esquerdo
» LL (Left Leg) - Em qualquer lugar abaixo do torso esquerdo e acima do tornozelo esquerdo

Figura 3.7: Derivações Periféricas ou das Extremidades
Cables and Sensors

nas coxas. Entretanto, deve haver uniformidade na sua colocação. Não permitindo, por exemplo, que se coloque um eletrodo no pulso direito e outro no braço esquerdo. Já para pacientes do sexo feminino, os eletrodos V3-V6 devem ser postos sob a mama esquerda. A localização do mamilo nunca deve ser referência para onde posicionar os eletrodos, uma vez que a posição deste muda de pessoa pra pessoa.

No ECG, as derivações são o registro da diferença de potencial elétrico entre dois pontos. Nas derivações bipolares são a diferença entre dois eletrodos, e nas derivações monopoloares a diferença entre um ponto virtual e um eletrodo. Dependendo do plano elétrico do coração que registrem, temos as derivações periféricas (plano frontal) e as derivações precordiais (horizontal). (Feitosa et al. (2002))

Se denominam **derivações periféricas** as derivações do ECG obtidas a partir dos eletrodos colocados nos membros(Figura 3.7. Estas derivações fornecem dados eletrocardiográficos do plano frontal (não proporcionam dados sobre potenciales dirigidos para a frente ou para trás). Existem dois tipos de derivações periféricas: as derivações bipolares, ou de Einthoven, e as derivações unipolares aumentadas.

Derivações bipolares do eletrocardiograma são as derivações clássicas do eletrocardiograma, descritas por Einthoven, registram a diferença de potencial entre dois eletrodos localizados em diferentes membros.

- D1: diferença de potencial entre o braço direito e o braço esquerdo. O vector é em direcção de 0º
- D2: diferença de potencial entre o braço direito e a perna esquerda. O vector é em direcção de 60º.
- D3: diferença de potencial entre o braço esquerdo e a perna esquerda. O vector é em direcção de 120º.

No **Triângulo e lei de Einthoven** (Figura 3.8) as três derivações bipolares formam o triângulo de Einthoven (inventor do eletrocardiograma). Estas derivações mantêm uma proporção matemática refletida na lei de Einthoven, que diz: D2=D1+D3. Esta lei é muito útil na interpretação de um eletrocardiograma. Permite determinar se os eletrodos periféricos estão correctamente posicionados, porque, se a posição de um eletrodo é variada, esta lei não se cumpre, deixando-nos saber que o ECG é mal feito.

Derivações unipolares aumentadas registram a diferença de potencial entre um ponto teórico no centro do triângulo de Einthoven, com um valor de 0, e os eletrodos em cada extremidade. Estas derivações, inicialmente, foram denominadas VR, VL e VF. A letra V significa Vector, e as letras R, L, F significam direita, esquerda e pé (em inglês). Posteriormente foi adicionada a letra a minúscula, que significa amplificada (as derivações unipolares atuais são amplificados com relação ao inicial).

- aVR: potencial absoluto do braço direito. O vector é em direcção de -150º.

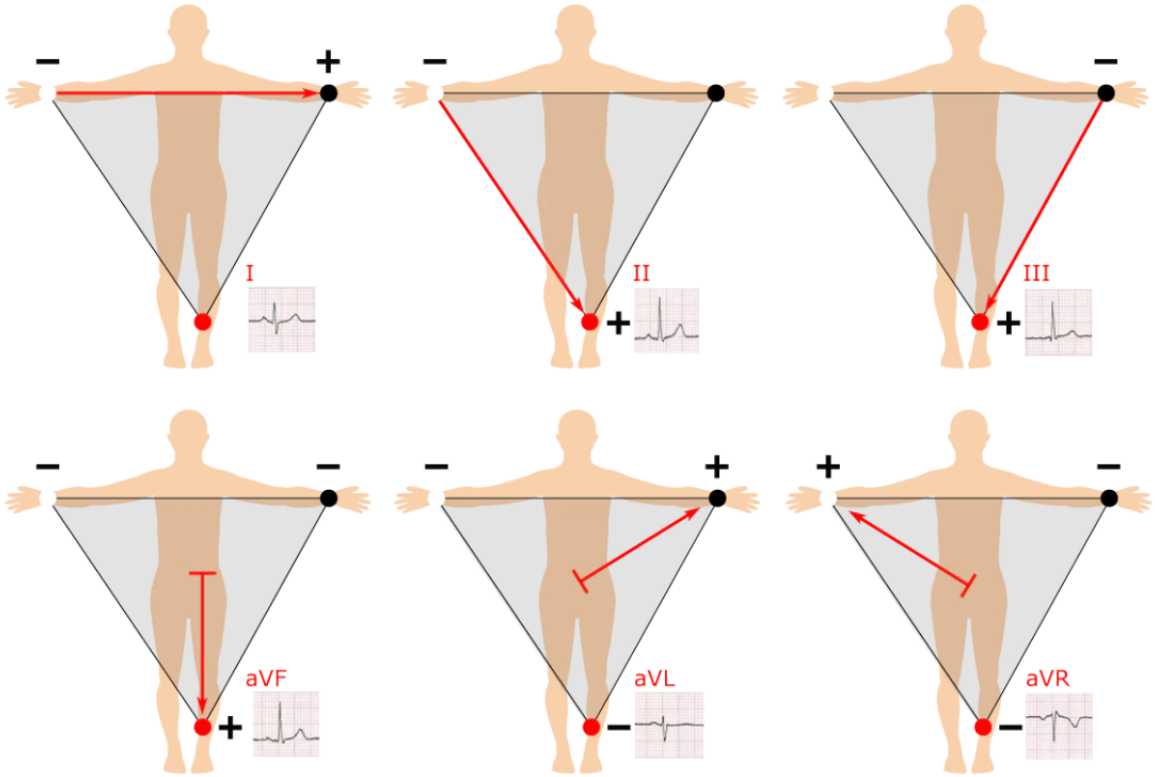


Figura 3.8: Triângulo de Einthoven
Cables and Sensors

- aVL: potencial absoluto do braço esquerdo. O vector é em direcção de -30º.
- aVF: potencial absoluto da perna esquerda. O vector é em direcção de 90º.

As **derivações precordiais** (Figura 3.6) do eletrocardiograma são seis, e são denominadas com uma letra V maiúscula e um número de 1 a 6. São derivações unipolares e registram o potencial do ponto em que o eletrodo de mesmo nome é posicionado. São as melhores derivações do ECG para determinar alterações do ventrículo esquerdo, especialmente das paredes anterior e posterior. No eletrocardiograma normal os complexos QRS são predominantemente negativos nas derivações V1 e V2 (morfologia rS) e predominantemente positivo nas derivações V4, V5, e V6 (morfologia Rs). Segue abaixo as derivações precordiais:

- V1: esta derivação do eletrocardiogra registra os potenciais dos átrios, de uma parte do septo interventricular e da parede anterior do ventrículo direito. O complexo QRS tem uma pequena onda R (despolarização do septo interventricular), seguido por uma onda S profunda, ver morfologia do complexo QRS.
- V2: esta derivação precordial está acima da parede do ventrículo direito, por conseguinte, a onda R é ligeiramente maior do que em V1, seguida por uma onda S profunda (activação do ventrículo esquerdo).

- V3: derivação de transição entre os potenciais esquerdos e direitos do ECG. O eletrodo é localizado sobre o septo interventricular. A onda R e a onda S são praticamente iguais (QRS isodifásico).
- V4: o eletrodo desta derivação está localizado no ápice do ventrículo esquerdo, onde a espessura é maior. Tem uma onda R alta seguida por uma onda S pequena (ativação do ventrículo direito).
- V5 e V6: estas derivações estão localizadas no miocárdio ventricular esquerdo, cuja espessura é menor do que em V4. Por conseguinte, a onda R é menor do que em V4, embora é alta. A onda R é precedida de uma onda q pequena (despolarização do septo).

Outras derivações ocorrem em pacientes com uma síndrome coronariana aguda com suspeita de infarto posterior ou de ventrículo direito é aconselhável colocar os eletrodos do eletrocardiograma em posições incomuns. No caso de suspeita de infarto posterior, devem ser colocados três eletrodos nas costas para obter as derivações posteriores. Quando existem dados de afetação do ventrículo direito devem ser colocados os eletrodos precordiais no lado direito, para obter as derivações direitas. Também quando o paciente apresenta dextrocardia ou situs inversus, se devem colocar os eletrodos no lado direito.

Interferências leves em um ECG não são incomuns, porém podem ser reduzidos ao adotar algumas medidas:

- Desligando os dispositivos e equipamentos elétricos não essenciais nas proximidades.
- Verificando se há loops de cabos e evite passar cabos adjacentes a objetos metálicos, pois eles podem afetar o sinal.
- Inspecionando os fios e cabos quanto a rachaduras ou quebras. Substitua conforme necessário.
- Usando supressores de picos com a fonte de alimentação.
- Certifiando que os filtros e pré-amplificadores estejam ajustados adequadamente.
- Garantindo a conexão segura entre o cabo do paciente e o dispositivo de ECG.
- Verificando se há folgas entre os conectores.

3.2 Arritmias

A arritmia ocorre quando há uma alteração no batimento cardíaco. Seja esta alteração de forma descompassada, lenta ou acelerada. A frequência do batimento cardíaco em uma pessoa saudável está entre 60 e 100 bpm (batimentos por minuto). A arritmia cardíaca pode ser classificada

como benigna ou maligna. No primeiro caso, a alteração do ritmo cardíaco não altera o desempenho cardíaco. Pode ser controlada com a administração de medicamentos e com exercícios físicos. Já a maligna concentra o maior número de mortes, já que o desgaste físico contribui para o comprometimento do coração. A arritmia pode ser assintomática ou sintomática. Quando sintomática, dentre os seus sintomas, temos palpitações no coração, que duram de segundos a semanas, queda de pressão, fadiga, falta de ar, desmaios, enjoos e vertigem. Dentre os fatores de risco temos tabagismo, sedentarismo, apneia do sono, exageros na ingestão de bebidas alcoólicas, distúrbios de tireoide, hipertensão, diabetes, estresse ou predisposição genética. Com o eletrocardiograma, holter, teste ergométrico e ecocardiograma é possível diagnosticar com precisão essa disfunção, que está entre as principais doenças cardiovasculares. Cada uma das arritmias possui seu tratamento, como por exemplo, a inserção de marca-passo em pacientes portadores de bradicardia (i.e., batimentos inferiores aos 60 bpm), uso de farmácos, alterações nos hábitos, alimentação saudável, prática esportiva, dentre outros. A prevenção da rritmia está associada aos hábitos saudáveis, evitar consumo em excesso de alcool, cigarro e outras drogas.

3.2.1 Ritmo Sinusal

As arritmias cardíacas podem ainda ser classificadas de acordo com o local de origem: Ritmos Sinusais, Ritmos Atriais e Ritmos Ventriculares. O **Ritmo Sinusal** é o ritmo normal do coração, disparando de 60 a 100 vezes por minuto (bpm). Suas variações incluem: a bradicardia sinusal (<60 bpm), quando o nódulo dispara de forma mais lenta, com menos de 60 batimentos por minuto, tornando o intervalo entre os sinais mais espaçados, como podemos observar na Figura 3.9. O fluxogrma de abordagem da bradicardia sinusal pode ser observado na Figura 3.10. Já a taquicardia sinusal (>100 bpm) gera uma frequência cardíaca mais rápida, como podemos observar na Figura 3.11. O tratamento mais usado na taquicardia sinusal inapropriada são os bloqueadores dos receptores beta-adrenérgicos (betabloqueadores) e os antagonistas do cálcio não diidropiridínicos (diltiazem e verapamilo). É costume iniciar com doses baixas e ir aumentando de acordo com a sintomatologia. O aumento do consumo de sal e água também pode diminuir a clínica dos pacientes. Recentemente tem aparecido a ivabradina (inibidor seletivo), fármaco que atua a nível do nó sinusal e que está indicado na cardiopatia isquêmica e na insuficiência cardíaca, que se sugere como uma alternativa no tratamento da taquicardia sinusal inapropriada.

Tanto a bradicardia sinusal, quanto a taquicardia sinusal, podem ser normais ou clínicas, uma vez que a bradicardia sinusal pode ser considerada normal durante o sono, por estarmos com os batimentos mais baixos que o normal. Assim como a taquicardia sinusal pode ser considerada normal durante a prática esportiva, por naturalmente aumentar a frequência dos batimentos.

Figura 3.9: ECG Característico da Bradicardia Sinusal
Sanarmed (2019)

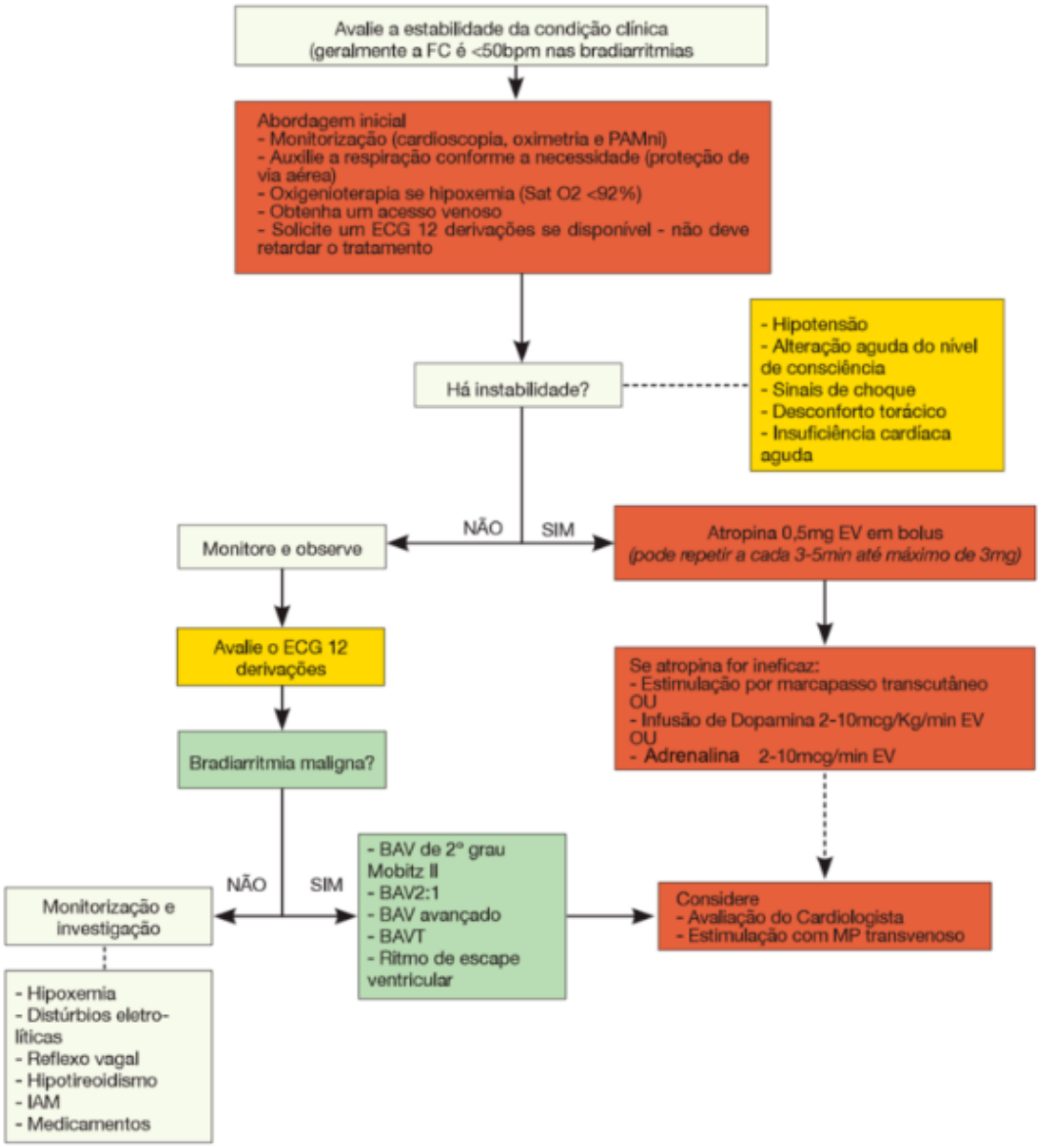


Figura 3.10: Fluxograma de Abordagem da Bradicardia
Sanarmed (2019)

Figura 3.11: ECG Característico da Taquicardia Sinusal
Sanarmed (2019)

Figura 3.12: ECG Característico do Flutter Atrial
Mitchell (2021)

3.2.2 Ritmo Atrial

As arritmias cardíacas originárias dentro de outras partes do átrio são sempre patológicas, classificadas como ritmos atriais. Os tipos de Ritmos Atriais são Flutter atrial, Fibrilação Atrial e Taquicardia Atrioventricular Reentrante Nodal. Estas são formas de Taquicardia Supraventricular (TSV). O flutter atrial ou auricular, é causado por um impulso elétrico que gira em torno de um ciclo, localizado no átrio direito (Figura 3.1), que se auto perpetua. Em um ECG, flutter atrial é caracterizado pela ausência de ondas P normais (Figura 3.12), ao invés disto são observadas as ondas flutter, também chamadas de ondas f, que possuem um padrão serrilhado.

A fibrilação atrial é causada por múltiplos impulsos elétricos que são iniciados aleatoriamente, dentro e ao redor dos átrios, geralmente perto das raízes das veias pulmonares. A variação da fibrilação pode ser inferior a 60bpm ou superior a 100bpm. Em um ECG a fibrilação atrial se caracteriza pela ausência de ondas P e presença de complexos QRS estreitos e irregulares, com a linha de base aparentando ser ondulada ou totalmente plana, como podemos observar na Figura 3.13.

A Taquicardia Atrioventricular Reentrante Nodal (TAVRN) é causada por um pequeno circuito de reentrada, neste caso, a frequência ventricular e a atrial são idênticas, e a frequência cardíaca é regular e rápida, variando de 150 a 250 bpm. Seu comportamento em um ECG pode

Figura 3.13: ECG Característico da Fibrilação Atrial (FA)
Alila (2020)

Figura 3.14: ECG Característico da TAVRN
Alila (2020)

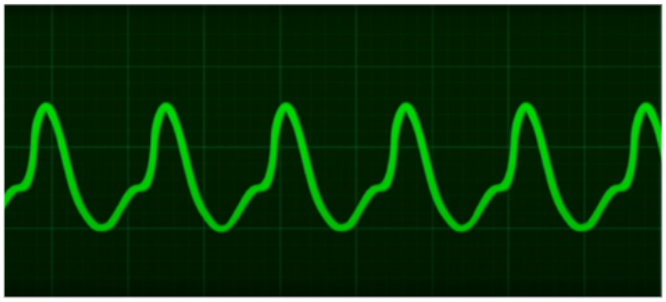


Figura 3.15: ECG Característico da Taquicardia Ventricular
Alila (2020)

ser observado na Figura 3.14.

3.2.3 Ritmo Ventricular

Os ritmos ventriculares são os mais perigosos sendo ameaça para vida. A taquicardia ventricular é causada por um foco de disparo ou circuito, sendo forte e único em um dos ventrículos, geralmente ocorrem em pessoas com problemas cardíacos estruturais, como cicatrizes de um ataque cardíaco prévio, ou anormalidades nos músculos cardíacos. Os batimentos são regulares e rápidos, variam de 100 a 250 bpm. Em um ECG a taquicardia ventricular é caracterizada por complexos QRS largos e estranhos, e a onda P ausente, como podemos observar na Figura 3.15. Pode ocorrer em episódios curtos com menos de 30 segundos, já sua forma sustentada dura mais de 30 segundos, e neste caso requer tratamento imediato para prevenção de uma parada cardíaca.

A taquicardia ventricular pode evoluir para Fibrilação Ventricular (FV), nesta o coração bombeia pouco ou nenhum sangue, podendo rapidamente levar a uma parada cardíaca. Em um ECG, a Fibrilação Ventricular (FV) é caracterizada por formas de onda irregular e aleatória de amplitude variável, não sendo possível identificar a onda P, o complexo QRS ou a onda T, como observado na Figura 3.16.

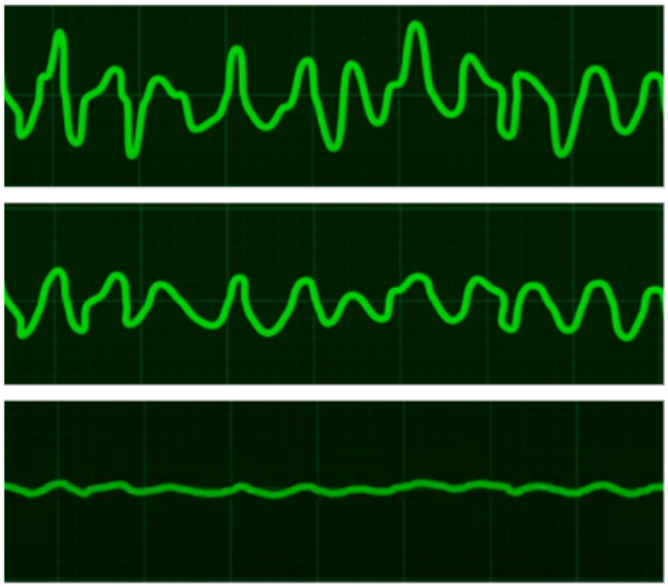


Figura 3.16: ECG da Fibrilação Ventricular evoluindo para uma Parada Cardíaca
Alila (2020)

3.3 Fatores de Risco

As arritmias cardíacas podem ser congênitas, como dito anteriormente, ou adquiridas ao longo da vida. No segundo caso, fatores de risco em conjunto com os hábitos do paciente possuem grande impacto para levar um paciente a adquirir uma arritmia, evita-la ou conviver melhor com a mesma. Alguns artigos da literatura apontam a relação das arritmias com determinados fatores de risco, como listamos abaixo, em conjunto com fatores não modificáveis e outros fatoes. (Gatzoulis et al. (2000), Silva et al. (2004))

- Os fatores de risco reconhecidos pela literatura atualmente incluem fumo, colesterou e triglicérides elevados, hipertensão arterial, falta de atividades físicas, diabetes, sobrepeso, obesidade, estresse, álcool, além da proteína C reativa ou hemocisteína, substâncias estas recentemente associadas com o aumento do risco de eventos cardiovasculares.
- Fatores não modificáveis incluem a predisposição genética advinda do histórico familiar (episódios da doença em familiares de primeiro grau), e o gênero.
- Outros fatores incluem certas medicações, tabagismo, cafeína e drogas, baixo teor de oxigênio no sangue, apneia do sono.

3.4 CHA_2DS_2-VASc

No contexto das arritmias, o escore CHA_2DS_2-VASc (Lip, Joundi et al. (2016)) é um dos mais disseminados na atual literatura. Seu uso visa melhorar a estratificação do risco de AVC entre

Tabela 3.1: Indicadores e pontuações do CHA2DS2-VASc

	Descrição	Pontos
C	Insuficiência Cardíaca	1
H	Hipertensão	1
A_2	Idade (≥ 75)	2
D	Diabetes Mellitus	1
S_2	AVC ou AIT Prévio	2
V	Doença Vascular	1
A	Idade (65-74 anos)	1
Sc	Sexo (se feminino)	1

pacientes com fibrilação atrial, para que seja possível administrar os medicamentos de anticoagulação da melhor maneira. Durante a pandemia estudos indicaram seu uso na predição de mortalidade em pacientes hospitalizados com a COVID-19 (Cetinkal et al. (2020)), provendo ao corpo médico informações que possibilitaram adotar as melhores abordagens e tratamentos dado o contexto em que o paciente estivera inserido. Uma série de estudos recentes estão sendo desenvolvidos para otimizar o seu uso, trazendo uma maior precisão e consequente ganhos para os pacientes avaliados com este escore. (Friberg et al. (2012b), Friberg et al. (2012b)).

O escore CHA_2DS_2-VASc é composto de é calculado com base em 7 indicadores: (i) Insuficência Cardíaca, (ii) Hipertensão, (iii) Idade, (iv) Diabetes Mellitus, (v) Ocorrência prévia de Ataque Vascular Cerebral (AVC) ou Ataque Isquêmico Transitório (AIT, também chamado Mini-AVC), (vi) Doença Vascular e (vii) Sexo. Alguns destes indicadores tem um maior peso, são eles a idade quando maior ou igual a 75 anos, e a ocorrência de AVC ou AIT prévio. Sua pontuação varia em um intervalo que pode ir de zero até 9. (Tabela 3.1).

As principais diretrizes usaram o risco de AVC anual fixo como uma diretriz para iniciar o tratamento anticoagulante; onde o risco de acidente vascular cerebral isquêmico de mais de 1% a 2% deve ser uma indicação para iniciar uma terapia anticoagulante. No entanto, o risco real de contrair AVC varia de acordo com o método de amostragem e regiões geográficas, bem como o uso de metodologia de análise de estudo apropriada. (Nielsen et al. (2016)). Uma meta-análise de vários estudos em 2015 mostrou que o risco anual de AVC é inferior a 1% em 13 dos 17 estudos para pontuação CHA_2DS_2-VASc, 6 de 15 estudos relataram risco de 1 a 2% e 5 de 15 estudos relataram risco de mais de 2% para pontuação CHA_2DS_2-VASc. (Quinn et al. (2017)). No entanto, as taxas de AVC variam de acordo com a população (amostragem), etnia, cenário do estudo (hospital ou comunidade), dentre outros fatores. Alguns estudos incluídos na meta-análise incluem mulheres com pontuação 1 em virtude do sexo (que são de baixo risco), no agregado cotações; outros incluídos não levam em conta o uso de anticoagulação de acompanhamento (reduzindo assim as taxas).

O escore CHA_2DS_2-VASc tem mostrado popularidade crescente ao longo do tempo, en-

Tabela 3.2: Risco Anual de **AVC** Baseado no CHA_2DS_2-VASc

CHA_2DS_2-VASc Score	Friberg et al. (2012a)	Lip et al. (2010)	Intervalo de Confiança (95%)
0	0.2%	0.0%	0.0 - 0.0
1	0.6%	0.6%	0.0 - 3.4
2	2.2%	1.6%	0.3 - 4.7
3	3.2%	3.9%	1.7 - 7.6
4	4.8%	1.9%	0.5 - 4.9
5	7.2%	3.2%	0.7 - 9.0
6	9.7%	3.6%	0.4 - 12.3
7	11.2%	8.0%	1.0 - 26.0
8	10.8%	11.1%	0.3 - 48.3
9	12.2%	100%	2.5 - 100.0

Tabela 3.3: Risco Anual de **AIT** ou **Embolia Sistemica** Baseado no CHA_2DS_2-VASc

CHA_2DS_2-VASc Score	Friberg et al. (2012a)
0	0.3%
1	0.9%
2	2.9%
3	4.6%
4	6.7%
5	10.0%
6	13.6%
7	15.7%
8	15.2%
9	17.4%

quanto o CHADS2 tem mostrado popularidade decrescente, (Habboushe et al. (2019)) o que pode estar relacionado à introdução de diretrizes que recomendam o uso do escore CHA_2DS_2-VASc para estratificação de risco do AVC. O percentual de risco de AVC baseado no escore do CHA_2DS_2-VASc possuí algumas validações, e seus valores possuem leves alterações de acordo com os autores de cada uma delas, como podemos obervar na Tabela 3.2 (Friberg et al. (2012c), Lip et al. (2010)). Friberg ainda fornece o risco anual de ocorrência do ataque isquemico transitóro (ou simplesmente mini AVC), ou embolia sistêmica, como podemos observar na Tabela 3.3.

