Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Programa de Pós-Graduação em Estatística

ADENICE GOMES DE OLIVEIRA FERREIRA

CLASSIFICAÇÃO BINÁRIA PARA PRESENÇA DE OCORRÊNCIA DE CARDIOPATIAS USANDO CARACTERÍSTICAS CLÁSSICAS E NOVOS PARÂMETROS

Recife

2019

ADENICE GOMES DE OLIVEIRA FERREIRA

CLASSIFICAÇÃO BINÁRIA PARA PRESENÇA DE OCORRÊNCIA DE CARDIOPATIAS USANDO CARACTERÍSTICAS CLÁSSICAS E NOVOS PARÂMETROS

Dissertação apresentada ao Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial à obtenção do título de mestre em Estatística.

Área de Concentração: Estatística Aplicada

Orientador: Prof^{o} Dr. Raydonal Ospina Martínez

Recife

2019

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

F383c Ferreira, Adenice Gomes de Oliveira
Classificação binária para presença de ocorrência de cardiopatias usando características clássicas e novos parâmetros / Adenice Gomes de Oliveira Ferreira. – 2019.
98 f.: il., fig., tab.

Orientador: Raydonal Ospina Martínez.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2019.
Inclui referências e anexo.

1. Estatística. 2. Classificação binária. 3. Cardiopatia. I. Martínez, Raydonal Ospina (orientador). II. Título.

310 CDD (23. ed.) UFPE- MEI 2019-127

ADENICE GOMES DE OLIVEIRA FERREIRA

CLASSIFICAÇÃO BINÁRIA PARA PRESENÇA DE OCORRÊNCIA DE CARDIOPATIAS USANDO CARACTERÍSTICAS CADÍACAS CLÁSSICAS E NOVOS PARÂMETROS

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovada em: 30 de julho de 2019.

BANCA EXAMINADORA

Prof.(º) Raydonal Ospina Martínez
UFPE

Prof.(º) Marcelo Rodrigo Portela Ferreira
UFPB

Prof.(º) Hélio Magalhães de Oliveira
UFPE

À família que tanto amo e aos amigos que tanto prezo.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelo seu imenso amor e misericórdia, pelo seu carinho e cuidado comigo e minha família, pela força que nos tem dado para transpor os obstáculos e continuar crescendo em graça e sabedoria.

Agradeço aos meus pais por sempre me apoiarem e orientarem em todas as decisões importantes que tomei. Vocês são meu tesouro e exemplos vivos de amor, união e, sobretudo, ética. Tenho orgulho de tê-los como meus principais professores, pois "amar a Deus sobre todas as coisas e ao próximo como a si mesmo" foi uma lição que aprendi primeiro com vocês.

Agradeço ao meu irmão, João Marcos, pelas conversas e desabafos, pela cumplicidade. Te admiro muito pela sua dedicação, disciplina, perseverança, e pelo seu coração enorme que acolhe a todos, sem distinção. Você é um exemplo a ser seguido.

Agradeço ao meu doce Diogo, que me acompanha desde antes do início dessa caminhada. Você é um dos meus maiores incentivadores, meu companheiro de tantos momentos, é o presente que Deus me deu. Poder continuar crescendo contigo é um privilegio para mim.

Agradeço aos amados amigos e familiares que de forma direta ou indireta contribuíram com o aprendizado nesta etapa da minha vida. Em especial aos meus sogros, Sandra e Ginaldo, minha tia Adriana, minha prima Elisa e seu esposo Elton, minha amiga Ellen Winny, meu amigo Wesley e às amoras da minha conexão. Obrigada por todo apoio, oração e torcida.

Agradeço à Anny, que nesse período dividiu não só o apartamento comigo, mas também o estudo, as alegrias, as tristezas. Você é um ser humano incrível e tem mais força e potencial do que imagina.

Agradeço aos amigos que trouxe comigo desde a graduação - André, Zé, Jodavid e Saul - e às novas amizades que cultivei - Anabeth, Lucas, Eduardo, João Eudes Miqueias, Jairo, César, Pedro, Jonas, Daniel, Joás, Jordan, Cristine, Ranah, Larissa, seu Luis, Fernando, Yuri, Ana Cristina e Bruna - com certeza vocês tornaram essa jornada mais leve e prazerosa. Desejo que nossa amizade siga para toda vida e que possamos continuar nos encontrando, colocando a

conversa em dia e rindo dos momentos que vivemos.

Agradeço à secretária Valéria Bittencourt, que sempre disposta, me orientou em todos os processos administrativos, me tratando com muita humanidade desde o primeiro dia que pisei no departamento de Estatística.

Agradeço a psicóloga Dra. Cássia. A senhora foi um presente que Deus colocou na minha vida no momento certo. Existe um pedacinho da senhora refletido não só neste trabalho, mas também na minha história de vida.

Agradeço ao prof^{o} Raydonal pela confiança, por me orientar, por me corrigir quando necessário, por me incentivar e dar forças para continuar.

Meus sinceros agradecimentos ao prof^{o} Hélio Magalhães, pela sua significativa contribuição intelectual ao trabalho, e ao prof^{o} Marcelo Ferreira, pela solicitude em participar da banca que contribuiu no enriquecimento deste trabalho.

Por fim, meus agradecimentos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

RESUMO

Essa dissertação tem como objetivo utilizar características cardíacas clássicas e os novos parâmetros introduzidos por Campello de Souza (2010) [O Apoio ao Diagnóstico Médico: o que se pode fazer com um tensiômetro e um relógio. 2. ed. Recife] no processo de classificação de indivíduos cardiopatas. Considerando distintos cenários de análise e baseados em quatro diferentes bancos de dados, os parâmetros de Campello de Souza foram incluídos no processo de seleção dos atributos mais informativos e no processo de classificação binária dos cardiopatas. Foram testados cinco classificadores bem consolidados na literatura a saber: *Naive Bayes*, Florestas Aleatórias, Regressão Logística, *Adaboost* e Máquinas de Vetores de Suporte. Os desempenhos destes classificadores foram avaliados com base nas acurácias e respectivos desvios padrões (DPs). Dada a alta dimensionalidade das matrizes de características contínuas usadas e sob ausência de ortogonalidade, as classificações foram também avaliadas utilizando Componentes Principais. Nessa fase é adicionando um sexto método de classificação: as Redes Neuronais Artificiais. Os resultados empíricos indicam que dentre os parâmetros introduzidos por Campello de Souza, a Pressão Arterial Média (PAM), que aparece em 8 dos 12 modelos selecionados pelo fator de inflação de variância VIF melhora o desempenho dos classificadores, apresentando acurácias que variaram entre 78.77% (DP = 4.54%) e 99.20% (DP = 1.17%), respectivamente. Considerando os classificadores, a Regressão Logística e o *Adaboost* foram os métodos com maiores médias de acurácias, cada classificador presente em um terço dos 12 modelos selecionados pelo VIF. Dominic, Gupta e Khare (2015) obtiveram 98% de acurácia com o classificador Adaboost, Umamaheswuari et al. (2017) obtiveram 91.89% com o classificador *Stacking*, enquanto que neste trabalho e para o mesmo banco de dados encontrou-se resultados mais competitivos na classificação dos cardiopatas, sendo a Regressão Logística o modelo contendo dentre suas variáveis explicativas a PAM, o Índice Pulsátil da Pressão Arterial (IPPA) e o parâmetro RC (Resistência $\times$ Complacência), obtendo uma acurácia média nas bases de teste foi igual a 99.20% (DP = 1.17%).

Palavras-chave: Cardiopatia. Classificação binária. Parâmetros de Campello de Souza.

ABSTRACT

The aim of this work is to use classic cardiac characteristics and new parameters introduced by Campello de Souza (2010) [Support of Medical Diagnosis: what can be done with a tensiometer and a clock?. 2. Ed. Recife] in the process of classification of individuals with heart disease. Considering different scenarios, and based on four different databases, the parameters of Campello de Souza were included in the selection process of the most informative attributes and added in the binary classification process of the cardiac patients. Five well-consolidated classifiers were tested: Naive Bayes, Random Forests, Logistic Regression, Adaboost and Support Vector Machine. The performances of these classifiers were evaluated based on the accuracy and their respective standard deviations (SDs). Given the high dimensionality of the matrices of continuous features used and in the absence of orthogonality, the classifiers were also evaluated using Principal Components. In this phase, we are adding the Artificial Neural Networks as a sixth classification method. The empirical results indicate that among the parameters introduced by Campello de Souza, the mean arterial pressure (PAM), which appears in 8 of the 12 models selected by the VIF variance inflation factor, improves the performance of the classifiers, with accuracy ranging from 78.77% (SD = 4.54%) and 99.20% (SD = 1.17%), respectively. Considering the classifiers Logistic Regression and Adaboost were obtained the highest average of accuracy, each present in a third of the 12 models selected by FIV. Dominic, Gupta and Khare (2015) obtained 98% accuracy with the Adaboost classifier, Umamaheswuari et al. (2017) obtained 91.89% with the Stacking classifier, whereas in this study and for the same database, we found more competitive results in the classification of presence of heart diseases, with Logistic Regression are being the model containing, among its explanatory variables, the Pulsed Index (IPPA) and RC (Resistance x Complacency), where the mean accuracy of the test bases was 99.20% (SD = 1.17%).

Keywords: Cardiopathy. Binary classification. Campello de Souza's parameters.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ACP	ANÁLISE DE COMPONENTES PRINCIPAIS
ADABOOST	*ADAPTATIVE BOOSTING*
AIC	CRITÉRIO DE INFORMAÇÃO DE AKAIKE
ANOVA	ANÁLISE DE VARIÂNCIA
BPM	BATIMENTOS POR MINUTO
C	COMPLACÊNCIA
DNT	DOENÇA NÃO TRANSMISSÍVEL
DT	ÁRVORE DE DECISÃO - FLORESTAS ALEATÓRIAS
F	FREQUÊNCIA CARDÍACA
FN	FALSOS NEGATIVOS
FP	FALSOS POSITIVOS
HAS	HIPERTENSÃO ARTERIAL SISTÊMICA
HM	HARMONIA
IHME	INSTITUTO DE MÉTRICAS DE SAÚDE E AVALIAÇÃO
INFOGAIN	INFORMAÇÃO DE GANHO
IPPA	ÍNDICE PULSÁTIL DA PRESSÃO ARTERIAL
IPPARC	RELAÇÃO ENTRE IPPA E RC
LR	REGRESSÃO LOGÍSTICA
NA	OBSERVAÇÃO NÃO DISPONÍVEL
NB	NAIVE BAYES
P	PRESSÃO ARTERIAL
PAM	PRESSÃO ARTERIAL MÉDIA
PS	PRESSÃO SISTÓLICA
PD	PRESSÃO DIASTÓLICA
R	RESISTÊNCIA PERIFÉRICA
RN	REDES NEURONAIS
SVM	MÁQUINA DE VETORES DE SUPORTE
V	VOLUME
VIF	FATOR DE INFLAÇÃO DE VARIÂNCIA
VN	VERDADEIROS NEGATIVOS
VP	VERDADEIROS POSITIVOS
VPP	VERDADEIROS PREDITIVOS POSITIVOS
WHO	ORGANIZAÇÃO MUNDIAL DA SAÚDE
#	NÚMERO, QUANTIDADE

SUMÁRIO

1 INTRODUÇÃO

Uma identificação correta da *causa mortis* pode constituir um modo de medir a eficácia do sistema de saúde de um país, bem como verificar quais áreas necessitam de maior atenção para aumento na expectativa de vida, prevenção e/ou diminuição da mortalidade humana (WHO, 2018).

As doenças cardíacas, mais precisamente as cardiopatias isquêmicas, têm sido as principais causas de mortes no mundo, de acordo com as pesquisas mais recentes conduzidas pela Organização Mundial da Saúde (WHO, 2016). As cardiopatias isquêmicas[1] se apresentam clinicamente desde angina estável até a morte súbita, incluindo os infartos agudos do miocárdio. O diagnóstico de tais patologias nos pacientes não é tão simples e requer uma anamnese[2] cautelosa, investigação da história clínica (considerando os fatores de risco), experiência do especialista que faz a avaliação, bem como resultados de exames (CARVALHO e SOUSA, 2001).

Patel, Tejalupadhyay e Patel (2015) relatam que uma das dificuldades em determinar o diagnóstico de cardiopatia com rapidez está na pluralidade dos sintomas associados, que são confundíveis com os de outras doenças. Além disso, há também a questão de invasividade e dispendiosidade dos exames necessários para diagnosticar o comprometimento cardíaco. Como consequência, esses fatores contribuem indiretamente no aumento do risco de morte súbita ou acometimento por cardiopatias que resultam em sequelas, interferindo na qualidade de vida do indivíduo.

Diversas pesquisas têm sido realizadas com a finalidade de identificar fatores que possam apontar indícios de doenças cardíacas de maneira precisa e precoce. Nesse contexto, bancos de dados aprimorados, atualizados e completos com informações de pacientes cardiopatas, servem como um reflexo populacional e constituem instrumentos importantes no estudo dos atributos condicionantes de tais doenças (PATEL, TEJALUPADHYAY e PATEL, 2015).

Por outro lado, este tipo de pesquisa serve como fonte complementar para tomada de decisões na área da Saúde. Técnicas da mineração de dados (*data mining*) associadas as do aprendizado de máquinas (*machine learning*) vêm se consolidando na construção do conhecimento adequado para analisar os dados no contexto de doenças cardíacas e fornecer informações relevantes para o diagnóstico de cardiopatias. De acordo com Dominic, Gupta e Khare (2015), a redução do número de testes clínicos durante o processo diagnóstico é um dos

[1] doenças cardíacas resultantes do desequilíbrio entre a oferta e demanda de oxigênio para as células do músculo cardíaco, o miocárdio.

[2] Exame meticuloso da saúde de um paciente, que vai desde os sintomas iniciais até o momento da observação clínica.

benefícios trazidos pelos resultados gerados através das técnicas de aprendizado de máquinas e mineração de dados, por isso é relevante que sejam desenvolvidos sistemas que minimizem a probabilidade de erro de diagnóstico.

Em seu estudo, Campello de Souza (2010) sumarizou, o que o autor denominou de parâmetros, algumas medidas utilizadas no contexto da cardiologia, bem como propôs novos indicadores e analisou o comportamento de ambos em aplicações realizadas em bases de dados reais. Dentre estes parâmetros temos a Pressão Arterial Média (PAM), que corresponde ao valor esperado da pressão arterial durante um ciclo cardíaco; o Índice Pulsátil da Pressão Arterial (IPPA), que se refere à razão da diferença entre pressão sistólica e diastólica com a pressão diastólica; o RC, que corresponde à resistência periférica multiplicada pela complacência; o IPPARC, que resulta da razão entre IPPA e RC; a harmonia (HM), parâmetro fruto de um paralelo realizado entre a lei de Kepler e o ciclo cardíaco; e o α, que equivale a uma proporção resultante da relação entre o tempo de ejeção e o período do ciclo cardíaco; o α_2, que resulta de uma transformação cologaritma aplicada em α.

Considerando o diagnóstico de cardiopatas, e os parâmetros descritos e discutidos por Campello de Souza (2010), esta dissertação teve como objetivo principal incluir tais parâmetros como variáveis explicativas em modelos para classificação de cardiopatas, e avaliar o desempenho, bem como a acurácia dos mesmos. Outro objetivo consistiu em identificar qual dos parâmetros Campello de Souza mostra-se mais relevante no contexto da classificação de doentes cardíacos.

Para tanto, utilizou-se o diretório *heart-disease* do Repositório de Aprendizado de Máquina da UC Irvine (DUA e GRAFF, 2019), o qual contém quatro bases de dados com informações de indivíduos cardiopatas e não cardiopatas. Os parâmetros Campello de Souza foram computados com os dados de tais bancos, foram descritos e analisados, e, posteriormente, incluídos nas bases de dados como variáveis explicativas para implementação da classificação dos cardiopatas.

Foram aplicadas técnicas para seleção das variáveis explicativas relevantes na predição dos cardiopatas, e definidos modelos[3] que posteriormente foram usados na aplicação dos classificadores: *Naive Bayes*, Florestas Aleatórias, Regressão Logística, *Adaboost* e Máquinas de Vetores de Suporte.

O método *holdout* múltiplo (KOUROU et al., 2015) foi implementado com a finali-

[3] No âmbito desta dissertação, definiu-se por modelo um dado conjunto de variáveis explicativas que tenham passado por algum dos processos de seleção aqui mencionados.

dade de minimizar a possibilidade de ocorrer problemas de *overfitting*[4] e com o objetivo de se obter medidas de média e desvio padrão das acurácias na classificação dos cardiopatas. Outras métricas de desempenho calculadas também foram Sensibilidade média, Especificidade média e Valor Preditivo Positivo médio.

Durante a pesquisa surgiu um questionamento relacionado à dependência entre algumas variáveis explicativas, essa identificada por meio da presença de multicolinearidade em alguns modelos selecionados, e à grande dimensionalidade da matriz explicativa utilizada nas predições, que pode resultar em *overfitting*. Assim, considerou-se avaliar o desempenho dos modelos de classificação utilizando variáveis oriundas de uma Análise de Componentes Principais (ACP), no intuito de aproveitar o máximo de informações possíveis dos atributos de natureza contínua, diminuir a dimensão da matriz explicativa e mitigar a presença de multicolinearidade (FERRÉ, 1995; ZERBA e COLLINS, 1992; CATTELL, 1966). Nesta etapa, adicionou-se a Rede Neuronal (RN) na lista dos classificadores avaliados (RAHMAN et al., 2017).

No intuito de verificar se a localidade da coleta dos bancos de dados tem alguma relevância sobre a variável resposta, que se refere à presença de cardiopatia, os bancos de dados foram unificados e uma nova variável foi adicionada: `locality`. As análises realizadas nesta base de dados unificada está descrita na metodologia desta dissertação.

Este trabalho inicia-se com um breve referencial teórico sobre o sistema cardiológico, mineração de dados e classificação no contexto de aprendizado de máquinas. Foi apresentada a ideia de cada técnica de classificação aqui utilizada, bem como a da Análise de Componentes Principais (ACP). Posteriormente, a metodologia é descrita, seguida da apresentação e discussão dos resultados obtidos. Por fim, são realizadas as considerações finais, e logo depois seguem as intenções de trabalhos futuros e referências bibliográficas.

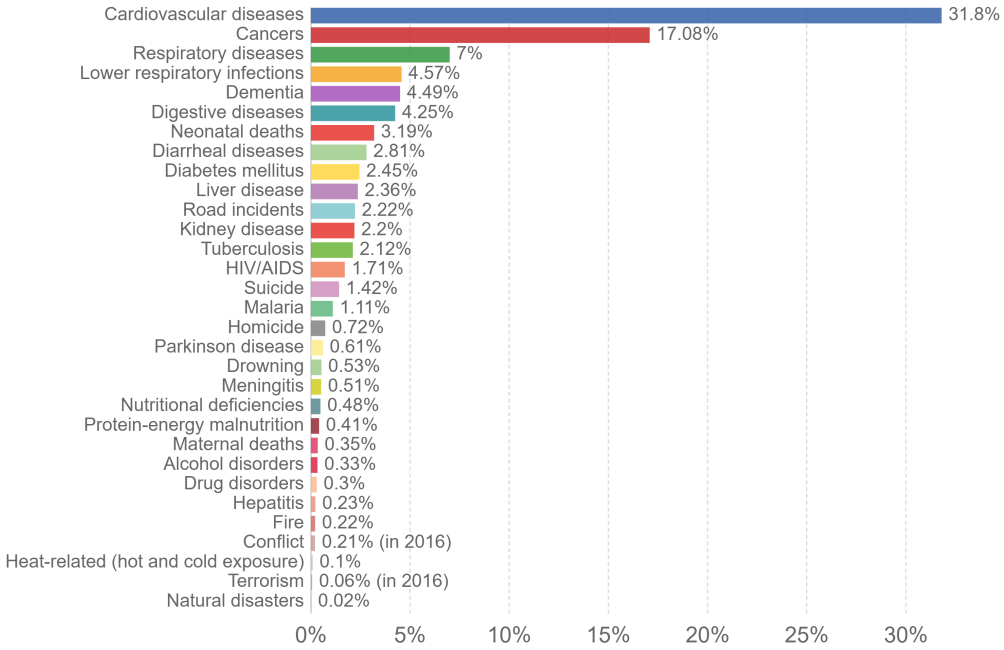
[4] *Overfitting* ocorre quando um método de classificação se ajusta bem aos padrões dos dados utilizados no treinamento do classificador, entretanto não categoriza bem dados usados na etapa de teste do classificador.