3.5 Classificador NYHA (*New York Heart Association*)

A Insuficiência Cardíaca é uma síndrome clínica, caraterizada pela intolerância ao exercício e/ou sinais de congestão na presença de uma condição cardíaca (congênita ou adquirida). A utilização da classificação funcional da New York Heart Association (NYHA)(Bredy et al. (2018),

Tabela 3.4: Classificação NYHA - Classes Funcionais

NYHA	Sintomas do Paciente
Classe I	Assintomático
Classe II	Levemente Assintomática
Classe III	Sintomático
Classe IV	Sintomático em Repouso

Tabela 3.5: Mortalidade a partir da Classificação NYHA

NYHA	Sem Tratamento	Com Tratamento	Padrão
I	5-19%	5-10%	10-15%
II	15-40%	5-10%	10-15%
III	15-40%	10-15%	15-20%
IV	44-64%*	30-40%	20-50%

Bennett et al. (2002)) é recomendada no contexto clínco para avaliar a gravidade da insuficiência cardíaca. Esta avaliação subjetiva do estado funcional dos pacientes é importante para o seu diagnóstico e gestão, todavia também é importante por razões prognósticas. Existem vários estudos na literatura voltados para o aprimoramento deste classificador (Kubo et al. (2004)).

Na Tabela 3.4 podemos observar a classificação NYHA com base nos sintomas do paciente. A Classe I referente aos pacientes assintomáticos, ou seja, aqueles com cardiopatia estrutural definida e diagnósticada, porém sem sintomas e sem limitações para atividades físicas. A Classe II referente aos pacientes levemente assintomáticos, i.e., pacientes que apresentam sintomas desencadeados por atividades habituais, como subir lances de escada por exemplo. Classe III dos sintomáticos, pacientes que apresentam sintomas com atividades menores que as habituais, como tomar banho, falar ou comer. Por fim, a Classe IV, dos pacientes sintomáticos em repouso.

A sobrevida do paciente está de acordo com a classe funcional na qual ele está inserido. Sobrevida esta que varia acentuadamente não só com a classificação recebida, quanto com outros fatores, como idade, gênero, raça, dentre outros. Tais fatores quando levados em consideração aumentam a precisão da mortalidade predita. Na tabela 3.5 iremos adotar as convenções da literatura que mostram os intervalos de mortalidade em percentuais, dentro de um intervalo que varia entre 1 e 4 anos, variando os cenários onde nenhum tratamento está sendo feito, com tratamento adotado(Bredy et al. (2018)).

3.6 Ciência de Dados

Embora não seja novo, o termo ciência de dados, os significados e suas conotações mudaram ao longo do tempo. A palavra apareceu pela primeira vez na década de 1960 como um nome

alternativo para estatísticas. No final da década de 1990, profissionais de ciência da computação formalizaram o termo. Uma definição proposta para ciência de dados o viu como um campo separado com três aspectos: design de dados, coleta e análise. Ainda levou mais uma década para que o termo fosse usado fora da academia.

A ciência de dados (Dhar (2013)) é o estudo dos dados para extrair informações significativas. É uma abordagem multidisciplinar que combina princípios e práticas das áreas de matematica, estatística, inteligência artificial e engenharia da computação para analisar grandes quantidades de dados. Essa análise ajuda os cientistas de dados a fazer e responder perguntas como o que aconteceu, por que aconteceu, o que acontecerá e o que pode ser feito com os resultados. Sua importância é associada ao fato de combinar ferramentas, métodos e tecnologias para gerar siginificado com base nos dados avaliados. Cientistas de dados podem converter dados brutos em recomendações para tomadas de decisão. A ciência de dados é utilizada para estudar dados de quatro maneiras principais: Análise descritiva, análise diagnóstica, análise preditiva e análise prescritiva.

- **Análise descritiva**: Analisa os dados para obter insights sobre o que aconteceu ou o que está acontecendo no ambiente de dados. Ela é caracterizada por visualizações de dados, como gráficos de pizza, gráficos de barras, gráficos de linhas, tabelas ou narrativas geradas.

- **Análise diagnóstica**: Análise aprofundada ou detalhada de dados para entender por que algo aconteceu. Ela é caracterizada por técnicas como drill-down, descoberta de dados, mineração de dados e correlações. Várias operações e transformações de dados podem ser realizadas em um determinado conjunto de dados para descobrir padrões exclusivos em cada uma dessas técnicas.

- **Análise preditiva**: Usa dados históricos para fazer previsões precisas sobre padrões de dados que podem ocorrer no futuro. Ela é caracterizada por técnicas como machine learning, previsão, correspondência de padrões e modelagem preditiva. Em cada uma dessas técnicas, os computadores são treinados para fazer engenharia reversa de conexões de causalidade nos dados.

- **Análise prescritiva**: A análise prescritiva leva os dados preditivos a um novo patamar. Ela não só prevê o que provavelmente acontecerá, mas também sugere uma resposta ideal para esse resultado. Ela pode analisar as potenciais implicações de diferentes escolhas e recomendar o melhor plano de ação. A análise prescritiva usa análise de gráficos, simulação, processamento de eventos complexos, redes neurais e mecanismos de recomendação de machine learning.

3.6.1 Processo de Ciência de Dados - OSEMN

Um cientista de dados trabalha com as partes interessadas do negócio para compreender quais são as necessidades de cada negócio. Uma vez definido o problema, o cientista de dados pode solucioná-lo usando o processo de ciência de dados OSEMN:

- **O: Obter dados:** Os dados podem ser pré-existentes, recém-adquiridos ou um repositório de dados que pode ser baixado da Internet. Os cientistas de dados podem extrair dados de bancos de dados internos ou externos, software de CRM da empresa, logs de servidores da Web, mídias sociais ou comprá-los de fontes confiáveis de terceiros.

- **S: Suprimir dados** A supressão de dados, ou limpeza de dados, é o processo de padronização dos dados de acordo com um formato predeterminado. Ela inclui lidar com a ausência de dados, corrigir erros de dados e remover quaisquer dados atípicos.

- **E: Explorar dados** A exploração de dados é uma análise de dados preliminar que é usada para planejar outras estratégias de modelagem de dados. Os cientistas de dados obtêm uma compreensão inicial dos dados usando estatísticas descritivas e ferramentas de visualização de dados. Em seguida, eles exploram os dados para identificar padrões interessantes que podem ser estudados ou acionados.

- **M:Modelar dados** Os algoritmos de software e machine learning são usados para obter insights mais profundos, prever resultados e prescrever o melhor plano de ação. Técnicas de machine learning, como associação, classificação e clustering, são aplicadas ao conjunto de dados de treinamento. O modelo pode ser testado em relação a dados de teste predeterminados para avaliar a precisão dos resultados. O modelo de dados pode ser ajustado várias vezes para melhorar os resultados.

- **N: Interpretar resultados** Os cientistas de dados trabalham em conjunto com analistas e empresas para converter insights de dados em ação. Eles fazem diagramas, gráficos e tabelas para representar tendências e previsões. A sumarização de dados ajuda as partes interessadas a entender e implementar os resultados de forma eficaz.

3.6.2 Ténicas de Ciência de Dados

O princípio básico por trás das técnicas de ciência de dados, busca (i) Ensinar uma máquina a classificar dados com base em um conjunto de dados conhecido. (ii) Fornecer dados desconhecidos à máquina e permitir que o dispositivo classifique o conjunto de dados de forma independente. E (iii) Permitir imprecisões de resultados e lidar com o fator de probabilidade do resultado.

Classificação

Trata-se da ordenação de dados em grupos ou categorias específicos. Os computadores são treinados para identificar e classificar dados. Conjuntos de dados conhecidos são usados para criar algoritmos de decisão em um computador que processa e categoriza rapidamente os dados. Exemplo: Classificar sentimento de tweets como positivos ou negativos sobre determinado tema.

Regressão

Método de encontrar uma relação entre dois pontos de dados aparentemente não relacionados. A conexão geralmente é modelada em torno de uma fórmula matemática e representada como um gráfico ou curvas. Quando o valor de um ponto de dados é conhecido, a regressão é usada para prever o outro ponto de dados. Exemplo: A relação entre o sentimento dos tweets de determinado tema, e uma notícia sobre o mesmo.

Clustering

Método de agrupar dados intimamente relacionados para procurar padrões e anomalias. O clustering é diferente da classificação porque os dados não podem ser classificados com precisão em categorias fixas. Portanto, os dados são agrupados em relações mais prováveis. Novos padrões e relações podem ser descobertos com o clustering. Por exemplo: Agrupar artigos em diversas categorias de notícias diferentes e usar essas informações para encontrar conteúdo de notícias falsas.

3.6.3 Big Data

Big Data é o termo que trata sobre grandes conjuntos de dados que precisam ser processados e armazenados. O conceito do *Big Data* se iniciou com 3 Vs : Velocidade, Volume e Variedade. Mas já pode ser extendido para 5 Vs, considerando Valor e Veracidade. Atualmente, em decorrência dos avanços tecnologicos, o volume de dados gerado é monstruoso, todos os dias bilhões de novas informações são geradas globalmente, a partir de Apps, Sistemas, TVs, Celulares, aparelhos com IoT (*Internet of Things* ou Internet das Coisas) capturam, processm e armazenam novos dados.

3.7 Aprendizagem de Máquina

O Aprendizado de Máquina (*Machine Learning* - ML) (Mitchell et al. (1990)) é uma área da Inteligência Artificial que tem como objetivo o desenvolvimento de técnicas computacionais sobre

o aprendizado e construção de sistemas capases de adquirir conhecimento de forma automatizada. Um sistema de aprendizado é um programa computacional capaz de tomar decisões com base em experiências acumuladas, tomando com base soluções anteriores bem sucedidas. Sistemas de Aprendizagem de Máquina possuem peculiaridades que permitem classificações quanto a linguagem de descrição, modo, paradigma e forma de aprendizado utilizado. As principais formas de aprendizado são classificadas como (i) Supervisionado, (ii) Não Supervisionado e (iii) Por Reforço.

3.7.1 Aprendizado Supervisionado

O Aprendizado Supervisionado costuma ser aplicado em cenários de classificação. Trata-se de uma tarefa preditiva (Figura 3.17), uma forma de aprendizagem de máquina que consiste em ensinar o modelo o que ele deve fazer, fornecendo para cada entrada a saída desejada, também chamado de rótulo. A disponibilidade destes rótulos no conjunto de dados, permite que os algoritmos supervisionados selecionem as características mais efetivas na distinção de instâncias de diferentes rótulos. Neste cenário, as características são selecionadas a partir conjunto de dados de treinamento. Dessa forma, não são utilizados todas as propriedades no treinamento do modelo de aprendizagem supervisionada, inicialmente seleciona-se um subconjunto de características, após isso aplica-se o conjunto de dados resultantes sobre o modelo de aprendizagem utilizado. Em resumo, ocorre quando apresentamos ao algoritmo dados de entrada e as respectivas saídas.

3.7.2 Aprendizado Não Supervisionado

Os modelos de aprendizado não supervisionados são voltados para problemas de clusterização. Trata-se de uma tarefa descritiva. Neste cenário não há informações sobre os dados esperados enquanto saída (rótulos). Sem esta informação para elencar a efetividade das características, este tipo de modelo (não supervisionado) utiliza de outros critérios para analisar a relevância de cada característica, obtendo uma melhor percepção e descrição sobre a estrutura tratada, detectando agrupamentos implícitos, e características úteis para categorização. O critério adotado

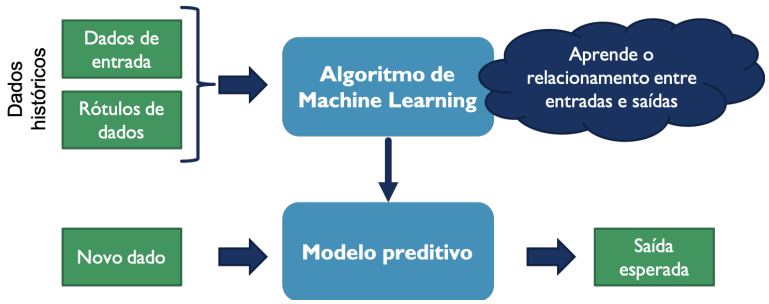


Figura 3.17: Aprendizado Supervisionado - Modelo Preditivo
Escovedo (2020)

deve preservar a estrutura múltipla dos dados originais. Em resumo, ocorre quando apresentamos somente os dados de entrada, e o algoritmo descobre as saídas. Existem duas formas de utilizar algoritmos de agrupamento: (i) buscando indicadores de agrupamento paralelo à seleção de caracteristicas supervisionadas dentro de uma estrutura unificada. (ii) Busca inicialmente o indicador de agrupamento, depois utiliza uma característica de inclusão ou exclusão de outras características, repetindo o processo até que determinado critério seja satisfatório.

Modelos Semi-supervisionados

O Modelo Semi-supervisionado é indicado para cenários onde a base de dados apresente parte rotulada, e parte não rotulada. Todavia, a maioria dos modelos semi-supervisionados existentes dependem da construção de uma matriz de similaridade, com o objetivo de selecionar as características mais relevantes. Tanto informações dos rótulos quanto medidas de similaridade entre os dados são considerados para a construção da matriz de similaridade. A partir disto, as informações rotuladas disponibilizam aspectos que irão diferenciar as características relevantes, e os dados não rotulados iram prover informações complementares. (Sammut and Webb, 2017).

3.7.3 Aprendizado Por Reforço

O Aprendizado por Reforço (Figura 3.19) é baseado na tentativa e erro, sob a exploração de um espaço desconhecido, e evolução baseada em recompensa. Os algoritmos aprendem políticas para agir através da interação repetida com o ambiente, o qual provê recompensa positiva ou negativa para o algoritmo, assim adquire experiência em decisões ou passos subsequentes O agente aprende a atingir uma meta em um ambiente incerto e potencialmente complexo. No aprendizado por reforço, o sistema de inteligência artificial enfrenta uma situação

3.8 Algoritmos

Nesta seção iremos apresentar o cojunto dos algoritmos que foram utilizados na elaboração do modelo de prognóstico proposto. Ánalise dos Principais Componentes (PCA), K Vizinhos Mais Próximos (KNN) e Árvore de Decisão (DT).

3.8.1 Ánalise dos Principais Componentes (PCA)

O objetivo do processo de Análise dos Principais Componentes é reduzir a quantidade de variáveis que são aplicadas em um modelo de Aprendizado de Máquina. A etapa de redução de características pode ser feita através da análise e processamento das características, com o objetivo de reconhecer quais são as mais relevantes dentre as existentes. As aplicações de

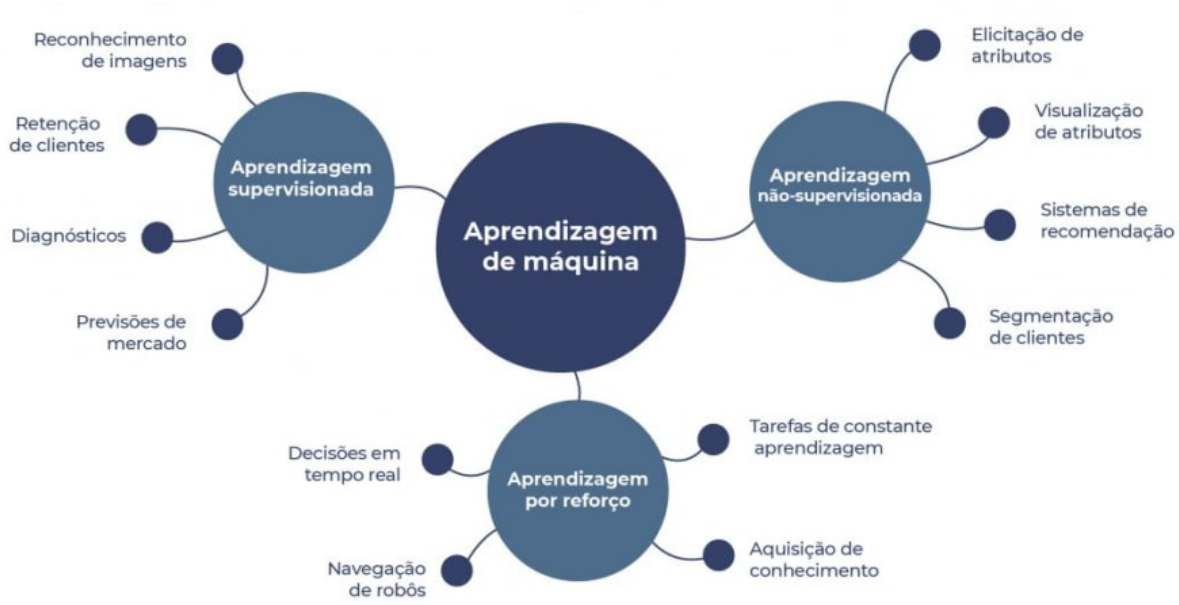


Figura 3.18: Exemplos de Aplicações dos Tipos de Aprendizagem
Maia (2020)

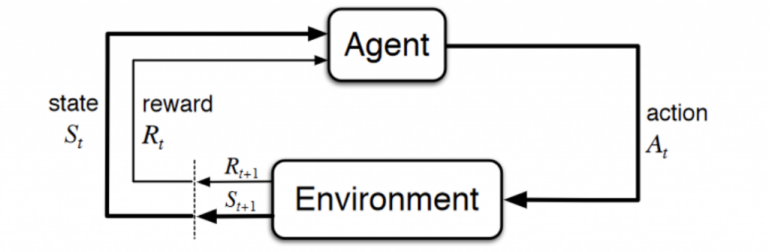


Figura 3.19: Aprendizagem por Reforço
Maia (2020)

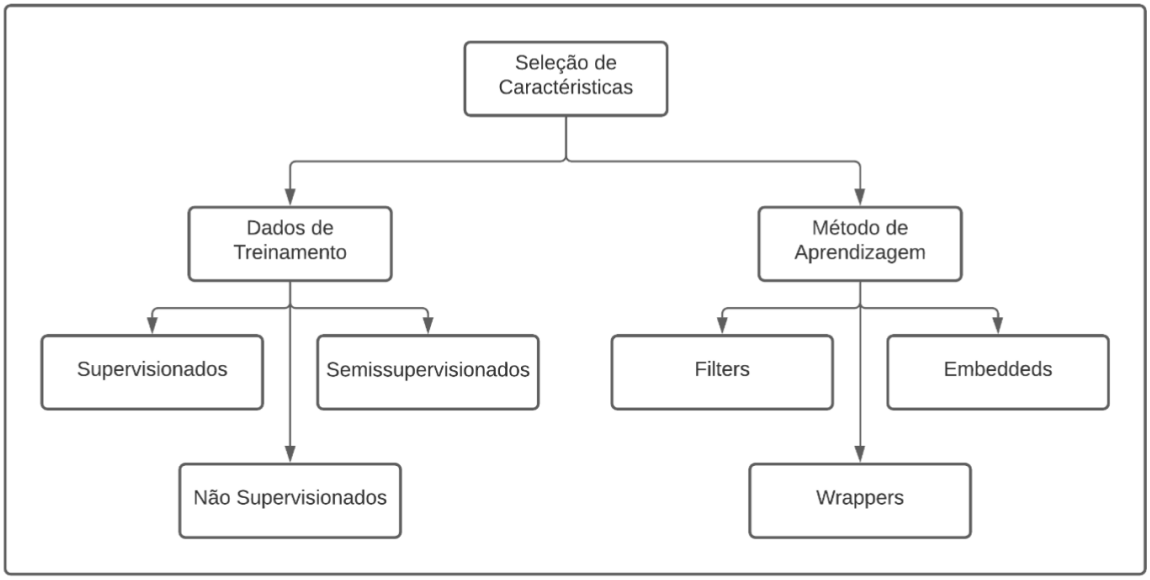


Figura 3.20: Seleção de características - Categorias
(Sammut and Webb, 2017)

seleção de características está presente nas mais diversas áreas. Identificar a correlação e dependência entre as variáveis, possibilita eliminar variáveis irrelevantes, trazendo melhorias no desempenho do modelo implementado.

Faz-se necessária a definição de um critério para remover características que não tenham um nível de relevância aceitável com a variável alvo. Entretanto, é necessário ressaltar que a remoção de características irrelevantes não tem relação com métodos de redução de dimensionalidade de um conjunto de dados, tais como o *Principal Component Analysis* (PCA). A eliminação de propriedades através de técnicas de seleção de características não cria novas propriedades. Pois uma vez que é selecionado um critério de seleção de características, é iniciado um procedimento para encontrar um subconjunto de características úteis dentro do conjunto analisado (CHA, 2014).

Sua aplicação está presente em diversas áreas, tais como análise de dados, reconhecimento de imagens, mineração de texto, seleção de indicadores de maior peso para determinada doença, etc. Podem ser classificados de acordo com o padrão utilizado. Quando feito de acordo com os dados de treinamento, estes podem ser classificados como modelos supervisionados, não supervisionados e semi-supervisionados. Quando analisados de acordo os métodos de aprendizagem, podem ser classificados como *Filters, Wrappers e Embedded*.(Cai et al., 2018). A Figura 3.20 demonstra um diagrama sobre a categorização dos modelos de seleção de características.

- **Filter**: Estabelece um ranqueamento das características utilizando critérios pré-definidos. Sua execução occore anterior à aplicação dos modelos de *Machine Learning*, e são independentes do modelo aplicado no estudo. Este tipo de técnica pode ser classificado de acordo com os parâmetros de filtragem que empregam, tais como nível de dependên-

cia entre as propriedades ou grau de similidade. Porém, este tipo de modelo não leva em consideração o viés e as heurísticas dos algoritmos de aprendizagem de máquina. Dessa forma, podem ser que faltem recursos que são relevantes para o algoritmo de aprendizagem alvo. Alguns dos critérios utilizados neste tipo de classificação das características são: a capacidade efetiva de separar amostras de diferentes classes considerando a variação de classes e a variância entre elas, além de analisar a dependência entre uma característica e o rótulo da instância analisada, verificar também a correção entre as classes dos dados e entre as características presentes. Porém, vale ressaltar que a principal desvantagem deste tipo de abordagem é que é ignorado totalmente o efeito do subconjunto de características selecionados na execução de algoritmos de classificação ou agrupamento. (Sammut and Webb, 2017).

- **Wrappers**: Cria uma quantidade de subconjuntos pré-definido de acordo com a quantidade total de características. Posteriormente calcula um score para cada subconjunto utilizando uma função objetiva específica. O subconjunto com a maior performance é selecionado, desconsiderando os demais subconjuntos. Geralmente este tipo de modelo possuem um maior custo de processamento computacional quando comparado com os algoritmos filters.O subconjunto de características considerado ideal depende do viés e heurísticas específicas dos algoritmos de aprendizagem. Com base nisto, modelos wrappers usam algoritmos de aprendizagem específicos na avaliação da qualidade das características (Sammut and Webb, 2017).

- **Embedded**: Combinação dos dois métodos supracitados. Sua vantagem é ser mais performático em relação ao método wrappers, uma vez que realiza a integração de forma paralela entre o modelo e as características selecionadas. Pode ser realizado através de uma função objetiva dividida em duas partes, um termo de adequação e, uma penalidade para um subconjunto com um número elevado de características (Sammut and Webb, 2017).

3.8.2 Máquinas de Vetores Suporte (SVM)

Máquina de vetores suporte, ou simplesmente SVM, é um algoritmo de aprendizado de máquina supervisionado que pode ser utilizado para classificação ou regressão. Seu foco está no treinamento e classificação de um dataset. Neste algoritmo, cada item é disposto como um ponto no espaço n-dimensional, onde n é o número de recursos disponíveis. O valor de cada um destes recursos será uma coordenada, e a classificação ocorre identificando o hiperplano que melhor divide os dados em duas classes.

Os chamados vetores de suporte são apenas as coordenadas de observação individual. SVM é a fronteira que melhor segrega as duas classes (hiperplano, ou linha). As SVMs podem ter suas margens em três casos, as margens rígidas (lineares), margens suaves, e as margens

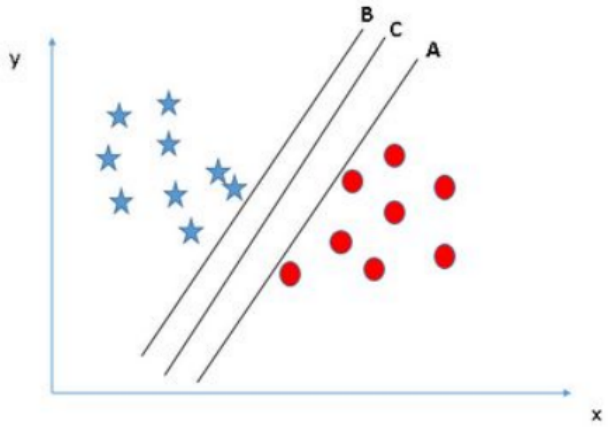


Figura 3.21: Hiperplano - SVM - Linear
(Addan, 2019)

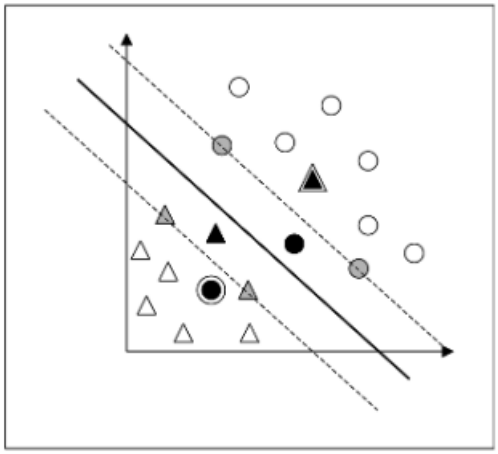


Figura 3.22: Hiperplano - SVM - Suave
(Lorena and Carvalho, 2007)

não lineares. (i) As margens rígidas, também chamadas de lineares, definem fronteiras lineares a partir de dados linearmente separáveis, isto é, torna-se possível separar completamente os 2 conjuntos de dados com o hiperplano identificado, como observa-se na Figura 3.21. (ii) Já as margens suaves, não conseguem separar completamente os dados em seu hiperplano, permitindo então que alguns dados violem o hiperplano, sendo então o melhor hiperplano aquele que somar o menor número de dados no campo errado do hiperplano, podemos observar o exemplo na Figura 3.22. (iii) Por fim, as SVMs não lineares, onde se faz necessária a alteração da dimensão dos dados para a melhor separação dos dados através do hiperplano, como podemos observar na Figura 3.23, onde é necessária a mudança da segunda para a terceira dimensão.

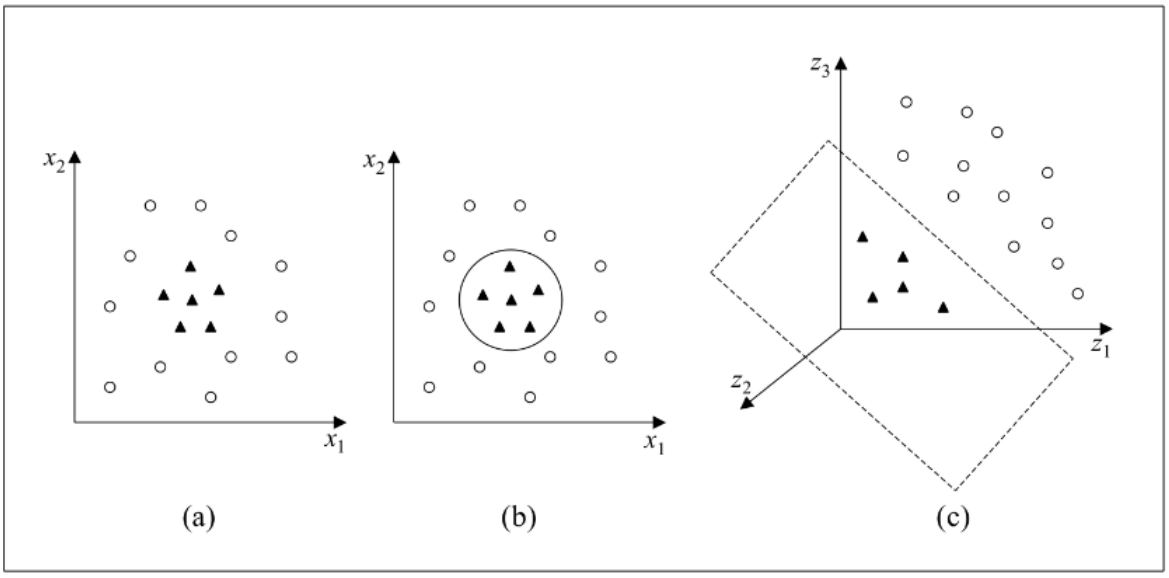


Figura 3.23: Hiperplano - SVM - Não Linear
(Lorena and Carvalho, 2007)

3.8.3 Redes Neurais (RN)

Dentre os algoritmos mais utilizados para o diagnóstico e prognóstico de arritmia, encontram-se as redes neurais e suas variações, como redes neurais profundas, recorrentes e convolucionais. Neste tópico iremos introduzir a ideia de rede neural, que será a base para ideias dos próximos tópicos.

Em 1986 foi feita a publicação do clássico Parallel Distributed Processing, editado por Rummelhart e McClelland (PDP Research Group da Universidade da Califórnia em San Diego). Após esta publicação a área de redes neurais desenvolveu-se exponencialmente entre os Journal's, associações internacionais, teses e papers científicos. Naturalmente surgiram também várias empresas para exploração de produtos de software, desenvolvimento de aplicações ou simulações acadêmicas que envolviam redes neurais.

As Redes Neurais Artificiais são técnicas da computação que representam um modelo matemático inspirado na estrutura neural de organismos inteligentes, e que adquirem conhecimento através da experiência. Uma rede neural robusta pode ter centenas ou milhares de unidades de processamento. Já o cérebro de um mamífero pode ter bilhões de neurônios. (Figura 3.25)

Os neurônios se comunicam através de sinapses. Sinapse é a região onde dois neurônios entram em contato e através da qual os impulsos nervosos são transmitidos entre eles. Os impulsos recebidos por um neurônio A, em um determinado momento, são processados, e atingindo um dado limiar de ação, o neurônio A dispara, produzindo uma substância neurotransmissora que flui do corpo celular para o axônio, que pode estar conectado a um dendrito de um outro neurônio B. O neurotransmissor pode diminuir ou aumentar a polaridade da membrana pós-sináptica, inibindo ou excitando a geração dos pulsos no neurônio B. Este processo depende de vários fatores, como a geometria da sinapse e o tipo de neurotransmissor.

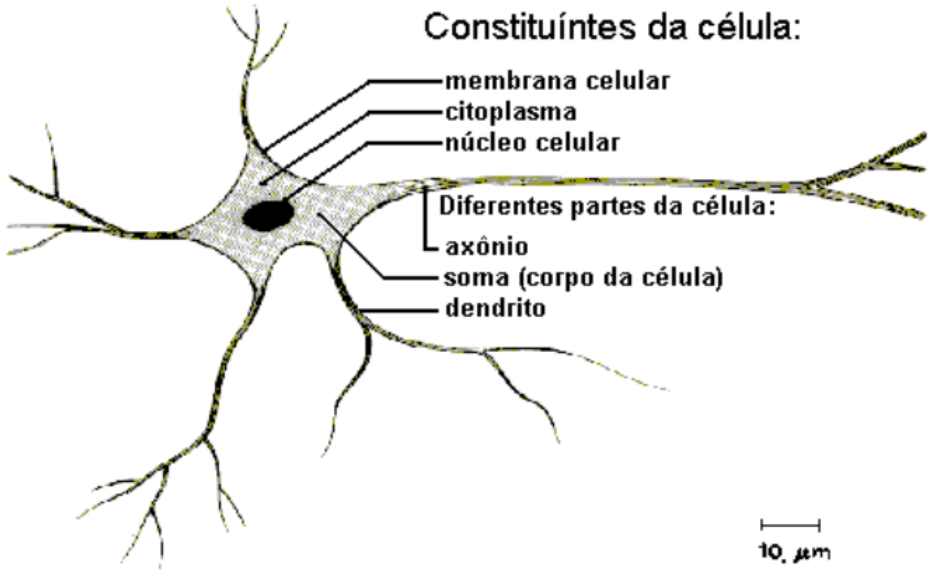


Figura 3.24: Constituintes da célula neuronal
(USP)

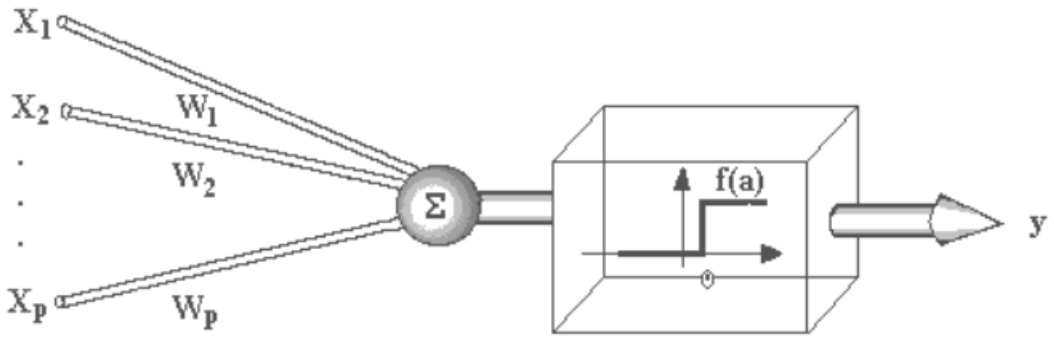


Figura 3.25: Esquema de Unidade McCullock - Pitts
(USP)

Uma rede neural artificial é composta por várias unidades de processamento, cujo funcionamento é bastante simples. Essas unidades, geralmente, são conectadas por canais de comunicação que estão associados a determinado peso. As unidades fazem operações apenas sobre seus dados locais, que são entradas recebidas pelas suas conexões. O comportamento inteligente de uma Rede Neural Artificial vem das interações entre as unidades de processamento da rede.

A operação de uma unidade de processamento, proposta por McCullock e Pitts em 1943, pode ser resumida da seguinte maneira: (1) Sinais são apresentados à entrada; (2) Cada sinal é multiplicado por um número, ou peso, que indica a sua influência na saída da unidade; (3) É feita a soma ponderada dos sinais que produz um nível de atividade; (4) Se este nível de atividade exceder um certo limite (threshold) a unidade produz uma determinada resposta de saída.

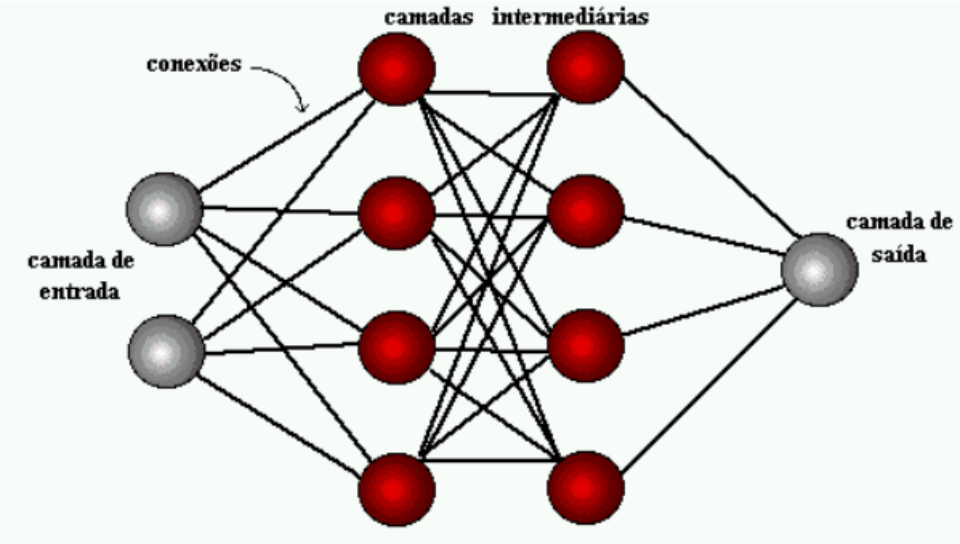


Figura 3.26: Organização em Camadas - Rede Neural
(USP)

A maioria dos modelos de redes neurais possui alguma regra de treinamento, onde os pesos de suas conexões são ajustados de acordo com os padrões apresentados. Em outras palavras, elas aprendem através de exemplos. Arquiteturas neurais são tipicamente organizadas em camadas, com unidades que podem estar conectadas às unidades da camada posterior.

Usualmente as camadas são classificadas em três grupos (Figura 3.26): (i) Camada de Entrada: onde os padrões são apresentados à rede; (ii) Camadas Intermediárias ou Escondidas: onde é feita a maior parte do processamento, através das conexões ponderadas; podem ser consideradas como extratoras de características; (iii) Camada de Saída: onde o resultado final é concluído e apresentado.

Uma rede neural é especificada, principalmente pela sua topologia, pelas características dos nós e pelas regras de treinamento. A propriedade mais importante das redes neurais é a habilidade de aprender de seu ambiente e com isso melhorar seu desempenho. Isso é feito através de um processo iterativo de ajustes aplicado a seus pesos, o treinamento. O aprendizado ocorre quando a rede neural atinge uma solução generalizada para uma classe de problemas. Denomina-se algoritmo de aprendizado a um conjunto de regras bem definidas para a solução de um problema de aprendizado. Existem muitos tipos de algoritmos de aprendizado específicos para determinados modelos de redes neurais, estes algoritmos diferem entre si principalmente pelo modo como os pesos são modificados.

Outro fator importante é a maneira pela qual uma rede neural se relaciona com o ambiente. Nesse contexto existem os seguintes paradigmas de aprendizado: (i) Aprendizado Supervisionado, quando é utilizado um agente externo que indica à rede a resposta desejada para o padrão de entrada; (ii) Aprendizado Não Supervisionado (auto-organização), quando não existe uma agente externo indicando a resposta desejada para os padrões de entrada; (iii) Reforço, quando um crítico externo avalia a resposta fornecida pela rede.

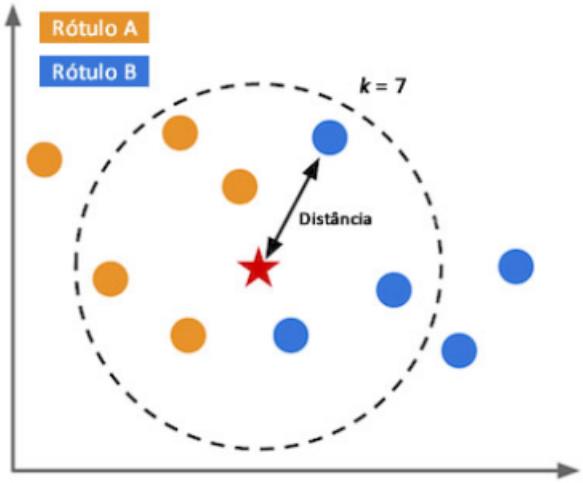


Figura 3.27: K Vizinhos Mais Próximos (KNN)
(Fukunaga and Narendra, 1975)

Denomina-se ciclo uma apresentação de todos os N pares (entrada e saída) do conjunto de treinamento no processo de aprendizado. A correção dos pesos num ciclo pode ser executada de dois modos: (1) Modo Padrão: A correção dos pesos acontece a cada apresentação à rede de um exemplo do conjunto de treinamento. Cada correção de pesos baseia-se somente no erro do exemplo apresentado naquela iteração. Assim, em cada ciclo ocorrem N correções. (2) Modo Batch: Apenas uma correção é feita por ciclo. Todos os exemplos do conjunto de treinamento são apresentados à rede, seu erro médio é calculado e a partir deste erro fazem-se as correções dos pesos.

3.8.4 K Vizinhos Mais Próximos (KNN)

O KNN foi proposto por Fukunaga e Narendra (Fukunaga and Narendra, 1975) em 1975. É um dos classificadores mais simples de ser implementado, de fácil compreensão e ainda hoje pode obter bons resultados dependendo de sua aplicação. A ideia central do KNN é determinar o rótulo de classificação de uma amostra com base nas amostras vizinhas derivadas de um conjunto de treinamento. Na Figura 3.27 há um problema de classificação com dois rótulos de classe e com k = 7. Neste caso, são aferidas as distâncias de uma nova amostra, representada por uma estrela, e as demais amostras de treinamento, representadas por esferas azuis e laranjas. A variável K representa a quantidade de vizinhos mais próximos que serão adotados para verificar a que classe esta nova amostra pertence. Desta forma, das sete amostras de treinamento mais próximas da nova amostra, 4 são do rótulo A (laranja) e 3 do rótulo B (azul). Logo, uma vez que existem mais vizinhos com rótulo A, a nova amostra recebera o mesmo rótulo deles, i.e., A. As duas métricas determinantes que devem ser defininas na aplicação do KNN são métrica de distância e o valor de k zinhos.

Quanto ao cálculo da distância existem várias métricas adotadas, variando de acordo com o problema. Todavia, a mais utilizada é a distância Euclidiana, descrita na equação abaixo.

$$(1) D_E(\mathbf{p},\mathbf{q}) = \sqrt{(p_1-q_1)^2+\cdots+(p_n-q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}$$

Outros exemplos incluem Minkowsky e Chebyshev, equações descritas abaixo.

$$(2) D_M(\mathbf{p},\mathbf{q}) = \left(\sum_{i=1}^{n}|p_i-q_i|^r\right)^{\frac{1}{r}} \; Minkowsky$$

$$(3) D_C(\mathbf{p},\mathbf{q}) = max_i(|p_i,q_i|) \; Chebyshev$$

Em todos os casos, $\mathbf{p} = (p_1,\cdots,p_n)$ e $\mathbf{q} = (q_1,\cdots,q_n)$ são dois pontos n-dimensionais e na Equação 2, r é uma constante que deve ser escolhida. Em nosso exemplo (Figura 3.27), estas distâncias seriam calculadas entre as bolas (azuis e laranjas) e a estrela (nova entrada).

Sobre a escolha do valor k, não existe um valor único para a constante, a mesma vai variar de acordo com a base de dados em que está sendo aplicada. A recomendação é pelo uso de valores primos (ou ímpares), mas o valor considerado ótimo vai variar de acordo com a base de dados.

3.8.5 Florestas Aleatória

Floresta Aleatória ou *Random Forest* (RF) (Breiman, 2001) é um algoritmo de aprendizado de máquina, especificamente o aprendizado supervisionado. Um dos algoritmos mais utilizados por usa precisão, flexibilidade e simplicidade. Em virtude de poder ser utilizado tanto para tarefas de classificação quanto para tarefas de regressão, mesclando sua caraterística linear, faz deste um algoritmo altamente adaptável a uma variedade de dados e contextos.

Proposto em 1995 por Tin Kam Ho, desenvolvendo uma fórmula para usar dados aleatórios para criar previsões. Em 2006, Leo Breiman e Adele Cutler estenderam o algoritmo e criaram florestas aleatórias como conhecido atualmente. Dito isso, essa tecnologia, a matemática e a ciência por trás dela, são relativamente novas.

O termo floresta faz referência ao desenvolvimento de uma árvore de decisão. Os dados dessa árvore são mesclados visando garantir as previsões mais precisas. Enquanto uma árvore de decisão individual tem um resultado e uma gama restrita de grupos, a floresta garante um resultado mais preciso, com um número maior de grupos e respectivas decisões (Figura 3.28. Seu benefício é adicionar aleatoriedade ao modelo, encontrando o melhor recurso e um subconjunto aleatorio de recursos.

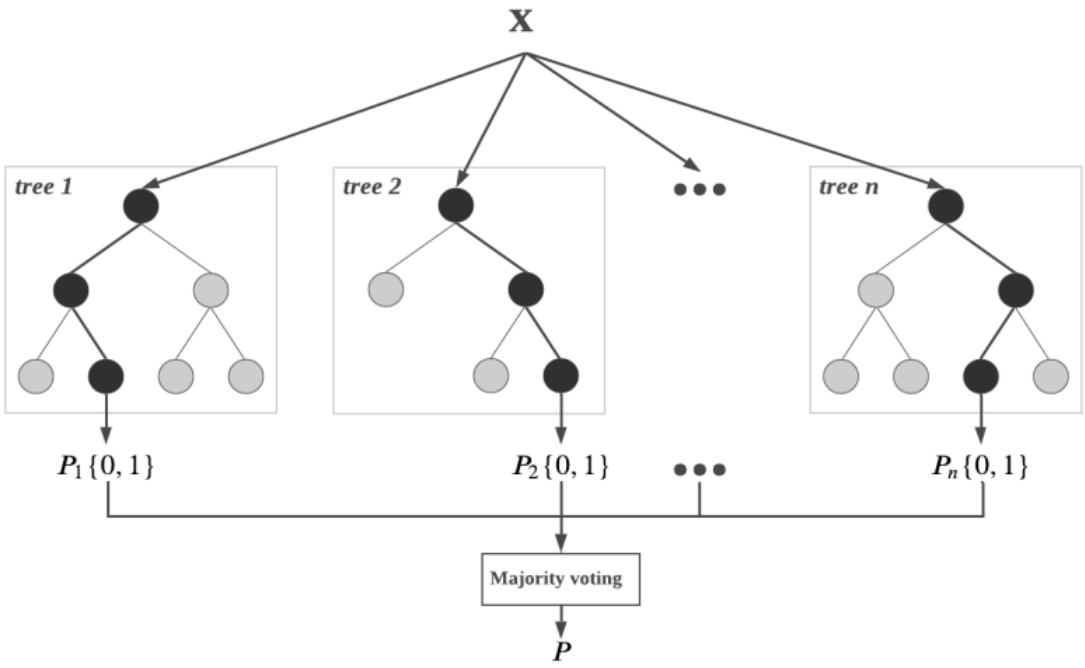


Figura 3.28: Floresta Aleatória (Random Forest)
(Shahhosseini)

3.8.6 Naïve Bayes (NB)

O Naïve Bayes (Murphy et al., 2006) é um dos modelos mais populares no aprendizado de máquina. Ao tomar como premissa a suposição de independência entre as variáveis do problema, o modelo de Naïve Bayes realiza uma classificação probabilística de observações, caracterizando-as em classes pré-definidas. Trata-se de um modelo adequado para classificação de atributos discretos, tem aplicações na análise de crédito, diagnósticos médicos ou busca por falhas em sistemas mecânicos. O termo *naïve*, (do inglês, ingênuo) se refere à premissa central do algoritmo de que os atributos considerados são não correlacionados entre si. Vale ressaltar que este modelo é um dos mais conhecidos a aplicar o conceito de probabilidade, fazendo uso do teorema de Bayes como princípio fundamental. Teorema este desenvolvido pelo estatístico, filósofo e ministro presbiteriano Thomas Bayes (1701 - 1761).

$$P(E)P(E) = P(H)P(H)$$

Onde $P(H)$ é a probabilidade de que uma hipótese (H) seja verdadeira, antes de qualquer evidência (E) ser conhecida. O termo $P(H)$ é a probabilidade de observarmos uma evidência (E), dado que a hipótese (H) é verdadeira. Por outro lado, $P(E)$ é a probabilidade de observação da evidência (E) e $P(E)$ é a probabilidade de que a hipótese (H) seja verdadeira, dada a evidência observada (E).

4

Modelo Proposto

Nesta seção será apresentado o modelo proposto para o prognóstico de morbidade e mortalidade em pacientes diagnosticadoscom taquicardia. Descrição de seu objetivo, relevância da proposta, contexto de aplicação, definição do experimento, descrição dos cenários utilizados.

4.1 Modelo

A proposta deste trabalho visa realizar predições quanto à morbidade e mortalidade em pacientes diagnosticadoscom taquicardia, caso onde os batimentos por minuto (bpm) ficam acima do normal. Tomamos como hipótese, que o paciente diagnosticado com arritmia, seja do tipo taquicardia ventricular, com agravamento para insuficiência cardíaca e consequente fibrilação ventricular, como podemos observar na Figura 4.2. A predição da morbidade será baseada em uma relação com os scores CHA_2DS_2-VASc(Figura 3.1) e Hábitos do paciente (Tabagismo, IMC e estresse). Os scores serão normalizados e por definição categorizados por quartis, de acordo com a classificação NYHA(Bredy et al., 2018). A predição da morbidade irá retornar como os sintomas se apresentam no paciente, desde assintomáticos (NYHA I), variando para levemente assintomático (NYHA II), sintomáticos (NYHA III), até os sintomáticos em repouso (NYHA IV). A mortalidade por sua vez será obtida a partir das classificações do NYHA, que possui taxas de mortalidade definidas por padrão, em casos onde não se é adotado nenhum tratamento, e em casos onde o tratamento adequado é adotado. Trazendo aos cardiologistas uma ferramenta que possa auxiliar no contexto de tomadas de decisão dado o prognóstico obrido.

4.1.1 Definição do Experimento

A estratégia utilizada no desenvolvimento da experimentação deste modelo consiste em:

- Identificar indicadores do score CHA_2DS_2-VASc.

- Identificar fatores de risco das arritmias, em específico, taquicardias, associados aos hábitos.
- Identificar uma base de dados satisfatória para o cenário, ou seja, uma base que contenha o maior número possível de variáveis equivalentes aos indicadores dos scores adotados.
- Tratar a base selecionada, desde o pré-processamento, com limpeza e enriquecimento dos dados dispostos, até a transformação dos mesmos. Dispondo assim de variáveis nos parâmetros do modelo (Dhar, 2013).

  Etapa de limpeza deve eliminar colunas que não possuam informações relevantes, como tipo de trabalho, estado civil, id, dentre outros. Eliminar as linhas que possuam alguma coluna sem valores (nan).

  O enriquecimento dos dados, visa gerar, a partir dos dados já existentes, informações que alimentem a necessidade dos parâmetros do modelo. Como, por exemplo, assumir o ambiente enquanto urbano ou rural, como um indicador de estresse, uma vez que estudos apontam que a vida urbana tende a ser mais estressante que a vida em ambiente rural. (Jaidka et al., 2021).

  Transformar os dados para se adequar aos parâmetros do modelo. Por exemplo, transformar as informações sobre peso e altura, em indice de massa corporal (imc), e posteriormente utilizar o valor adquirido para classificar entre sobrepeso, obeso ou outros, com seus respectivos pesos os pesos adotados no modelo.
- Uma vez gerados os scores CHA_2DS_2-VASc e Hábitos, atribuir pesos para uma média ponderada entre estes valores, seguindo sua representatividade no cenário. Por exemplo, (CHA_2DS_2-VASc x 0.7) + (HÁBITOS x 0.3).
- Categorizar classe NYHA com base no score obtido, a partir de uma distribuição pré definida das 4 categorias. Por exemplo, distribuição em quartis (0-25%, 26-50%, 51-75%, 75-100%).
- Uma vez classificados por NYHA, será possível predizer a mortalidade do mesmo, de acordo com a categoria e com os tratamentos adotados.
- O valor do score CHA_2DS_2-VASc fornece o risco de AVC, AIT ou embolia sistemica, baseados em Lip et al. (2010) e Friberg et al. (2012a).

4.1.2 Relevância da Proposta

As informações obtidas pelo modelo proposto, extraem a informação de insuficiência cardíaca com base nos demais dados do CHA_2DS_2-VASc, gerando scores, que associados com os hábitos vão fornecer ao médico cardiologista, e ao próprio paciente, uma maior nitidez no prognóstico da taquicardia, dado um cenário específico, onde os portadores da taquicardia evoluam

para taquicardia ventricular (TV), e avancem para insuficiência cardíaca (IC) e consequente fibrilação. Com a visão do prognóstico, o médico cardiologista será capaz de tomar decisões acerca de que medicamentos adotar, em que dosagens, em que nível deve exigir as mudanças de hábitos do paciente avaliado, bem como acompanhar progresso ou regresso em novas execuções/avaliações. Portanto, uma vez validado por um corpo médico composto por cardiologistas, tende a se tornar uma ferramenta de auxílio importante para aquisição de prognósticos, inicialmente de taquicardia, mas adaptável a outros tipos, desde que assumidos os scores, riscos, e peculiaridades de cada um deles.

4.2 Descrição do Cenário Utilizado

O campo da cardiologia é vasto, composto por diversas doenças, cabendo ao modelo em questão, restringir o campo de estudo. Optou-se pelas arritmias, o conjunto de doenças cuja as alterações nos ritmos dos batimentos geram problemas ao organismo do portador. Seus três tipos clássicos incluem a bradicardia (batimentos inferiores aos 60bpm), taquicardia (batimentos superiores aos 100bpm) e fibrilação (batimentos fora de padrão). A escolha pela taquicardia, surge pela busca de um contexto de prognóstico em um intervalo de médio a longo prazo. Uma vez que pacientes com bradicardia, em sua maioria, possuem um prognóstico positivo, convivendo com aquela doença por longos prazos. As fibrilações já consideram cenários mais críticos, com impactos a curto prazo. Em geral, ocorre em unidades de tratamento intensivo (UTI), onde pacientes com a FA precisam de abordagens mais imediatas. Por fim, a taquicardia, casos que são abordados por fármacos, mudanças de hábitos, e que precisam ser acompanhados de perto em seu progresso ou regresso. Podemos visualizar o fluxo destas etapas na Figura 4.1.

Em nosso cenário, consideramos um paciente diagnosticado com taquicardia, sendo esta de tipo maligno, como uma Taquicardia Ventricular (TV). A TV tem grandes chances de evoluir para uma fibrilação ventricular (FV), onde exatamente se aplica o score CHA_2DS_2-VASc, mensurando o risco de ataque vascular cerebral (AVC), ataque isquêmico transitório (AIT) e embolia sistêmica. Considerando que a FV pode evoluir para uma insuficiência cardíaca (IC), utilizamos dos dados dispostos do CHA_2DS_2-VASc, onde a própria Insuficiência Cardíaca está relacionada, para mensurar o risco do paciente desenvolver a IC (Figura 4.2. Assumindo o subconjunto dos pacientes que, pelos nossos calculos, irão desenvolver a IC, relacionamos seu score CHA_2DS_2-VASc ao score dos Hábitos e demais fatores de risco, gerando um score que será categorizado de acordo com o NYHA, aplicável exatamente aos casos de IC. Tal categorização estará a prover o nível de morbidade do paciente, em relação a como os sintomas serão apresentados: assintomático (I), levemente assintomático (II), sintomático (III) e sintomático em repouso (IV). Cada uma destas categorias citadas anteriormente já carregam sua taxa de mortalidade definidas na literatura.