2 SISTEMA CARDIOVASCULAR

De acordo com a World Health Organization (WHO, 2018), do total das 56,9 milhões de mortes no mundo em 2016, 54% foram atribuídas às 10 principais causas, das quais, a doença isquêmica do coração e o derrame são responsáveis por 15,2 milhões de mortes no mesmo ano. Este cenário perdura desde o ano 2000 a 2017.

Ritchie e Roser (2019), analisaram os dados disponibilizados pelo *Institute for Health Metrics and Evaluation* (IHME, 2017) e fizeram considerações sobre a majoritariedade das mortes por doenças não transmissíveis (DNTs), que correspondem a mais de 70% das mortes no mundo, dentre as quais destacam-se doenças cardiovasculares (incluindo derrame cerebral), doenças respiratórias, cânceres e diabetes.

Figura 1 – Gráfico das principais causas de morte no mundo em 2017, segundo dados do IHME.



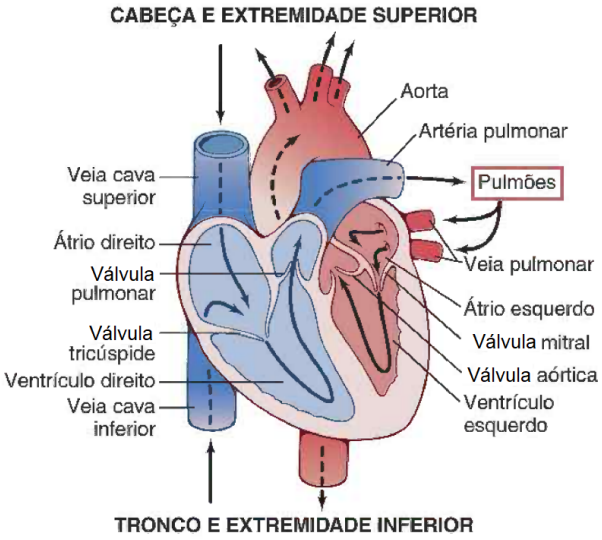
Fonte: Dados do IHME (2017) apresentados por RITCHIE e ROSER (2019).

As estatísticas são preocupantes e trazem reflexão sobre a importância das políticas públicas direcionadas à saúde, bem como investimento tecnológico para prevenção e promoção da mesma. Assim, se faz necessário compreender a fisiologia do sistema cardiovascular com a finalidade de identificar fatores que possam contribuir na rapidez e precisão do diagnóstico da presença de cardiopatias.

Compreender a fisiologia do sistema cardiológico é o ponto de partida para identificar quando há desequilíbrio da homeostase, e consequentemente, o desenvolvimento de patologias.

Como mostra a Figura 2, o coração possui quatro cavidades: átrio e ventrículo direito, separados pela válvula tricúspide, átrio e ventrículo esquerdo, separados pela válvula mitral.

Figura 2 – Estrutura do coração e fluxo do sangue pelas câmaras e válvulas cardíacas.



Fonte: GUYTON, HALL e GUYTON, 2006, pg. 107.

Comumente se define o coração com uma bomba muscular rítmica e ajustável, que impulsiona o sangue pelo corpo através de artérias e veias, com a principal função de manter o *"suprimento adequado de sangue a uma pressão suficiente para atender às demandas dos tecidos por nutrientes e remoção de resíduos em todos os órgãos do corpo"* (KOGAN, 2009).

As etapas que o coração percorre para suprir as necessidades do corpo acontece de maneira cíclica, desta forma, o ciclo cardíaco pode ser dividido em duas fases: relaxamento e contração. A fase do relaxamento é conhecida como diástole, que corresponde à distensão do coração ao receber o sangue na cavidade ventricular. Já na fase de contração, chamada sístole, as fibras do músculo cardíaco se contraem, ejetando o sangue para as artérias. O intervalo de tempo que vai desde a sístole até a diástole é denominado Período, que aqui é denotado por τ (GUYTON e HALL, 2006).

Esse movimento periódico do coração resulta na pressão arterial (força exercida pelo sangue na parede dos vasos), a qual é fundamental para o equilíbrio do organismo e está intimamente ligada à frequência cardíaca, força de contração do coração, complacência do vaso e volume de sangue bombeado por minuto. Quanto mais sangue o coração bombear, maior será a pressão arterial, a qual apresenta dois valores: máximo (sistólico) e mínimo (diastólico).

Na fisiologia, complacência é uma medida da resistência que um órgão tem para

retornar a sua estrutura inicial após a aplicação de uma força externa. No contexto da cardiologia, é a capacidade que os vasos têm de retomarem o raio original de sua circunferência após a passagem do fluxo de sangue (CAMPELLO DE SOUZA, 2010). Complacência é um termo recíproco à elastância e tem seu cálculo através da seguinte equação:

$$C = \frac{\Delta V}{\Delta P},$$

em que ΔV é a variação do volume e ΔP é a variação da pressão.

Outra medida de relevância para o estudo é a frequência cardíaca, que se trata da quantidade de batimentos cardíacos por minuto (bpm), estando assim, diretamente relacionada com o Período τ. De acordo com as necessidades físicas de cada organismo, a absorção de oxigênio e a excreção de gás carbônico, a frequência cardíaca pode sofrer variação. Fatores como exercícios físicos, sono, ansiedade, estresse, doença ou ingestão de drogas são responsáveis por alterar esta frequência (GUYTON e HALL, 2006).

Todas as condições que favorecem as chances de uma pessoa desenvolver doenças no coração ou nos vasos sanguíneos são consideradas fatores de risco. Existem os fatores irreversíveis, onde não há como controlá-los, assim como há fatores que podem ser prevenidos, ou até evitados, que são chamados de modificáveis. Fatores como a idade, o sexo e a hereditariedade são os principais fatores de risco para doença cardiovasculares. Estes não podem ser modificados. Todos os demais podem ser evitados, ou contornados, desde que haja as devidas precauções. São exemplos de fatores contornáveis: sedentarismo, obesidade, hipertensão arterial sistêmica (HAS), estresse, colesterol alto, triglicerídeos elevados e diabetes (POLANCZYK et al., 2005).

Os aspectos fisiológicos apresentados devem ser considerados quando se trata do diagnóstico de cardiopatia, e portanto, da classificação de cardiopatas. Os próximos capítulos trazem um breve referencial teórico sobre as técnicas de mineração de dados e classificação binária que foram aqui executadas com a finalidade de categorizar os cardiopatas utilizando parâmetros que envolvem complacência, resistência periférica, frequência cardíaca, pressão sistólica e pressão diastólica.

3 CIÊNCIA DE DADOS, MINERAÇÃO DE DADOS E CLASSIFICAÇÃO BINÁRIA

Este capítulo traz uma revisão bibliográfica sobre os classificadores que aqui foram utilizados para prever presença ou ausência de cardiopatia nos indivíduos das amostras (descritas no capítulo de Materiais e Métodos desta dissertação), bem como a técnica de análise de componentes principais, também utilizada nos dados de natureza contínua, seguida das Redes Neuronais, que, juntamente com demais classificadores (*Naive Bayes*, Florestas Aleatórias, *Adaboost*, Regressão Logística e Máquina de Vetores de Suporte), foram aplicadas às componentes principais para predizer cardiopatas. Com o objetivo de introduzir tais métodos, segue uma breve revisão sobre Ciência de Dados, mineração de dados e classificação binária.

De acordo com Van Der Aalst (2016), Ciência de Dados (*Data Science*) "*pode ser vista como uma fusão de disciplinas clássicas como Estatística, mineração de dados, bancos de dados e sistemas distribuídos*". Tal área de estudo tornou-se importante devido o desenvolvimento das tecnologias computacionais aliadas à internet, as quais promoveram, e continuam promovendo, a produção e o armazenamento de grande quantidade de informações diversas. O autor ainda relata que tal exorbitante numero de dados trouxe questionamentos sobre como tratá-los corretamente e obter informações úteis dos mesmos, e é nesse ponto que introduz-se a mineração de dados.

Na mineração de dados (*Data Mining*), o objetivo principal é extrair o máximo de informação possível de bancos de dados extensos, se utilizando de técnicas que vão desde análise exploratória básica (DA CUNHA e CARVAJAL, 2009), particionamento recursivo (STEINER et al., 2004), análise de agrupamentos (KAUFMAN e ROUSSEEUW, 2009), modelagem de regressão (BISHOP, 2006; FOX, 1997), segmentação de séries temporais (HIMBERG et al., 2001) até métodos complexos de reconhecimento de padrões (BISHOP, 2006). Essa área tem ganhado espaço na mesma proporção em que os avanços tecnológicos favorecem maior armazenamento de todo tipo de dados (HAND, 2006).

A mineração de dados encontra-se intimamente relacionada com a área de aprendizado de máquinas (*Machine Learning*), dispondo de ferramentas para caracterizar, identificar e localizar padrões em dados de resposta multivariada. Bishop (2006) relata que, no âmbito do reconhecimento de padrões, há uma preocupação em descobrir automaticamente regularidades em dados através do uso de algoritmos e, com o uso dessas regularidades, realizar ações como classificar os dados em diferentes categorias.

Um exemplo da mineração de dados associada ao aprendizado de máquinas é a área

de classificação binária, na qual tem-se interesse em extrair as informações de um conjunto de dados para predizer duas classes distintas predefinidas (KUMARI e SRIVASTAVA, 2017). Para tanto, comumente utilizam-se técnicas de subdivisão das bases de dados em grupos de treinamento, que são utilizados para estimar os coeficientes dos modelos visando minimizar o erro de classificação, e grupos de teste, os quais geram predições das classes de interesse ao receber a aplicação do modelo previamente estimado nos grupos treinamento, fornecendo informação sobre a adequação do modelo estimado.

Um excelente modelo de classificação deve se adequar bem ao grupo treinamento e classificar corretamente os indivíduos do grupo teste. Quando um modelo estima muito bem no grupo treinamento e tem um péssimo desempenho no grupo teste, acontece o que, no contexto de classificação, é chamado de *overfitting*. Tal problema está relacionado com a baixa complexidade do modelo ajustado, no que diz respeito à identificação e definição dos padrões da amostra, que resulta na inabilidade de classificar novos dados, que não foram utilizados no ajuste inicial (KOUROU et al., 2015). Os próximos tópicos deste capítulo trazem uma ideia geral dos classificadores binários que foram utilizados neste trabalho.

3.1 CLASSIFICADORES BINÁRIOS

Para introdução desta seção é importante definir o problema de classificação binária no âmbito de estudo desta dissertação: categorização de indivíduos cardiopatas e não cardiopatas. Assim, nossa variável de interesse, aqui denotada por Y, refere-se à presença de cardiopatia, em que Y poderá assumir dois valores distintos, $\{0,1\}$, tal que para $\{Y=0\}$ o indivíduo não possui doença cardíaca e para $\{Y=1\}$, possui doença cardíaca.

As categorias estimadas de Y serão denotadas por $\widehat{Y}$, e as informações utilizadas no processo de predição, denominadas variáveis explicativas, são denotadas de forma matricial, em que X representa a matriz explicativa de dimensão $n \times p$, na qual n corresponde à quantidade de indivíduos na amostra e p equivale ao número de variáveis explicativas. Adicionalmente, o objeto x_{ij}, para $i = 1,2,...,n$ e $j = 1,2,...,p$, denotará a i-ésima observação da j-ésima coluna da matrix X; $x_{i.}$ representa o vetor de valores das p variáveis explicativas que caracterizam o i-ésimo indivíduo da amostra; $x_{.j}$ corresponde ao vetor de valores dos indivíduos da amostra para a variável X_j; Y_i e $\widehat{Y}_i$, para $i = 1,2,...,n$, equivalem à classe real e a estimado para variável resposta Y, respectivamente, ambas referentes ao i-ésimo indivíduo da amostra.

Para tal problema de predição, na literatura, é comum utilizar classificadores binários

a exemplo do *Bayes* ingênuo - *Naive Bayes* (NB) (DOMINIC, GUPTA e KHARE, 2015), Árvore de Decisão (DT) - *Random Forest* (UMAMAHESWUARI et al., 2017; DOMINIC, GUPTA e KHARE, 2015; PATEL et al., 2015; TU et al., 2009), Máquina de Suporte Vetorial (SVM) (UMAMAHESWUARI et al., 2017; DOMINIC, GUPTA e KHARE, 2015), Regressão Logística (LR) e Adaboost (DOMINIC, GUPTA e KHARE, 2015). As próximas subseções definem a ideia de cada um destes métodos no contexto da classificação da presença de cardiopatias.

3.1.1 Regressão Logística

A regressão logística (*Logistic Regression* - LR) é uma técnica utilizada nos casos em que a variável resposta Y é de natureza binária, ou seja, possui duas categorias, comumente definidas como $\{Y = 0\}$ e $\{Y = 1\}$, onde $\{Y = 1\}$ geralmente é a categoria de interesse. O objetivo de tal método é "encontrar um modelo clinicamente interpretável, mais adequado e mais parcimonioso, para descrever a relação entre uma variável resposta e um conjunto de variáveis independentes" (HOSMER, LEMESHOW e STURDIVANT, 2013, pg. 1).

Se tratando de um problema de regressão linear, o interesse é predizer o valor esperado (médio) da variável resposta condicionado às variáveis explicativas, o qual pode receber o nome de "esperança de Y, dado X", aqui denotada por $E(Y|X)$, de modo que $E(Y|X) = \beta X$, no qual β é um vetor de coeficientes da reta de regressão associados à matriz explicativa X. Para melhor entendimento, consideremos o modelo de regressão linear simples (DRAPER e SMITH, 1981; MONTGOMERY e PECK, 1982):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{para } i = 1, 2, ..., n, \tag{3.1}$$

na qual y_i corresponde ao i-ésimo valor da variável resposta Y; x_i é o i-ésimo valor de uma variável independente X; β_0 corresponde ao intercepto da reta de regressão; β_1 é o coeficiente de angulação; e ε_i, o i-ésimo valor do vetor de erros não observados, numa amostra de tamanho igual a n.

Tal modelo exige que a variável resposta pertença ao conjunto dos números reais $\mathbb{R}$, entretanto, no caso em que a variável de interesse na predição é binária se faz necessário implementar uma transformação para que tal condição seja satisfeita. Dentre as transformações disponíveis na literatura, se popularizam a *logit*, *probit* e *cauchy* (SHI e RENTON, 2013). Aqui denotada por $g(\cdot)$, a transformação que é de interesse nesta dissertação é a *logit*, definida como:

$$g(x_i) = \ln\left[\frac{\pi(x_i)}{1 - \pi(x_i)}\right] = \beta_0 + \beta_1 x_i, \tag{3.2}$$

na qual $\pi(X) = E(Y|X)$ (HOSMER, LEMESHOW e STURDIVANT, 2013) representa a esperança de Y, dado X, para o caso em que Y segue uma distribuição de Bernoulli (devido sua natureza dicotômica), e que $E(Y|X)$ está compreendida no intervalo $(0,1)$. Como é possível visualizar na equação abaixo,

$$\pi(x_i) = E(Y_i|x_i) = \frac{\exp\{\beta_0 + \beta_1 x_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}}, \tag{3.3}$$

$\pi(X)$ é uma função de um preditor linear. Note que a relação entre $\pi(x_i)$ e x_i não é linear, o que justifica a necessidade de uma transformação de linearização (HOSMER, LEMESHOW e STURDIVANT, 2013; SHI e RENTON, 2013).

Em termos de interpretabilidade, $g(x_i)$ representa o logaritmo natural da razão de chance (*odds ratio*) de ocorrer o evento $\{Y_i = 1\}$ sob $\{Y_i = 0\}$, considerando que $\pi(x_i)$ fornece a probabilidade condicional da ocorrência de $\{Y_i = 1\}$, dado x_i, e que a quantidade $1 - \pi(x_i)$ corresponde à probabilidade condicional de $\{Y_i = 0\}$, dado x_i (HOSMER, LEMESHOW e STURDIVANT, 2013).

Considerando agora, que $g(x_i)$ tem relação linear com x_i, é possível definir um erro de estimação ε_i, para $i = 1, 2, ..., n$, que refere-se à diferença entre o i-ésimo valor estimado e o valor real para $g(x_i)$. Seja

$$\widehat{g(x_i)} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i, \tag{3.4}$$

o valor estimado para $g(x_i)$, então

$$\varepsilon_i = g(x_i) - \widehat{g(x_i)}. \tag{3.5}$$

Utilizando a técnica de Mínimos Quadrados Ponderados Iterativamente (*Iteratively Reweighted Least Squares* - IWLS), vide Firth (1992) para mais detalhes, tomando o vetor de erros ε como base para minimização, é possível estimar os coeficientes β_0 e β_1, e, consequentemente, definir o modelo estimado para predizer $g(\cdot)$.

Como $\widehat{g(x_i)}$ equivale à estimativa do logaritmo natural da *odds ratio* de $\{\widehat{Y}_i = 1\}$ sob $\{\widehat{Y}_i = 0\}$, condicionado a x_i, então $\exp\{\widehat{g(x_i)}\}$ corresponde à própria *odds ratio* estimada. Por fim, definindo um limiar igual à 0.5, é possível estimar Y_i, considerando que *odds ratio* estimada menor que 0.5 indica também menor probabilidade de $\{\widehat{Y}_i = 1\}$ ocorrer sob $\{\widehat{Y}_i = 0\}$, nas condições de x_i, portanto $\{\widehat{Y}_i = 0\}$, e *odds ratio* estimada maior que 0.5 indica maior probabilidade de $\{\widehat{Y}_i = 1\}$ ocorrer sob $\{\widehat{Y}_i = 0\}$, nas condições de x_i, resultando assim em $\{\widehat{Y}_i = 1\}$. Tal compreensão de $\widehat{g(x_i)}$ permite atribuir interpretabilidade ao coeficiente β_1 associado à variável resposta X, quando aplicada a transformação adequada.

Friedman, Hastie e Tibshirani (2001) relatam que dentre algumas vantagens da LR podemos destacar a facilidade de classificação dos indivíduos em categorias, bem como o fornecimento de resultados em termos da probabilidade que um evento tem de ocorrer sob o outro, além da possibilidade de interpretabilidade dos coeficientes do modelo, frente aos demais métodos de classificação.

3.1.2 *Naive Bayes*

Este classificador surgiu com base no Teorema de *Bayes* (D'AGOSTINI, 1995) que estima a probabilidade de um dado evento ocorrer considerando informações a priori associadas com o mesmo.

Como sugere o nome "*Naive*" (ingênuo), o método de classificação *Naive Bayes* (NB) torna-se simplório no sentido de desconsiderar qualquer relação de dependência que haja entre as variáveis explicativas e estimar as probabilidades condicionais, utilizando suas respectivas informações a priori (RISH et al., 2001).