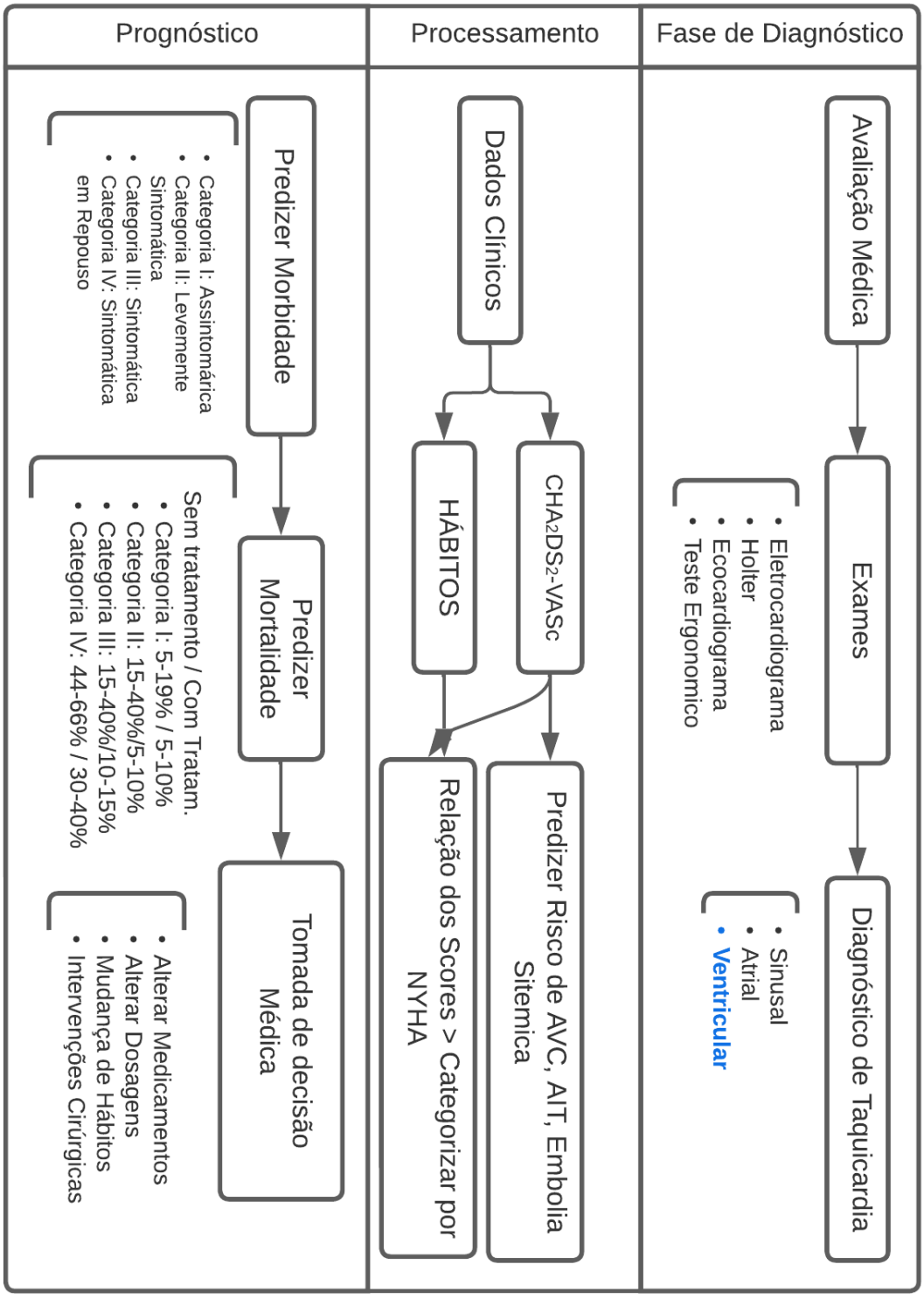


Figura 4.1: Fluxo de Etapas/Estágios do Modelo

CHA_2DS_2-VASc score	Annual stroke risk (%) Friberg 2012	Lip 2010
0	0.2	0.0
1	0.6	0.6
2	2.2	1.6
3	3.2	3.9
4	4.8	1.9
5	7.2	3.2
6	9.7	3.6
7	11.2	8.0
8	10.8	11.1
9	12.2	100

Figura 4.2: Fluxo de Prognóstico da Morbidade e Mortalidade Taquicardia Ventricular

5

Metodologia

Nesta seção será descrita a metodologia do experimento, desde as ferramentas utilizadas, linguagens de implementação, frameworks e bibliotecas, até a base de dados utilizada, transformações realizadas, alterações no na estrutura do dataset, informações obtidas, descrição dos parâmetros e métricas utilizadas.

5.0.1 Base de Dados

Na busca pela base de dados ideal houveram algumas dificuldades. Algumas tinham seus dados com acesso restrito, outras tinham muitos dados faltantes em relação aos parâmetros necessários ao modelo. Uma abordagem pessoal nos hospitais da cidade trouxe uma realidade com a presença de muitos dados físicos, anotações em papéis, e afins. Bem como os prontuários eletrônicos dos pacientes (PEPs), apresentados pelo corpo médico que participou da pesquisa, se apresentou em uma estrutura muito genérica, sem as especifidades necessárias para o nosso modelo. A base adotada no modelo proposto e que mais se aproximou da necessidade de nosso cenário, neste momento, foi a base de Liu et al. (2019), um conjunto de dados médicos não estruturados que foi utilizado para predição de ataque vascular cerebral (AVC) a partir de uma abordagem híbrida de aprendizagem de máquina. A base dedados consiste em 12 atributos de 5110 pacientes. Os atributos incluem:

- Hipertensão
- Idade
- Doença Cardíaca
- Nível de Glicose
- AVC prévio

- Tipo de Residência: Rural ou Urbana
- Gênero: Masculino, Feminino e Outro
- Fumante: Fuma, Ex Fumante, Nunca Fumou
- IMC: Índice de Massa Corporal
- Tipo de Trabalho
- Se já foi casado
- Identificador

5.0.2 Ferramentas

Durante o desenvolvimento do modelo proposto foram utilizadas as seguintes ferramentas:

- Linguagem de programação **Python** (versão 3.9) (Van Rossum and Drake Jr, 1995). Trata-se de uma linguagem de programação de alto nível, criada pelo holandês Van Russum. Dentre suas características estão a liberdade por ser um código aberto e gratuito, a disponibilidade em diversos sistemas operacionais (Windows, Linux, MacOS, Smartphones), clareza com sua sintaxe limpa e direta, tipagem dinâmica, multiparadigma (Orientado à objetos, funcional e procedural), compilada e interpretada, costuma aumentar produtividade justamente pelo conjunto de bibliotecas disponíveis.
- IDE **Spyder** (versão 5.2.2) (Raybaut, 2009),é um ambiente de desenvolvimento integrado de plataforma cruzada de código aberto para programação científica na linguagem Python
- **Numpy** é uma biblioteca para linguagem de programação python que suporta o processamento de grandes, multi-dimensionais arranjos e matrizes, juntamente com uma grande coleção de funções matemáticas de alto nível para operar sobre estas matrizes.
- **Pandas** (pandas development team, 2020) é uma biblioteca de software criada para a linguagem Python para manipulação e análise de dados. Em particular, oferece estruturas e operações para manipular tabelas numéricas e séries temporais. É software livre sob a licensa licença BSD.
- **Scikit-learn** (Pedregosa et al., 2011) é uma biblioteca de aprendizado de máquina de código aberto para a linguagem de programação Python.
- **Orange Data Mining** (Demšar et al., 2013) é um kit de ferramentas de visualização de dados, aprendizado de máquina e mineração de dados de código aberto. Ele apresenta um front-end de programação visual para análise exploratória de dados qualitativos rápida e visualização interativa de dados.

Index	gender	age	hypertension	heart disease	stress	avg glucose level	bmi	smoking status	stroke
0	Male	67	0	1	1	228.69	36.6	formerly smoked	1
2	Male	80	0	1	0	105.92	32.5	never smoked	1
3	Female	49	0	0	1	171.23	34.4	smokes	1
4	Female	79	1	0	0	174.12	24	never smoked	1
5	Male	81	0	0	1	186.21	29	formerly smoked	1
6	Male	74	1	1	0	70.09	27.4	never smoked	1
7	Female	69	0	0	1	94.39	22.8	never smoked	1
10	Female	81	1	0	0	80.43	29.7	never smoked	1
11	Female	61	0	1	0	120.46	36.8	smokes	1
12	Female	54	0	0	1	104.51	27.3	smokes	1
14	Female	79	0	1	1	214.09	28.2	never smoked	1
15	Female	50	1	0	0	167.41	30.9	never smoked	1
16	Male	64	0	1	1	191.61	37.5	smokes	1
17	Male	75	1	0	1	221.29	25.8	smokes	1

Figura 5.1: Base de dados após remoção de elementos irrelevantes e aquisição da variável estresse

5.0.3 Tratamento dos Dados

Dada a base original foram necessários tratamentos, como limpeza, enriquecimento, transformação e parametrização dos dados. Com o objetivo não só de descartar as variáveis sem relevância, mas também com o intuito de adquirir novas informações baseados na literatura e no que tínhamos a nossa disposição. Por fim, adequar os dados aos parâmetros dos scores e técnicas adotadas (Figura 5.2)(Holanda).

Remoção de Dados Irrelevantes

Inicialmente a base conta com 12 descritores (colunas) de 5110 pacientes (linhas). Durante a primeira etapa foram removidas as linhas que tinham qualquer dado faltante entre as 12 colunas, onde foram removidos um total de 201 pacientes (linhas), restando agora 4909 pacientes. Nesse momento removemos as colunas que tinham dados irrelevantes para nossa finalidade. Foram os casos das colunas 'id' (identificador), '*ever-merried*'(já foi casado) e 'work-type' (tipo de trabalho). Apóps a remoção destas 3 colunas temos como resultado a estrutura de (linhas, colunas) = (4909, 9). Optou-se por remover elementos cujo genero atribuido fosse 'other', apenas 1 caso existente, restando (4908, 9). Foram removidos um total de 1483 paciente, que estavam com o status de 'smoke'(fumante) como 'unknow' (desconhecido), restando (3425, 9).

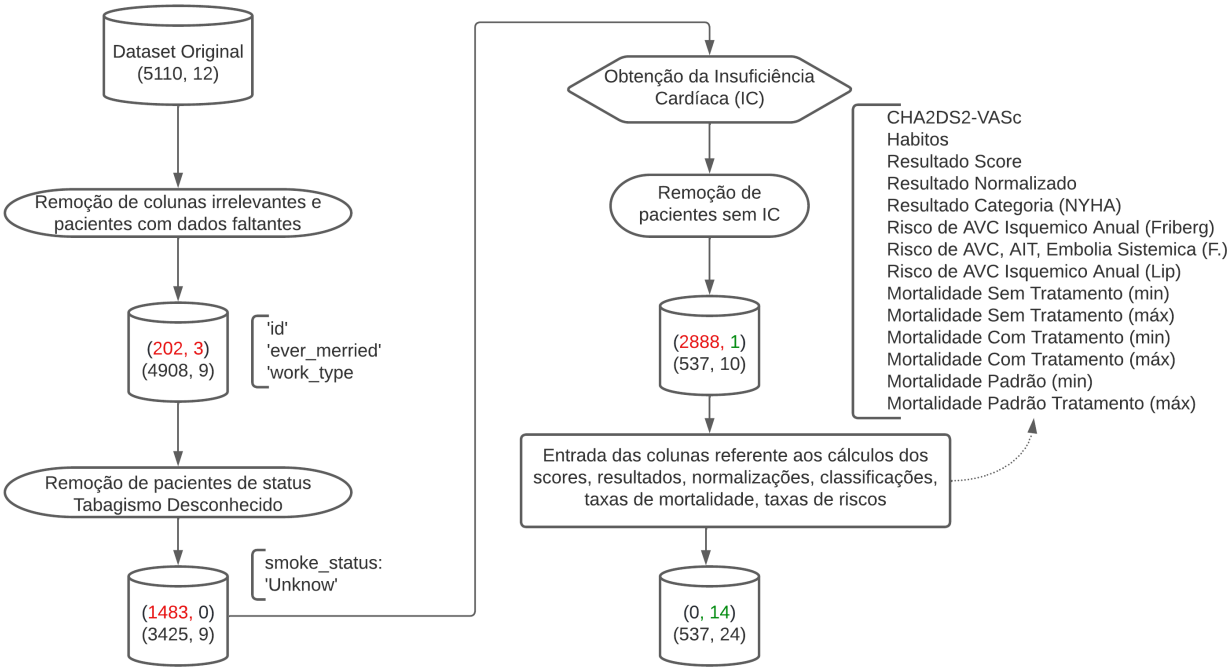


Figura 5.2: Tratamento dos Dados

Algoritmo 5.1: Limpeza dos Dados

```
1  #remove rows with any nan value
2  df = df.dropna()
3
4  # Remove id, ever-married, work-type
5  df = df.drop(['id'], axis=1)
6  df = df.drop(['ever_married'], axis=1)
7  df = df.drop(['work_type'], axis=1)
8
9  # remove rows with gender value 'Other'
10 df = df[df['gender'] != 'Other']
11
12 # Remove smoking-status 'Unknow'
13 df = df[df['smoking_status'] != 'Unknown']
```

Transformação dos Dados

Uma vez finalizada a limpeza dos dados, faz-se necessário a transformação de alguns dados, como tornar dados descritivos em dados binários, ou categóricos. Toda transformação é baseada na necessidade do modelo, consequentemente na necessidade dos scores utilizados: CHA_2DS_2-VASc (3.1) e Hábitos(Feitosa et al., 2002).

Embasados no artigo de Webb and Collette (1979), inferimos da variável 'residence-type', que recebia 'Urbano' ou 'Rural', em 1 e 0 respectivamente, como parâmetros da nova varíavel estresse. Neste momento, temos a base de dados como representada na Figura Transformamos o gênero 'Male' em zero, e 'Female' em 1, uma vez que no score CHA_2DS_2-VASc é atribuído um peso maior quando se trata de um paciente do gênero feminino. Transforma-se a idade, onde pacientes com 75 ou mais anos de idade recebem peso 2, pacientes entre 65 e 74 recebem peso 1, e aqueles que tiverem menos de 65 anos, tem peso zero. Transformamos a média do nível de glicose, em diabetes, uma vez que glicose maior ou igual a 126mg/dl indica diagnóstico de diabetes(1), e valores abaixo deste (0).

Renomeamos as variáveis para o português, e para o indicador do score, quando necessário: *bmi (body mass index)* para imc (índice de massa corporal), *stroke* para avc, *age* para idade, *gender* para sexo, *hypertension* para hipertensao, *heart-disease* para doenca-vascular, *stress* para estresse, *smoking-status* para tabagismo, e *avg-glucose-level* para diabetes.

Algoritmo 5.2: Transformação dos Dados

```
# Gender
df = df.replace('Male', 0)
df = df.replace('Female', 1)

# Set score according age of pacient cha2ds2-vasc
df.loc[df.age < 65, 'age'] = 0
df.loc[((df.age >= 65) & (df.age < 75)), 'age'] = 1
df.loc[df.age >= 75, 'age'] = 2


#categoriza diabetes de acordo com o n vel de glicose
df['avg_glucose_level'] = np.where((df.avg_glucose_level < 126),0,
df.avg_glucose_level)
df['avg_glucose_level'] = np.where((df.avg_glucose_level >= 126),1,
df.avg_glucose_level) #Diabetes


# Set score according smoking of pacient
df['smoking_status'] = np.where((df.smoking_status=='smokes'),
2,df.smoking_status) #Fumante
df['smoking_status'] = np.where((df.smoking_status=='formerly␣smoked'),
1,df.smoking_status) #Ex Fumante
df['smoking_status'] = np.where((df.smoking_status == 'never␣smoked'),
0,df.smoking_status)


```

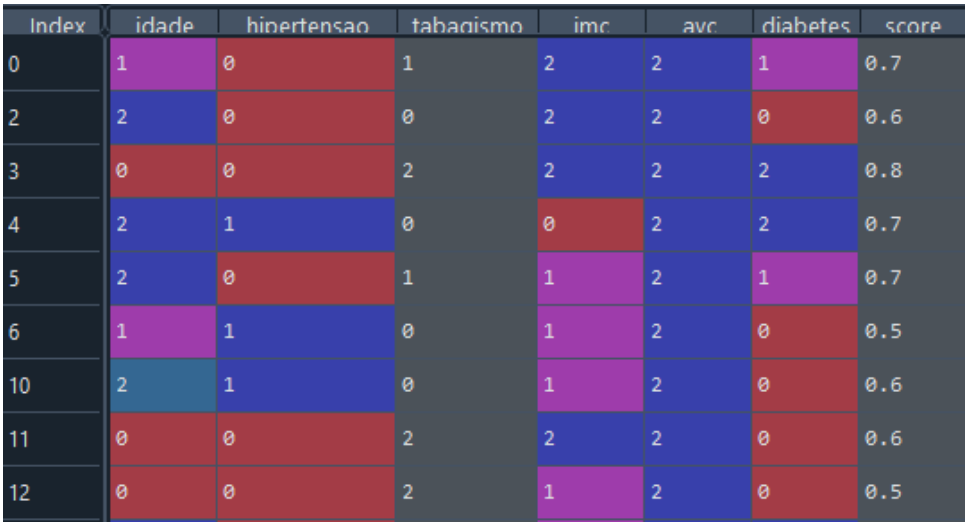
```
27 # categorizar bmi para se o valor for entre 25 e 30    sobrepeso
28 df['bmi'] = np.where((df.bmi < 25),0,df.bmi)
29 #Sobrepeso
30 df['bmi'] = np.where(((df.bmi >= 25) & (df.bmi <= 30)),1,df.bmi)
31 #Obesidade
32 df['bmi'] = np.where((df.bmi > 30),2,df.bmi)
```

Aquisição da Varíavel Insuficiência Cardíaca

Dentre todas as variáveis necessárias para o CHA_2DS_2-VASc, apenas a Insuficiência Cardíaca (C) não está presente. Entretanto, a literatura aponta que existe forte correlação entre as demais variáveis do CHA_2DS_2-VASc com a própria IC. A própia idade e gênero (Masoudi et al., 2003; Nieminen et al., 2008). Gênero este que relacionado à diabetes tem forte impacto (Kenchaiah et al., 2002), uma vez que homens com diabetes tem duas vezes mais chances de desenvolver a IC, enquanto que as mulheres que possuem diabetes, tem quatro vezes mais chances de desenvolver a IC. Também existe correlação entre outras doenças cardíacas (Silverberg et al., 2006) e a insuficiência cardíaca. O tabagismo (Aune et al., 2020; Ahmed et al., 2015)por causar o estreitamento das válvulas, também tem forte correlação com a insuficiência cardíaca. A hipertensão (Lenfant, 1996)também está relacionada ao risco de desenvolvimento de uma IC. Por fim a relação entre o ataque vascular cerebral (AVC) (Mozaffarian et al., 2016) e a insuficiência cardíaca (IC). Com todas estas variáveis em mãos, podemos inferir a IC no paciente avaliado. E efetuamos isto com as seguintes variáveis e pesos:

- **Idade**: 0, 1 ou 2; Seguindo CHA_2DS_2-VASc
- **Gênero**: 1 se feminino, 0 se masculino
- **Hipertensão**: 1 se tiver, se não 0
- **Tabagismo**: se fumante 2, se ex fumante 1, se nunca fumou 0
- **IMC**: 1, se IMC entre 25 e 30, por indicar **sobrepeso**. 2 se IMC maior que 30, o que indica **obesidade**
- Ataque Vascular Cereberal (**AVC**), se já teve 1, se não 0
- **Diabetes correlacionada ao gênero**, possui maior peso quando mulher. Se mulher 2, se homem 1

O índice de corte definido foi **0.5**. Nesta etapa foi utilizado PCA para identificar as variáveis principais, para definir a distribuição dos pesos. Foram mantidos os pesos pré definidos na Tabela 3.1 para os indicadores do CHA_2DS_2-VASc, com exceção da diabetes, que teve seu peso



Index	idade	hipertensao	tabagismo	imc	avc	diabetes	score
0	1	0	1	2	2	1	0.7
2	2	0	0	2	2	0	0.6
3	0	0	2	2	2	2	0.8
4	2	1	0	0	2	2	0.7
5	2	0	1	1	2	1	0.7
6	1	1	0	1	2	0	0.5
10	2	1	0	1	2	0	0.6
11	0	0	2	2	2	0	0.6
12	0	0	2	1	2	0	0.5

Figura 5.3: Base para Cálculo da Insuficiência Cardíaca

associado ao gênero (chances dobram para pacientes homens com diabetes, e chances quadruplicam para pacientes do gênero feminino com diabetes). Podemos observar os indicadores para o cálculo da IC, e o valor resultante na Figura 5.3. Com os resultados do cálculo normalizado em relação à insuficiência cardíaca, foi feito um filtro removendo todos os pacientes que estivessem abaixo do índice de corte (0.5). Para os 537 pacientes que restaram, com valores iguais ou acima do índice de corte, foi atribuída uma nova coluna 'ic' (insuficiência cardíaca) com valor 1. Com isso, a nova estrutura dos dados apresenta 537 linhas e 7 colunas (Figura 5.2).

Algoritmo 5.3: Insuficiência Cardíaca: Cálculo

```
1  #Soma variaveis para calcular ic
2  ic_calc['score'] = ic_calc['hipertensao'] + ic_calc['idade'] +
3  ic_calc['avc'] + ic_calc['diabetes'] + ic_calc['imc']
4  + ic_calc['tabagismo']
5
6  column = 'score'
7  ic_calc[column] = (ic_calc[column] - ic_calc[column].min()) /
8  (ic_calc[column].max()- ic_calc[column].min())
9  #verifica peso das variaveis
10 pca = PCA(n_components=1)
11 pca.fit(ic_calc)
12 pca_resultado = pca.components_
13
14 #Define corte para classificar IC
15 ic_index = 0.5
16 #remove aqueles com index inferior a 0.5
17 ic_calc = ic_calc.drop(ic_calc[ic_calc.score < ic_index].index)
```

```
18
19 new_df_index_array = ic_calc.index.tolist()
20 df = df.filter(items = new_df_index_array, axis=0)
21 #atribui novo campo de ic com valor 1,
22 #sem ic ja foi removido
23 df['ic'] = 1
```

5.0.4 Score CHA_2DS_2-VASc

A insuficiência cardiaca tem como uma de suas consequências a evolução para Fibrilação Ventricular (FV), e é justamente este o cenário que tomamos como hipotese (Figura 4.2).Uma vez que temos todas as variáveis do CHA_2DS_2-VASc transformadas para os seus respectivos pesos (Tabela 3.1), podemos atribuir o score de cada paciente, e respaldados pela literatura, podemos atribuir as taxas de risco de AVC, AIT ou Embolia Sitemica de acordo com Friberg et al. (2012a) e Lip et al. (2010). Podemos ver as taxas de risco nas Tabelas 3.2 e 3.3. Abaixo segue atribuição das Taxas referente ao risco de AVC Isquêmico de acordo com Friberg et al. (2012a). O mesmo foi feito para AIT/Embolia Sistêmica, também por Friberg, e para AVC Isquêmico por Lip et al. (2010).

Algoritmo 5.4: Taxas referente ao risco de AVC Isquêmico de acordo com Friberg et al. (2012a)

```
1  #CLASSIFICA O CHA2DS2-VASc (Friberg 2012)
2  df['f_risco_avc_Isqu mico_anual'] = 0
3  df['f_risco_avc_Isqu mico_anual'] = np.where
4  ((df.cha2ds2_vasc == 0),0.2,df.f_risco_avc_Isqu mico_anual)
5  df['f_risco_avc_Isqu mico_anual'] = np.where
6  ((df.cha2ds2_vasc == 1),0.6,df.f_risco_avc_Isqu mico_anual)
7  df['f_risco_avc_Isqu mico_anual'] = np.where
8  ((df.cha2ds2_vasc == 2),2.2,df.f_risco_avc_Isqu mico_anual)
9  df['f_risco_avc_Isqu mico_anual'] = np.where
10 ((df.cha2ds2_vasc == 3),3.2,df.f_risco_avc_Isqu mico_anual)
11 df['f_risco_avc_Isqu mico_anual'] = np.where
12 ((df.cha2ds2_vasc == 4),4.8,df.f_risco_avc_Isqu mico_anual)
13 df['f_risco_avc_Isqu mico_anual'] = np.where
14 ((df.cha2ds2_vasc == 5),7.2,df.f_risco_avc_Isqu mico_anual)
15 df['f_risco_avc_Isqu mico_anual'] = np.where
16 ((df.cha2ds2_vasc == 6),9.7,df.f_risco_avc_Isqu mico_anual)
17 df['f_risco_avc_Isqu mico_anual'] = np.where
18 ((df.cha2ds2_vasc == 7),11.2,df.f_risco_avc_Isqu mico_anual)
19 df['f_risco_avc_Isqu mico_anual'] = np.where
20 ((df.cha2ds2_vasc == 8),10.8,df.f_risco_avc_Isqu mico_anual)
```

```
21 df['f_risco_avc_Isqu mico_anual'] = np.where
22 ((df.cha2ds2_vasc == 9),12.2,df.f_risco_avc_Isqu mico_anual)
```

5.0.5 Aplicação do Classificador NYHA

Uma vez que estamos em posse de uma base com portadores de IC (Insuficiência Cardíaca), podemos então aplicar o classificador NYHA (Bredy et al., 2018)(Figura 4.2), uma vez que este foi feito para pacientes portadores de IC, sendo capaz de predizer os níveis de morbidade e mortalidade. Nosso modelo propõe utilizar uma relação entre os scores CHA_2DS_2-VASc e hábitos para posterior categorização por NYHA. Com a necessidade de definir uma equação capaz de relacionar estes scores dentro de sua representatividade, foi definido que o score CHA_2DS_2-VASc teria peso 7, enquanto o score dos hábitos peso 3. A justificativa para tal peso vem do fato de o score CHA_2DS_2-VASc ser gerado a partir de todos os seus indicadores (Tabela 3.1), que se fazem presentes na base processada, enquanto que o score dos hábitos, é composto apenas com base nos indicadores estresse, tabagismo, sobrepeso e obesidade, ficando desfalcado de indicadores como o alcoolismo, predisposição genética, sedentarismo, apneia do sono, distúrbios da tireoide, dentre outros. Segue abaixo a equação definida relacionando os scores e seus respectivos pesos.

$$NYHA_Category = (CHA_2DS_2\text{-VASc x 0,7}) + (\text{HABITS x 0,3})$$

O resultado obtido nesta equação foi normalizado, dentro de um valor entre 0 e 10. Foi definido que a distribuição para posterior classificação de NYHA, seria feita em quartis. Onde a categorização segue o definido abaixo:

- De 0 a 2.5% categoria NYHA I
- De 2.6 a 5.0% categoria NYHA II
- De 5.1 a 7.5% categoria NYHA III
- De 7.6 a 10.0% categoria NYHA IV

Cada uma destas categorias possui sua taxa de mortalidade definida na literatura. Em cenários com tratamento, sem tratamento, e um cenário "padrão", onde não se assume se o tratamento está sendo adotado ou não. As taxas podem ser observadas na Tabela 3.5 (Dekerlegand, 2007; Uretsky and Sheahan, 1997). No Algoritmo 5.5 podemos observar a execução dos passos na linguagem de programação Python(Van Rossum and Drake Jr, 1995), aplicando a equação que relaciona os 2 scores, normalizando o valor obtido, e categorizando cada um deles por quartil.

Algoritmo 5.5: Cálculo Normalização e Categorização por NYHA

```
# resultado da soma
df['resultado_score'] =
(df['cha2ds2_vasc'] * 0.7) + (df['habitos'] * 0.3)
df['resultado_normalizado'] = df['resultado_score']

# # normalizar resultados <<<<
column = 'resultado_normalizado'
min_value = 0;
max_value = 13; #atribuir maior obtido pelos scores
df[column] = (df[column] - df[column].min()) /
(df[column].max()- df[column].min())
df[column] = df[column] * 10

pca.fit(df)
pca_resultado_df = pca.components_

# cria novas categorias para inferencia
df.loc[df['resultado_normalizado'] <= 2.5,
'resultado_categoria'] = 'I'

df.loc[(df['resultado_normalizado'] > 2.5)
& (df['resultado_normalizado'] <= 5),
'resultado_categoria'] = 'II'

df.loc[(df['resultado_normalizado'] > 5)
& (df['resultado_normalizado'] <= 7.5), '
resultado_categoria'] = 'III'

df.loc[df['resultado_normalizado'] > 7.5,
'resultado_categoria'] = 'IV'
```

Abaixo temos a atribuição em Python, para o intervalo de mortalidade padrão de acordo com as categorias NYHA (Bredy et al., 2018). O mesmo foi feito para atribuir taxas de mortalidade com e sem tratamento e podem ser observadas no código completo presente no Apendice.

Algoritmo 5.6: Taxas de Mortalidade com Base na Classificação NYHA - Caso Padrão

```
# Atribui mortalidade com base na categoria (Padr o)
df['nyha_mortality_min'] = 0
df['nyha_mortality_max'] = 0
df['nyha_mort_com_tra_min'] = np.where((df.resultado_categoria
== 'I'),0.10,df.mort_com_tra_min)
df['nyha_mortality_max'] = np.where((df.resultado_categoria
```

```
7  == 'I'),0.15,df.mort_com_tra_max)
8
9  df['nyha_mortality_min'] = np.where((df.resultado_categoria
10 == 'II'),0.10,df.mort_com_tra_min)
11 df['nyha_mortality_max'] = np.where((df.resultado_categoria
12 == 'II'),0.15,df.mort_com_tra_max)
13
14 df['nyha_mortality_min'] = np.where((df.resultado_categoria
15 == 'III'),0.15,df.mort_com_tra_min)
16 df['nyha_mortality_max'] = np.where((df.resultado_categoria
17 == 'III'),0.20,df.mort_com_tra_max)
18
19 df['nyha_mortality_min'] = np.where((df.resultado_categoria
20 == 'IV'),0.20,df.mort_com_tra_min)
21 df['nyha_mortality_max'] = np.where((df.resultado_categoria
22 == 'IV'),0.50,df.mort_com_tra_max)
```

5.0.6 Orange Data Mining

Uma vez que os dados já foram tratados, estamos sob posse de um modelo que é capaz de, com base em dados sobre gênero, idade, hipertensão, doença vascular, tipo de moradia, nível de glicose, índice de massa corporal, hábito de fumar e AVC prévio, gerar um índice de potencial desenvolvimento de Insuficiência Cardíaca, com base nos scores adotados (CHA_2DS_2-VASc e Hábitos). A partir disto, os eminentes portadores da IC, são classificados na categoria NYHA, com base na relação dos dois scores citados. Logo, temos um modelo que apesar de prover outras informações, como risco de AVC/AIT ou embolia sistêmica anual, por CHA_2DS_2-VASc, e a própria mortalidade, por NYHA, tem como alvo de validação, a categoria NYHA obtida.

Para validar o modelo, foi adoto o framework **Orange Data Mining** (Demšar et al., 2013)(Holanda), um kit de ferramentas de visualização de dados, aprendizado de máquina e mineração de dados de código aberto. Ele apresenta um front-end de programação visual para análise exploratória de dados qualitativos rápida e visualização interativa de dados. Nesta ferramenta foi possível observar as distribuições dos dados, tanto na base original, quanto na base processada, observando graficamente a distribuições de cada indicador tratado. Com o objetivo de validar o modelo proposto, dentro da base de dados obtida após processamento, definimos o objetivo do modelo (*target*), que em nosso caso se trata da categoria NYHA (Figura 5.4). E atribuimos aos modelos, as variáveis que estiveram dispostas para aquisição dessa informação (variáveis estas carregando seus pesos de acordo com os scores adotados). A validação do modelo foi feita utilizando os seguintes algoritmos e hiperparâmetros da inteligência artificial:

- kNN (K Vizinhos Mais Próximos): Os hiperparâmetros definidos utilizam n = 7 e distância

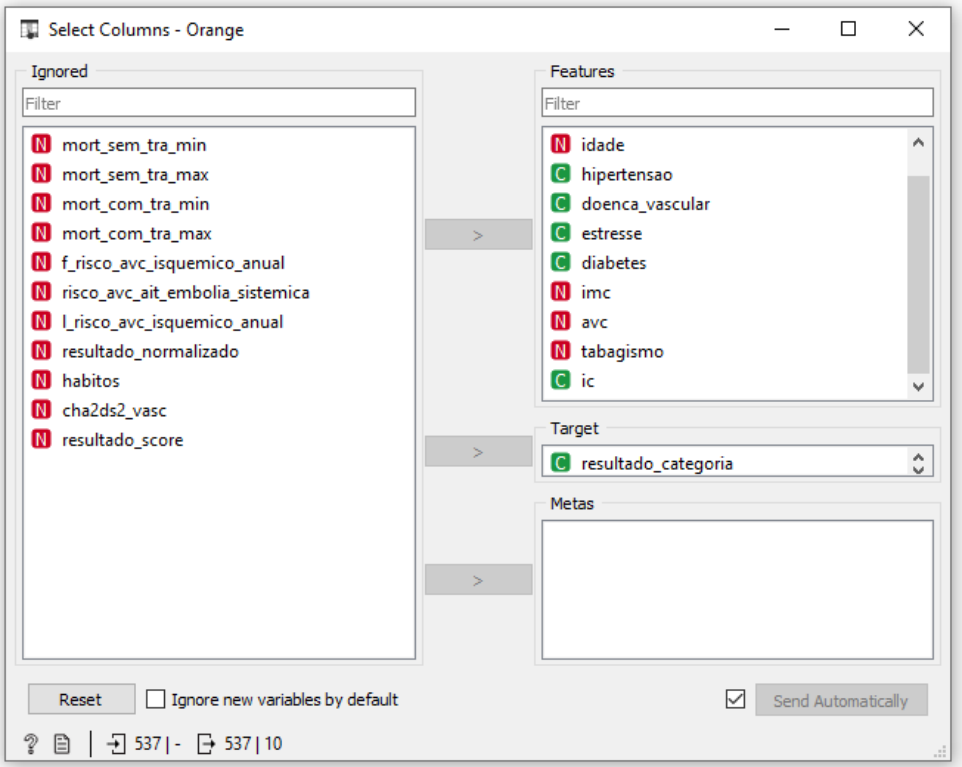


Figura 5.4: Orange Data Mining: Seleção da Variável Alvo do Modelo

Euclidiana.

- *Randon Forest* (Floresta Aleatória): Seus hiperparâmetro utilizam 100 árvores, e 4 como quantidade mínima de amostras que um nó interno deve conter para se dividir em outros nós.
- *Neural Network* (Redes Neurais): Função de ativação ReLu, 100 camadas ocultas, e limitação de 200 iterações.
- *SVM - Support Vector Machine* (Máquinas de Vetores Suporte): Custo 1, Kernel Linear, limitado a 100 iterações.
- Naïve Bayes: Sem hiperparâmetros.

Em todos os modelos foi realizada a validação cruzada, assumindo número de folds igual a 10, e amostra extratificada. Na Figura 5.5 podemos observar a estrutura elaborada no Orange Data Mining, desde a distribuição dos dados, até os dados sendo submetidos aos modelos de IA, passando pelo processo de aprendizagem, testes e scores obtidos, e apresentados sob as métricas da área sob a curva ROC (AUC), acurácia (AC), F1, Precision, Recall e Matriz de Confusão. Na seção seguinte iremos apresentar os resultados obtidos, tanto em função dos dados, quanto em função da validação do modelo em sua capacidade de predizer a categoria NYHA.

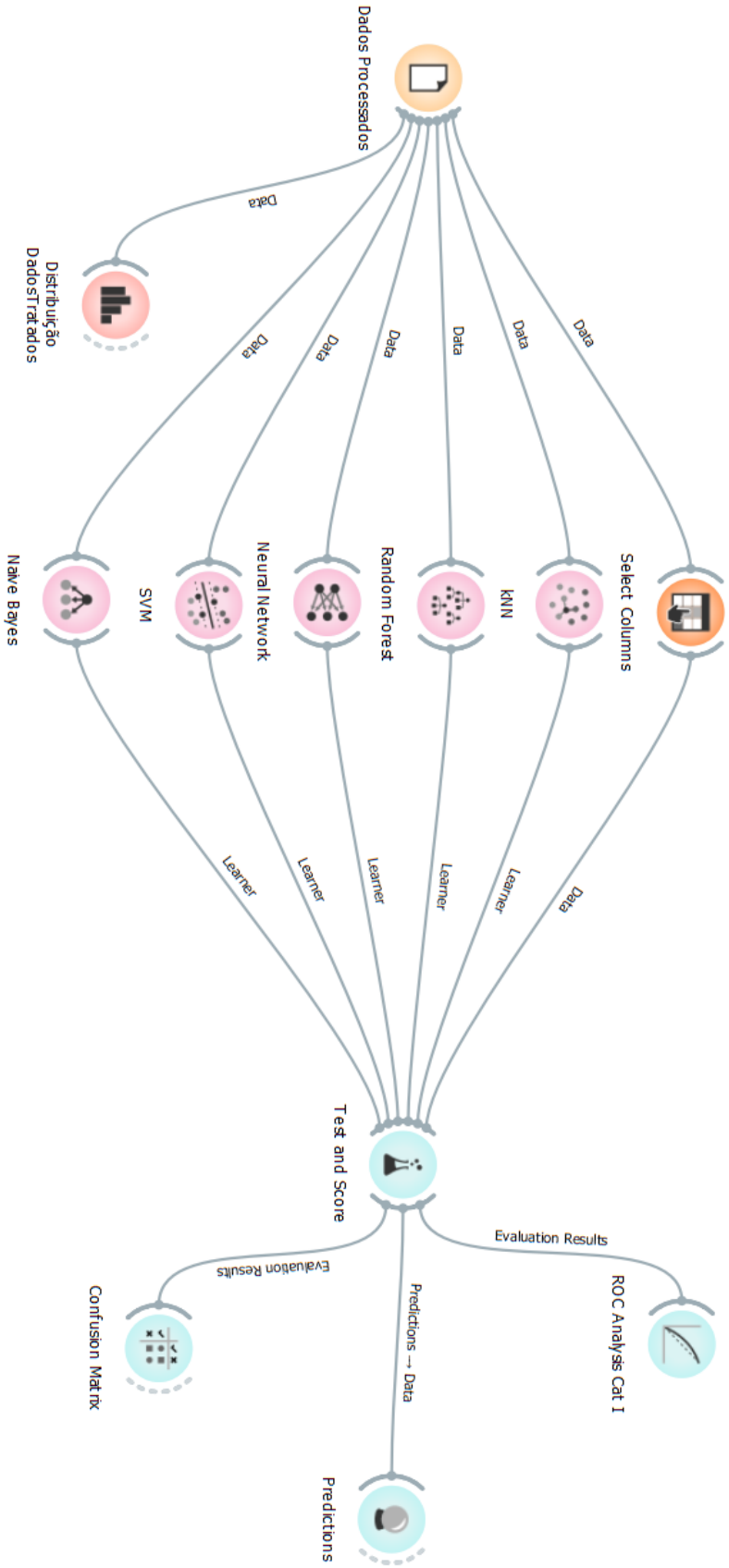


Figura 5.5: Orange Data Mining: Validação do Modelo

6

Resultados

Nesta seção serão expostos os resultados obtidos, como a distribuição dos dados originais e após processamento, validação do modelo nos algoritmos kNN, Redes Neurais, SVM, Naive Bayes e Árvores Aleatórias, além dos resultados de cada uma das métricas de validação, como a área sob a curva ROC (AUC), acurácia (CA), F1, *Precision*, *Recall* e Matriz de Confusão.

6.1 Distribuição dos Dados Originais

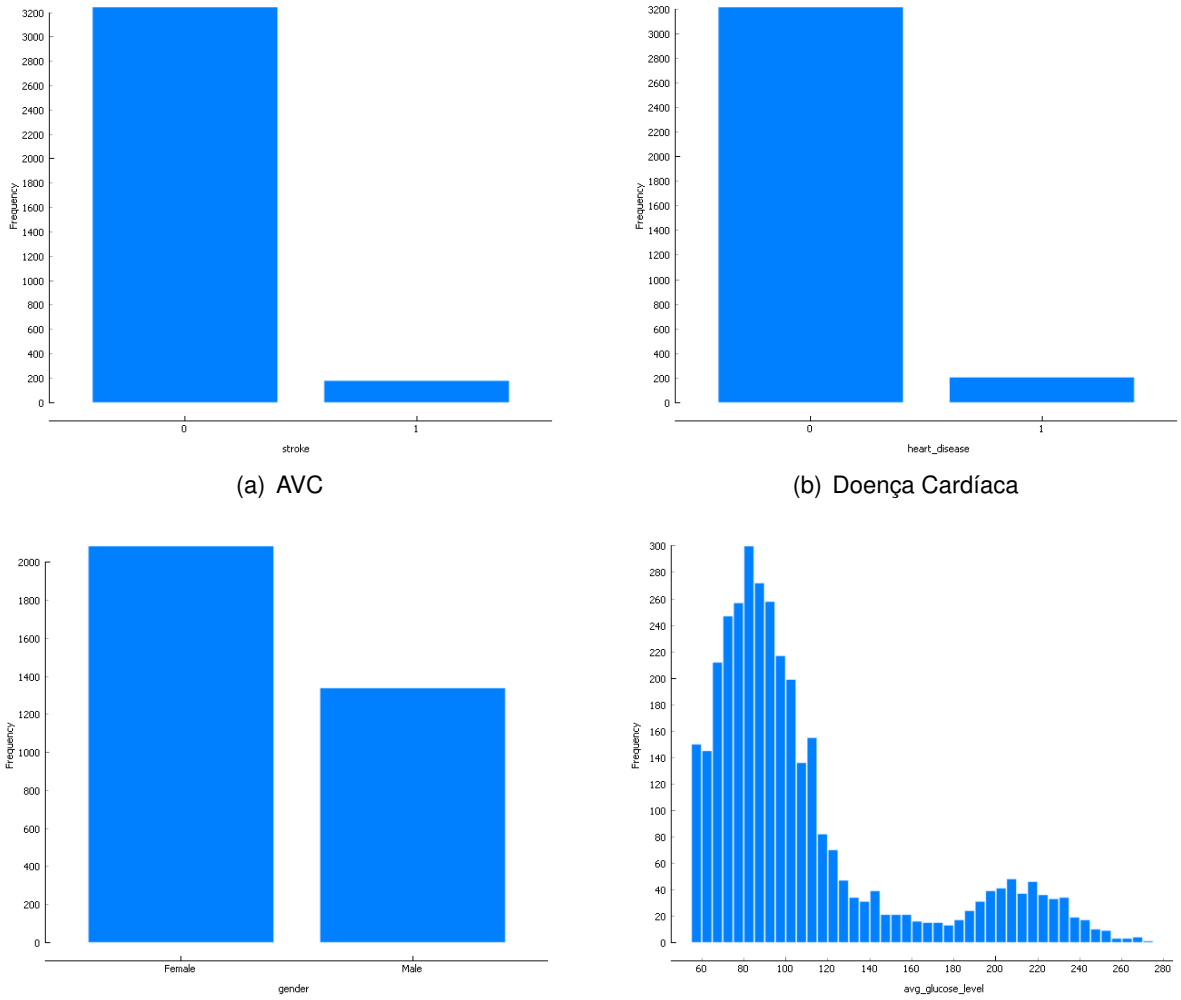
É possível observar a distribuição da base de dados original, apenas com a remoção dos dados irrelevantes, e pacientes com dados faltantes. Os dados incluem AVC, Doença Cardíaca, Gênero, Glicose, Hipertensão, Idade e Tabagismo. (Figuras 6.1 e 6.2)

6.2 Distribuição dos Dados Processados

É possível observar a distribuição dos dados processados, e classificados de acordo com as categorias de NYHA (Bredy et al., 2018). As distribuições incluem os indicadores AVC, Diabetes, Doença Vascular, Estresse, Gênero, Hipertensão, Idade, IMC e Tabagismo. Além dos scores CHA_2DS_2-VASc e Hábitos.(Figuras 6.3, 6.4 e 6.5)

6.3 Base de Dados Gerada

Aqui é possível observar a base de dados final, contendo os dados de entrada já transformados e parametrizados, os dados gerados como a Insuficiência Cardíaca, score CHA_2DS_2-VASc, score Hábitos, o risco de AVC Baseado no CHA_2DS_2-VASc, a classificação da categoria NYHA com base no cálculo ($resultado\_score \rightarrow resultado\_normalizado$), e as taxas de mortalidade com base na categoria NYHA e tratamentos adotados. (Figuras 6.6 e 6.7).



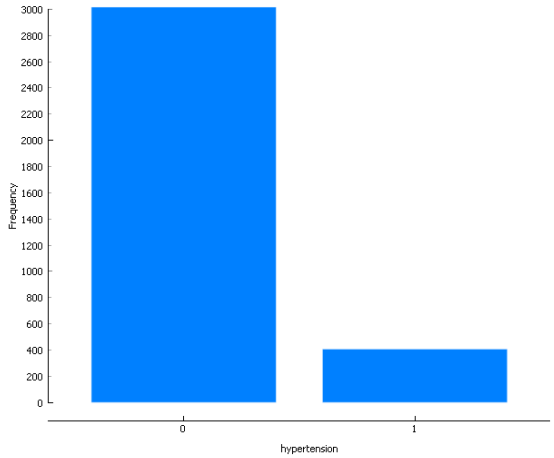
(a) AVC

(b) Doença Cardíaca

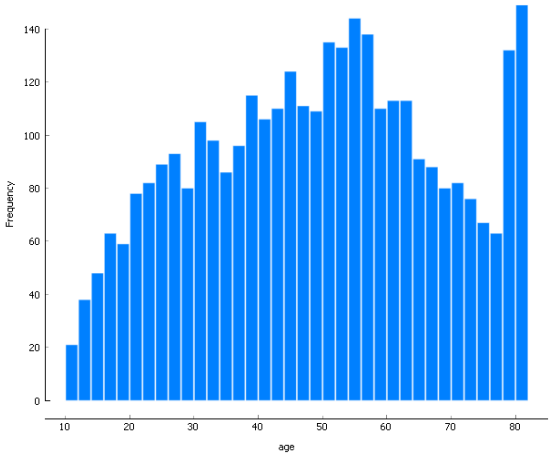
(c) Gênero

(d) Glicose

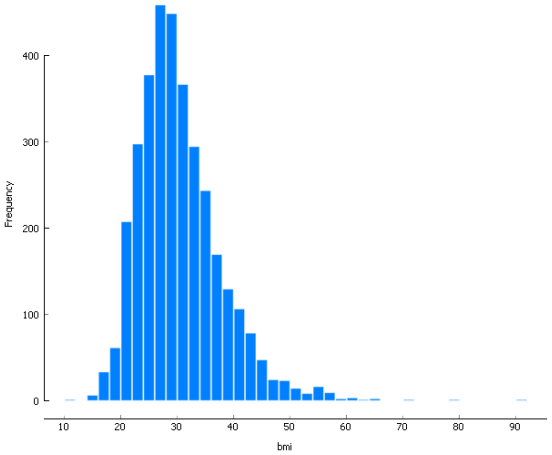
Figura 6.1: Distribuição Dados Originais Parte 1/2



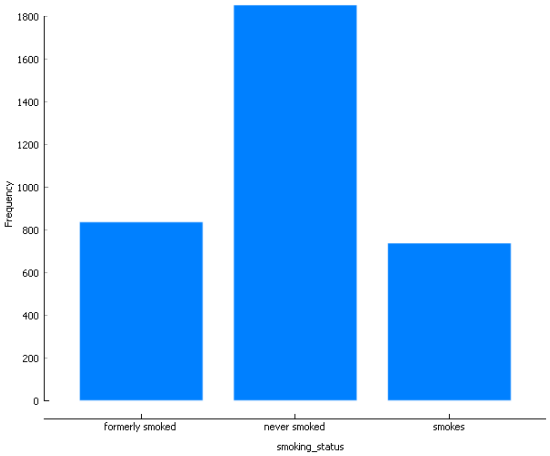
(a) Hipertensao



(b) Idade

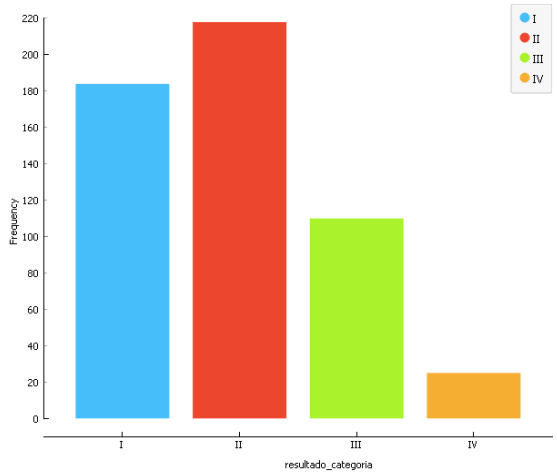


(c) IMC

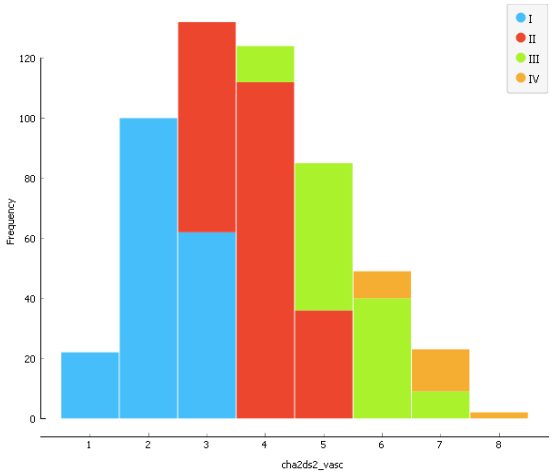


(d) Tabagismo

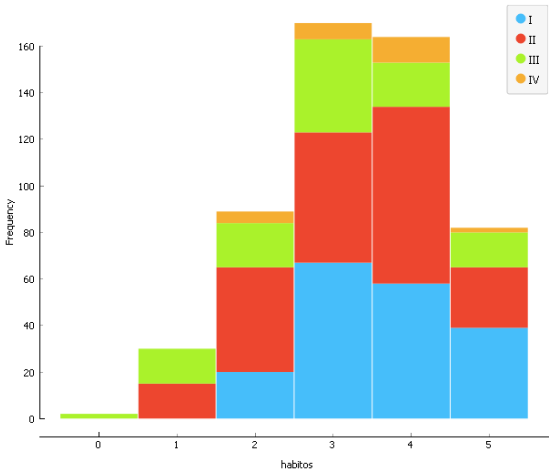
Figura 6.2: Distribuição Dados Originais Parte 2/2



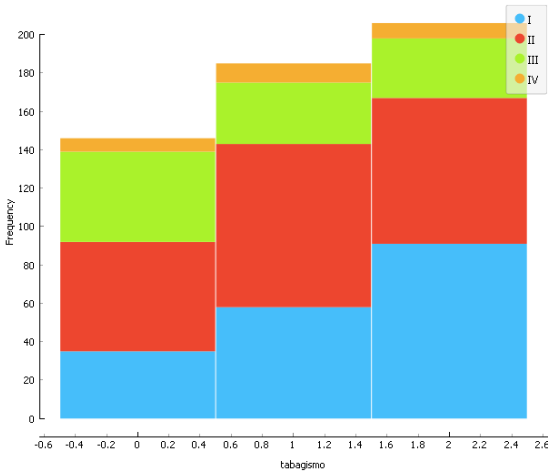
(a) Categoria NYHA



(b) Score CHA_2DS_2-VASc

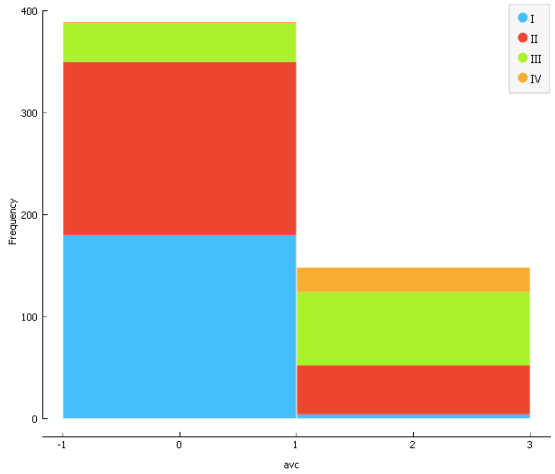


(c) Score Hábitos com Classificação NYHA

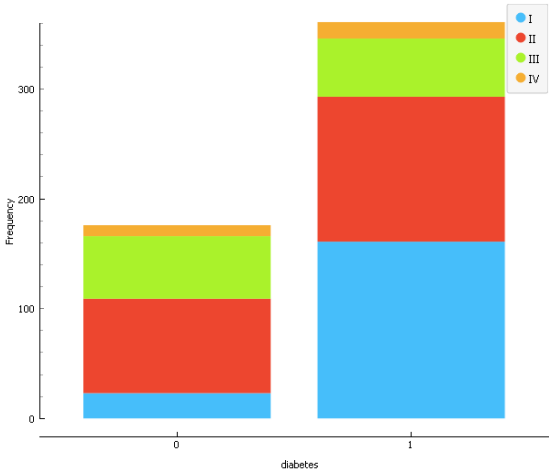


(d) Tabagismo com classificação NYHA

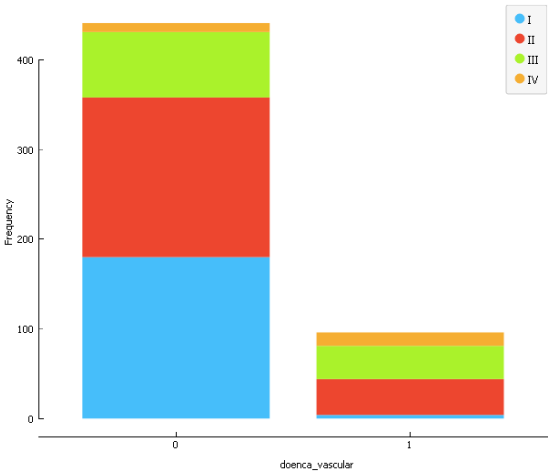
Figura 6.3: Distribuição Dados Processados e Classificados por Nyha Parte 1/3