Para definir o método de classificação NB no âmbito objetivado nesta dissertação, assumimos $\omega = \{Y = 0, Y = 1\}$ como sendo o espaço de decisão que determinará a classificação final, considerando que Y pode assumir tais valores. Tomando $X_i = (X_{i1}, X_{i2}, ..., X_{ip})$, para $i = 1, 2, ..., n$, como vetor de variáveis aleatórias explicativas que caracterizam o i-ésimo indivíduo da amostra, e partindo do Teorema de *Bayes*, a probabilidade de $Y_i = 1$ condicionado a X_i é dada por (MORAES e MACHADO, 2012; RISH et al., 2001):

$$\begin{aligned} \Pr(Y_i = 1|X_i = x_{i.}) &= \frac{\Pr(X_i = x_{i.}|Y_i = 1)\Pr(Y_i = 1)}{\Pr(X_i = x_{i.})} \\ &= \frac{\Pr(X_1 = x_{1.}, X_2 = x_{2.}, ..., X_{ip} = x_{ip}|Y_i = 1)\Pr(Y_i = 1)}{\Pr(X_1 = x_{1.}, X_2 = x_{2.}, ..., X_{ip} = x_{ip})}. \end{aligned} \tag{3.6}$$

Supor independência entre as variáveis $\{X_1, X_2, ..., X_p\}$, e aplicando um fator de escala S (MORAES e MACHADO, 2012), é possível reescrever (3.6) como sendo:

$$\Pr(Y_i = 1|X_i = x_{i.}) = \frac{1}{S}\Pr(Y_i = 1)\prod_{j=1}^{p}\Pr(X_i = x_{ij}|Y_i = 1). \tag{3.7}$$

Como regra de decisão, tem-se que $\{\widehat{Y}_i = 1\}$, se $\Pr(Y_i = 1|X_1 = x_{1.}, X_2 = x_{2.}, ..., X_{ip} = x_{ip}) > \Pr(Y_i = 0|X_1 = x_{1.}, X_2 = x_{2.}, ..., X_{ip} = x_{ip})$ (MORAES e MACHADO, 2012).

A suposição de independência entre as variáveis $\{X_1, X_2, ..., X_p\}$ não é verdadeira quando partimos para muitos casos reais, além de resultar em estimativas enviesadas. Entretanto, ainda assim, esse viés pode não interferir significativamente na predição e, gerar bons resultados.

Desta forma, o método de *Naive Bayes* surpreende pela sua simplicidade, resultante dessa assumptível independência entre as variáveis explicativas, e desempenho competitivo com outros classificadores (FRIEDMAN, HASTIE e TIBSHIRANI, 2001).

3.1.3 Florestas Aleatórias - *Random Forest*

Antes de definir o método de Florestas Aleatórias se faz necessário falar um pouco sobre Árvores de Decisão. O método da Árvore de Decisão (*Decision Tree* - DT) baseia-se na ideia de subdividir o espaço gerado pelos dados de maneira recursiva, em retângulos, onde é possível ajustar modelos com baixa complexidade (SAFAVIAN e LANDGREBE, 1991). Apesar de serem conceitualmente simples, estas técnicas baseadas em árvores têm se mostrado ferramentas competitivas para classificação, no contexto do aprendizado de máquinas supervisionado (FRIEDMAN, HASTIE e TIBSHIRANI, 2001).

Genericamente, na computação, árvores são sistemas de informações organizados em conjuntos e armazenados em estruturas chamadas "nós". Toda árvore possui o "nó raiz", que é de onde se originam os demais, e o "nó folha" ou "nó terminal", que corresponde a um último elemento da árvore. Tomando os dados fornecidos pela matriz explicativa X e o conhecimento prévio dos verdadeiros valores de Y, regras de decisão são construídas para cada nó e, deste modo, a decisão final é fruto do percorrimento que parte do nó raiz ao nó folha, o qual retorna o resultado estimado da classificação, $\widehat{Y} = 1$ ou $\widehat{Y} = 0$ (SAFAVIAN e LANDGREBE, 1991).

Uma extensão dos modelos de Árvore de Decisão são as Florestas Aleatórias (*Random Forest* - RF), que correspondem a um conjunto de árvores não correlacionadas, com as quais se toma decisão partindo de uma votação simples, resultando na estimativa/classificação ($\widehat{Y}$).

Segundo Friedman, Hastie e Tibshirani (2001), a ideia das Florestas Aleatórias consiste em construir árvores de decisões inúmeras vezes, utilizando a técnica de *bootstrap* (DAVISON e HINKLEY, 1997) para reamostrar aleatoriamente e com reposição o vetor Y, e suas respectivas características em X, e, posteriormente, computar as categorias estimadas por cada árvore, elegendo a classe majoritária nas predições (BREIMAN, 2001).

Algumas vantagens que resultam na competitividade das Florestas Aleatórias com outros classificadores - como o Adaboost, por exemplo -, residem na robustez relativa aos *outliers* - que correspondem aos valores discrepantes nas amostras, como explica Guo et al. (2004)-, além de apresentar baixo viés e conseguirem captar relações complexas de interação dos dados (FRIEDMAN, HASTIE e TIBSHIRANI, 2001; BREIMAN, 2001).

3.1.4 *Adaboost*

Freund e Schapire (1996) definiram o método de aprendizado de máquina *Adaboost* (nome derivado de *Adaptative Boosting* - impulso ou estímulo adaptativo) em 1996, descrevendo-o como um algoritmo que utiliza a combinação dos resultados de vários algoritmos "fracos" de aprendizagem, gerando uma resposta conjunta mais consistente - principal característica dos métodos *boost*. Os chamados algoritmos "fracos" de aprendizagem são algoritmos que possuem um vetor de características (informações que configuram a amostra), um limiar (valor que subdivide os grupos) e uma paridade (regra de decisão que define a classe considerando o limiar).

O método *Adaboost* mais popular é o *AdaBoost*.M1, que considera um problema de classificação dicotômica, com variável de saída codificada como $Z \in \{-1, 1\}$ (FREUND e SCHAPIRE, 1997). No contexto desta pesquisa, a saída é recodificada para atender o espaço amostral das categorias de Y, no qual $Z = \{-1\}$ equivalerá a $Y = \{0\}$, e $Z = \{1\}$ equivalerá a $Y = \{1\}$.

O *AdaBoost*.M1 utiliza o algoritmo "fraco" de aprendizagem denominado *WeakLearn*, com a finalidade de minimizar a taxa do erro de classificação - ambos encontram-se detalhados no artigo de Freund e Schapire (1996).

Por se tratar de um método *boost*, o *AdaBoost*.M1 basicamente implementa o *WeakLearn* em versões ponderadas dos dados, ou seja, cada indivíduo do banco recebe um peso, que inicialmente é igual para todas as observações, e a cada iteração esse peso muda. Os indivíduos que não foram bem classificados na iteração anterior recebem um peso maior na próxima iteração, e a taxa de erro da classificação serve como medida para avaliar se tal mudança na ponderação melhora ou piora a categorização. O principal objetivo é melhorar a classificação baseando-se nos indivíduos mais difíceis de categorizar corretamente (FREUND, SCHAPIRE e ABE, 1999).

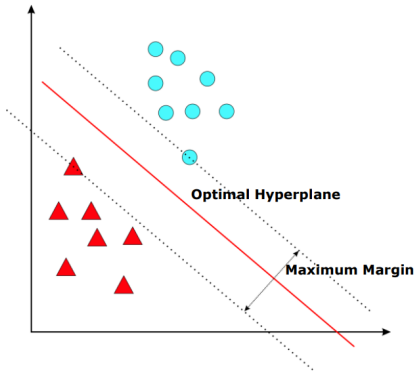
3.1.5 Máquina de Vetores de Suporte

Considere X a matriz de variáveis explicativas (contendo os vetores de atributos) de um banco de dados com informações para classificar cardiopatas/não cardiopatas e Y o vetor correspondente à variável resposta (presença de cardiopatia).

No âmbito da Máquina de Vetores de Suporte (*Support Vector Machines* - SVM), o processo de categorização da resposta baseia-se em utilizar as informações fornecidas pela matriz X para encontrar um hiperplano ótimo, que apresente margem máxima separando a resposta em

duas classes, $\{0, 1\}$, e tenha melhor desempenho na classificação, como exemplifica a Figura 3, na qual os dois grupos são representados por cores diferentes (HUSSAIN et al., 2011).

Figura 3 – Gráfico que exemplifica o método de classificação SVM definindo o hiperplano ideal que tem margem máxima entre duas classes.



Fonte: HUSSAIN et al., 2011, pg. 146.

Basicamente, o SVM classifica separando os grupos de maneira linear, mas quando partimos para a realidade, dificilmente encontram-se dados com informações que possibilitem tal forma de separação. Para tanto, uma solução viável consiste em realizar uma transformação no espaço de entrada original (variáveis explicativas), na intenção de obter um espaço de características em dimensões superiores (em geral, dimensão infinita), no qual seja possível aplicar a separação linear por meio dos vetores de suporte, os quais correspondem aos limiares que definem a margem máxima de separação dos grupos (SMITS e JORDAAN, 2002).

Assim, embora a margem máxima que diferencia as classes seja linear no plano modificado, ela pode ser não linear no espaço original. A exemplo dessa implementação temos o uso do truque da distância de *Kernel* (*"Kernel trick"*), o qual possui algumas variações que podem ser utilizadas, como o *kernel polynomial*, *radial basis* e *sigmoid*, que encontram-se melhor detalhados nos artigos de Hussain et al. (2011) e Shawe-Taylor et al. (2004).

O classificador SVM tem natureza robusta (desconsidera os *outliers*, se necessário) e atua bem em domínios complexos que possuam nítidas margens de separação das classes. O ponto fraco deste método está na utilização em bancos de dados que contenham inúmeras variáveis categóricas, mas, em contrapartida, tem se mostrado uma técnica promissora no tratamento de dados de natureza contínua (KOTSIANTIS, ZAHARAKIS e PINTELAS, 2007; FRIEDMAN, HASTIE e TIBSHIRANI, 2001).

3.2 ANÁLISE DE COMPONENTES PRINCIPAIS

O principal propósito em Análise de Componentes Principais (ACP) é explicar a estrutura de variância/covariância de um conjunto de dados, reduzindo a dimensionalidade das variáveis e gerando novos atributos (denominadas componentes) que sejam combinações lineares dos originais e independentes (ortogonais). A ACP é apropriada quando as variáveis sob investigação são de natureza contínua e possuem algum grau de dependência entre si (FRIEDMAN, HASTIE e TIBSHIRANI, 2001).

A ideia é que se algumas das variáveis originais são correlacionadas, elas estão contribuindo para que haja um *overfitting*, o qual se caracteriza pelo ótimo ajuste no grupo de dados separados para treinamento do classificador e péssimo desempenho no grupo de dados separado para realização do teste de predição das classes de interesse (RAHMAN et al., 2017). Desta forma, entende-se que um conjunto menor de variáveis, não-correlacionadas, pode ser tão eficaz quanto o conjunto de variáveis originais para explicar a estrutura de variância/covariância dos dados. Assim, as componentes principais apresentam-se como uma técnica que pode contornar este problema em algumas aplicações (FERRÉ, 1995; ZERBA e COLLINS, 1992; CATTELL, 1966).

Basicamente, a técnica se fundamenta em realizar um processo de transformação ortogonal (RAHMAN et al., 2017) nas variáveis de entrada (matriz explicativa X), de maneira que as componentes geradas contenham toda a informação da variabilidade desses dados. Em geral, espera-se que as primeiras componentes compreendam maior parte da variação total do conjunto de dados original, de forma que a dimensionalidade efetiva dos dados possa ser reduzida sem que haja perda significante de informação.

Com o intuito de selecionar uma quantidade de componentes considerada suficiente para análise, ou seja, que apresentem o máximo de representabilidade da variação dos dados sem deixar de lado a redução da dimensionalidade dos dados de entrada, é comum utilizar o gráfico *screeplot* (CATTELL, 1966), que mostra a representação acumulada da variância dos dados explicada por cada componente principal. A ideia é que, partindo das primeiras componentes principais (as que possuem maior representabilidade da informação dos dados de entrada), seja traçado um ponto de corte no local onde há porcentagem de variabilidade acumulada desejada (FERRÉ, 1995; ZERBA e COLLINS, 1992). Jolliffe e Cadima (2016) relatam que comumente se usa o valor de 70% de variância total explicada como ponto de corte na definição da quantidade de componentes principais a serem utilizadas.

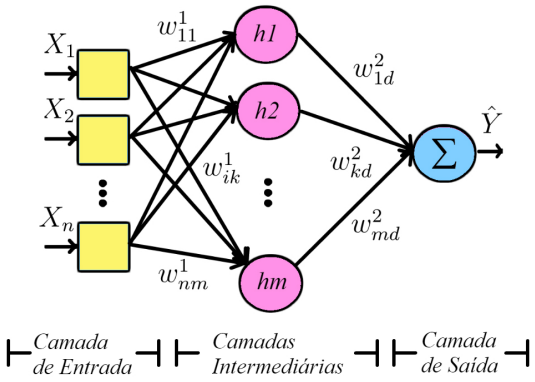
3.2.1 Redes Neuronais

Considerando a natureza contínua das componentes principais, além dos classificadores já descritos neste capítulo (Regressão Logística, *Naive Bayes*, Florestas Aleatórias, *Adaboost* e Máquina de Vetores de Suporte), é possível aplicar também a técnica de Redes Neuronais, a qual encontra-se descrita nesta subseção.

A técnica de Redes Neuronais (RN) foi inspirada na estrutura e funcionamento dos neurônios de um sistema nervoso, as quais recebem informações, processam, transmitem e aprendem com elas. Nesse contexto, os dados observados na amostra (dados de entrada) são utilizados para treinar as unidades de processamento que compõem a RN. Nesse processo, o objetivo é encontrar uma aproximação da relação entre os dados de entrada, adaptando iterativamente os parâmetros da RN (MAGLOGIANNIS, 2007), de maneira que minimize o erro de predição da categoria de saída (variável resposta) e seja possível obter uma decisão final ($\widehat{Y}$), resultando assim na aprendizagem (GÜNTHER e FRITSCH, 2010).

Uma rede neuronal é um modelo de regressão ou classificação, geralmente binária, que se organiza em camadas compostas por nós, onde ocorrem os processamentos dos dados. As camadas podem ser classificadas em camada de entrada, onde os dados são inseridos na rede; camadas intermediárias ou escondidas, nas quais são aplicadas as funções de ativação que realizam as conexões ponderadas (RAHMAN et al., 2017); e camada de saída, onde o resultado final é apresentado, como exemplifica a Figura 4 (ROSA, 2011; MAGLOGIANNIS, 2007; FRIEDMAN, HASTIE e TIBSHIRANI, 2001).

Figura 4 – Diagrama de uma Rede Neuronal artificial.



Fonte: Própria autora, 2019.

No diagrama acima, $X_1, X_2, ..., X_n$ correspondem às variáveis explicativas (dados de entrada); $h1, h2, ..., hm$ são as funções de ativação implementadas no k-ésimo nó, para $k =$

$1, 2, ..., m$, da camada 2 (camada intermediária ou camada escondida) da rede; w_{ik}^{1} equivalem aos pesos atribuídos do nó i, para $ki = 1, 2, ..., n$, da camada 1 (camada de entrada), para o nó k, da camada 2; w_{kd}^{2} equivalem aos pesos atribuídos do nó k, da camada 2, para o nó d, da camada 3 (camada de saída); $\widehat{Y}$ refere-se à variável resposta, que no caso desta dissertação diz respeito à presença de cardiopatia; e $\sum$ corresponde à soma das ponderações realizadas na camada anterior, na qual aplica-se um limiar pré-determinado e obtém-se a tomada de decisão: $\widehat{Y} = 0$ ou $\widehat{Y} = 1$. O algoritmo, bem como a generalização de suas funções, erro e limiar, podem ser vistos com mais detalhes no artigo de Rahman et al. (2017).

4 MATERIAIS E MÉTODOS

Este capítulo detalha a metodologia usada neste trabalho, bem como as bases de dados utilizadas e a forma de apresentação dos dados. O primeiro tópico descreve os atributos - que, para facilitar a discussão e apresentação dos resultados, aqui serão denominados parâmetros de Campello de Souza - que foram incluídos como variáveis nas análises, seguidos pela medida de consistência. Posteriormente, há uma caracterização das bases de dados, dos cenários definidos, descrição das etapas de seleção dos modelos preditivos, classificação e avaliação por meio das medidas de desempenho.

4.1 PARÂMETROS INCLUÍDOS NAS BASES DE DADOS

Os parâmetros aqui descritos encontram-se detalhados por Campello de Souza (2010) em seu livro sobre apoio ao diagnóstico médico. A finalidade para utilizá-los foi dispor de informações que apresentem estabilidade em seus valores, baixo custo de coleta por meio de métodos não invasivos e que, coerentemente, possam ser relevantes no diagnóstico de cardiopatia.

Esses parâmetros são classificados como indiretos por serem resultantes de cálculos realizados entre medidas obtidas diretamente do paciente (variáveis diretas). Desta maneira, foram utilizadas as relações entre frequência cardíaca (F), pressão sistólica (PS), pressão diastólica (PD) e período do batimento cardíaco (τ), medidas tais quais podem ser aferidas com apenas um esfigmomanômetro simples e um estetoscópio (Figura 5). As próximas subseções trazem um resumo dos parâmetros.

Figura 5 – Fotografia de Esfigmomanômetro e estetoscópio.

Fonte: WIKIPÉDIA, 2018.

4.1.1 PAM

A Pressão Arterial Média (PAM) corresponde à pressão média (valor esperado da pressão arterial) durante o ciclo cardíaco, que vai de $[0, \tau]$, em que τ representa um período e

$F = 1/\tau$, para F equivalente à Frequência cardíaca. A PAM é resultante do modelo apresentado na equação a seguir:

$$\text{PAM} = \frac{1}{\tau}\int_0^{\tau} P(t)dt \tag{4.1}$$

no qual $P(\cdot)$ corresponde à pressão arterial no período τ. Campello de Souza (2010) demonstra em seu livro que $P(\cdot) = PS\exp\left\{-\frac{1}{\text{RC}}t\right\}$, na qual RC equivale ao produto entre a resistência periférica (R) e a complacência (C) e (PS) à Pressão Sistólica. Portanto,

$$\text{PAM} = \frac{1}{\tau}\int_0^{\tau} PS\exp\left\{-\frac{1}{\text{RC}}t\right\}dt = \frac{1}{\log\left(\frac{PS}{PD}\right)}(PS-PD), \tag{4.2}$$

para PD referente à Pressão Diastólica. Logo,

$$\text{PAM} = \frac{PS-PD}{\log PS - \log PD}. \tag{4.3}$$

A expressão dada pela equação (4.3) foi a utilizada nesta dissertação, entretanto, na prática médica, a aproximação utilizada para estimar a PAM, devido a simplicidade no cálculo, é dada por

$$\text{PAM} \approx PD + \frac{PS-PD}{3} \approx \frac{PS+2PD}{3}, \tag{4.4}$$

pois considera-se que a sístole está compreendida no intervalo $(0, p\tau)$ e a diástole em $(p\tau, \tau)$, para $0 < p < 1$ (MORAN et al., 1995). No caso da aproximação dada pela Eq. (4.4), assume-se $p = 1/3$, que como é possível visualizar na Figura 6, corresponde à proporção do período equivalente à fase de sístole. A Figura 6 representa os sinais de fluxo e pressão no Ventrículo Esquerdo do coração (VE) e na Aorta com relação ao tempo dado em segundos (t[s]).

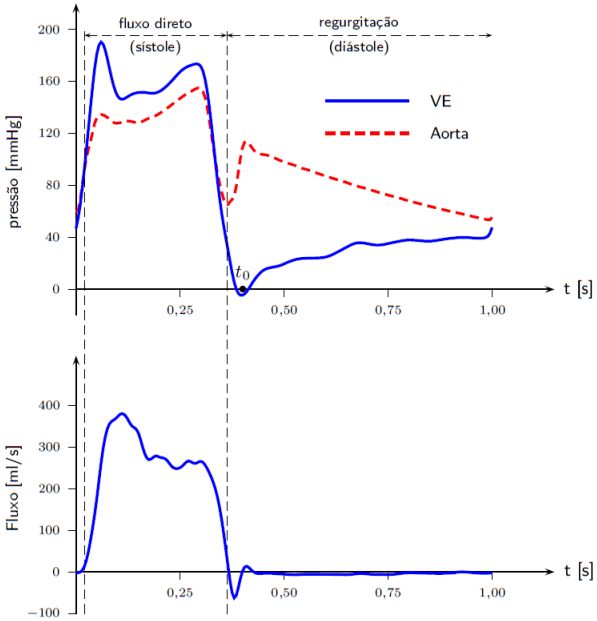
4.1.2 RC

A medida RC é o resultado do produto entre a resistência periférica (R) e a complacência (C) e pode ser estimada partindo da aproximação de primeira ordem (HOPPENSTEADT e PESKIN, 2013):

$$\frac{dP}{dt} + \frac{1}{\text{RC}}P = \frac{1}{C}\sum_i \delta(t-\tau_i), \tag{4.5}$$

em que τ_i, $i = 1,2,3,...$, é uma sequencia de períodos do ciclo cardíaco, $\tau_i = \frac{1}{F_i}$, F_i é a frequência cardíaca correspondente ao impulso sistólico i e $\delta(\cdot)$ é a função impulso delta de Dirac (MCQUEEN e PESKIN, 2001; FRANZ et al., 1991).