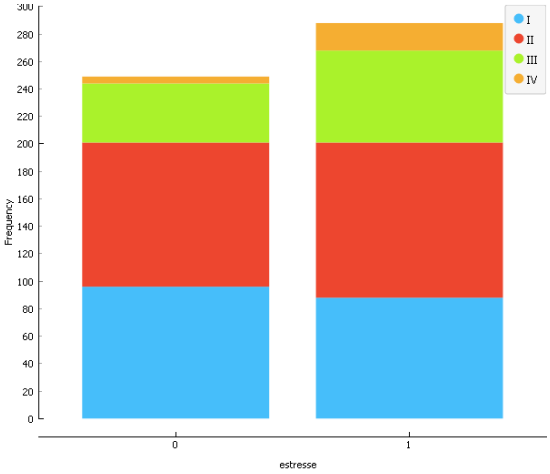
(a) AVC com classificação NYHA



(b) Diabetes com Classificação NYHA

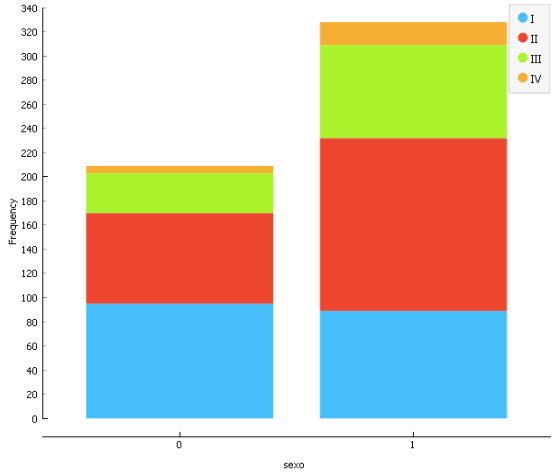


(c) Doença Vascular com classificação NYHA

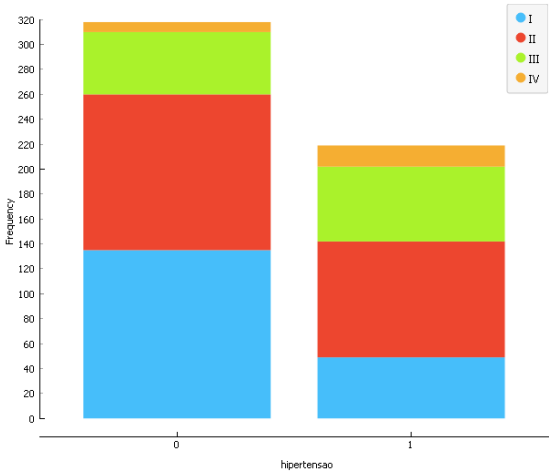


(d) Estresse com classificação NYHA

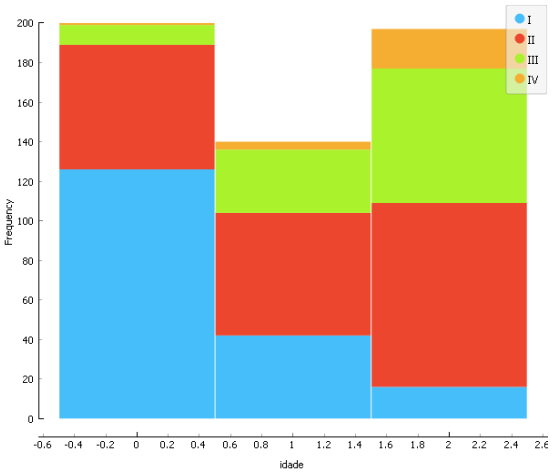
Figura 6.4: Distribuição Dados Processados e Classificados por Nyha Parte 2/3



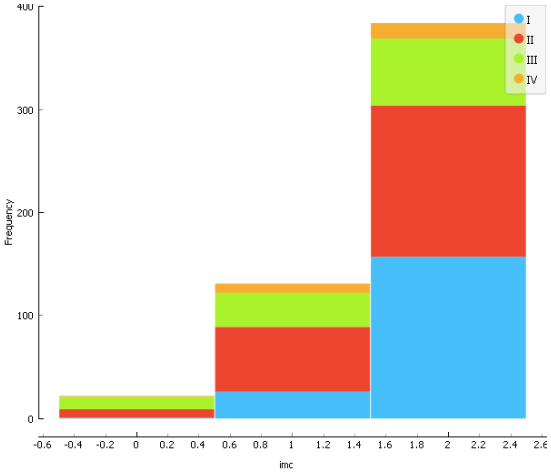
(a) Gênero com classificação NYHA



(b) Hipertensão com classificação NYHA



(c) Idade com classificação NYHA



(d) IMA com classificação NYHA

Figura 6.5: Distribuição Dados Processados e Classificados por Nyha Parte 3/3

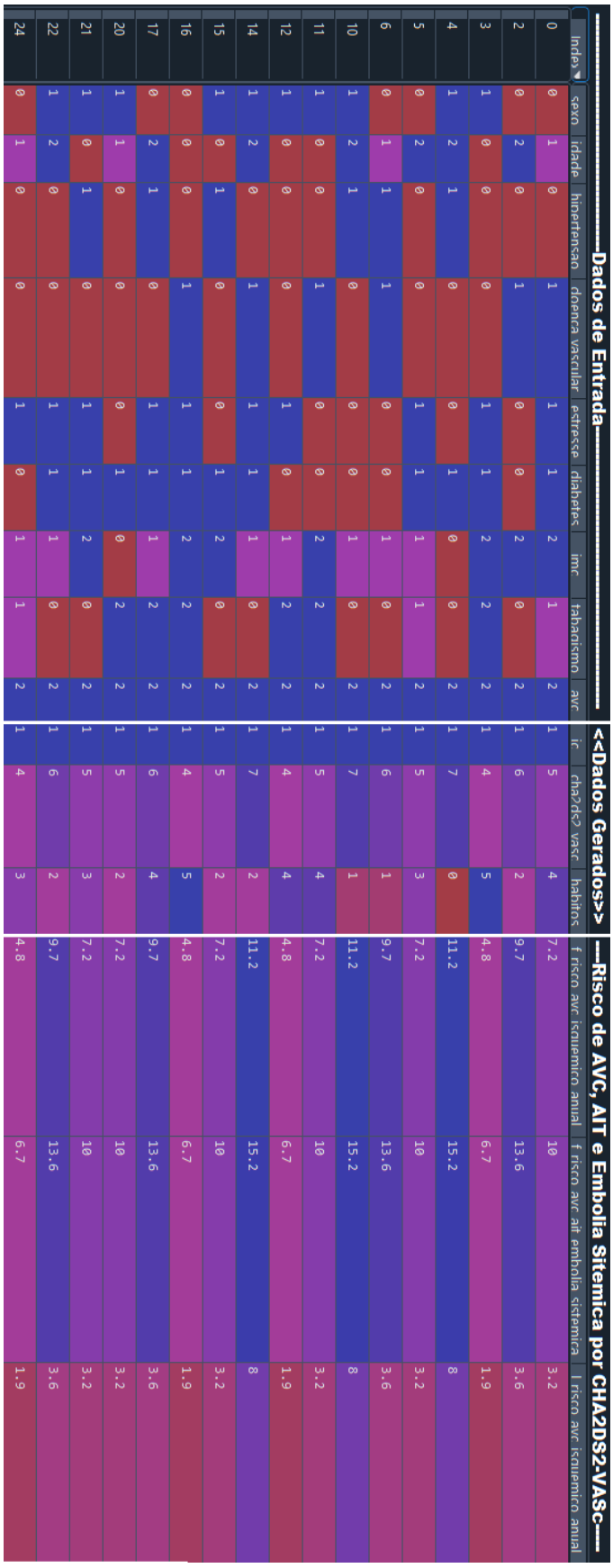


Figura 6.6: Dataset Final 1/2

<<<Cálculo e Categorização NYHA>>>				---Intervalo de Mortalidade Baseada em NYHA (Sem/Com Tratamento e Padrão)---						
resultado score	resultado normalizado	resultado categoria	mort sem tra min	mort sem tra max	mort com tra min	mort com tra max	nyha mortality min	nyha mortality max		
4.7	6.08696	III	0.15	0.4	0.1	0.15	0.1	0.15		
4.8	6.30435	III	0.15	0.4	0.1	0.15	0.1	0.15		
4.3	5.21739	III	0.15	0.4	0.1	0.15	0.1	0.15		
4.9	6.52174	III	0.15	0.4	0.1	0.15	0.1	0.15		
4.4	5.43478	III	0.15	0.4	0.1	0.15	0.1	0.15		
4.5	5.65217	III	0.15	0.4	0.1	0.15	0.1	0.15		
5.2	7.17391	III	0.15	0.4	0.1	0.15	0.1	0.15		
4.7	6.08696	III	0.15	0.4	0.1	0.15	0.1	0.15		
4	4.56522	II	0.15	0.4	0.05	0.1	0.05	0.1		
5.5	7.82609	IV	0.44	0.66	0.3	0.4	0.2	0.5		
4.1	4.78261	II	0.15	0.4	0.05	0.1	0.05	0.1		
4.3	5.21739	III	0.15	0.4	0.1	0.15	0.1	0.15		
5.4	7.6087	IV	0.44	0.66	0.3	0.4	0.2	0.5		
4.1	4.78261	II	0.15	0.4	0.05	0.1	0.05	0.1		
4.4	5.43478	III	0.15	0.4	0.1	0.15	0.1	0.15		
4.8	6.30435	III	0.15	0.4	0.1	0.15	0.1	0.15		
3.7	3.91304	II	0.15	0.4	0.05	0.1	0.05	0.1		

Figura 6.7: Dataset Final 2/2

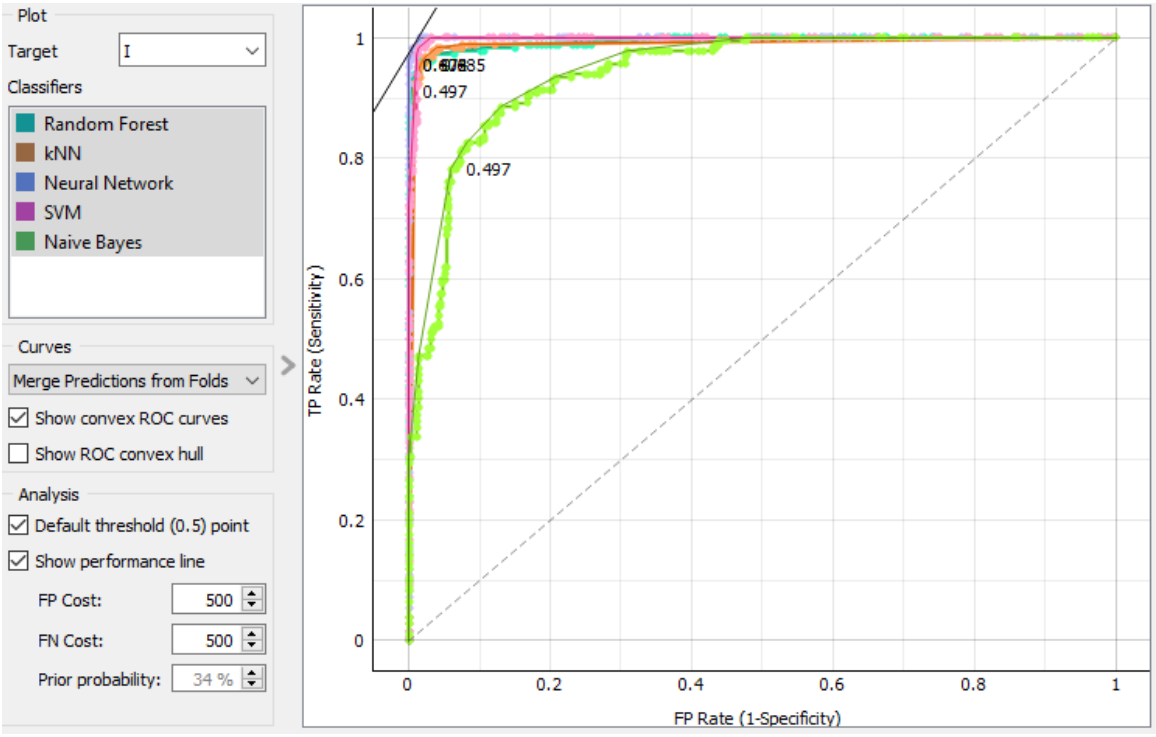


Figura 6.8: AUC Categoria NYHA I

6.4 Validação do Modelo

Aqui são expostos os scores atingidos durante a validação do modelo, considerando que foram utilizados os algoritmos kNN, Árvores Aleatórias, Redes Neurais, SVM e Naive Bayes. As métricas coletadas consideram a área sob a curva ROC (AUC) (Figuras 6.8, 6.9, 6.10 e 6.11), Acurácia (CA), F1, Precision, Recall (Figuras 6.12 e 6.13) e Matriz de Confusão (Figuras: 6.14,6.15, 6.16, 6.17, 6.18). Em todos os casos foi adotada a validação cruzada assumindo o número de folders igual à 10.

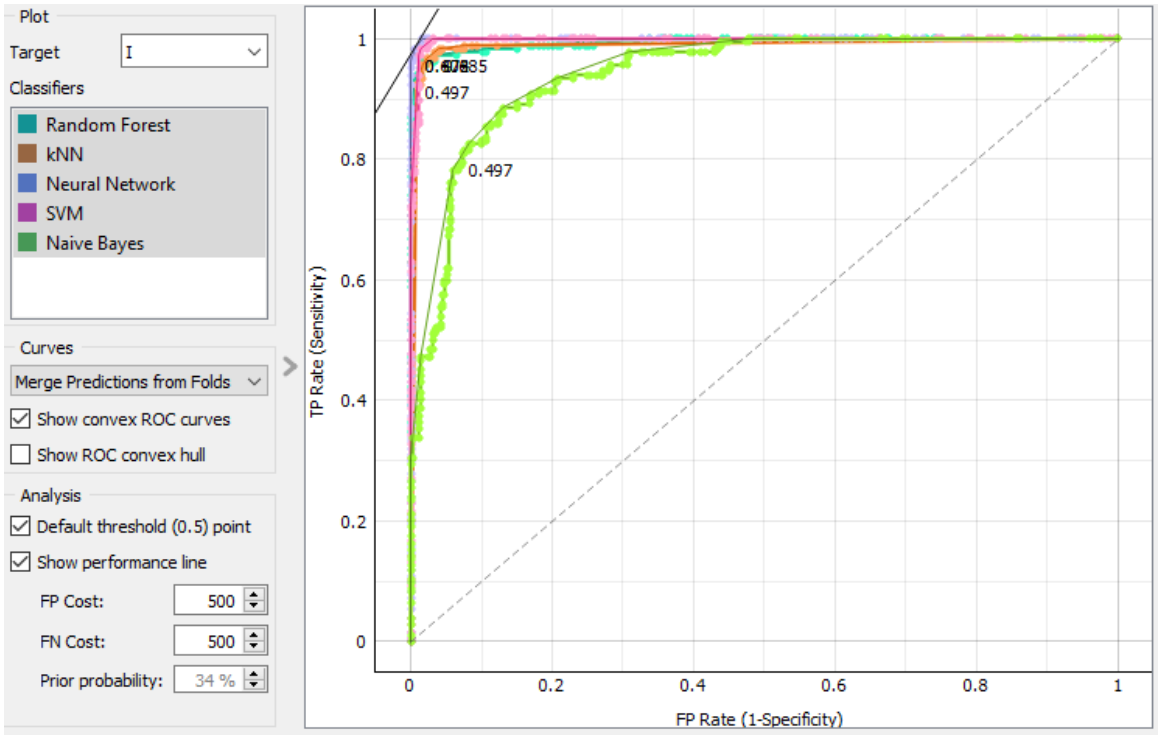


Figura 6.9: AUC Categoria NYHA II

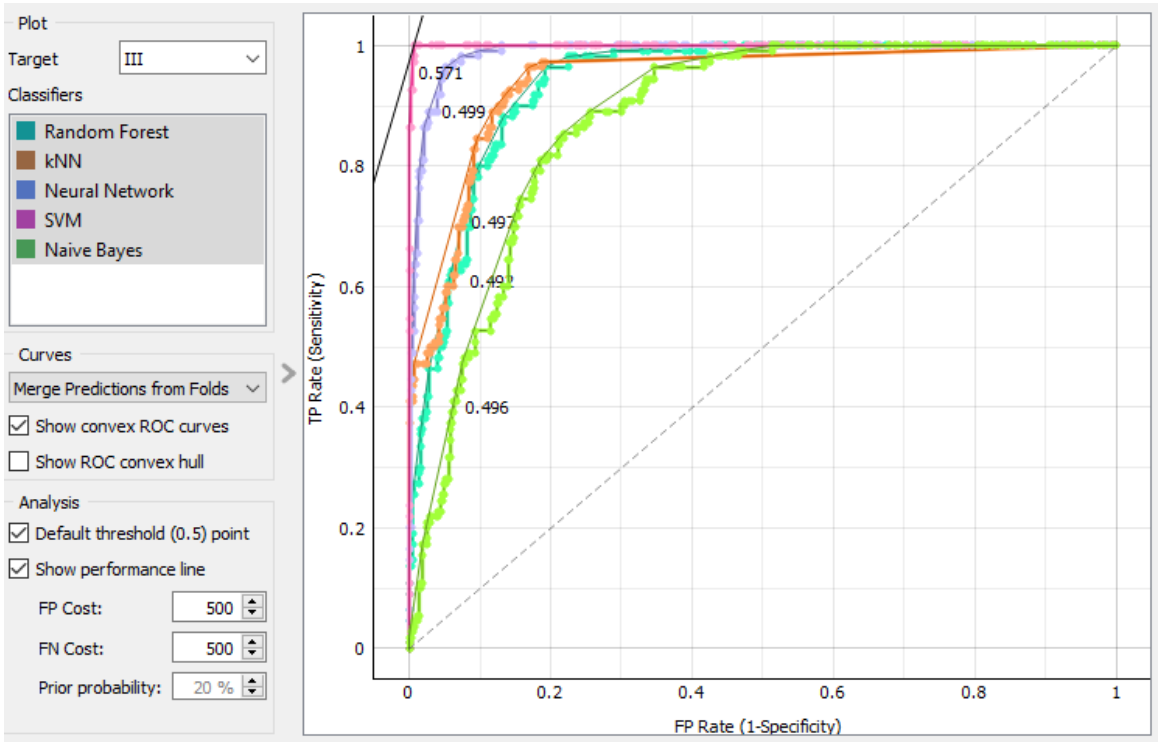


Figura 6.10: AUC Categoria NYHA III

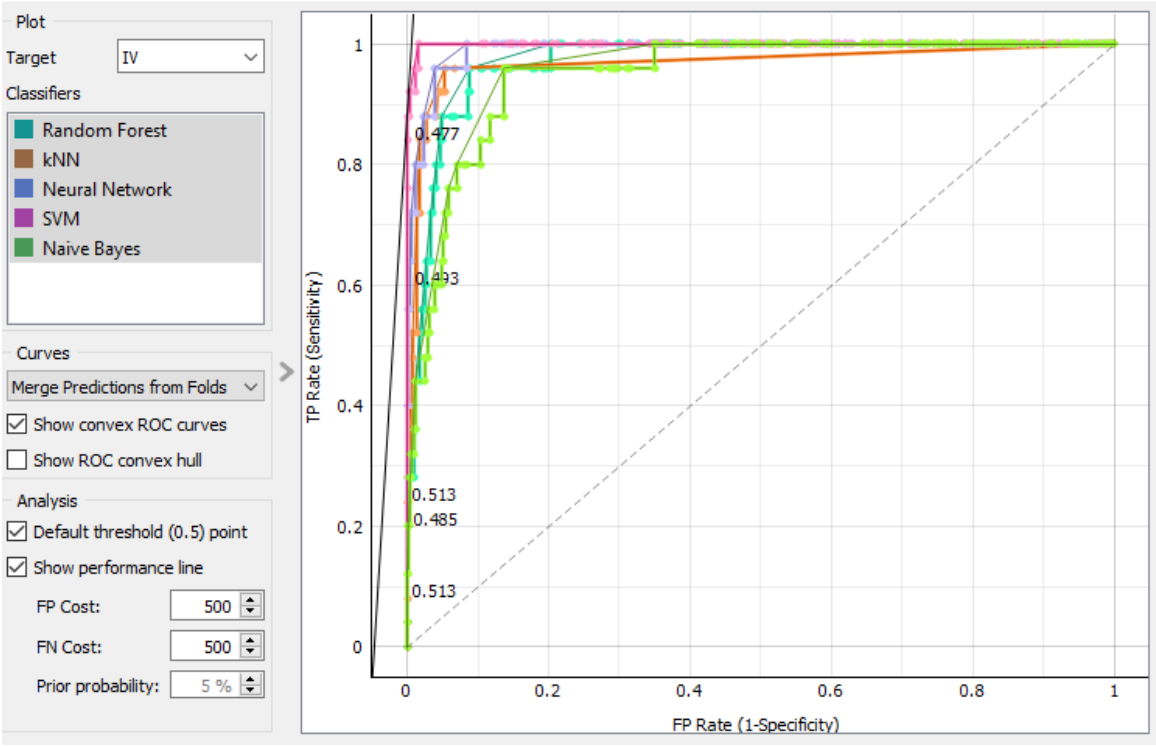


Figura 6.11: AUC Categoria NYHA IV

Model	AUC	CA	F1	Precision	Recall
SVM	0.997	0.976	0.976	0.976	0.976
Neural Network	0.995	0.942	0.941	0.943	0.942
Random Forest	0.965	0.845	0.841	0.845	0.845
kNN	0.964	0.845	0.838	0.842	0.845
Naive Bayes	0.875	0.715	0.704	0.716	0.715

Figura 6.12: AUC, CA, F1, Precision e Recall - Média das 4 Categorias NYHA

Model	AUC	CA	F1	Precision	Recall
Neural Network	1.000	0.989	0.984	0.984	0.984
SVM	0.998	0.985	0.978	0.973	0.984
Random Forest	0.995	0.972	0.959	0.972	0.946
kNN	0.989	0.965	0.950	0.923	0.978
Naive Bayes	0.942	0.890	0.837	0.849	0.826

(a) Scores Categoria NYHA I

Model	AUC	CA	F1	Precision	Recall
SVM	0.997	0.985	0.982	0.986	0.977
Neural Network	0.998	0.966	0.959	0.963	0.954
Random Forest	0.957	0.886	0.862	0.849	0.876
kNN	0.959	0.886	0.858	0.872	0.844
Naive Bayes	0.792	0.754	0.726	0.663	0.803

(b) Scores Categoria NYHA II

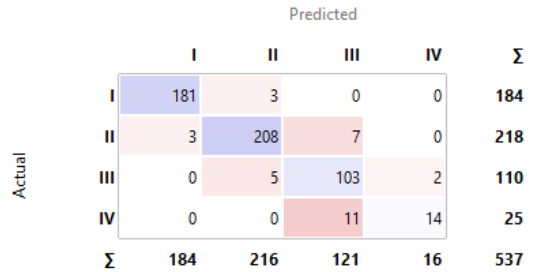
Model	AUC	CA	F1	Precision	Recall
SVM	1.000	0.991	0.977	0.973	0.982
Neural Network	0.990	0.953	0.892	0.851	0.936
kNN	0.943	0.881	0.724	0.689	0.764
Random Forest	0.939	0.873	0.707	0.672	0.745
Naive Bayes	0.875	0.827	0.528	0.598	0.473

(c) Scores Categoria NYHA III

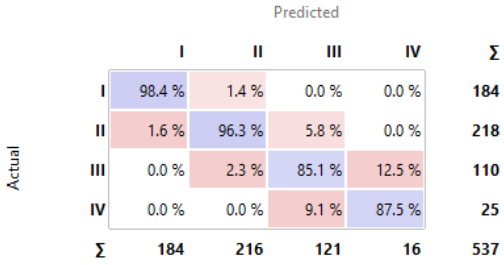
Model	AUC	CA	F1	Precision	Recall
SVM	1.000	0.991	0.898	0.917	0.880
Neural Network	0.993	0.976	0.683	0.875	0.560
Random Forest	0.971	0.959	0.389	0.636	0.280
kNN	0.975	0.959	0.353	0.667	0.240
Naive Bayes	0.950	0.959	0.312	0.714	0.200

(d) Scores Categoria NYHA IV

Figura 6.13: AUC, CA, F1, Precision e Recall das Categorias NYHA individualmente

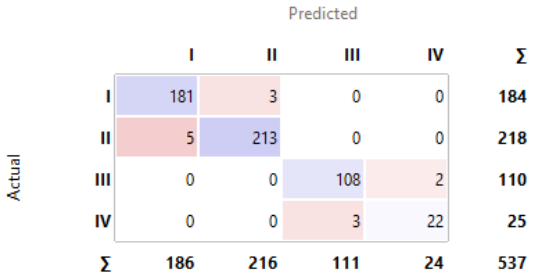


(a) Nº de Instancias

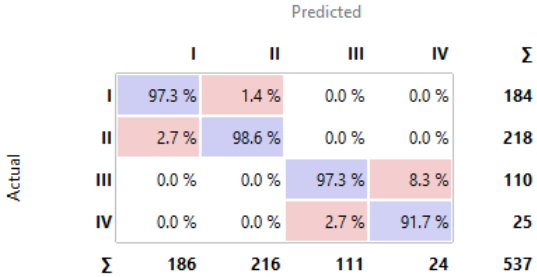


(b) Proporção de Predições

Figura 6.14: Matriz de Confusão Redes Neurais

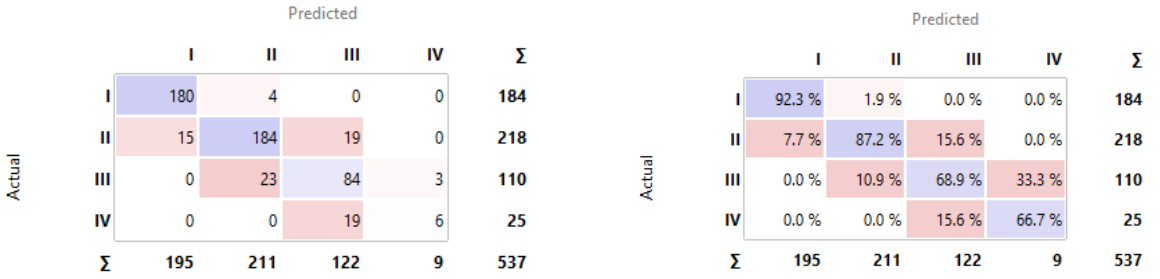


(a) Nº de Instancias



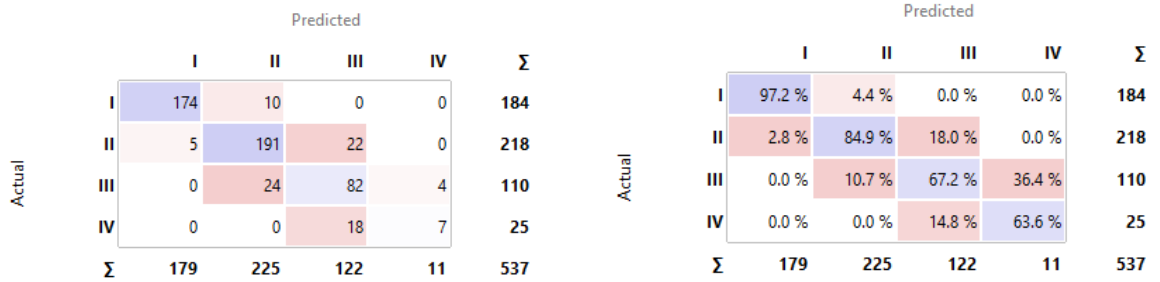
(b) Proporção de Predições

Figura 6.15: Matriz de Confusão SVM



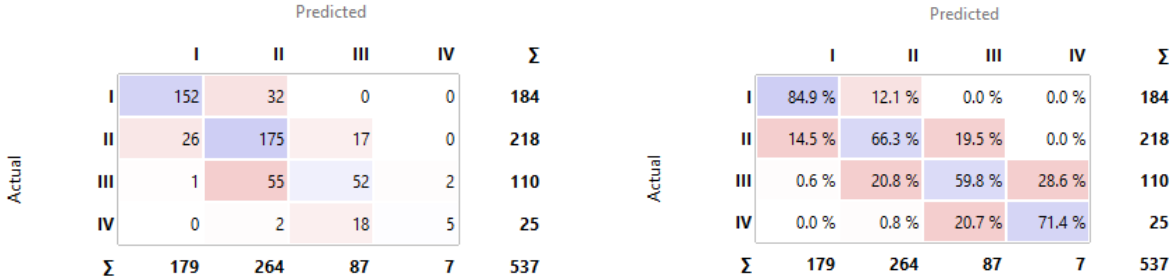
(a) Nº de Instancias (b) Proporção de Predições

Figura 6.16: Matriz de Confusão kNN



(a) Nº de Instancias (b) Proporção de Predições

Figura 6.17: Matriz de Confusão Florestas Aleatórias



(a) Nº de Instancias (b) Proporção de Predições

Figura 6.18: Matriz de Confusão Naive Bayes

7

Discussão

Nesta seção iremos discutir os resultados obtidos, expor as ameaças a validade do modelo proposto, resumir o trabalho, elencar os benefícios obtidos, expor as limitações do trabalho, por fim, sugerir trabalhos futuros a serem realizados a partir do que foi proposto e validado neste trabalho.