Figura 6 – Os eventos do ciclo cardíaco.



Fonte: WANDERLEY, 2005, pg. 22.

A solução para a equação diferencial (4.5) é dada por:

$$P(t) = P(0)\exp\left\{-\frac{1}{\mathrm{RC}}t\right\}, \tag{4.6}$$

em que t varia de 0 até τ_i. Assim,

$$PD = PS\exp\left\{-\frac{1}{\mathrm{RC}}\tau_i\right\} = PS\exp\left\{-\frac{1}{\mathrm{RC}}\frac{1}{F_i}\right\}, \tag{4.7}$$

e por fim:

$$\mathrm{RC} = \frac{1}{F\log\left(\frac{PS}{PD}\right)}, \tag{4.8}$$

em que F corresponde à frequência cardíaca média, comumente dada em número de batimentos cardíacos por minuto (bpm).

Rego e Campello de Souza (2002) relatam que RC *"variará ao longo do ciclo circadiano*[1]*, devido essencialmente aos ajustes fisiológicos em F. Ele é mais estável, entretanto, do que PS, PD e F"*, ou seja, é menos propenso à mudança brusca devido oscilações fisiológicas resultante de atividades diárias, como a mudança postural, por exemplo. Isto acontece porque fisiologicamente a resistência periférica (R) e a complacência (C) têm comportamentos inversamente proporcionais, ou seja, quando R cresce, C decresce, ou vice-versa, promovendo maior estabilidade nos valores de RC, o que se torna a principal vantagem deste índice.

[1] De acordo com Maximiano (2008), ciclo circadiano se refere ao período de 24 horas durante um dia.

Um valor mais elevado de RC sugere que a Resistência periférica ou a Complacência estão elevadas, indicando comprometimento no equilíbrio fisiológico que deveria existir entre R e C (REGO e CAMPELLO DE SOUZA, 2002).

4.1.3 IPPA e IPPARC

O Índice Pulsátil da Pressão Arterial (IPPA), que pode ser interpretado como uma medida adimensional (CHARNOV e BERRIGAN, 1991), muito utilizado na fisiologia, possui valores tão estáveis quanto as do RC, quando se trata de alterações fisiológicas. Jan et al. (2000) relatam que valor de IPPA elevado sugere falha no sistema de regulação do sistema cardiovascular, reflexa à alteração das pressões sistólicas e diastólicas, e comumente indica presença de hipertensão, a qual é fator de risco para doenças cardíacas (CARVALHO e SOUSA, 2001).

Para o cálculo do IPPA precisamos apenas da pressão sistólica PS e diastólica PD, como mostra a equação a seguir:

$$\text{IPPA} = \frac{PS - PD}{PD}. \tag{4.9}$$

É possível ir além e relacionar o IPPA com RC, resultando no IPPARC:

$$\text{IPPARC} = \frac{\text{IPPA}}{\text{RC}}. \tag{4.10}$$

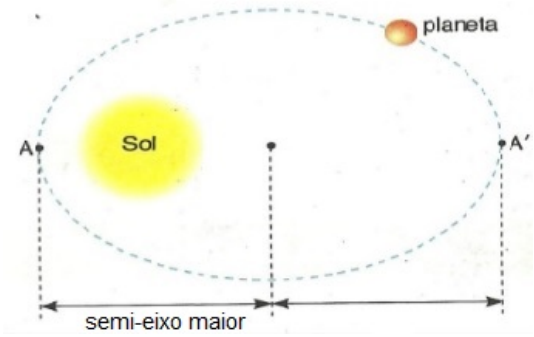
De acordo com Jan et al. (2000), tanto o IPPA quanto a PAM elevados são reflexo da alteração das pressões sistólica e diastólica, podendo indicar falha no sistema de regulação do sistema cardiovascular, e possibilidade da presença de hipertensão arterial.

4.1.4 HM

Campello de Souza (2010) define um parâmetro denominado harmonia (HM), baseado na expressão fundamental para o movimento de translação dos planetas (lei harmônica), proposta por Kepler em 1618 (RONAN, 1994). Partindo da premissa de que os planetas realizam um percurso elíptico ao redor do sol, nessa lei considera-se a proporcionalidade entre os quadrados dos períodos orbitais e os cubos dos semi-eixos maiores das órbitas.

Campello de Souza (2010) faz um paralelo do ciclo de translação da terra (Figura 7) com o ciclo cardíaco, no qual o período τ equivale ao período do movimento de translação, considera-se $(PS - \text{PAM})$ como sendo o semi-eixo maior da elipse e $(\text{PAM} - PD)$ o semi-eixo menor.

Figura 7 – Movimento de Translação da terra.



Fonte: Google Imagens, 2019.

Partindo da equação proposta por Kepler (RONAN, 1994):

$$\text{constante} = \frac{(\text{semi-eixo maior})^2}{(\text{período de translação})^3}, \tag{4.11}$$

em que a relação entre o semi-eixo maior da elipse com o período de translação é igual a uma constante, Campello de Souza apresenta com detalhes os cálculos, e por fim define HM como:

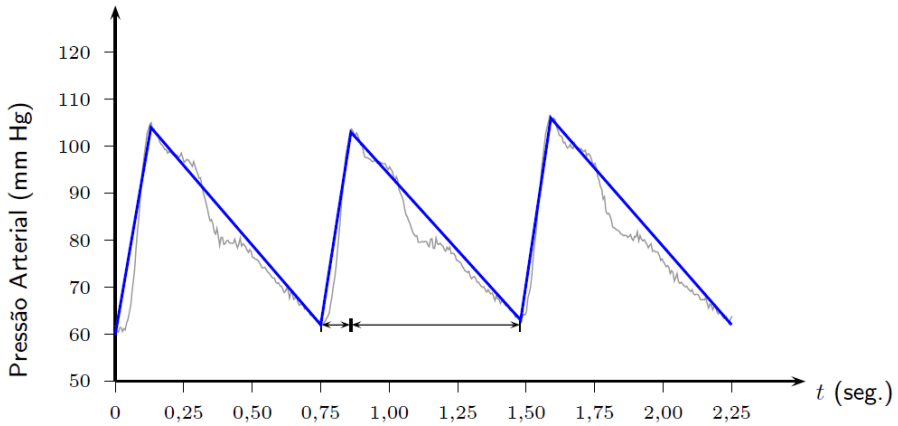
$$\text{HM} = \frac{\left(\frac{1000}{F/60}\right)^2}{(PS - \text{PAM})^3}, \tag{4.12}$$

na qual F é a frequência cardíaca, PS é a pressão sistólica e PAM corresponde à pressão arterial média.

4.1.5 Parâmetros α e α_2

No modelo Triangular da Onda de Pressão (WANDERLEY, 2005) utiliza-se uma onda triangular para estimar a curva de pressão arterial, como exemplifica a Figura 8. Neste

Figura 8 – Contorno da curva de pressão arterial.



Fonte: WANDERLEY, 2005, pg. 42.

caso a pressão arterial média é dada pela média aritmética entre *PS* e *PD*, e α é uma proporção resultante da relação entre o tempo de ejeção e o período do ciclo cardíaco.

De acordo com Wanderley (2005, pg. 42), considera-se a hipótese de que *"o ângulo de subida (entre a curva da pressão e o eixo das pressões) é igual ao ângulo de descida, formado a partir de uma reta horizontal partindo da pressão sistólica e a curva de pressão"*. Assim, é possível chegar à relação

$$\frac{\alpha\tau}{PS-PD} = \frac{PS-PD}{(1-\alpha)\tau}. \tag{4.13}$$

Resolvendo a equação anterior, com a finalidade de encontrar o valor de α, tem-se como resultado a seguinte expressão:

$$\alpha = \frac{1}{2} - \frac{1}{2\tau}\sqrt{\tau^2 - 4(PS-PD)^2}. \tag{4.14}$$

Ainda é possível realizar uma transformação de cologaritmo nos dados no intuito de modificar a escala do α e facilitar a análise dos valores. Esta transformação é apresentada na equação a seguir.

$$\alpha_2 = \log\left(\frac{1}{\alpha}\right) = \text{colog}(\alpha). \tag{4.15}$$

4.2 MEDIDA DE CONSISTÊNCIA ADAPTADA

Em seu artigo, Antal e Szabó (2017) utilizam uma medida de consistência $d(\cdot)$, proposta por Lee, Berger e Aviczer (1996), com a finalidade de diferenciar falsas assinaturas de sujeitos em um banco de dados. Aqui adaptou-se esta medida para comparar os valores dos parâmetros novos em cada grupo (cardiopatas e não cardiopatas) e determinar valores que possam dar indícios da relevância de tais indicadores na classificação de pessoas com doenças cardíacas.

A medida é um cálculo intuitivo da distância entre as médias de cada grupo, levando em consideração as suas respectivas variâncias, como mostra a seguinte equação:

$$d(S) = \frac{|\hat{\mu}_0 - \hat{\mu}_1|}{\sqrt{\hat{\sigma}_0^2 + \hat{\sigma}_1^2}}, \tag{4.16}$$

em que S corresponde ao novo atributo inserido na base de dados, que no contexto desta dissertação se refere aos parâmetros Campello de Souza; $\hat{\mu}_0$ e $\hat{\sigma}_0^2$ são a média e a variância estimadas para o parâmetro Campello de Souza (S), do grupo de não cardiopatas, respectivamente;

e por fim, $\hat{\mu}_1$ e $\hat{\sigma}_1^2$ referem-se à média e variância estimadas para o parâmetro S, referente ao grupo de pessoas com doença cardíaca.

Assim, cada parâmetro S recebeu um valor calculado de $d(S)$. Espera-se que o parâmetro com maior valor de $d(S)$ tenha indícios de ser mais influente na classificação de indivíduos cardiopatas.

4.3 BASES DE DADOS

Os bancos de dados aqui utilizados formam o diretório *heart-disease*, disponível gratuitamente na plataforma do Repositório de Aprendizado de Máquina da UC Irvine, no endereço eletrônico *https://archive.ics .uci.edu/ml/datasets/Heart+Disease*.

Esse diretório contém 4 bases de dados referentes ao diagnóstico de doenças cardíacas, e cada banco possui 75 variáveis (descritas no Anexo desta dissertação), das quais 14 são realmente utilizadas na maioria das pesquisas, sendo uma variável resposta e 13 explicativas. As bases recebem o nome do local da coleta dos dados: *Cleveland*, *Hungarian*, *Long Beach*, *Switzerland*.

A variável de interesse, intitulada como `V58`, aqui foi denominada Y. De natureza nominal, possui cinco categorias (de 0 à 4) as quais, comumente, são dicotomizadas de maneira que {0} indica ausência de cardiopatia (<50% de redução do diâmetro na angiografia) e {1,2,3,4} se referem aos cardiopatas da amostra (>50% de redução do diâmetro na angiografia). Sendo assim, na re-codificação dicotomizada, 0 refere-se aos não cardiopatas e 1, aos cardiopatas.

Tabela 1 – Diretório *heart-disease* - Tamanho amostral (n) e variáveis resposta Y.

Base de dados	n	Y	
		0	1
Cleveland	282	157	125
Hungarian	294	188	106
Long Beach	200	51	149
Switzerland	123	8	115

Os parâmetros Campello de Souza foram computados utilizando as variáveis `V10` (Pressão Sistólica), `V33` (Frequência Cardíaca) e `V37` (Pressão Diastólica). Posteriormente foram calculadas médias, medianas, desvios padrões, dos respectivos parâmetros. A nível de notação: $\hat{\mu}_0$, $\hat{m}_0$ e $\hat{\sigma}_0$ correspondem aos valores estimados para média, mediana e desvio padrão, respectivamente, do grupo de não cardiopatas; e $\hat{\mu}_1$, $\hat{m}_1$ e $\hat{\sigma}_1$ correspondem aos valores estimados para média, mediana e desvio padrão, respectivamente, para o grupo dos cardiopatas.

Com a finalidade de identificar se houve diferença estatisticamente significativa entre os valores dos parâmetros Campello de Souza dos cardiopatas e não cardiopatas, aplicou-se o teste Wilcoxon-Mann-Whitney (HOLLANDER e WOLFE, 1999), usado para testar diferença entre as medianas de tais parâmetros nos grupos em questão (cardiopatas e não cardiopatas). Espera-se que os parâmetros Campello de Souza que apresentem diferença significativa entre os grupos cardiopatas/não cardiopatas sejam também relevantes na classificação da variável resposta Y.

4.4 CONTEXTO E ETAPAS DE SELEÇÃO DAS VARIÁVEIS

Neste trabalho, para cada banco foram definidos dois cenários:

- Cenário 1, com as variáveis `V3` (idade), `V4` (sexo), `V11` (histórico de hipertensão), os parâmetros PAM, IPPA, RC, IPPARC, HM, α, α_2 e a variável resposta Y. Estas variáveis foram escolhidas para tal cenário no intuito de eleger modelos que utilizem variáveis de baixa complexidade no quesito coleta;
- Cenário 2, com as 75 variáveis do diretório *heart-disease* mais os parâmetros PAM, IPPA, RC, IPPARC, HM, α, α_2 e a variável resposta Y.

Inicialmente, realizou-se uma análise dos bancos de dados, dos quais foram excluídas as variáveis consideradas como inválidas para o estudo, ou seja, as que apresentavam apenas uma opção categórica em suas respostas ou que não tivessem valor algum em suas observações (NA - *Not Available*). Posteriormente, foram aplicados quatro critérios de seleção de variáveis com a finalidade de determinar os modelos a serem introduzidos em cada classificador. No contexto desta dissertação, denominou-se "modelo" um conjunto específico de variáveis que tenha passado pelo processo de seleção de atributos.

- Ganho de Informação (*Information Gain*), aqui denotado por InfoGain: método que utiliza o ganho entrópico de cada variável da matrix explicativa X, calculado com base na entropia de Shannon (SHANNON, 1948), para selecionar as mais significantes com relação à variável resposta Y (BORLAND, PLASTINO e TSALLIS, 1998);
- Modelo saturado selecionado pelo VIF, aqui denotado por Saturado+VIF (O'BRIEN, 2007): consiste em utilizar o modelo de regressão logística, selecionando as variáveis pelo VIF (fator de inflação de variância), com a finalidade de obter todas as variáveis válidas sem presença de multicolinearidade (forte correlação entre duas ou mais variáveis explicativas). De acordo com Menard (1995), valores de VIF maiores que 10 indicam

forte multicolinearidade, e no contexto de regressão, o problema de multicolinearidade afeta as estimativas do modelo (NETER et al., 1996). Assim, foram retiradas do modelo saturado (modelo de regressão logística com todas as variáveis da base de dados), de maneira sequencial, todas as variáveis que apresentaram VIF maior que 10;

- ANOVA (HAIR et al., 2009): modelo selecionado pela Análise de Variância (ANOVA), da qual foram escolhidas as variáveis que obtiveram significância estatística, mostrando influência na variável resposta;
- ANOVA + VIF: modelo selecionado pela Análise de Variância (ANOVA) e pelo VIF, ou seja, partindo da seleção de variáveis utilizando a ANOVA, seguiu-se para análise do VIF de tal modelo, retirando as variáveis com VIF maior que 10;
- AIC: modelo selecionado pela minimização do critério de informação de Akaike (AKAIKE, 1974), o qual utiliza em seu cálculo a verossimilhança do modelo e o número de variáveis explicativas utilizadas;
- AIC + VIF: modelo selecionado pela minimização do critério de informação e Akaike, seguido pela seleção por VIF (retirada das variáveis com VIF maior que 10).

4.5 ETAPA DE CLASSIFICAÇÃO E AVALIAÇÃO

Após a seleção das variáveis dos modelos, segue a etapa de classificação propriamente dita. Para cada modelo, em cada cenário, foram aplicados cinco classificadores comuns na literatura: *Naive Bayes* (NB), Florestas Aleatórias (RF), Máquina de Suporte Vetorial (SVM), Regressão Logística (LR) e *Adaboost*. A escolha de tais classificadores baseia-se na pesquisa de Dominic, Gupta e Khare (2015), os quais utilizaram as quatro bases de dados completas (*Cleveland*, *Hungarian*, *Long Beach* e *Switzerland*), com 75 atributos, e realizaram um processo de seleção de variáveis, com subsequente classificação dos cardiopatas.

Para cada banco de dados foram realizadas amostragens aleatórias simples, separando em grupos de treinamento e teste. Tal técnica denomina-se *holdout* (KOUROU et al., 2015). O grupo de treinamento é com o qual são estimados os modelos e grupo de teste é o utilizado para, literalmente, testar o modelo gerado e avaliar suas respectivas classificações. No caso desta dissertação, subdividiu-se os bancos de dados em 70% para o grupo de treinamento e 30% para o de teste, porcentagens comumente usadas na literatura de mineração de dados (ZHOU, ZHANG e KARYPIS, 2012). A Tabela 1 mostra a quantidade de elementos a serem classificados em cada grupo, isto para cada base de dados.

Tabela 2 – Diretório *heart-disease* - Tamanhos amostrais nos grupos de treinamento e teste.

Base de dados	Total	treinamento	teste
Cleveland	282	197	85
Hungarian	294	206	88
Long Beach	200	140	60
Switzerland	123	86	37

Kourou et al. (2015) descrevem um método que aqui se chamou de *holdout* múltiplo. Consiste em realizar a partição da base de dados em treinamento e teste várias vezes (aqui utilizou-se a quantidade de 100 iterações), considerando que a semente[2] para geração de números aleatórios é diferente em cada rodada do *holdout* múltiplo, tem-se como resultado, a seleção dos indivíduos que compõem os grupos de treinamento e teste de maneira pseudoaleatória em cada iteração (VIEIRA, SOUZA e RIBEIRO, 2004). Entendendo que, ao aplicar o *holdout* uma única vez, existe a possibilidade de se obter grupos de treinamento/teste que não representem bem os padrões da amostra total (isto a depender da semente geradora utilizada), o objetivo do *holdout* múltiplo é obter uma estimativa média das acurácias dos modelos e suas respectivas dispersões. A quantidade de 100 iterações foi utilizada apenas para se ter noção da variabilidade das classificações, facilitando os cálculos e interpretabilidade de algumas medidas em termos de porcentagens.

Com a finalidade de avaliar o desempenho dos modelos aplicados aos classificadores, em cada uma das 100 iterações, foram geradas matrizes de confusão (GLAROS e KLINE, 1988), como segue o exemplo dado na Tabela 3, na qual VN (Verdadeiros Negativos) corresponde aos valores da classe 0 classificados corretamente, FN (Falsos Negativos), aos valores da classe 0 classificados como 1; FP (Falsos Positivos), se refere aos valores da classe 1 classificados como 0, e VP (Verdadeiros Positivos), aos valores da classe 1 classificados corretamente.

Tabela 3 – Matriz de Confusão - Exemplo.

Valor Observado Y	Valor Estimado $\widehat{Y}$	
	0	1
0	VN	FP
1	FN	VP

A ideia da matriz de confusão é que quanto menor a quantidade de observações

[2] Sementes, no contexto dos algoritmos geradores de números aleatórios, refere-se ao valor que inicia uma sequência de números que aparentemente são aleatórios, mas seguem uma sequência de forma determinística, portanto, denominados números pseudoaleatórios.

fora da diagonal principal, melhor está sendo a classificação. **Desta forma, um classificador perfeito gera uma matriz de confusão onde todos os elementos fora da diagonal principal são iguais a zero** (GLAROS e KLINE, 1988).

Em cada iteração λ, para $\lambda = 1, 2, ..., 100$, foram geradas matrizes de confusões, e por fim, computada a matriz de confusão média, dada por:

$$\begin{bmatrix} \overline{\text{VN}} & \overline{\text{FP}} \\ \overline{\text{FN}} & \overline{\text{VP}} \end{bmatrix} = \frac{1}{100} \begin{bmatrix} \sum_{\lambda=1}^{100} \text{VN}_\lambda & \sum_{\lambda=1}^{100} \text{FP}_\lambda \\ \sum_{\lambda=1}^{100} \text{FN}_\lambda & \sum_{\lambda=1}^{100} \text{VP}_\lambda \end{bmatrix}, \tag{4.17}$$

na qual $\overline{\text{VN}}$, $\overline{\text{FP}}$, $\overline{\text{FN}}$ e $\overline{\text{VP}}$ equivalem às médias de VN, FP, FN e VP, respectivamente.

A partir da matriz de confusão (Tabela 3) é possível computar a acurácia do classificador, a qual pode ser dada em porcentagem, e é calculada da seguinte forma (TU et al., 2009; KAHRAMANLI e ALLAHVERDI, 2008):

$$\begin{aligned} \text{Acurácia} &= \frac{\text{\# de indivíduos classificados corretamente}}{\text{\# de indivíduos totais da amostra}} \times 100 \\ &= \frac{\text{VN+VP}}{\text{VN+VP+FN+FP}} \times 100. \end{aligned} \tag{4.18}$$

No contexto do *holdout* múltiplo, a acurácia média estimada foi dada por:

$$\hat{\mu}_{\text{Acurácia}} = \sum_{\lambda=1}^{100} \frac{\text{VN}_\lambda + \text{VP}_\lambda}{\text{VN}_\lambda + \text{VP}_\lambda + \text{FN}_\lambda + \text{FP}_\lambda}, \tag{4.19}$$

na qual VN_λ, FP_λ, FN_λ e VP_λ correspondem aos valores de VN, FP, FN e VP na iteração λ do *holdout* múltiplo, respectivamente. O desvio padrão estimado da acurácia foi dado por:

$$\hat{\sigma}_{\text{Acurácia}} = \sum_{\lambda=1}^{100} \frac{(\text{Acurácia}_\lambda - \hat{\mu}_{\text{Acurácia}})^2}{99}, \tag{4.20}$$

no qual Acurácia_λ equivale ao valor da acurácia calculada na iteração λ.

Os valores de médias de acurácia e seus respectivos desvios padrões foram organizados e apresentados em gráficos de barra para melhor comparação do desempenho dos modelos/classificadores.

Considerando que, em nível computacional, a presença de dados faltantes em algumas bases resulta na exclusão de observações na classificação - em alguns casos específicos (detalhes na subseção 4.8) - e, para possibilitar comparação entre as matrizes de confusão média, construiu-se, o que aqui denominou-se matriz relativa de confusão (Tabela 4).

Simplesmente, toma-se o valor médio total de indivíduos classificados nas 100 iterações do *holdout* múltiplo para cada modelo e classificador: $\overline{\text{U}} = \overline{\text{VN}} + \overline{\text{FP}} + \overline{\text{FN}} + \overline{\text{VP}}$. Posteriormente, divide-se os valores médios de VN, FP, FN e VP por esse total, obtendo assim

uma porcentagem de indivíduos classificados naquela categoria. Diante disto, $\frac{\overline{\text{VN}}}{\overline{\text{U}}}$ equivale à porcentagem média de não cardiopatas que foram classificados corretamente; $\frac{\overline{\text{FP}}}{\overline{\text{U}}}$ é a porcentagem média de não cardiopatas que foram classificados incorretamente; $\frac{\overline{\text{FN}}}{\overline{\text{U}}}$ corresponde à porcentagem média de indivíduos com cardiopatia que foram classificados incorretamente; e, $\frac{\overline{\text{VP}}}{\overline{\text{U}}}$ se refere à porcentagem média de cardiopatas que foram classificados corretamente.

Tabela 4 – Matriz Relativa de Confusão - Exemplo.

$\text{Y}/\widehat{\text{Y}}$	0	1
0	$\frac{\overline{\text{VN}}}{\overline{\text{U}}}$	$\frac{\overline{\text{FP}}}{\overline{\text{U}}}$
1	$\frac{\overline{\text{FN}}}{\overline{\text{U}}}$	$\frac{\overline{\text{VP}}}{\overline{\text{U}}}$

A interpretação é análoga a da matriz de confusão convencional, ou seja, os melhores classificadores apresentam maiores valores na diagonal principal da matriz, e valores próximos de zero fora dela. Neste caso, as células de cor cinza representam a diagonal principal da matriz relativa.

Adicionalmente foi realizada a contagem das quantidades de modelos aqui selecionados, e computadas as porcentagens de vezes que os parâmetros Campello de Souza foram escolhidos como relevantes pelos critérios de seleção. O objetivo foi ter uma visão geral da importância/influência destes novos atributos.

Por fim, foram selecionados os modelos com maiores valores de acurácia média no cenário 2, grupo teste, para cada base de dados, e calculadas algumas métricas do desempenho da predição dos cardiopatas, dadas em porcentagem (BALDI et al., 2000; GLAROS e KLINE, 1988). Foram elas:

- Sensibilidade média - porcentagem média de verdadeiros positivos, ou seja, de cardiopatas que foram classificados corretamente dentro do grupo de pessoas com doença cardíaca em cada iteração do *holdout* múltiplo. O cálculo desta medida é dado por:

$$\sum_{\lambda=1}^{100} \frac{\text{VP}_\lambda}{\text{VP}_\lambda + \text{FN}_\lambda}.$$

- Especificidade média - porcentagem média de verdadeiros negativos, ou seja, de não cardiopatas classificados corretamente dentro do grupo de pessoas sem doença cardíaca em cada iteração do *holdout* múltiplo. Seu cálculo é dado por:

$$\sum_{\lambda=1}^{100} \frac{\text{VN}_\lambda}{\text{VN}_\lambda + \text{FP}_\lambda}.$$

- Verdadeiro Preditivo Positivo (VPP) médio - porcentagem média de verdadeiros positivos em relação a todas as predições positivas, ou seja, de cardiopatas que foram corretamente classificados dentro do grupo de pessoas que estimou-se terem doença cardíaca em cada iteração do *holdout* múltiplo. O cálculo desta medida é dado por:

$$\sum_{\lambda=1}^{100} \frac{\mathrm{VP}_\lambda}{\mathrm{VP}_\lambda + \mathrm{FP}_\lambda}.$$

Espera-se que os melhores modelos, dentre os encontrados nesta pesquisa, sejam aqueles que possuem validade, no que diz respeito ao critério de independência das variáveis explicativas, com maiores valores médios de acurácia, sensibilidade, especificidade e VPP.

4.6 VARIÁVEL `LOCALITY`

Além das análises nos quatro bancos separadamente, surgiu a curiosidade de verificar a significância da localidade na classificação dos cardiopatas em questão. Assim, os quatro bancos foram unidos, gerando uma base de dados com 899 indivíduos, e a variável qualitativa `locality` foi incluída, a qual apresenta quatro categorias: Cleveland, Hungarian, Long Beach e Switzerland.

A partir da Tabela de contingência da variável resposta Y com `locality` foi realizado um teste Qui-Quadrado de Pearson (MORETTIN e BUSSAB, 2017), para verificar se havia dependência entre Cardiopatas e localidade. Além disso, o teste Qui-Quadrado de Pearson também foi aplicado nos grupos da variável `locality`, dois a dois, com a finalidade de identificar entre quais cidades/países existe diferença significativa da proporção de cardiopatas. Também fora aplicados os métodos de seleção de variáveis aqui descritos, a fim de investigar em quantos deles `locality` mostrou-se relevante.

4.7 COMPONENTES PRINCIPAIS NA CLASSIFICAÇÃO

No decorrer desta pesquisa foi observada dependência (por meio do teste de fator de inflação de variância - VIF) entre algumas variáveis explicativas das bases de dados, o que se torna um problema na validação de alguns modelos construídos devido à possibilidade de *overffiting* (o modelo estima muito bem no grupo treinamento e tem um péssimo desempenho no grupo teste). Além disso, as bases de dados definidas no cenário 2 envolvem grande número de variáveis explicativas, 75 no total, o que aumenta a probabilidade de serem selecionadas variáveis dependentes entre si. Com o objetivo de diminuir a informação redundante, diminuir a

dimensionalidade da matriz de variáveis explicativas e eliminar a dependência entre as mesmas, utilizou-se os atributos de natureza contínua, inclusive os parâmetros de Campello de Souza, para construir Componentes Principais e implementá-las na classificação.

Por meio do Screeplot (CATTELL, 1966) foi possível visualizar a variabilidade cumulativa dos dados explicada pelas Componentes, definindo assim a quantidade que seria utilizada nos modelos. Chakraborty et al. (2010) constatam que quanto maior a quantidade de componentes utilizadas na predição binária, maior será a acurácia do método, considerando que maior número de componentes indica maior variabilidade dos dados representada na análise. Desta forma, o objetivo foi encontrar a quantidade de componentes de maneira parcimoniosa, de modo que se tivesse uma boa representabilidade dos dados nas componentes e que a dimensionalidade da matriz explicativa fosse reduzida.

Em dados reais, é comum selecionar a quantidade de componentes principais através da porcentagem da variância total explicada, e, usualmente implementa-se o valor de 70% como ponto de corte (JOLLIFFE e CADIMA, 2016). Nesta pesquisa foi considerada porcentagem de variabilidade explicada maior que 80% e padronizou-se a quantidade de componentes principais utilizadas em cada banco para facilitar a programação computacional, bem como o comparativo entre as bases de dados.

Posteriormente, foram realizados ajustes com os classificadores NB, RF, SVM, LR, RN e *Adaboost*, e apresentadas as métricas de desempenho em cada banco de dados.

4.8 AMBIENTE E CONDIÇÕES COMPUTACIONAIS

Todos os programas para implementação computacional foram construídos e executados no ambiente `R`, em uma máquina com processador `Intel(R) Core(TM) i5-5200U CPU 2.20GHz` e `8.00GB` de memória `RAM`, com sistema operacional de `64 bits`. Os códigos computacionais para reprodutibilidade desta pesquisa podem ser acessados pelo endereço eletrônico: https://github.com/Raydonal/Cardiac-Classification.

Inicialmente, para o cálculo das medidas descritivas dos parâmetros Campello de Souza, utilizou-se a função `summary` (CHAMBERS e HASTIE, 1992) para obter estimativas da média e mediana nos grupos cardiopatas e não cardiopatas, e a função `sd` (R Core Development Team, 2004) para computar os respectivos desvios padrões. Para implementação do teste Wilcoxon-Mann-Whitney, utilizou-se a função função `wilcox.test` (HOLLANDER e WOLFE, 1999).

Na etapa da seleção de variáveis utilizou-se a função `information_gain`, disponível no pacote `FSelectorRcpp` (ZYGMUNT et al., 2019) para implementar o InfoGain; foi usada a função `anova`, do pacote *stats* (CHAMBERS e HASTIE, 1992), para executar a Análise de Variância; e a função `vif`, do pacote *car* (FOX e WEISBERG, 2018) foi tomada para implementar o fator de inflação de variância.

A Tabela 5 traz as funções utilizadas nos ajustes dos modelos, bem como suas respectivas configurações e o nome dos pacotes que as contém, os quais correspondem à bibliotecas contendo funções e dados específicos para cada tipo de ajuste (ETAATI, 2019).

Tabela 5 – Pacotes, funções e seus respectivos parâmetros utilizados para implementação dos classificadores no ambiente *R*.

Método	Pacote	Função	Parâmetros utilizados
NB	e1071	`naiveBayes`	laplace = 0, na.action = na.pass
RF	randomForest	`randomForest`	ntree = 500, na.action = na.omit
SVM	e1071	`svm`	scale = F, kernel = "poly", cost = 100, epsilon = 1.0e-12, na.action = na.omit
LR	stats	`glm`	family = binomial(link = "logit"), na.action = na.omit
Adaboost	fastAdaboost	`adaboost`	nIter = 10
RN	nnet	`nnet`	size = 5, linout = T, rang = 0.1, decay = 5e-2, maxit = 1000

Os classificadores *Naive Bayes* (NB) e Florestas Aleatórias (RF) foram aplicados com os parâmetros padrões (*default*) de suas respectivas funções: `naiveBayes` (MEYER et al., 2019) e `randomForest` (FREUND e SCHAPIRE, 1996). Para implementação do classificador SVM utilizou-se a função `svm` (MEYER et al., 2019) com função *kernel polynomial*, a qual se baseia no algoritmo de Chang e Lin (2011). A Regressão Logística (LR) foi computada através da função `glm`, que foi baseada no livro de Dobson (1990), utilizando a função de ligação *logit* (Eq. (3.2)). O classificador *Adaboost* foi implementado por meio da função `adaboost` (FREUND e SCHAPIRE,1996), com o número de 10 iterações para execução do algoritmo *WeakLearn*. A RN aplicada nesta pesquisa (VENABLES e RIPLEY, 2002; RIPLEY e HJORT, 1996) conteve cinco nós na camada intermediária única, e número máximo de iterações igual à 1000.

Nos casos das funções dos classificadores RF, SVM e LR, há a implementação do argumento "`na.action = na.omit`", o qual promove a exclusão das observações da amostra que apresentam qualquer dado faltante (NA) nas variáveis do modelo, diminuindo assim o tamanho do conjunto amostral na classificação. Por esse motivo, julgou-se mais justo, em

nível de comparação, construir as matrizes relativas de confusão, as quais foram previamente exemplificadas e descritas na subseção 4.5 desta dissertação.

A função usada na predição das variáveis resposta foi o `predict`, disponível no pacote *stats* e, para implementação da análise de Componentes Principais utilizou-se a função `princomp`, também disponível no mesmo pacote, com parâmetro `cor=T`, que indica o uso da matriz de correlação no cálculo das componentes, considerando que os dados apresentavam unidades de medidas diferentes.

5 RESULTADOS E DISCUSSÕES

Este capítulo traz os resultados das aplicações computacionais descritas na metodologia desta dissertação, bem como a discussão e comparação com alguns resultados encontrados na literatura. Inicialmente, apresentam-se as características dos parâmetros de Campello de Souza (2010) para cada banco de dados, seguidas das variáveis selecionadas para compor os modelos aplicados aos classificadores e, posteriormente segue a discussão dos resultados referentes à utilização das Componentes Principais como variáveis na classificação e a inclusão da nova variável `locality`.

5.1 PARÂMETROS DE CAMPELLO DE SOUZA NAS BASES DE DADOS

A Tabelas 6, 7, 8 e 9 apresentam as principais descritivas dos parâmetros Campello de Souza computados com as informações dos bancos *Cleveland*, *Hungarian*, *Long Beach* e *Switzerland*, respectivamente. Como antes mencionado na metodologia desta dissertação, tais estatísticas descritivas são: $\hat{\mu}_0$, $\hat{m}_0$ e $\hat{\sigma}_0$, que correspondem aos valores estimados para média, mediana e desvio padrão, respectivamente, do grupo de não cardiopatas; $\hat{\mu}_1$, $\hat{m}_1$ e $\hat{\sigma}_1$, que correspondem aos valores estimados para média, mediana e desvio padrão, respectivamente, para o grupo dos cardiopatas.

Observa-se que os mesmos apresentam valores semelhantes aos encontrados por Campello de Souza (2010) em seus estudos sobre medidas de apoio ao diagnóstico médico. É perceptível também que o parâmetro HM apresenta desvio padrão elevado em todas as bases de dados, enquanto a PAM mostra-se mais precisa.

Tabela 6 – Parâmetros Campello de Souza - Estimativas de Média ($\hat{\mu}$), Mediana ($\hat{m}$), Desvio Padrão ($\hat{\sigma}$) e testes de diferença entre medianas (p-valor). Banco de dados *Cleveland*.

Parâmetros Campello de Souza	$\hat{\mu}_0$	$\hat{\mu}_1$	$\hat{m}_0$	$\hat{m}_1$	$\hat{\sigma}_0$	$\hat{\sigma}_1$	p-valor
PAM	105.1306	108.0998	105.9115	107.8869	11.6754	11.6864	**0.0793**·
IPPA	0.5439	0.5692	0.5294	0.5556	0.1824	0.1760	0.2434
RC	0.0339	0.0340	0.0329	0.0325	0.0115	0.0113	0.9683
IPPARC	19.3680	19.8608	15.7457	17.3187	12.5346	11.7759	0.5273
HM	88.2808	79.0296	51.6722	42.4812	118.2997	94.5672	0.3004
α	0.0038	0.0040	0.0029	0.0033	0.0029	0.0033	0.4837
α_2	5.8280	5.7834	5.8284	5.7121	0.7035	0.7473	0.4837

Teste aplicado: Teste Wilcoxon-Mann-Whitney.
· Significância estatística à 10%.

Dos testes de diferença de Medianas aplicados, os que obtiveram significância estatística foram da diferença das medianas de PAM entre cardiopatas e não cardiopatas, com p-valor igual à 0.0793 no banco *Cleveland* e 0.0262 no banco *Hungarian*, significativas à 10% e 5%, respectivamente.

Na base de dados *Switzerland* nenhum dos parâmetros novos obtiveram significância estatística nos testes de diferença das medianas, enquanto que no banco *Long Beach* houve significância à 10% em HM, com p-valor de 0.0886.

Tabela 7 – Parâmetros Campello de Souza - Estimativas de Média ($\hat{\mu}$), Mediana ($\hat{m}$), Desvio Padrão ($\hat{\sigma}$) e testes de diferença entre medianas (p-valor). Banco de dados *Hungarian*.

Parâmetros Campello de Souza	$\hat{\mu}_0$	$\hat{\mu}_1$	$\hat{m}_0$	$\hat{m}_1$	$\hat{\sigma}_0$	$\hat{\sigma}_1$	p-valor
PAM	105.4446	108.6070	102.9850	108.0512	11.4091	12.3450	**0.0262***
IPPA	0.5639	0.5971	0.5500	0.5556	0.1502	0.2195	0.1196
RC	0.0295	0.0300	0.0280	0.0291	0.0081	0.0080	0.5268
IPPARC	21.7992	22.9755	18.6552	19.2456	13.1868	18.9023	0.6145
HM	53.8264	55.4877	38.9873	37.2640	51.6480	77.3438	0.2751
α	0.0048	0.0047	0.0039	0.0038	0.0043	0.0035	0.7747
α_2	5.5739	5.5588	5.5552	5.5628	0.6453	0.6335	0.7747

Teste aplicado: Teste Wilcoxon-Mann-Whitney.
*Significância estatística à 5%.

Tabela 8 – Parâmetros Campello de Souza - Estimativas de Média ($\hat{\mu}$), Mediana ($\hat{m}$), Desvio Padrão ($\hat{\sigma}$) e testes de diferença entre medianas (p-valor). Banco de dados *Long Beach*.

Parâmetros Campello de Souza	$\hat{\mu}_0$	$\hat{\mu}_1$	$\hat{m}_0$	$\hat{m}_1$	$\hat{\sigma}_0$	$\hat{\sigma}_1$	p-valor
PAM	102.7137	106.1425	99.0182	104.4601	13.1111	11.7342	0.1067
IPPA	0.6404	0.6842	0.6085	0.6500	0.2085	0.2005	0.2863
RC	0.0327	0.0299	0.0307	0.0288	0.0106	0.0083	0.1720
IPPARC	23.5717	25.9667	20.7905	21.8479	15.9530	13.7833	0.2079
HM	67.7813	47.4181	44.0492	34.3783	69.1907	49.5149	**0.0886**·
α	0.0038	0.0044	0.0029	0.0039	0.0030	0.0028	0.1033
α_2	5.8083	5.8785	5.8440	5.5574	0.6754	3.0681	0.1033

Teste aplicado: Teste Wilcoxon-Mann-Whitney.
· Significância estatística à 10%.