7.1 Discussão dos Resultados

Tratamento dos Dados

O tratamento dos dados utilizando a linguagem Python na IDE Spyder, possibilitou uma redução considerável da dimensionalidade para iniciar as inferências sobre os dados disponíveis na base de dados adotada (Liu, 2019), mais precisamente saímos de (5110, 12) para (537, 10), como apresentado na Figura 5.2. Posteriormente, com as inferências, aquisição de novas variáveis, classificação, e disposição de taxas de riscos de AVC, AIC e Embolia Sistêmica, bem como as taxas de mortalidade baseadas no NYHA e os tratamentos adotados, tornaram a base muito mais completa (537, 24).

Aquisição de Variáveis

Ao passo que o dataset foi transformado, identificou-se a falta de um único indicador do score CHA_2DS_2-VASc, a Insuficiência Cardíaca. A partir de um levantamento da literatura foi possível identificar que grande parte dos indicadores presente em nossa base de dados possuía algum grau de correlação com as chances de desenvolver a IC. Prontamente levantamos estas variáveis correlacionadas, e com base na representatividade dos scores utilizados, conseguimos segregar os pacientes com maior potencial de desenvolver IC. As demais variáveis relacionadas às taxas de mortalidade e risco de episódios de AVC, AIT e Embolia Sistêmica, geradas em

nossa base dedos, estão apoiados nas relações diretas da literatura, tanto em relação ao score CHA_2DS_2-VASc, quanto à categoria NYHA.

7.2 Ameaças a validade

O fato do estudo assumir que o cenário adotado conta com o desenvolvimento de uma Fibrilação Ventricular (FV) evoluindo para uma Insuficiência Cardíaca (IC), que por sua vez avança para a Fibrilação Ventricular (FV), torna o estudo pouco, ou quase impossível de ser generalizável a outras frentes das arritmias. Todavia, este fluxo da taquicardia é comum, o que sustenta a representatividade e importância do modelo proposto.

O modelo não passou por validação de um corpo de médicos, sendo esta, uma etapa fundamental para o modelo ser avaliado e aprimorado a partir dos feedbacks de cardiologistas e outros profissionais da área da saúde, possibilitando assim, os primeiros testes em um cenário real. Vale salientar que todo sistema médico é um sistema de alta criticidade, pelo fato de lidar com vidas, logo, são sistemas que passam por validações mais exigentes. Por outro lado, a saúde vive momento de transição quanto à adesão de novas tecnologias, sendo comum a existência de certa resistência por parte de alguns médicos. O importante é que os passos estão sendo dados, como a própria iniciativa do Senado que aprovou o Projeto de Lei (PL) 3.814/2020, onde torna obrigatória por parte do Sistema Único de Saúde (SUS), a criação de uma plataforma digital para unificar informações de pacientes atendidos pelas redes de saúde pública e privada (CNM). Possibilitando assim que hajam dados dentro de uma estrutura, consequentemente mais "munição"para os sistema basados em aprendizado de máquina, que demandam bases de dados robustas, estruturadas e representativas para o contexto de aplicação.

7.3 Benefícios Obtidos

O sistema identifica o potencial de desenvolvimento de insuficiência cardíaca (IC) com base num conjunto de informações correlatas, assumindo scores CHA_2DS_2-VASc e Hábitos de um mesmo paciente. A partir destes scores, é capaz de determinar os riscos de ocorrência de AVC, AIT ou Embolia Sistêmica num intervalo de um ano, em pacientes que estão no quadro de Fibrilação Ventricular, decorrente da Taquicardia Ventricular e Insuficiência Cardíaca. Além disto, com a relação entre os scores, foi possível classificar os pacientes com Insuficiência Cardíaca dentro da classificação de NYHA, onde sua categoria determina nível de morbidade (em termos da intensidade na apresentação dos sintomas) e mortalidade em casos onde se adota ou não tratamentos.

7.4 Trabalhos Futuros

Como trabalhos futuros, há a necessidade de validar o atual modelo com o auxílio e conhecimento específico de médicos, preferencialmente cardiologistas. Para que, a partir dos feedbacks obtidos, seja possível realizar os ajustes necessários. Apesar de significativa, a base de dados final de pacientes com IC, foi de 537 indivíduos, e um sistema que carrega um nível de criticidade como este, requer bases mais robustas e representativas, logo, este também se torna mais um dos passos necessários. O modelo converge pelo caminho da taquicardia ventricular, seguida de insuficiência cardíaca e fibrilação ventricular. De modo, que fica em aberto as abordagens que considerem outros caminhos que as arritmias podem percorrer, desde suas classificações iniciais, como bradicardia e taquicardia, bem como seus focos, que podem ser atriais, sinusais ou ventriculares. Em outras palavras, existe uma gama enorme de possíveis abordagens na cardiologia, e este modelo pode ser escopo para uma nova doença alvo.

8

Conclusão

O modelo proposto foi capaz de predizer a morbidade e mortalidade de pacientes diagnosticados com taquicardia ventricular. Tomando como base os scores CHA_2DS_2-VASc e Hábitos, aplicados a uma base de dados representativa para o cenário de arritmias. Utilizou-se da linguagem de programação Pyhthon, para tratamento dos dados obtidos, e Orange Data Mining para validar o modelo proposto, a partir dos algoritmos SVM, Redes Neurais, kNN, Árvores Aleatórias e Naive Bayes. Como benefícios, o modelo proposta possuí o potencial de identificar o desenvolvimento de insuficiência cardíaca baseado no conjunto de doenças e dados clínicos correlatos, determinando a partir disto o risco anual de AVC, AIT ou Embolis Sistêmica. Além disto foi possível determinar a categoria de morbidade a partir da classificação NYHA, aplicável aos pacientes portadores de insuficiência cardíaca. Em relação aos trabalhos futuros, existe a necessidade da validação do modelo, feita por um corpo médico, especificamente cardiologistas. Apesar de significativa, a base de dados final conta com 537 indivíduos, e um sistemas de tal complexidade e criticidade requer bases mais robustas para fortalecer a representatividade. Uma vez que o modelo converge para pacientes com insuficiência cardíaca e fibrilação ventricular, surgem diversas possibilidades de abordar outros caminhos que envolvam, por exemplo, a bradicardia e taquicardia, além dos pontos focais (atriais, sinusais e ventriculares). Dito isso, o modelo aqui apresentado pode se tornar escopo para novas doenças alvo.

Acrónimos

TAVRN Taquicardia Atrioventricular Reentrante Nodal

TSV Taquicardia Supraventricular

DAC Doença Arterial Coronariana

IC Insuficiência Cardíaca

AVC Ataque Vascular Cerebral

NYHA New York Heart Association

TV Taquicardia Ventricular

IA Inteligência Artificial

ECG Eletrocardiograma

KNN K vizinhos mais próximos

SVM Máquinas de Vetores Suporte

ML Machile Learning

Referências

A survey on feature selection methods. *Computers Electrical Engineering*, 40(1):16–28, 2014. ISSN 0045-7906. **DOI** https://doi.org/10.1016/j.compeleceng.2013.11.024. URL https://www.sciencedirect.com/science/article/pii/S0045790613003066. 40th-year commemorative issue.

ACM. Acm digital library. https://dl.acm.org/. (Accessed on 08/14/2020).

Diego Addan. Ia07.pdf. https://www.inf.ufpr.br/dagoncalves/IA07.pdf, 2019. (Accessed on 08/19/2020).

Amiya A. Ahmed, Kanan Patel, Margaret A. Nyaku, Raya E. Kheirbek, Vera Bittner, Gregg C. Fonarow, Gerasimos S. Filippatos, Charity J. Morgan, Inmaculada B. Aban, Marjan Mujib, Ravi V. Desai, Richard M. Allman, Michel White, Prakash Deedwania, George Howard, Robert O. Bonow, Ross D. Fletcher, Wilbert S. Aronow, and Ali Ahmed. Risk of heart failure and death after prolonged smoking cessation. *Circulation: Heart Failure*, 8(4):694–701, July 2015. **DOI** 10.1161/circheartfailure.114.001885. URL https://doi.org/10.1161/circheartfailure.114.001885.

Alila. Alila medical media | vídeos em português. https://www.alilamedicalmedia.com/pt/-/galleries/videos-em-portugues, 2020. (Accessed on 08/17/2022).

Mehrdad Amir-Behghadami and Ali Janati. Population, Intervention, Comparison, Outcomes and Study (PICOS) design as a framework to formulate eligibility criteria in systematic reviews. *Emergency Medicine Journal*, 37(6):387–387, June 2020. ISSN 1472-0205, 1472-0213. **DOI** 10.1136/emermed-2020-209567. URL https://emj.bmj.com/lookup/doi/10.1136/emermed-2020-209567.

Syed Muhammad Anwar, Maheen Gul, Muhammad Majid, and Majdi Alnowami. Arrhythmia classification of ECG signals using hybrid features. *Computational and Mathematical Methods in Medicine*, 2018:1–8, November 2018. **DOI** 10.1155/2018/1380348. URL https://doi.org/10.1155/2018/1380348.

Dagfinn Aune, Sabrina Schlesinger, Teresa Norat, and Elio Riboli. Tobacco smoking and the risk of heart failure: A systematic review and meta-analysis of prospective studies. *European Journal of Preventive Cardiology*, 26(3):279–288, 08 2020. ISSN 2047-4873. **DOI** 10.1177/2047487318806658. URL https://doi.org/10.1177/2047487318806658.

BBC. O que mais mata as pessoas ao redor do mundo? URL https://www.bbc.com/portuguese/internacional-47471602.

Jill A. Bennett, Barbara Riegel, Vera Bittner, and Joyce Nichols. Validity and reliability of the NYHA classes for measuring research outcomes in patients with cardiac disease. *Heart & Lung*, 31(4):262–270, July 2002. ISSN 01479563. **DOI** 10.1067/mhl.2002.124554. URL https://linkinghub.elsevier.com/retrieve/pii/S0147956302000031.

Giuseppe Boriani, Giovanni Luca Botto, Luigi Padeletti, Massimo Santini, Alessandro Capucci, Michele Gulizia, Renato Ricci, Mauro Biffi, Tiziana De Santo, Giorgio Corbucci, Gregory Y.H. Lip, and for the Italian AT-500 Registry Investigators. Improving Stroke Risk Stratification Using the CHADS $_2$ and CHA $_2$ DS $_2$ -VASc Risk Scores in Patients With Paroxysmal Atrial Fibrillation by Continuous Arrhythmia Burden Monitoring. *Stroke*, 42(6):1768–1770, June 2011. ISSN 0039-2499, 1524-4628. **DOI** 10.1161/STROKEAHA.110.609297. URL https://www.ahajournals.org/doi/10.1161/STROKEAHA.110.609297.

Charlene Bredy, Margherita Ministeri, Alexander Kempny, Rafael Alonso-Gonzalez, Lorna Swan, Anselm Uebing, Gerhard-Paul Diller, Michael A Gatzoulis, and Konstantinos Dimopoulos. New York Heart Association (NYHA) classification in adults with congenital heart disease: relation to objective measures of exercise and outcome. *European Heart Journal - Quality of Care and Clinical Outcomes*, 4(1):51–58, January 2018. ISSN 2058-5225, 2058-1742. **DOI** 10.1093/ehjqcco/qcx031. URL http://academic.oup.com/ehjqcco/article/4/1/51/4083514.

Leo Breiman. *Machine Learning*, 45(1):5–32, 2001. **DOI** 10.1023/a:1010933404324. URL https://doi.org/10.1023/a:1010933404324.

Cables and Sensors. Guia de colocação de ecg de 12 derivações com ilustrações. https://www.cablesandsensors.com/pages/12-lead-ecg-placement-guide-with-illustrations. (Accessed on 08/17/2020).

Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79, 2018. ISSN 0925-2312. **DOI** https://doi.org/10.1016/j.neucom.2017.11.077. URL https://www.sciencedirect.com/science/article/pii/S0925231218302911.

A. John Camm. 2012 focused update of the ESC guidelines for the management of atrial fibrillation. *European Heart Journal*, 33(21):2719–2747, August 2012. **DOI** 10.1093/eurheartj/ehs253. URL https://doi.org/10.1093/eurheartj/ehs253.

Gokhan Cetinkal, Betul Balaban Kocas, Ozgur Selim Ser, Hakan Kilci, Kudret Keskin, Safiye Nur Ozcan, Yildiz Verdi, Mustafa Ismet Zeren, Tolga Demir, and Kadriye Kilickesmez. Assessment of the Modified CHA2DS2VASc Risk Score in Predicting Mortality in Patients Hospitalized With COVID-19. *The American Journal of Cardiology*, 135:143–149, November 2020. ISSN 00029149. **DOI** 10.1016/j.amjcard.2020.08.040. URL https://linkinghub.elsevier.com/retrieve/pii/S0002914920308973.

CNM. Cnm - confederação nacional de municípios | comunicação. https://www.cnm.org.br/comunicacao/noticias/senado-aprova-plataforma-para-unificar-dados-do-sus-e-da-rede-privada. (Accessed on 08/22/2022).

Jennifer Dekerlegand. Congestive heart failure. In *Physical Rehabilitation*, pages 669–688. Elsevier, 2007. **DOI** 10.1016/b978-072160361-2.50028-4. URL https://doi.org/10.1016/b978-072160361-2.50028-4.

Janez Demšar, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, Miha Štajdohar, Lan Umek, Lan Žagar, Jure Žbontar, Marinka Žitnik, and Blaž Zupan. Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, 14:2349–2353, 2013. URL http://jmlr.org/papers/v14/demsar13a.html.

Vasant Dhar. Data science and prediction. *Commun. ACM*, 56(12):64–73, dec 2013. ISSN 0001-0782. **DOI** 10.1145/2500499. URL https://doi.org/10.1145/2500499.

Tatiana Escovedo. Estatidados. http://estatidados.com.br/machine-learning-conceitos-e-modelos-parte-i-aprendizado-supervisionado/, 06 2020. (Accessed on 08/18/2020).

Gilson Soares Feitosa, José Carlos Nicolau, Adalberto Lorga, Adalberto Lorga Filho, André D'Ávila, Anis Rassi Jr, Angelo A. V. de Paola, Anísio Pedrosa, Álvaro Barros da Costa, Ayrton Peres, César Grupi, Cláudio Cirenza, Dalmo Moreira, Dário Sobral, Denise Hachul, Eduardo D'Andréa, Eduardo Sosa, Epotamênides M. Good God, Fábio Sândoli de Brito, Fernando Cruz, Guilherme Fenelon, Gustavo Glotz Lima, Hélio Brito, Ivan G. Maia, Jacob Atié, José Carlos Moura Jorge, José Carlos de Andrade, José Carlos Pachón Mateos, José Carlos Ribeiro, João Pimenta, José Tarcísio de Vasconcelos, Júlio Gizzi, Leandro Zimerman, Luiz Antonio Teno Castilho, Márcio Fagundes, Márcio Figueiredo, Martino Martinelli Filho, Maurício I. Scanavacca, Ney Valente, Paulo Medeiros, Paulo Brofman, Reynaldo Castro

Miranda, Roberto Costa, Ricardo Kunyioshi, Roberto Sá, Sérgio G. Rassi, Sérgio Siqueira, Silas Galvão, Silvana Nishioka, Tereza Grillo, Thiago da Rocha Rodrigues, and Washington Maciel. Diretrizes para Avaliação e Tratamento de Pacientes com Arritmias Cardíacas. *Arquivos Brasileiros de Cardiologia*, 79, 2002. ISSN 0066-782X. **DOI** 10.1590/S0066-782X2002001900001. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0066-782X2002001900001&lng=en&nrm=iso&tlng=en.

Leif Friberg, Mårten Rosenqvist, and Gregory Y.H. Lip. Evaluation of risk stratification schemes for ischaemic stroke and bleeding in 182 678 patients with atrial fibrillation: the swedish atrial fibrillation cohort study. *European Heart Journal*, 33(12):1500–1510, January 2012a. **DOI** 10.1093/eurheartj/ehr488. URL https://doi.org/10.1093/eurheartj/ehr488.

Leif Friberg, Mårten Rosenqvist, and Gregory Y.H. Lip. Evaluation of risk stratification schemes for ischaemic stroke and bleeding in 182 678 patients with atrial fibrillation: the swedish atrial fibrillation cohort study. *European Heart Journal*, 33(12):1500–1510, January 2012b. **DOI** 10.1093/eurheartj/ehr488. URL https://doi.org/10.1093/eurheartj/ehr488.

Leif Friberg, Mårten Rosenqvist, and Gregory Y.H. Lip. Evaluation of risk stratification schemes for ischaemic stroke and bleeding in 182 678 patients with atrial fibrillation: the Swedish Atrial Fibrillation cohort study. *European Heart Journal*, 33(12):1500–1510, 01 2012c. ISSN 0195-668X. **DOI** 10.1093/eurheartj/ehr488. URL https://doi.org/10.1093/eurheartj/ehr488.

Keinosuke Fukunaga and Patrenahalli M. Narendra. A branch and bound algorithm for computing k-nearest neighbors. *IEEE Transactions on Computers*, C-24:750–753, 1975.

Taís Freire Galvao, Thais de Souza Andrade Pansani, and David Harrad. Principais itens para relatar revisõees sistemáticas e meta-análises: A recomendação prisma. *Epidemiologia e serviços de saúde*, 24:335–342, 2015.

Michael A Gatzoulis, Seshadri Balaji, Steven A Webber, Samuel C Siu, John S Hokanson, Christine Poile, Mark Rosenthal, Makoto Nakazawa, James H Moller, Paul C Gillette, Gary D Webb, and Andrew N Redington. Risk factors for arrhythmia and sudden cardiac death late after repair of tetralogy of fallot: a multicentre study. *The Lancet*, 356(9234):975–981, 2000. ISSN 0140-6736. **DOI** https://doi.org/10.1016/S0140-6736(00)02714-8. URL https://www.sciencedirect.com/science/article/pii/S0140673600027148.

Planilhas Google. Google workspace | apps empresariais e ferramentas de colaboração. https://workspace.google.com/. (Accessed on 08/15/2020).

Joseph Habboushe, Caroline Altman, and Gregory Y. H. Lip. Time trends in use of the chads2 and cha2ds2vasc scores, and the geographical and specialty uptake of these scores from a

popular online clinical decision tool and medical reference. *International Journal of Clinical Practice*, 73(2):e13280, 2019. **DOI** https://doi.org/10.1111/ijcp.13280. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ijcp.13280.

Victor Holanda. Github - holandavictor/tachycardia-prognosis: Model for the prognosis of morbidity and mortality in patients diagnosed with ventricular tachycardia. https://github.com/holandavictor/tachycardia-prognosis. (Accessed on 08/25/2022).

Kokil Jaidka, Sharath Chandra Guntuku, Jane H. Lee, Zhengyi Luo, Anneke Buffone, and Lyle H. Ungar. The rural–urban stress divide: Obtaining geographical insights through twitter. *Computers in Human Behavior*, 114:106544, 2021. ISSN 0747-5632. **DOI** https://doi.org/10.1016/j.chb.2020.106544. URL https://www.sciencedirect.com/science/article/pii/S0747563220302958.

Raed A. Joundi, Lauren E. Cipriano, Luciano A. Sposato, Gustavo Saposnik, and on behalf of the Stroke Outcomes Research Working Group. Ischemid stroke risk in patients with fa and cha2ds2-vasc score of 1: Systematic review and meta-analysis. *Stroke*, 47(5):1364–1367, May 2016. ISSN 0039-2499, 1524-4628. **DOI** 10.1161/STROKEAHA.115.012609. URL https://www.ahajournals.org/doi/10.1161/STROKEAHA.115.012609.

I. Kadi, A. Idri, and J.L. Fernandez-Aleman. Knowledge discovery in cardiology: A systematic literature review. *International Journal of Medical Informatics*, 97:12–32, January 2017. **DOI** 10.1016/j.ijmedinf.2016.09.005. URL https://doi.org/10.1016/j.ijmedinf.2016.09.005.

Satish Kenchaiah, Jane C. Evans, Daniel Levy, Peter W.F. Wilson, Emelia J. Benjamin, Martin G. Larson, William B. Kannel, and Ramachandran S. Vasan. Obesity and the risk of heart failure. *New England Journal of Medicine*, 347(5):305–313, August 2002. **DOI** 10.1056/nejmoa020245. URL https://doi.org/10.1056/nejmoa020245.

B. Kitchenham and S Charters. Guidelines for performing systematic literature reviews in software engineering, 2007.

Spencer H. Kubo, Steven Schulman, Randall C. Starling, Mariell Jessup, Deborah Wentworth, and Daniel Burkhoff. Development and validation of a patient questionnaire to determine New York heart association classification. *Journal of Cardiac Failure*, 10(3):228–235, June 2004. ISSN 10719164. **DOI** 10.1016/j.cardfail.2003.10.005. URL https://linkinghub.elsevier.com/retrieve/pii/S1071916403007450.

C. Lenfant. High blood pressure: some answers, new questions, continuing challenges. *JAMA: The Journal of the American Medical Association*, 275(20):1604–1606, May 1996. **DOI** 10.1001/jama.275.20.1604. URL https://doi.org/10.1001/jama.275.20.1604.

M. A. B. Lima. Ecgnow telecardiologia. https://www.ecgnow.com.br/blog/3-dicas-importantes-sobre-ecg-em-criancas/, 2018. (Accessed on 08/17/2020).

Gregory Y H Lip. Chads-vasc score for atrial fibrillation stroke risk - mdcalc. https://www.mdcalc.com/calc/801/cha2ds2-vasc-score-atrial-fibrillation-stroke-risk. (Accessed on 08/16/2020).

Gregory Y.H. Lip, Robby Nieuwlaat, Ron Pisters, Deirdre A. Lane, and Harry J.G.M. Crijns. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach. *Chest*, 137(2):263–272, February 2010. **DOI** 10.1378/chest.09-1584. URL https://doi.org/10.1378/chest.09-1584.

Tianyu Liu. Data for: A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical-datasets, November 2019. URL https://data.mendeley.com/datasets/x8ygrw87jw/1. Type: dataset.

Tianyu Liu, Wenhui Fan, and Cheng Wu. A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. *Artificial Intelligence in Medicine*, 101: 101723, November 2019. ISSN 09333657. **DOI** 10.1016/j.artmed.2019.101723. URL https://linkinghub.elsevier.com/retrieve/pii/S0933365719302295.

Ana Carolina Lorena and André C. P. L. F. De Carvalho. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, 14(2):43–67, December 2007. **DOI** 10.22456/2175-2745.5690. URL https://doi.org/10.22456/2175-2745.5690.

Beatriz Maia. Tipos de aprendizado de máquina #3 - dev community. https://dev.to/beatrizmaiads/tipos-de-aprendizado-de-maquina-3-5d66, 2020. (Accessed on 08/18/2020).

Frederick A Masoudi, Edward P Havranek, Grace Smith, Ronald H Fish, John F Steiner, Diana L Ordin, and Harlan M Krumholz. Gender, age, and heart failure with preserved left ventricular systolic function. *Journal of the American College of Cardiology*, 41(2):217–223, 2003. **DOI** 10.1016/S0735-1097(02)02696-7. URL https://www.jacc.org/doi/abs/10.1016/S0735-1097%2802%2902696-7.

Mendeley. Reference manager - mendeley | elsevier solutions. https://www.elsevier.com/solutions/mendeley. (Accessed on 08/14/2020).

Brent Mitchell. Bloqueio atrioventricular - doenças cardiovasculares - manuais msd edição para profissionais. https://www.msdmanuals.com/pt/profissional/doen%C3%A7as-cardiovasculares/arritmias-e-doen%C3%A7as-de-condu%C3%A7%C3%A3o/bloqueio-atrioventricular, 2021. (Accessed on 08/17/2020).

Tom Mitchell, Bruce Buchanan, Gerald DeJong, Thomas Dietterich, Paul Rosenbloom, and Alex Waibel. Machine learning. *Annual review of computer science*, 4(1):417–433, 1990.

George B Moody and Roger G Mark. MIT-BIH Arrhythmia Database, 1992. URL https://physionet.org/content/mitdb/. Type: dataset.