A Tabela 10 traz os valores das medidas de consistências adaptadas $d(S)$, na qual visualiza-se maior relevância em PAM, para os bancos de *Cleveland* (0.1797) e *Hungarian* (0.1881), e em HM, para *Long Beach* (0.2393) e *Switzerland* (0.2568). Este resultado corrobora

Tabela 9 – Parâmetros Campello de Souza - Estimativas de Média ($\hat{\mu}$), Mediana ($\hat{m}$), Desvio Padrão ($\hat{\sigma}$) e testes de diferença entre medianas (p-valor). Banco de dados *Switzerland*.

Parâmetros Campello de Souza	$\hat{\mu}_0$	$\hat{\mu}_1$	$\hat{m}_0$	$\hat{m}_1$	$\hat{\sigma}_0$	$\hat{\sigma}_1$	p-valor
PAM	98.9302	104.4039	102.2406	102.0779	16.6840	14.5477	0.5347
IPPA	0.5972	0.5976	0.6587	0.5714	0.2409	0.2411	0.7942
RC	0.0355	0.0358	0.0247	0.0341	0.0161	0.0134	0.5839
IPPARC	22.1667	21.8520	26.9187	15.5101	13.9135	19.4161	0.6276
HM	197.6817	112.8445	33.3719	56.5868	285.4129	166.4005	0.9293
α	0.0039	0.0038	0.0042	0.0026	0.0031	0.0036	0.9293
α_2	5.9888	5.9288	5.4949	5.9362	1.1305	0.8388	0.9293

Teste aplicado: Teste Wilcoxon-Mann-Whitney.

com os testes de diferença das médias/medianas aplicados às classes de interesse da pesquisa (cardiopatas e não cardiopatas).

Tabela 10 – Medidas de consistências adaptadas $d(S)$.

Banco de dados	PAM	IPPA	α	HM	α_2	IPPARC	RC
Cleveland	**0.1797**	0.0996	0.0648	0.0611	0.0435	0.0286	0.0083
Hungarian	**0.1881**	0.1249	0.0091	0.0179	0.0167	0.0510	0.0473
Long Beach	0.1949	0.1514	0.1522	**0.2393**	0.0223	0.1136	0.2037
Switzerland	0.2473	0.0011	0.0204	**0.2568**	0.0426	0.0132	0.0163

5.2 SELEÇÃO DAS VARIÁVEIS E DESEMPENHO DOS MODELOS APLICADOS AOS CLASSIFICADORES

Nesta seção são apresentados os resultados da seleção de variáveis em ambos os cenários, para todos os quatro bancos de dados (*Cleveland*, *Hungarian*, *Long Beach* e *Switzerland*), bem como os gráficos das médias estimadas das acurácias dos modelos delineados e seus respectivos desvios padrões.

5.2.1 Banco de dados Cleveland

A Tabela 11 mostra as variáveis eleitas por cada critério de seleção aqui designado e aplicado à base de dados *Cleveland*, no cenário 1. Verifica-se que a variável referente ao sexo (`V4`) foi determinada como relevante em todos os critérios de seleção, seguida pela idade (`V3`), que apenas não teve relevância quando utilizado o InfoGain.

Quando no cenário 2, a relevância das variáveis `V4` (sexo), `V44` (número de grandes

Tabela 11 – Dados *Cleveland* - Variáveis selecionadas no cenário 1.

Modelo	Critério de seleção	# de variáveis	Lista das variáveis
A	InfoGain	2	V4,V11
B	Saturado+VIF	7	V4,V3,V11,PAM,IPPA,HM,α
C	ANOVA	4	V4,V3,PAM,IPPARC
D	ANOVA + VIF	3	V4,V3,PAM
E	AIC	5	V4,V3,IPPA,IPPARC,α
F	AIC + VIF	2	V4,V3

vasos coloridos por fluoroscopia), `V51` (um dos valores do teste físico) e `V40` (depressão do segmento ST induzida pelo exercício em relação ao repouso) manteve-se em todos os critérios de seleção (Tabela 12).

Tabela 12 – Dados Cleveland - Variáveis selecionadas no cenário 2.

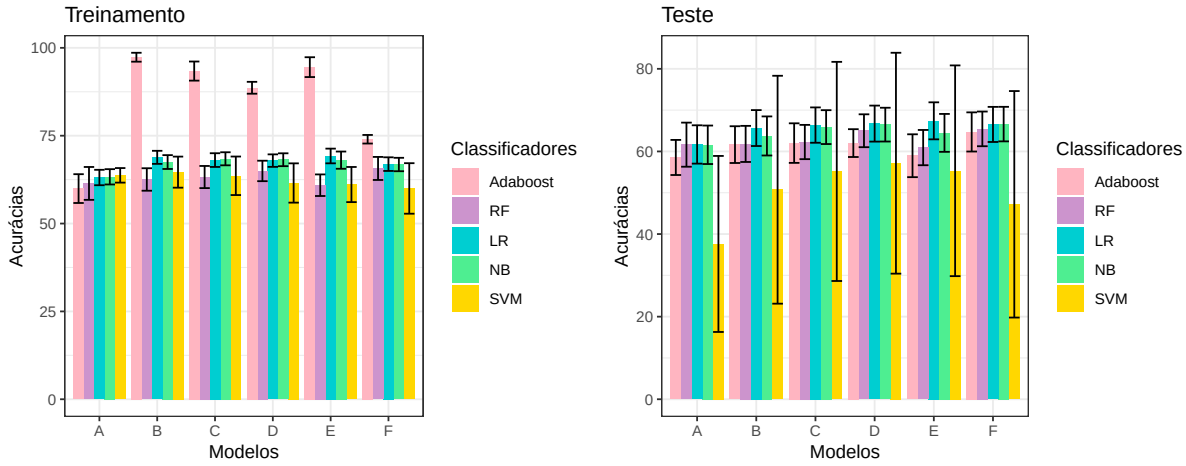
Modelo	Critério de seleção	# de variáveis	Lista das variáveis
A	InfoGain	27	V4,V9,V11,V16,V18,V23,V24, V25,V26,V27,V29,V30,V31, V32,V38,V39,V40,V41, V44,V51,V60,V61,V63,V65, V67,V68,V72
B	Saturado+VIF	41	V3,V4,V9,V11,V12,V14,V15,V16, V18,V19,V23,V24,V25,V26, V27,V29,V31,V32,V33, V34,V35,V38,V40, V41,V43,V44,V51,V59,V60, V61,V63,V65,V67,V68, V71,V72,V73,PAM,IPPA,HM,α
C	ANOVA	14	V3,V4,V9,V10,V23,V24,V32, V34,V38,V40,V44,V51, V60,V61
D	ANOVA + VIF	6	V4,V10,V34,V40,V44,V51
E	AIC	33	V3,V4,V10,V15,V16,V18, V19,V23,V25,V27,V29, V31,V33,V37,V38,V40, V43,V44,V51,V59,V60,V61, V63,V65,V67,V68,V71, V72,V73,PAM,IPPA,RC,α_2
F	AIC + VIF	27	V3,V4,V15,V16,V18,V19, V23,V25,V27,V29,V38, V40,V43,V44,V51,V60, V61,V63,V65,V67,V68, V71,V72,V73,PAM,IPPA,RC

A Figura 9 mostra as médias de acurácias, e seus respectivos desvios padrões, para cada classificador (NB,DT,SVM,LR e *Adaboost*) aplicado aos modelos (A,B,C,D,E e F), no

cenário 1, tanto nos grupos de treinamento quanto nos de teste. Observa-se que o método *Adaboost* se ajusta bem aos dados, nos modelos B (acurácia de 97.33% e DP= $\pm$ 1.27%), C (acurácia de 93.39% e DP= $\pm$ 2.72%), D (acurácia de 88.64% e DP= $\pm$ 1.69%) e E (acurácia de 94.51% e DP= $\pm$ 2.82%), no grupo treinamento. Entretanto, no grupo teste, estas médias de acurácias decaem, variando entre 58.96% e 62.02%, o que pode caracterizar um problema de *overfitting* (RAHMAN et al., 2017).

Os maiores valores de médias de acurácia nos grupos teste correspondem aos modelos C (com classificador LR e acurácia média de 66.38%), D (com classificadores NB e LR, e acurácias médias de 66.49% e 66.74%, respectivamente), E (com classificador LR e acurácia média de 67.41%) e F (com classificadores NB e LR, e acurácias médias de 66.62% e 66.54%, respectivamente). As piores médias de acurácias nesses grupos corresponderam as do classificador SVM, variando de 37.60%-57.15%, bem como os desvios padrões mais elevados, com variação entre 21.31% e 27.59%.

Figura 9 – Dados Cleveland - Acurácia dos modelos de classificação no cenário 1.



As Tabelas 13 e 14 trazem as matrizes relativas de confusão das classificações realizadas nos dados *Cleveland*, para os grupos treinamento e teste, respectivamente. Como já descrito na metodologia desta dissertação, o objetivo é que haja maior porcentagem de indivíduos classificados corretamente, ou seja, que as células de cor cinza possuam maiores valores de porcentagem quando comparadas às células de cor branca da matriz.

Nota-se que, no grupo treinamento, o classificador *Adaboost* apresenta melhor desempenho, quando comparado aos demais. Ainda nesse grupo, o modelo que apresentou menor porcentagem de erro (3% no total) foi o B, com classificador *Adaboost*, o qual encontra-se destacado na Tabela 13. Entretanto, no grupo de teste esse modelo decai em acurácia e apresenta

porcentagem de erro de 38% (Tabela 14).

Tabela 13 – Matrizes relativas de confusão, dadas em porcentagem. Grupos de treinamentos, banco *Cleveland*, cenário 1. Modelos A, B, C, D, E e F.

		A		B		C		D		E		F	
	Y/$\widehat{Y}$	0	1	0	1	0	1	0	1	0	1	0	1
NB	0	30	26	39	17	39	16	39	16	40	15	37	18
	1	11	33	15	29	14	29	16	29	17	28	15	30
RF	0	33	23	38	18	37	19	37	18	37	18	40	16
	1	16	28	19	25	18	26	17	28	21	24	18	26
SVM	0	34	22	38	18	36	20	35	21	36	20	37	19
	1	14	30	17	27	16	28	17	27	19	25	21	23
LR	0	28	27	41	15	40	16	40	16	41	15	39	17
	1	10	35	16	28	16	28	16	28	16	28	16	28
Adaboost	0	20	35	**53**	**2**	49	6	47	8	51	5	37	19
	1	5	40	**1**	**44**	1	44	3	42	0	44	7	37

NB-*Bayes* ingênuo; RF-Florestas Aleatórias; SVM-Máquina de Vetores de Suporte; LR-Regressão Logística; Y-variável resposta observada na amostra; $\widehat{Y}$-variável resposta estimada.

Tabela 14 – Matrizes relativas de confusão, dadas em porcentagem. Grupos de testes, banco *Cleveland*, cenário 1. Modelos A, B, C, D, E e F.

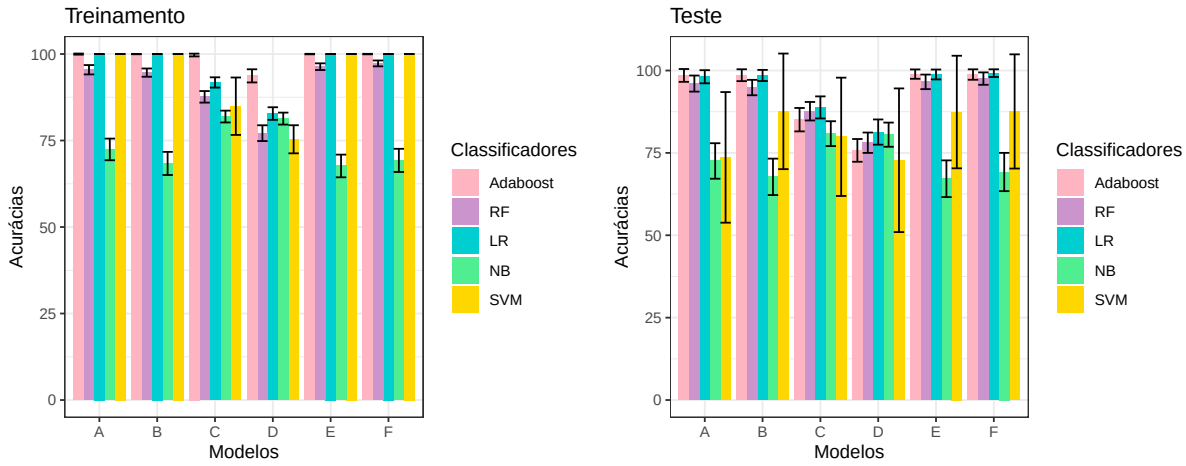
		A		B		C		D		E		F	
	Y/$\widehat{Y}$	0	1	0	1	0	1	0	1	0	1	0	1
NB	0	30	26	37	19	39	17	39	17	39	17	37	19
	1	12	32	17	27	17	27	16	28	18	26	15	29
RF	0	34	22	38	18	37	19	38	18	38	18	40	16
	1	16	28	20	24	19	25	17	27	21	23	19	25
SVM	0	4	14	10	9	9	9	9	10	9	10	10	9
	1	47	35	40	41	35	47	34	49	36	45	44	37
LR	0	28	28	39	17	39	17	40	16	40	16	39	17
	1	11	33	18	26	17	27	17	27	16	28	16	28
Adaboost	0	20	36	**35**	**21**	32	24	34	22	31	25	33	23
	1	6	38	**17**	**27**	14	30	16	28	16	28	12	32

NB-*Bayes* ingênuo; RF-Florestas Aleatórias; SVM-Máquina de Vetores de Suporte; LR-Regressão Logística; Y-variável resposta observada na amostra; $\widehat{Y}$-variável resposta estimada.

Para o cenário 2, como mostra a Figura 10, no grupo treinamento, as médias de acurácia de classificação mostraram melhora em todos os modelos, quando comparadas com as do grupo treinamento no cenário 1, e em alguns casos, alcançou-se acurácias de 100.00%, com desvio padrão nulo, como por exemplo: modelos A, B, E e F, ambos com classificador SVM, LR

e *Adaboost*.

Figura 10 – Dados Cleveland - Acurácia dos modelos de classificação no cenário 2.



Partindo para o grupo teste, destacam-se os classificadores *Adaboost* e LR, aplicados aos modelos A, B, E e F, com médias de acurácias variando de 98.11% à 99.20%, e desvios padrões entre 1.17% e 1.99%. Dentre estes modelos, apenas o A não possui nenhum dos parâmetros de Campello de Souza, e o que obteve maior média de acurácia (99.20%) e menor desvio padrão (1.17%) foi o F, com classificador LR. Este mesmo modelo, no grupo treinamento obteve acurácia média de 100.00% e desvio padrão nulo.

Assim como no cenário 1, o classificador SVM foi o que apresentou maiores desvios padrões no grupo teste, variando de 17.08% à 21.80%. Já o classificador NB, no grupo de teste do cenário 2, obteve baixas médias de acurácias, quando comparado aos demais, principalmente nos modelos A, B, E e F, com valores variando de 67.16% à 72.55%.

Tais resultados podem ser corroborados pelas matrizes relativas de confusão apresentadas nas Tabelas 15 e 16, nas quais, os modelos que acertaram mais na classificação, a exemplo dos modelos B, E e F com classificador LR e os modelos A, B, E e F, com classificador *Adaboost*. Tais modelos apresentaram valores de porcentagens próximos de zero nas células da matriz que não fazem parte da diagonal principal.

Tabela 15 – Matrizes relativas de confusão, dadas em porcentagem. Grupos de treinamentos, banco *Cleveland*, cenário 2. Modelos A, B, C, D, E e F.

		A		B		C		D		E		F	
	Y/$\widehat{Y}$	0	1	0	1	0	1	0	1	0	1	0	1
NB	0	56	0	56	0	56	0	46	10	56	0	55	1
	1	27	17	31	13	18	26	9	35	32	12	30	14
RF	0	50	1	55	1	51	5	44	11	55	1	56	1
	1	4	45	4	40	8	36	12	33	3	41	2	42
SVM	0	**51**	**0**	**56**	**0**	48	8	45	11	**56**	**0**	**56**	**0**
	1	**0**	**49**	**0**	**44**	8	36	14	30	**0**	**44**	**0**	**44**
LR	0	**51**	**0**	**56**	**0**	53	3	48	8	**56**	**0**	**56**	**0**
	1	**0**	**49**	**0**	**44**	6	38	10	34	**0**	**44**	**0**	**44**
Adaboost	0	**56**	**0**	**56**	**0**	**56**	**0**	51	5	**56**	**0**	**56**	**0**
	1	**0**	**44**	**0**	**44**	**0**	**44**	2	42	**0**	**44**	**0**	**44**

NB-*Bayes* ingênuo; RF-Florestas Aleatórias; SVM-Máquina de Vetores de Suporte; LR-Regressão Logística; Y-variável resposta observada na amostra; $\widehat{Y}$-variável resposta estimada.

Tabela 16 – Matrizes relativas de confusão, dadas em porcentagem. Grupos de testes, banco *Cleveland*, cenário 2. Modelos A, B, C, D, E e F.

		A		B		C		D		E		F	
	Y/$\widehat{Y}$	0	1	0	1	0	1	0	1	0	1	0	1
NB	0	56	0	55	0	56	0	46	10	55	0	55	0
	1	27	17	32	13	19	25	9	35	33	12	31	14
RF	0	50	1	55	1	51	5	45	11	55	1	55	1
	1	3	46	4	40	8	36	11	33	3	41	2	42
SVM	0	12	7	18	1	16	2	18	0	19	0	19	0
	1	19	62	11	70	18	64	28	54	13	68	12	69
LR	0	51	0	**56**	**0**	52	4	47	8	**56**	**0**	**56**	**0**
	1	2	47	**1**	**43**	8	36	11	34	**1**	**43**	**1**	**43**
Adaboost	0	**56**	**0**	**56**	**0**	49	7	42	13	**56**	**0**	**56**	**0**
	1	**1**	**43**	**1**	**43**	8	36	11	34	**1**	**43**	**1**	**43**

NB-*Bayes* ingênuo; RF-Florestas Aleatórias; SVM-Máquina de Vetores de Suporte; LR-Regressão Logística; Y-variável resposta observada na amostra; $\widehat{Y}$-variável resposta estimada.

5.2.2 Banco de dados Hungarian

Quanto à base de dados Hungarian, as Tabelas 17 e 18 mostram as variáveis selecionadas no cenário 1 e 2, respectivamente. Neste grupo de variáveis verifica-se que o sexo (`V4`) permanece como variável relevante em todos os critérios de seleção, seguida pela idade (`V3`), que mais uma vez, não teve relevância apenas de acordo com a técnica de seleção de variáveis InfoGain.

Tabela 17 – Dados Hungarian - Variáveis selecionadas no cenário 1.

Modelo	Critério de seleção	# de variáveis	Lista das variáveis
A	InfoGain	2	V4,V11
B	Saturado+VIF	7	V4,V3,V11,PAM,IPPA,HM,α
C	ANOVA	3	V4,V3,IPPA
D	AIC	7	V4,V3,V11,PAM,RC,IPPARC,α_2
E	AIC + VIF	5	V4,V3,V11,PAM,RC,IPPARC

O modelo selecionado pela ANOVA já não possuía multicolinearidade de acordo com o VIF, desta forma apenas cinco modelos foram definidos.