Dariush Mozaffarian, Emelia J. Benjamin, Alan S. Go, Donna K. Arnett, Michael J. Blaha, Mary Cushman, Sandeep R. Das, Sarah de Ferranti, Jean-Pierre Després, Heather J. Fullerton, Virginia J. Howard, Mark D. Huffman, Carmen R. Isasi, Monik C. Jiménez, Suzanne E. Judd, Brett M. Kissela, Judith H. Lichtman, Lynda D. Lisabeth, Simin Liu, Rachel H. Mackey, David J. Magid, Darren K. McGuire, Emile R. Mohler, Claudia S. Moy, Paul Muntner, Michael E. Mussolino, Khurram Nasir, Robert W. Neumar, Graham Nichol, Latha Palaniappan, Dilip K. Pandey, Mathew J. Reeves, Carlos J. Rodriguez, Wayne Rosamond, Paul D. Sorlie, Joel Stein, Amytis Towfighi, Tanya N. Turan, Salim S. Virani, Daniel Woo, Robert W. Yeh, and Melanie B. Turner. Heart disease and stroke statistics—2016 update. *Circulation*, 133(4), January 2016. **DOI** 10.1161/cir.0000000000000350. URL https://doi.org/10.1161/cir.0000000000000350.

Kevin P Murphy et al. Naive bayes classifiers. *University of British Columbia*, 18(60):1–8, 2006.

Peter Brønnum Nielsen, Torben Bjerregaard Larsen, Flemming Skjøth, Thure Filskov Overvad, and Gregory Y. H. Lip. Stroke and thromboembolic event rates in atrial fibrillation according to different guideline treatment thresholds: A nationwide cohort study. *Scientific Reports*, 6 (1), June 2016. **DOI** 10.1038/srep27410. URL https://doi.org/10.1038/srep27410.

Markku S. Nieminen, Veli-Pekka Harjola, Matthias Hochadel, Helmut Drexler, Michel Komajda, Dirk Brutsaert, Kenneth Dickstein, Piotr Ponikowski, Luigi Tavazzi, Ferenc Follath, and Jose Luis Lopez-Sendon. Gender related differences in patients presenting with acute heart failure. results from EuroHeart failure survey II. *European Journal of Heart Failure*, 10(2): 140–148, February 2008. **DOI** 10.1016/j.ejheart.2007.12.012. URL https://doi.org/10.1016/j.ejheart.2007.12.012.

Shu Lih Oh, Eddie Y.K. Ng, Ru San Tan, and U. Rajendra Acharya. Automated diagnosis of arrhythmia using combination of cnn and lstm techniques with variable length heart beats. *Computers in Biology and Medicine*, 102:278–287, 2018. ISSN 0010-4825. **DOI** https://doi.org/10.1016/j.compbiomed.2018.06.002. URL https://www.sciencedirect.com/science/article/pii/S0010482518301446.

OMS. Cardiovascular diseases. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1. (Accessed on 06/14/2020).

The pandas development team. pandas-dev/pandas: Pandas, February 2020. URL https://doi.org/10.5281/zenodo.3509134.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Rafael Pereira, Rodrigo Varejão Andreão, and Gabriel Tozatto Zago. Identificação de fibrilação atrial em registros eletrocardiograma utilizando redes neurais recorrentes do tipo long short-term memory. In *Anais do 14º Simpósio Brasileiro de Automação Inteligente*. Galoa, 2019. **DOI** 10.17648/sbai-2019-111461. URL https://doi.org/10.17648/sbai-2019-111461.

Artzai Picon, Unai Irusta, Aitor Álvarez-Gila, Elisabete Aramendi, Felipe Alonso-Atienza, Carlos Figuera, Unai Ayala, Estibaliz Garrote, Lars Wik, Jo Kramer-Johansen, and Trygve Eftestøl. Mixed convolutional and long short-term memory network for the detection of lethal ventricular arrhythmia. *PLOS ONE*, 14(5):e0216756, May 2019. **DOI** 10.1371/journal.pone.0216756. URL https://doi.org/10.1371/journal.pone.0216756.

PubMed. Pubmed. https://pubmed.ncbi.nlm.nih.gov/. (Accessed on 08/14/2020).

Sai Manoj Pudukotai Dinakarrao and Axel Jantsch. Addhard: Arrhythmia detection with digital hardware by learning ecg signal. In *Proceedings of the 2018 on Great Lakes Symposium on VLSI*, GLSVLSI '18, page 495–498, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357241. **DOI** 10.1145/3194554.3194647. URL https://doi.org/10.1145/3194554.3194647.

Gene R. Quinn, Olivia N. Severdija, Yuchiao Chang, and Daniel E. Singer. Wide variation in reported rates of stroke across cohorts of patients with atrial fibrillation. *Circulation*, 135(3): 208–219, January 2017. **DOI** 10.1161/circulationaha.116.024057. URL https://doi.org/10.1161/circulationaha.116.024057.

Pierre Raybaut. Spyder-documentation. *Available online at: pythonhosted. org*, 2009.

Hannah Ritchie. O que mais mata as pessoas ao redor do mundo? *BBC News Brasil*, 03 2019. URL https://www.bbc.com/portuguese/internacional-47471602.

Claude Sammut and Geoffrey I. Webb. *Encyclopedia of Machine Learning and Data Mining*. Springer Publishing Company, Incorporated, 2nd edition, 2017. ISBN 148997685X.

Sanarmed. Bradicardia (bradiarritmia): Resumo completo com fluxograma! - sanar medicina. https://www.sanarmed.com/bradicardia-bradiarritmia-resumo-fluxograma-yellowbook, Setembro 2019. (Accessed on 08/17/2020).

SANTOS. Coração: função, camadas, cavidades e válvulas - mundo educação. https://mundoeducacao.uol.com.br/biologia/coracao.htm, 2020. (Accessed on 08/17/2020).

ScienceDirect. Sciencedirect.com | science, health and medical journals, full text articles and books. https://www.sciencedirect.com/. (Accessed on 08/14/2020).

Mohsen Shahhosseini. Microsoft word - mohsenshahhosseini.docx. https://arxiv.org/ftp/arxiv/papers/2009/2009.00534.pdf. (Accessed on 08/20/2022).

Tariq Jamal Siddiqi, Muhammad Shariq Usman, Izza Shahid, Jawad Ahmed, Safi U Khan, Lina Ya'qoub, Charanjit S Rihal, and Mohamad Alkhouli. Utility of the CHA2DS2-VASc score for predicting ischaemic stroke in patients with or without atrial fibrillation: a systematic review and meta-analysis. *European Journal of Preventive Cardiology*, 29(4):625–631, March 2022. ISSN 2047-4873, 2047-4881. **DOI** 10.1093/eurjpc/zwab018. URL https://academic.oup.com/eurjpc/article/29/4/625/6163157.

Rogério Gomes da Silva, Gustavo Glotz de Lima, Andréia Laranjeira, Altamiro Reis da Costa, Edemar Pereira, and Rubem Rodrigues. Fatores de risco e morbimortalidade associados à fibrilação atrial no pós-operatório de cirurgia cardíaca. *Arquivos Brasileiros de Cardiologia*, 83(2):99–104, August 2004. ISSN 0066-782X. **DOI** 10.1590/S0066-782X2004001400002. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0066-782X2004001400002&lng=pt&nrm=iso&tlng=pt.

Donald S. Silverberg, Dov Wexler, Adrian Iaina, and Doron Schwartz. The interaction between heart failure and other heart diseases, renal failure, and anemia. *Seminars in Nephrology*, 26 (4):296–306, 2006. ISSN 0270-9295. **DOI** https://doi.org/10.1016/j.semnephrol.2006.05.006. URL https://www.sciencedirect.com/science/article/pii/S0270929506000726. The Kidney and Anemia.

Springer. Brazil. https://www.springer.com/br. (Accessed on 08/14/2020).

UFMG. Morte por doenças cardiovasculares aumentaram durante pandemia - faculdade de medicina da ufmg. https://www.medicina.ufmg.br/morte-por-doencas-cardiovasculares-aumentaram-durante-pandemia/. (Accessed on 08/17/2020).

Barry F Uretsky and Richard G Sheahan. Primary prevention of sudden cardiac death in heart failure: Will the solution be shocking? *Journal of the American College of Cardiology*, 30(7): 1589–1597, December 1997. **DOI** 10.1016/s0735-1097(97)00361-6. URL https://doi.org/10.1016/s0735-1097(97)00361-6.

USP. Redes neurais artificiais. https://sites.icmc.usp.br/andre/research/neural/. (Accessed on 08/20/2022).

Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

Stephen D. Webb and John Collette. Rural-urban stress: New data and new conclusions. *American Journal of Sociology*, 84(6):1446–1452, May 1979. **DOI** 10.1086/226945. URL https://doi.org/10.1086/226945.

IEEE Xplore. Ieee xplore. https://ieeexplore.ieee.org/Xplore/home.jsp. (Accessed on 08/14/2020).

Zotero. Zotero | your personal research assistant. https://www.zotero.org/. (Accessed on 08/14/2020).

Özal Yıldırım, Paweł Pławiak, Ru-San Tan, and U. Rajendra Acharya. Arrhythmia detection using deep convolutional neural network with long duration ecg signals. *Computers in Biology and Medicine*, 102:411–420, 2018. ISSN 0010-4825. **DOI** https://doi.org/10.1016/j.compbiomed.2018.09.009. URL https://www.sciencedirect.com/science/article/pii/S0010482518302713.

Apêndice A

Código em Python

Algoritmo A.1: Tratamento dos dados e Geração dos Scores e Predições

```
1
2  ###LEITURA E LIMPEZA DOS DADOS ORIGINAIS:
3  from sklearn.model_selection import cross_val_score
4  import numpy as np
5  import pandas as pd
6  from sklearn.preprocessing import LabelEncoder
7  from time import time
8  from sklearn.preprocessing import MinMaxScaler
9  from sklearn.decomposition import PCA
10
11 df = pd.read_csv("healthcare-dataset-stroke-data.csv",
12 low_memory=False)
13
14 print(len(df.index), 'dataset original')
15
16 #remove rows with any nan value
17 df = df.dropna()
18
19 print(len(df.index), 'sem nan')
20 print(len(df.columns), 'com todas colunas')
21
22
23 # Exclus o da coluna de id dos pacientes, se foi casado,
24 #tipo de trabalho
25 df = df.drop(['id'], axis=1)
26 df = df.drop(['ever_married'], axis=1)
27 df = df.drop(['work_type'], axis=1)
28
```

```
print(len(df.columns),'sem colunas removidas{id, married, work type}')

# remove rows with gender value 'Other'
df = df[df['gender'] != 'Other']


print(len(df.index), 'removendo genero other')

# Exclus o de inst ncia que n o tem a info de smoke
df = df[df['smoking_status'] != 'Unknown']

print(len(df.index),'removendo pacientes sem i n f o r m a  o sobre
tabagismo')



###INFERE ESTRESSE COM BASE NO TIPO DE RESID NCIA
# R e l a  o entre o tipo de residencia e o n v e l de estresse
df.rename(columns = {'Residence_type':'stress'}, inplace = True)
df = df.replace('Urban', 1)
df = df.replace('Rural', 0)




#TRANSFORMA O DOS DADOS PARA OS PAR METROS NECESS RIOS
### preprocess original data ####
original_df = df

# Gender
df = df.replace('Male', 0)
df = df.replace('Female', 1)


# Set score according age of pacient
# categoriza idade de acordo com os parametros do hasvasc
df.loc[df.age < 65, 'age'] = 0
df.loc[((df.age >= 65) & (df.age < 75)), 'age'] = 1
df.loc[df.age >= 75, 'age'] = 2

#categoriza diabetes de acordo com o n v e l de glicose
```

```
df['avg_glucose_level'] = np.where((df.avg_glucose_level < 126),
0,df.avg_glucose_level)
df['avg_glucose_level'] = np.where((df.avg_glucose_level >= 126),
1,df.avg_glucose_level) #Diabetes

# Set score according smoking of pacient
df['smoking_status'] = np.where((df.smoking_status == 'smokes'),
2,df.smoking_status) #Fumante
df['smoking_status'] =np.where((df.smoking_status=='formerly␣smoked'),
1,df.smoking_status) #Ex Fumante
df['smoking_status'] =np.where((df.smoking_status == 'never␣smoked'),
0,df.smoking_status)

# categorizar bmi (imc) para sobrepeso e obesidade
df['bmi'] = np.where((df.bmi < 25),0,df.bmi)

#Sobrepeso
df['bmi'] = np.where(((df.bmi >= 25) & (df.bmi <= 30)),1,df.bmi)

#Obesidade
df['bmi'] = np.where((df.bmi > 30),2,df.bmi)



# renomear nomes de colunas
df.rename(columns = {'bmi':'imc'}, inplace = True)
df.rename(columns = {'stroke':'avc'}, inplace = True)
df.rename(columns = {'age':'idade'}, inplace = True)
df.rename(columns = {'gender':'sexo'}, inplace = True)
df.rename(columns = {'hypertension':'hipertensao'}, inplace = True)
df.rename(columns = {'heart_disease':'doenca_vascular'},inplace=True)
df.rename(columns = {'stress':'estresse'}, inplace = True)
df.rename(columns = {'smoking_status':'tabagismo'}, inplace = True)
df.rename(columns = {'avg_glucose_level':'diabetes'}, inplace = True)

# avc to 2
df['avc'] = np.where((df.avc == 1),2,df.avc)


#extrair vari vel insuficiencia card aca a partir das informacoes
#obtidas no dataset
```

```
#(Hipertens o, Idade >= 70, Tabagismo, Diabetes (genero(2x homem,
#4x mulher)), Obesidade, AVC) = % ic (insuficiencia cardiaca)

# cria copia das colunas de interesse, com excecao da coluna age
ic_calc = df[['sexo', 'idade', 'hipertensao', 'tabagismo',
'diabetes', 'imc', 'avc']].copy()


# processamento do score de diabetes baseado no sexo
ic_calc['novo_diabetes'] = 0
ic_calc['novo_diabetes'] = np.where(((ic_calc.sexo == 1) &
(ic_calc['diabetes'] == 1)),2,ic_calc.novo_diabetes)
# se mulher e diabetes = 2
ic_calc['novo_diabetes'] = np.where(((ic_calc.sexo == 0) &
(ic_calc['diabetes'] == 1)),1,ic_calc.novo_diabetes)
# se homem e diabetes = 1

# remove as colunas que nao vao ser mais usadas
ic_calc = ic_calc.drop(['sexo'], axis=1)
ic_calc = ic_calc.drop(['diabetes'], axis=1)

# renomear coluna diabetes
ic_calc.rename(columns = {'novo_diabetes':'diabetes'},
inplace = True)


ic_calc['score'] = ic_calc['hipertensao'] + ic_calc['idade'] +
ic_calc['avc'] + ic_calc['diabetes'] + ic_calc['imc'] +
ic_calc['tabagismo']

column = 'score'
ic_calc[column] = (ic_calc[column] - ic_calc[column].min()) /
(ic_calc[column].max()- ic_calc[column].min())

#PCA para indicadores do ic calc
pca = PCA(n_components=1)
pca.fit(ic_calc)
pca_resultado = pca.components_

#Define corte para classificar IC
ic_index = 0.5
```

```
ic_calc = ic_calc.drop(ic_calc[ic_calc.score < ic_index].index)

new_df_index_array = ic_calc.index.tolist()

df = df.filter(items = new_df_index_array, axis=0)

df['ic'] = 1

# converte columns to numeric
df["sexo"] = pd.to_numeric(df["sexo"])
df["tabagismo"] = pd.to_numeric(df["tabagismo"])

# reordenar colunas
# df = df[['ic', 'hipertensao', 'idade', 'diabetes', 'avc',
'doenca_vascular', 'sexo', 'estresse', 'imc', 'tabagismo']]


###CALCULO E CLASSIFICACAO DOS RISCOS POR CHA2DS2VASC (Lip, 2010;
#Friberg, 2012)
# create new column adapted chadvasc
df['cha2ds2_vasc'] = df['hipertensao'] + df['idade'] + df['avc'] +
df['doenca_vascular'] + df['sexo'] + df['ic']
df['habitos'] = df['estresse'] + df['tabagismo'] + df['imc']


#novas colunas relacionados ao risco de acordo com chadvasc


#CLASSIFICA O CHA2DS2-VASc (Friberg 2012)
df['f_risco_avc_isquemico_anual'] = 0
df['f_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 0),
0.2,df.f_risco_avc_isquemico_anual)
df['f_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 1),
0.6,df.f_risco_avc_isquemico_anual)
df['f_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 2),
2.2,df.f_risco_avc_isquemico_anual)
df['f_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 3),
3.2,df.f_risco_avc_isquemico_anual)
df['f_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 4),
4.8,df.f_risco_avc_isquemico_anual)
df['f_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 5),
```

```
7.2,df.f_risco_avc_isquemico_anual)
df['f_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 6),
9.7,df.f_risco_avc_isquemico_anual)
df['f_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 7),
11.2,df.f_risco_avc_isquemico_anual)
df['f_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 8),
10.8,df.f_risco_avc_isquemico_anual)
df['f_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 9),
12.2,df.f_risco_avc_isquemico_anual)

df['risco_avc_ait_embolia_sistemica'] = 0
df['risco_avc_ait_embolia_sistemica'] = np.where((df.cha2ds2_vasc
== 0),0.3,df.risco_avc_ait_embolia_sistemica)
df['risco_avc_ait_embolia_sistemica'] = np.where((df.cha2ds2_vasc
== 1),0.9,df.risco_avc_ait_embolia_sistemica)
df['risco_avc_ait_embolia_sistemica'] = np.where((df.cha2ds2_vasc
== 2),2.9,df.risco_avc_ait_embolia_sistemica)
df['risco_avc_ait_embolia_sistemica'] = np.where((df.cha2ds2_vasc
== 3),4.8,df.risco_avc_ait_embolia_sistemica)
df['risco_avc_ait_embolia_sistemica'] = np.where((df.cha2ds2_vasc
== 4),6.7,df.risco_avc_ait_embolia_sistemica)
df['risco_avc_ait_embolia_sistemica'] = np.where((df.cha2ds2_vasc
== 5),10.0,df.risco_avc_ait_embolia_sistemica)
df['risco_avc_ait_embolia_sistemica'] = np.where((df.cha2ds2_vasc
== 6),13.6,df.risco_avc_ait_embolia_sistemica)
df['risco_avc_ait_embolia_sistemica'] = np.where((df.cha2ds2_vasc
== 7),15.2,df.risco_avc_ait_embolia_sistemica)
df['risco_avc_ait_embolia_sistemica'] = np.where((df.cha2ds2_vasc
== 8),15.7,df.risco_avc_ait_embolia_sistemica)
df['risco_avc_ait_embolia_sistemica'] = np.where((df.cha2ds2_vasc
== 9),17.4,df.risco_avc_ait_embolia_sistemica)



#CLASSIFICA O CHA2DS2-VASc (Lip 2010)
df['l_risco_avc_isquemico_anual'] = 0
df['l_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 0),
0.0,df.l_risco_avc_isquemico_anual)
df['l_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 1),
0.6,df.l_risco_avc_isquemico_anual)
df['l_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 2),
```

```
1.6,df.l_risco_avc_isquemico_anual)
df['l_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 3),
3.9,df.l_risco_avc_isquemico_anual)
df['l_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 4),
1.9,df.l_risco_avc_isquemico_anual)
df['l_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 5),
3.2,df.l_risco_avc_isquemico_anual)
df['l_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 6),
3.6,df.l_risco_avc_isquemico_anual)
df['l_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 7),
8.0,df.l_risco_avc_isquemico_anual)
df['l_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 8),
11.1,df.l_risco_avc_isquemico_anual)
df['l_risco_avc_isquemico_anual'] = np.where((df.cha2ds2_vasc == 9),
100,df.l_risco_avc_isquemico_anual)


df['ic'] = df['ic'].astype("int")
df['avc'] = df['avc'].astype("int")
df['imc'] = df['imc'].astype("int")
df['idade'] = df['idade'].astype("int")
df['sexo'] = df['sexo'].astype("int")
df['hipertensao'] = df['hipertensao'].astype("int")
df['doenca_vascular'] = df['doenca_vascular'].astype("int")
df['estresse'] = df['estresse'].astype("int")
df['tabagismo'] = df['tabagismo'].astype("int")
df['diabetes'] = df['diabetes'].astype("int")


### CALCULO E NORMALIZACAO DO SCORE CHA2DS2-VASc E HABITOS
# resultado da soma
df['resultado_score'] =
(df['cha2ds2_vasc'] * 0.7) + (df['habitos'] * 0.3)

df['resultado_normalizado'] = df['resultado_score']


# # normalizar resultados <<<<
column = 'resultado_normalizado'
min_value = 0;
max_value = 13; #atribuir maior obtido pelos scores
```

```
df[column] =
(df[column] - df[column].min()) / (df[column].max()- df[column].min())
df[column] = df[column] * 10

pca.fit(df)
pca_resultado_df = pca.components_


###CATEGORIZA O NYHA POR QUARIL
# cria novas categorias para inferencia
df.loc[df['resultado_normalizado'] <= 2.5, 'resultado_categoria']= 'I'
df.loc[(df['resultado_normalizado'] > 2.5) &
(df['resultado_normalizado'] <= 5), 'resultado_categoria'] = 'II'
df.loc[(df['resultado_normalizado'] > 5) &
(df['resultado_normalizado'] <= 7.5), 'resultado_categoria'] = 'III'
df.loc[df['resultado_normalizado'] > 7.5, 'resultado_categoria']= 'IV'

###MORTALIDADE DAS CATEGORIAS NYHA SEM E COM TRATAMENTO
# Atribui mortalidade com base na categoria (Sem tratamento)
df['mort_sem_tra_min']= 0
df['mort_sem_tra_max'] = 0
df['mort_sem_tra_min'] = np.where((df.resultado_categoria
== 'I'),0.05,df.mort_sem_tra_min)
df['mort_sem_tra_max'] = np.where((df.resultado_categoria
== 'I'),0.19,df.mort_sem_tra_max)

df['mort_sem_tra_min'] = np.where((df.resultado_categoria
== 'II'),0.15,df.mort_sem_tra_min)
df['mort_sem_tra_max'] = np.where((df.resultado_categoria
== 'II'),0.40,df.mort_sem_tra_max)

df['mort_sem_tra_min'] = np.where((df.resultado_categoria
== 'III'),
0.15,df.mort_sem_tra_min)
df['mort_sem_tra_max'] = np.where((df.resultado_categoria
== 'III'),
0.40,df.mort_sem_tra_max)

df['mort_sem_tra_min'] = np.where((df.resultado_categoria
== 'IV'),
0.44,df.mort_sem_tra_min)
```

```
316 df['mort_sem_tra_max'] = np.where((df.resultado_categoria
317 == 'IV'),
318 0.66,df.mort_sem_tra_max)
319
320
321 # Atribui mortalidade com base na categoria (Com tratamento)
322 df['mort_com_tra_min'] = 0
323 df['mort_com_tra_max'] = 0
324 df['mort_com_tra_min'] = np.where((df.resultado_categoria
325 == 'I'),0.05,df.mort_com_tra_min)
326 df['mort_com_tra_max'] = np.where((df.resultado_categoria
327 == 'I'),0.10,df.mort_com_tra_max)
328
329 df['mort_com_tra_min'] = np.where((df.resultado_categoria
330 == 'II'),0.05,df.mort_com_tra_min)
331 df['mort_com_tra_max'] = np.where((df.resultado_categoria
332 == 'II'),0.10,df.mort_com_tra_max)
333
334 df['mort_com_tra_min'] = np.where((df.resultado_categoria
335 == 'III'),0.10,df.mort_com_tra_min)
336 df['mort_com_tra_max'] = np.where((df.resultado_categoria
337 == 'III'),0.15,df.mort_com_tra_max)
338
339 df['mort_com_tra_min'] = np.where((df.resultado_categoria
340 == 'IV'),0.30,df.mort_com_tra_min)
341 df['mort_com_tra_max'] = np.where((df.resultado_categoria
342 == 'IV'),0.40,df.mort_com_tra_max)
343
344 # Atribui mortalidade com base na categoria (avg)
345 df['nyha_mortality_min'] = 0
346 df['nyha_mortality_max'] = 0
347 df['nyha_mort_com_tra_min'] = np.where((df.resultado_categoria
348 == 'I'),0.10,df.mort_com_tra_min)
349 df['nyha_mortality_max'] = np.where((df.resultado_categoria
350 == 'I'),0.15,df.mort_com_tra_max)
351
352 df['nyha_mortality_min'] = np.where((df.resultado_categoria
353 == 'II'),0.10,df.mort_com_tra_min)
354 df['nyha_mortality_max'] = np.where((df.resultado_categoria
355 == 'II'),0.15,df.mort_com_tra_max)
356
```

```
df['nyha_mortality_min'] = np.where((df.resultado_categoria
== 'III'),0.15,df.mort_com_tra_min)
df['nyha_mortality_max'] = np.where((df.resultado_categoria
== 'III'),0.20,df.mort_com_tra_max)

df['nyha_mortality_min'] = np.where((df.resultado_categoria
== 'IV'),0.20,df.mort_com_tra_min)
df['nyha_mortality_max'] = np.where((df.resultado_categoria
== 'IV'),0.50,df.mort_com_tra_max)

print('categoria␣I␣->␣', df[df['resultado_categoria'] == 'I']
.shape[0])
print('categoria␣II␣->␣', df[df['resultado_categoria'] == 'II']
.shape[0])
print('categoria␣III␣->␣', df[df['resultado_categoria'] == 'III']
.shape[0])
print('categoria␣IV␣->␣', df[df['resultado_categoria'] == 'IV']
.shape[0])

###ORGANIZA  O E EXPORTA  O DOS DADOS PARA O ORANGE DATA MINING
print(list(df.columns))
#ordenando colunas do df
df.loc[:, ['ic', 'hipertensao', 'idade', 'diabetes', 'avc',
            'doenca_vascular', 'sexo', 'cha2ds2_vasc',
           'f_risco_avc_isquemico_anual',
           'risco_avc_ait_embolia_sistemica',
           'l_risco_avc_isquemico_anual',

           'estresse', 'imc', 'tabagismo', 'habitos',
           'resultado_score', 'resultado_normalizado',
           'resultado_categoria',

           'nyha_mortality_min','nyha_mortality_max',
           'mort_sem_tra_min', 'mort_sem_tra_max',
           'mort_com_tra_min', 'mort_com_tra_max']]




df_saidas = df[['cha2ds2_vasc', 'habitos', 'resultado_normalizado',
'resultado_categoria','f_risco_avc_isquemico_anual',
```

```
'risco_avc_ait_embolia_sistemica', 'l_risco_avc_isquemico_anual',
'mort_sem_tra_min', 'mort_sem_tra_max', 'mort_com_tra_min',
'mort_com_tra_max']]
df_saidas.insert(0, 'n', df_saidas.index.tolist())
df_saidas.reset_index(drop=True, inplace = True)


# exportar dataframes
df.to_csv('preprocess_result.csv', sep=',', encoding='utf-8',
index=False)
original_df.to_csv('original_df.csv', sep=',', encoding='utf-8',
index=False)
ic_calc.to_csv('ic_calc.csv', sep=',', encoding='utf-8', index=False)
```