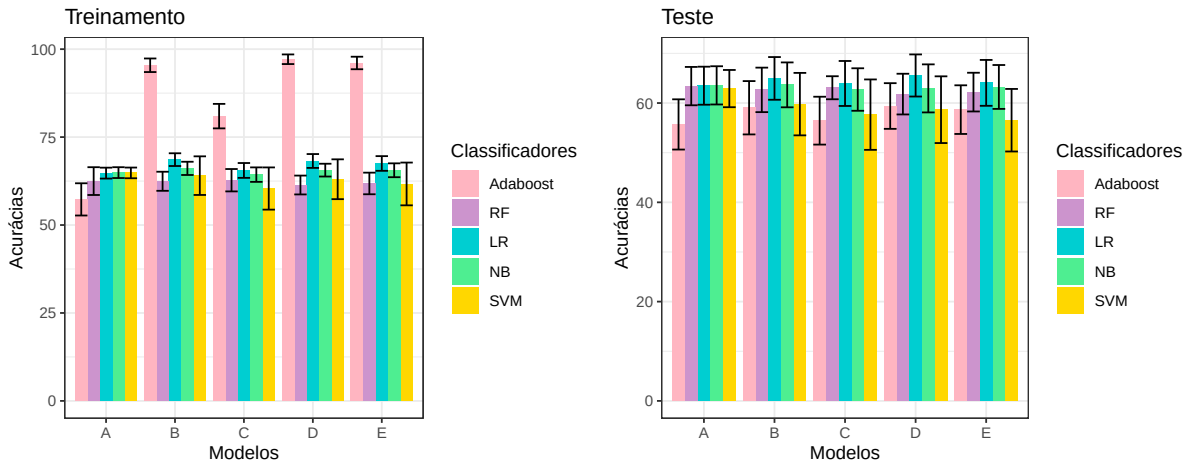
No cenário 2, como apresenta a Tabela 18, as variáveis relevantes em todos critérios de seleção foram `V4` (sexo feminino ou masculino) e `V6`, que se refere à presença de dor no peito ao fazer esforço.

Tabela 18 – Dados Hungarian - Variáveis selecionadas no cenário 2.

Modelo	Critério de seleção	# de variáveis	Lista das variáveis
A	InfoGain	15	V4,V5,V6,V7,V11, V16,V24,V25,V26,V32, V38,V39,V40,V72,V73
B	Saturado+VIF	20	V3,V4,V5,V6,V11,V12,V16, V19,V24,V25,V32, V35,V38,V40,V43, V72,V73,PAM,IPPA,HM
C	ANOVA	5	V4,V6,V11,V28,V29
D	ANOVA + VIF	4	V4,V6,V11,V28
E	AIC	35	V3,V4,V5,V6,V7, V10,V11,V12,V16,V19,V24, V25,V26,V28,V29,V30, V31,V32,V33,V34,V35, V37,V38,V40,V42, V43,V72,V73,PAM,IPPA,RC, IPPARC,HM,α,α_2
F	AIC + VIF	18	V4,V5,V6,V12,V16,V19, V24,V31,V32,V34, V35,V38,V40,V42, V72,V73,PAM,RC

No cenário 1, como mostra a Figura 11, observa-se que o *Adaboost*, assim como no banco *Cleveland*, apresenta problema de *overfittig* nos modelos B, C, D e E. Já o classificador que apresentou maiores médias de acurácias no grupo teste foi o LR, nos modelos B (64.96%), C (63.94%), D (65.55%) e E (64.06%), e mais uma vez, o SVM mostrou variabilidade alta nas estimativas das médias de acurácia, com desvios padrões variando entre 3.75% e 7.09%.

Figura 11 – Dados Hungarian - Acurácia dos modelos de classificação no cenário 1.



A Tabelas 19 e 20 trazem as matrizes relativas de confusão para ambos os grupos, treinamento e teste, no cenário 1. Observa-se que nenhum dos modelos classificou corretamente todos os indivíduos do grupo treinamento, e quando no grupo de teste, não houve porcentagens próximas de zero nas médias das classificações incorretas, as quais correspondem aos valores nas células da matrizes que não fazem parte da diagonal principal. Portanto, não registrou-se bom desempenho dos modelos neste cenário.

Tabela 19 – Matrizes relativas de confusão, dadas em porcentagem. Grupos de treinamento, banco *Hungarian*, cenário 1. Modelos A, B, C, D e E.

		A		B		C		D		E	
	$Y/\widehat{Y}$	0	1	0	1	0	1	0	1	0	1
NB	0	54	9	55	9	56	8	54	10	55	9
	1	26	11	25	11	28	8	25	11	25	11
RF	0	61	3	50	14	56	7	49	15	49	14
	1	35	1	24	12	30	7	24	12	24	13
SVM	0	56	8	46	18	47	17	47	16	44	20
	1	27	9	18	18	23	13	21	16	18	18
LR	0	54	10	54	9	54	10	54	10	53	10
	1	25	11	22	15	24	12	22	14	23	14
Adaboost	0	28	36	60	4	50	14	61	3	60	4
	1	7	29	0	36	5	31	0	36	0	36

NB-*Bayes* ingênuo; RF-Florestas Aleatórias; SVM-Máquina de Vetores de Suporte; LR-Regressão Logística; Y-variável resposta observada na amostra; $\widehat{Y}$-variável resposta estimada.

Partindo para o cenário 2, como mostra a Figura 12, visualiza-se melhora em quase todos os modelos aplicados aos classificadores nos grupos teste, com exceção do modelo E com classificador LR (acurácia média igual a 64.87% e desvio padrão de 9.35%) e os modelos

Tabela 20 – Matrizes relativas de confusão, dadas em porcentagem. Grupos de testes, banco *Hungarian*, cenário 1. Modelos A, B, C, D e E.

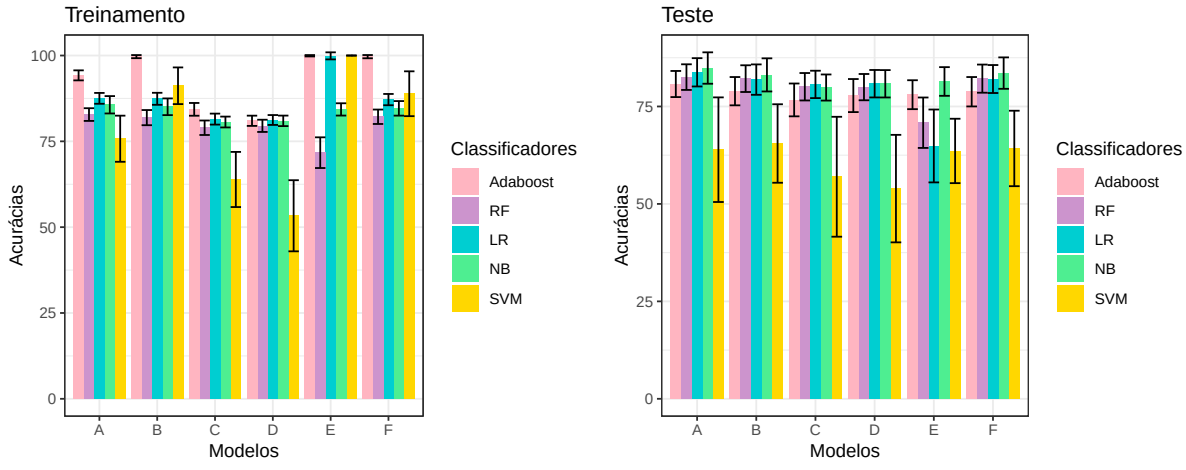
		A		B		C		D		E	
	$Y/\widehat{Y}$	0	1	0	1	0	1	0	1	0	1
NB	0	54	10	55	9	55	9	54	10	54	10
	1	27	9	27	9	29	7	27	9	27	9
RF	0	62	2	50	14	57	7	50	14	50	14
	1	35	1	24	12	30	6	24	12	24	12
SVM	0	55	9	44	20	45	19	45	19	41	23
	1	28	8	20	16	24	12	22	14	21	15
LR	0	54	10	53	11	53	11	53	11	53	11
	1	26	10	24	12	25	11	23	13	25	11
Adaboost	0	28	36	41	23	38	26	43	21	42	22
	1	8	28	18	18	18	18	20	16	19	17

NB-*Bayes* ingênuo; RF-Florestas Aleatórias; SVM-Máquina de Vetores de Suporte; LR-Regressão Logística; Y-variável resposta observada na amostra; $\widehat{Y}$-variável resposta estimada.

com classificador SVM, os quais apresentaram as piores médias de acurácias neste cenário, compreendidas entre 53.93% e 65.50%, além de desvios padrões elevados, variando entre 8.26% e 15.38%.

Observa-se que os modelos A, B, E e F, com classificadores *Adaboost* e SVM, bem como o modelo E com classificador LR, apresentaram problemas de *overfitting* expressivos, quando comparam-se os resultados dos grupos treinamento e teste. A acurácia média mais alta foi observada no modelo A com classificador NB, com valor de 84.86%.

Figura 12 – Dados Hungarian - Acurácia dos modelos de classificação no cenário 2.



A Tabelas 21 e 22 trazem as matrizes relativas de confusão para ambos os grupos, treinamento e teste, no cenário 2. Observa-se que, no grupo treinamento, os modelos que

classificaram corretamente todos os indivíduos nas iterações do *holdout* múltiplo foram os modelos B, com classificador *Adaboost*, modelo E, com classificadores SVM, LR e *Adaboost*, e modelo F, com classificador *Adaboost*. Contudo, nos grupos de teste não visualiza-se tal desempenho, o que decorre do problema de *overfitting*, antes mencionado.

Tabela 21 – Matrizes relativas de confusão, dadas em porcentagem. Grupos de treinamento, banco *Hungarian*, cenário 2. Modelos A, B, C, D, E e F.

		A		B		C		D		E		F	
	$Y/\widehat{Y}$	0	1	0	1	0	1	0	1	0	1	0	1
NB	0	60	4	60	3	54	10	53	10	59	5	61	3
	1	11	25	12	25	9	27	9	28	11	25	12	24
RF	0	57	6	55	7	53	11	53	11	6	22	55	7
	1	11	26	11	27	10	26	9	27	7	65	11	27
SVM	0	54	9	58	4	44	20	33	31	**28**	**0**	57	5
	1	15	22	5	33	16	20	16	20	**0**	**72**	6	32
LR	0	59	4	57	5	55	9	54	10	**28**	**0**	57	5
	1	8	29	8	30	10	26	9	27	**0**	**72**	8	30
Adaboost	0	60	3	**64**	**0**	55	9	51	12	**64**	**0**	**64**	**0**
	1	3	34	**0**	**36**	6	30	7	30	**0**	**36**	**0**	**36**

NB-*Bayes* ingênuo; RF-Florestas Aleatórias; SVM-Máquina de Vetores de Suporte; LR-Regressão Logística; Y-variável resposta observada na amostra; $\widehat{Y}$-variável resposta estimada.

Observa-se também que classificador SVM apresentou as maiores porcentagens de erro de classificação, variando de 11% à 34%.

Tabela 22 – Matrizes relativas de confusão, dadas em porcentagem. Grupos de testes, banco *Hungarian*, cenário 2. Modelos A, B, C, D, E e F.

		A		B		C		D		E		F	
	$Y/\widehat{Y}$	0	1	0	1	0	1	0	1	0	1	0	1
NB	0	61	4	60	4	54	10	54	11	58	7	60	4
	1	11	24	13	23	10	26	8	27	12	23	13	23
RF	0	58	6	56	7	54	10	54	11	7	23	56	7
	1	11	25	10	27	10	26	9	26	6	64	11	26
SVM	0	17	12	16	14	19	**11**	17	12	12	18	15	15
	1	24	47	20	50	32	38	**34**	37	19	51	20	50
LR	0	58	7	55	9	55	10	55	10	14	15	56	8
	1	10	26	10	26	10	25	9	26	20	51	10	26
Adaboost	0	55	10	54	11	52	13	50	14	53	12	53	12
	1	9	26	10	25	10	25	8	28	10	25	10	25

NB-*Bayes* ingênuo; RF-Florestas Aleatórias; SVM-Máquina de Vetores de Suporte; LR-Regressão Logística; Y-variável resposta observada na amostra; $\widehat{Y}$-variável resposta estimada.

5.2.3 Banco de dados Long Beach

A Tabela 23 apresenta os quatro modelos que foram utilizados na classificação de presença ou não de doença cardíaca. Mais uma vez o sexo (`V4`) aparece como variável relevante de acordo com todos os critérios de seleção aqui utilizados, e desta vez acompanhado pela `V11`, que diz respeito a presença ou ausência de hipertensão. A idade (`V3`) aparece com frequência na seleção de variáveis desta base dados, com exceção apenas do modelo selecionado pelo critério InfoGain.

Tabela 23 – Dados Long Beach - Variáveis selecionadas no cenário 1.

Modelo	Critério de seleção	# de variáveis	Lista das variáveis
A	InfoGain	2	V4,V11
B	Saturado+VIF	7	V4,V3,V11,PAM,IPPA,HM,α
C	ANOVA	5	V4,V3,V11,RC,α_2
D	AIC	5	V4,V3,V11,IPPARC,α_2

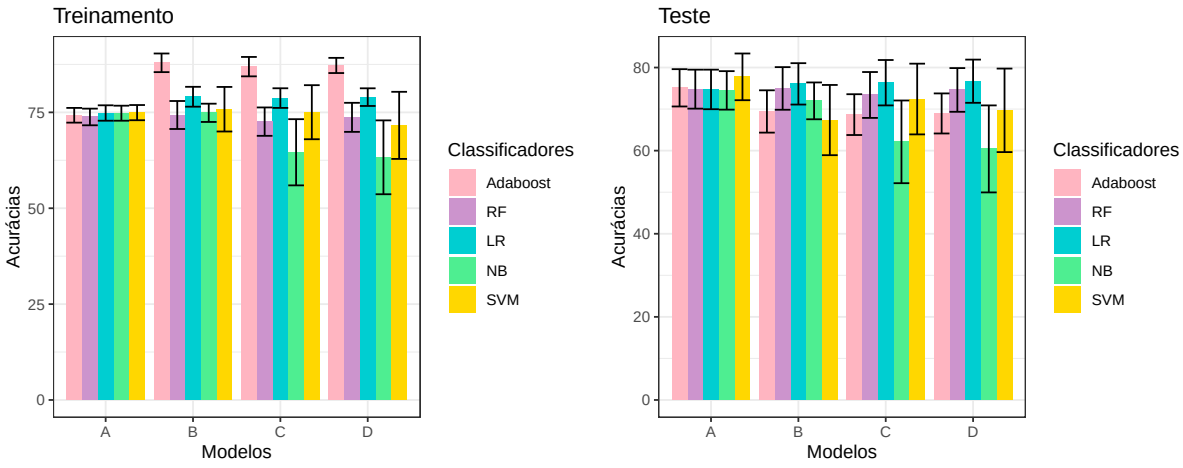
No cenário 2, como apresenta a Tabela 24, as variáveis relevantes em todos os critérios de seleção foram `V4` (sexo) e `V6`, que se refere à presença dor no peito provocada pelo esforço físico.

Tabela 24 – Dados Long Beach - Variáveis selecionadas no cenário 2.

Modelo	Critério de seleção	# de variáveis	Lista das variáveis
A	InfoGain	21	V4,V6,V7,V11,V13, V16,V18,V23,V24,V25, V26,V27,V38,V41,V60, V61,V63,V65,V67,V75
B	Saturado+VIF	19	V4,V6,V7,V11,V12,V14, V15,V19,V28,V59,V60,V62, V63,V64,V65,V68,V70,V71,α
C	ANOVA	5	V4,V6,V43,V60,V61
D	AIC	44	V3,V4,V6,V7,V10,V11, V12,V13,V14,V15,V16, V18,V19,V28,V29,V31, V32,V33,V37,V38,V40, V42,V43,V59,V60,V61,V62, V63,V65,V66,V67,V68,V70, V71,V72,V73,V74,PAM, IPPA,RC,IPPARC,HM,α,α_2
E	AIC + VIF	19	V4,V6,V7,V11,V14, V18,V28,V33,V42,V59, V61,V63,V65,V66,V67, V71,V73,PAM,IPPA

A Figura 13 traz as médias de acurácias e seus respectivos desvios padrões, considerando o contexto do cenário 1, os modelos B, C e D, com classificador *Adaboost*, apresentam problema de *overfitting*, quando se comparam as médias de acurácia do grupo treinamento com as do grupo teste.

Figura 13 – Dados Long Beach - Acurácia dos modelos de classificação no cenário 1.



As melhores classificações dos grupos teste, no cenário 1, foram observadas nos modelos A, com classificador SVM (acurácia média de 77.76% e desvio padrão de 5.62%); modelo B (acurácia média igual à 76.07% e desvio padrão de 4.98%), C (acurácia média de 76.35% e desvio padrão de 5.46%) e D (acurácia média igual à 76.71% e desvio padrão de 5.20%), ambos com classificador LR.

Observa-se, pelas matrizes relativas de confusão nas Tabelas 25 e 26, que os classificadores apresentam melhor predição dos indivíduos cardiopatas, a exemplo do *Adaboost*, no modelo A, que não classificou corretamente nenhum indivíduo que não possuía cardiopatia, tanto no grupo treinamento, quanto no teste.

Quando no cenário 2, como apresenta a Figura 14, observa-se *overfitting* no modelo A, B, D e E, com os classificadores *Adaboost* e LR. Visualiza-se também que o classificador SVM ajusta-se aos dados de treinamento, entretanto, quando aplicam-se os modelos no grupo teste não há predição para estes indivíduos.

Para este cenário, no grupo teste, o classificador NB obteve as menores médias de acurácia, variando de 34.72% à 60.32%. Já as maiores acurácias médias nesta base de dados, grupo teste e cenário 2, em geral, foram geradas pelo classificador *Adaboost*, com valores que variaram entre 78.77% e 94.28%.

Tabela 25 – Matrizes relativas de confusão, dadas em porcentagem. Grupos de treinamento, banco *Long Beach*, cenário 1. Modelos A, B, C e D.

		A		B		C		D	
	Y/$\widehat{Y}$	0	1	0	1	0	1	0	1
NB	0	1	25	6	20	10	16	10	16
	1	0	74	6	68	19	55	20	54
RF	0	0	25	2	20	2	20	2	20
	1	1	74	6	72	7	71	7	71
SVM	0	1	25	10	13	7	16	6	16
	1	0	74	12	65	9	68	12	66
LR	0	1	25	3	19	3	19	3	19
	1	0	74	2	76	2	76	2	76
Adaboost	0	**0**	**26**	15	11	14	12	15	11
	1	**0**	**74**	1	73	2	72	2	72

NB-*Bayes* ingênuo; RF-Florestas Aleatórias; SVM-Máquina de Vetores de Suporte; LR-Regressão Logística; Y-variável resposta observada na amostra; $\widehat{Y}$-variável resposta estimada.

Tabela 26 – Matrizes relativas de confusão, dadas em porcentagem. Grupos de testes, banco *Long Beach*, cenário 1. Modelos A, B, C e D.

		A		B		C		D	
	Y/$\widehat{Y}$	0	1	0	1	0	1	0	1
NB	0	1	25	4	21	8	17	7	18
	1	0	74	7	68	21	54	22	53
RF	0	0	25	3	19	2	19	3	19
	1	1	74	6	72	8	71	7	71
SVM	0	0	21	6	16	5	16	5	16
	1	1	78	16	62	12	67	14	65
LR	0	1	24	2	20	2	20	2	20
	1	1	74	4	74	3	75	3	75
Adaboost	0	**0**	**25**	4	21	4	21	4	21
	1	**0**	**75**	9	66	10	65	10	65

NB-*Bayes* ingênuo; RF-Florestas Aleatórias; SVM-Máquina de Vetores de Suporte; LR-Regressão Logística; Y-variável resposta observada na amostra; $\widehat{Y}$-variável resposta estimada.

As Tabelas 27 e 28 mostram as matrizes relativas de confusão para cada grupo, treinamento e teste, respectivamente, no cenário 2. Observa-se que, no grupo treinamento, os modelos que classificaram todos os indivíduos corretamente foram: modelos A, B, D e E, com classificador SVM; modelos A, D e E, com classificador LR; modelos A e D, com classificador *Adaboost*.

Figura 14 – Dados Long Beach - Acurácia dos modelos de classificação no cenário 2.

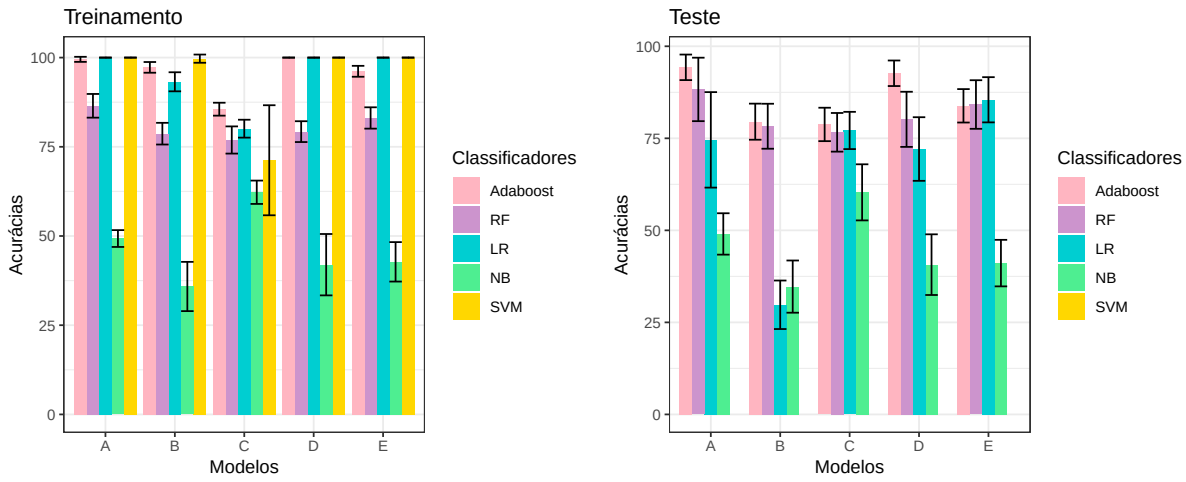


Tabela 27 – Matrizes relativas de confusão, dadas em porcentagem. Grupos de treinamentos, banco *Long Beach*, cenário 2. Modelos A, B, C, D e E.

		A		B		C		D		E	
	$Y/\widehat{Y}$	0	1	0	1	0	1	0	1	0	1
NB	0	26	0	26	0	26	0	26	0	26	0
	1	50	24	64	10	37	37	58	16	57	17
RF	0	10	11	6	16	4	18	3	19	10	12
	1	3	76	6	72	5	73	2	76	4	74
SVM	0	**21**	**0**	**22**	**0**	9	13	**22**	**0**	**22**	**0**
	1	**0**	**79**	**0**	**78**	16	62	**0**	**78**	**0**	**78**
LR	0	**21**	**0**	19	3	8	14	**22**	**0**	**22**	**0**
	1	**0**	**79**	4	74	6	72	**0**	**78**	**0**	**78**
Adaboost	0	**26**	**0**	24	2	17	9	**26**	**0**	23	2
	1	**0**	**74**	0	74	5	69	**0**	**74**	2	73

NB-*Bayes* ingênuo; RF-Florestas Aleatórias; SVM-Máquina de Vetores de Suporte; LR-Regressão Logística; Y-variável resposta observada na amostra; $\widehat{Y}$-variável resposta estimada.

Para esta base de dados, no cenário 2, grupo teste, o classificador SVM não retornou resultado de classificação no software utilizado, como mostra a Tabela 28. As menores porcentagens de erros de classificação foram observadas nos modelos A (6%) e D (7%), com classificador *Adaboost*.

Tabela 28 – Matrizes relativas de confusão, dadas em porcentagem. Grupos de testes, banco *Long Beach*, cenário 2. Modelos A, B, C, D e E.

		A		B		C		D		E	
	$Y/\widehat{Y}$	0	1	0	1	0	1	0	1	0	1
NB	0	25	0	25	0	25	0	25	0	25	0
	1	51	24	65	10	40	35	59	16	59	16
RF	0	11	10	6	16	4	18	4	17	10	12
	1	2	77	6	72	5	73	2	77	4	74
SVM	0	-	-	-	-	-	-	-	-	-	-
	1	-	-	-	-	-	-	-	-	-	-
LR	0	13	8	13	9	6	16	13	8	16	6
	1	18	61	11	67	7	71	20	59	9	69
Adaboost	0	23	**3**	14	11	12	13	21	**4**	15	10
	1	**3**	71	9	66	9	66	**3**	72	6	69

NB-*Bayes* ingênuo; RF-Florestas Aleatórias; SVM-Máquina de Vetores de Suporte; LR-Regressão Logística; Y-variável resposta observada na amostra; $\widehat{Y}$-variável resposta estimada.

5.2.4 Banco de dados Switzerland

Na base de dados *Switzerland*, para o cenário 1, a variável `V11`, referente à presença de hipertensão arterial sistêmica, mostrou-se relevante em quase todos os critérios de seleção, com exceção do InfoGain, o qual selecionou apenas sexo (`V4`) e idade (`V3`), como mostra a Tabela 29.

Tabela 29 – Dados Switzerland - Variáveis selecionadas no cenário 1.

Modelo	Critério de seleção	# de variáveis	Lista das variáveis
A	InfoGain	2	V4,V3
B	Saturado+VIF	5	V4,V3,V11,PAM,RC
C	ANOVA	3	V11,RC,IPPARC
D	AIC	4	V11,PAM,α,α_2
E	AIC + VIF	3	V11,PAM,α_2

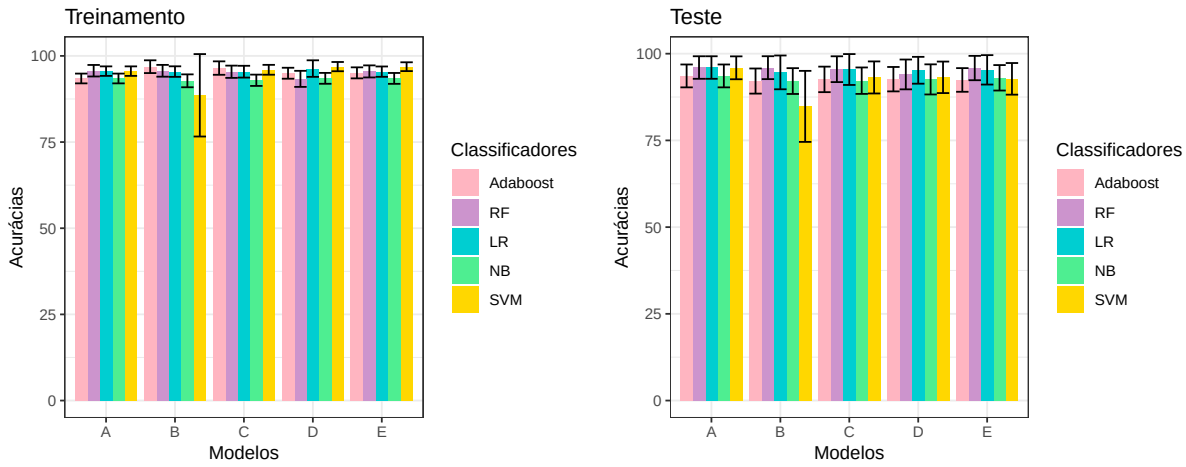
No cenário 2, como é possível visualizar na Tabela 30, a variável `V7` foi a única relevante em todos os critérios de seleção. Esta variável diz respeito ao alívio da dor após descanso.

Com exceção do modelo B, com classificador SVM, tanto nos grupos treinamento quanto nos teste, os demais modelos apresentaram médias elevadas de acurácias, variando de 84.81% à 96.89%, como mostra a Figura 15.

Tabela 30 – Dados Switzerland - Variáveis selecionadas no cenário 2.

Modelo	Critério de seleção	# de variáveis	Lista das variáveis
A	InfoGain	10	V4,V5,V6,V7,V11, V24,V25,V26,V27,V38
B	Saturado+VIF	13	V4,V7,V25,V27,V33, V38,V40,V59,V61, V62,V65,V67,PAM
C	ANOVA	4	V7,V61,V67,PAM
D	AIC	21	V4,V6,V7,V19,V24, V25,V27,V32,V33, V36,V38,V40,V60, V61,V62,V64,PAM, IPPA,RC,IPPARC,HM
E	AIC + VIF	11	V4,V7,V19,V27, V33,V38,V40, V61,V64,PAM,HM

Figura 15 – Dados Switzerland - Acurácia dos modelos de classificação no cenário 1.



Quando visualizadas as matrizes relativas de confusão, dispostas nas Tabelas 31 e 32, percebe-se que, no cenário 1, tanto para o grupo treinamento quanto para o de teste, o modelo A, para todos os classificadores aplicados, não categorizou corretamente os indivíduos não cardiopatas desta amostra em todas as iterações do método *holdout* múltiplo.

No grupo teste, os modelos que também não classificaram bem os não cardiopatas foram: modelo B, com classificadores NB, RF, LR e *Adaboost*; modelo C, com classificadores NB, RF, SVM e LR; modelo D, com classificadores NB, RF e *Adaboost*; modelo E, com classificadores NB, RF, SVM, LR e *Adaboost*.

Tabela 31 – Matrizes relativas de confusão, dadas em porcentagem. Grupos de treinamento, banco *Switzerland*, cenário 1. Modelos A, B, C, D e E.

		A		B		C		D		E	
	$Y/\widehat{Y}$	0	1	0	1	0	1	0	1	0	1
NB	0	**0**	7	0	6	0	6	1	6	0	7
	1	0	93	1	93	1	93	0	93	0	93
RF	0	**0**	4	0	4	0	5	0	5	0	5
	1	0	96	0	96	0	95	2	93	0	95
SVM	0	**0**	4	3	2	0	5	2	2	2	2
	1	0	96	9	86	0	95	1	95	1	95
LR	0	**0**	4	0	5	0	5	1	4	0	5
	1	0	96	0	95	0	95	0	95	0	95
Adaboost	0	**0**	7	4	3	4	3	2	5	2	5
	1	0	93	0	93	0	93	0	93	0	93

NB-*Bayes* ingênuo; RF-Florestas Aleatórias; SVM-Máquina de Vetores de Suporte; LR-Regressão Logística; Y-variável resposta observada na amostra; $\widehat{Y}$-variável resposta estimada.

Tabela 32 – Matrizes relativas de confusão, dadas em porcentagem. Grupos de testes, banco *Switzerland*, cenário 1. Modelos A, B, C, D e E.

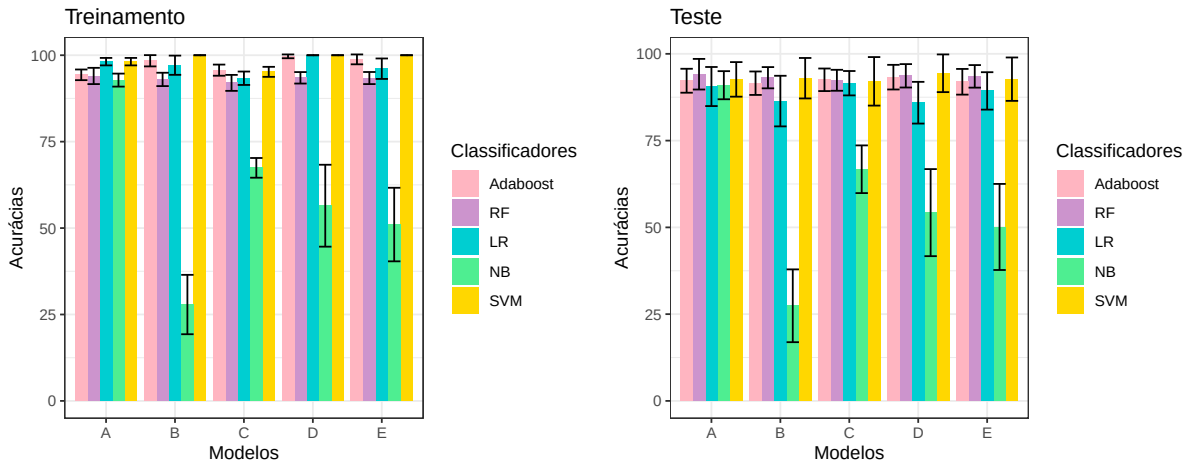
		A		B		C		D		E	
	$Y/\widehat{Y}$	0	1	0	1	0	1	0	1	0	1
NB	0	**0**	6	0	6	0	6	0	6	0	6
	1	0	94	2	92	2	92	1	93	1	93
RF	0	**0**	4	0	4	0	5	0	4	0	4
	1	0	96	0	96	0	95	2	94	0	96
SVM	0	**0**	4	1	3	0	4	1	3	0	4
	1	0	96	12	84	3	93	3	93	4	92
LR	0	**0**	4	0	4	0	4	1	4	0	4
	1	0	96	1	95	1	95	0	95	1	95
Adaboost	0	**0**	6	0	6	1	5	0	6	0	6
	1	0	94	2	92	2	92	1	93	1	93

NB-*Bayes* ingênuo; RF-Florestas Aleatórias; SVM-Máquina de Vetores de Suporte; LR-Regressão Logística; Y-variável resposta observada na amostra; $\widehat{Y}$-variável resposta estimada.

A Figura 16 mostra que, no cenário 2, o classificador SVM obtém acurácias médias altas no grupo treinamento, variando entre 95.19% e 100.00%. Tanto nos grupos treinamento quanto nos de teste, o classificador NB mostrou péssimo desempenho nos modelos B, C, D e E, com médias de acurácias variando entre 27.41% e 66.75%.

No grupo teste, a maior média de acurácia foi observada no modelo D aplicado ao classificador SVM, com valor médio de 94.40% e desvio padrão de 5.43%.

Figura 16 – Dados Switzerland - Acurácia dos modelos de classificação no cenário 2.



Partindo para as matrizes relativas de confusão, dispostas nas Tabelas 33 e 34, os modelos que classificaram melhor os cardiopatas não classificam bem os não cardiopatas, ao contrário do classificador NB, que classificou bem os não cardiopatas das amostras de teste, entretanto errou consideravelmente na predição dos cardiopatas. Se tratando do banco *Switzerland*, a desproporcionalidade entre quantidade total de cardiopatas (115 indivíduos) frente aos não cardiopatas (8 indivíduos) torna-se um problema no momento da adequação do modelo aos dados de treinamento, devido à pouca quantidade de informação da classe dos não cardiopatas. Isto pode justificar o péssimo desempenho do classificador NB nas médias de acurácia em ambos os grupos treinamento e teste.

Tabela 33 – Matrizes relativas de confusão, dadas em porcentagem. Grupos de treinamentos, banco *Switzerland*, cenário 2. Modelos A, B, C, D e E.

		A		B		C		D		E	
	Y/$\widehat{Y}$	0	1	0	1	0	1	0	1	0	1
NB	0	4	2	**7**	0	**7**	0	**7**	0	**7**	0
	1	5	89	**72**	21	**32**	61	**43**	50	**49**	44
RF	0	1	3	0	6	0	6	0	6	0	6
	1	3	93	1	93	2	92	0	94	0	94
SVM	0	3	2	6	0	3	4	6	0	6	0
	1	0	95	0	94	1	92	0	94	0	94
LR	0	3	2	4	2	2	5	6	0	3	3
	1	0	95	1	93	2	91	0	94	1	93
Adaboost	0	1	5	5	2	3	4	6	0	5	1
	1	1	93	0	93	0	93	0	94	0	94

NB-*Bayes* ingênuo; RF-Florestas Aleatórias; SVM-Máquina de Vetores de Suporte; LR-Regressão Logística; Y-variável resposta observada na amostra; $\widehat{Y}$-variável resposta estimada.

Tabela 34 – Matrizes relativas de confusão, dadas em porcentagem. Grupos de testes, banco *Switzerland*, cenário 2. Modelos A, B, C, D e E.

		A		B		C		D		E	
	$Y/\widehat{Y}$	0	1	0	1	0	1	0	1	0	1
NB	0	3	3	6	0	6	0	6	1	5	1
	1	6	88	**73**	21	**33**	61	**45**	48	**49**	45
RF	0	1	3	0	6	0	6	0	6	0	6
	1	3	93	1	93	2	92	0	94	0	94
SVM	0	1	2	2	1	0	3	1	2	1	2
	1	6	91	6	91	5	92	4	93	6	91
LR	0	1	3	1	5	0	6	2	4	1	5
	1	7	89	9	85	3	91	10	84	6	88
Adaboost	0	0	6	0	6	0	6	1	5	0	6
	1	2	92	3	91	2	92	2	92	2	92

NB-*Bayes* ingênuo; RF-Florestas Aleatórias; SVM-Máquina de Vetores de Suporte; LR-Regressão Logística; Y-variável resposta observada na amostra; $\widehat{Y}$-variável resposta estimada.

5.3 DISCUSSÃO GERAL SOBRE OS MODELOS E SEUS RESULTADOS

Inicialmente, vale ressaltar os casos espúrios de modelos, no que diz respeito à interpretação dos dados. Foram os casos do modelo A, no cenário 1 (Tabelas 11, 17, 23 e 29), onde utilizou-se o critério de seleção InfoGain: banco Cleveland (`V4`,`V11`), Banco Hungarian (`V4`,`V11`), banco Long Beach (`V4`,`V11`), banco Switzerland (`V4`,`V3`). Mesmo obtendo valores de acurácias médias que variaram de 37.60% à 96.01%, as variáveis selecionadas para tais modelos, por si só, não devem ser suficientes para firmar um diagnóstico. Sexo, idade e presença do histórico de HAS, coerentemente, são variáveis auxiliares que podem aumentar a probabilidade de se ter doença cardíaca ou não, entretanto é imprudente defini-las como objeto determinante do diagnóstico de tais patologias. O fato que pode justificar a seleção desses modelos é que, para todos os cenários, o critério de seleção InfoGain não elegeu nenhum dos novos parâmetros aqui descritos, e como foi definido na metodologia deste trabalho, no cenário 1 só haviam as variáveis `V3`, `V4`, `V11`, os parâmetros PAM, IPPA, RC, IPPARC, HM, α, α_2 e a variável resposta `V58`.

Dominic, Gupta e Khare (2015) relatam encontrar acurácias variando de 60-70% para as bases de dados *Cleveland* e *Hungarian*, em comparação com uma acurácia de 30-35% para os conjuntos de dados da *LongBeach* e *Switzerland*. Os autores trabalharam a classificação com as 5 categorias originais da variável resposta (valores variando de 0 à 4) e chegaram à conclusão de que esta não é a maneira mais eficaz de se classificar, haja vista que outros estudos utilizando a variável resposta dicotomizada obtiveram mais sucesso em suas classificações.

A exemplo disto temos a pesquisa de Patel et al. (2015), que apresenta resultados para variável resposta sem dicotomização, no qual o melhor algoritmo, um tipo de árvore de decisão denominada J48, obteve acurácia de apenas 56.76%, isto no banco de dados *Cleveland* (Tabela 35). Entretanto, um artigo recente, escrito por Putra et al. (2019), relata acurácia de 92.25% através da aplicação do classificador de Florestas Aleatórias na predição das classes originais da variável resposta Y, as quais variam de 0 à 4. Tal resultado de acurácia demonstra competitividade com os demais encontrados na literatura, todavia ainda não supera algumas das acurácias médias aqui encontradas nos grupos teste, cenário 2, dentre as quais se destacaram as obtidas pelos classificadores *Adaboost* e LR, aplicados aos modelos A, B, E e F, com médias de acurácias variando de 98.11% à 99.20%, e desvios padrões entre 1.17% e 1.99%

Em análises realizadas no banco de dados *Cleveland*, vários pesquisadores encontraram modelos com medidas de acurácia consideráveis, são exemplos deles: Kahramanli e Allahverdi (2008), que apresentaram acurácia de 87.4% ao implementar uma Rede Neuronal do tipo Fuzzy; Tu et al (2009), que relatam acurácia de 81.41% utilizando uma técnica de *Bagging* definida pelos mesmos autores; Sayad e Halkarnikar (2014), que alcançaram o valor 94% de acurácia através do uso de um classificador do tipo Rede Neuronal com múltiplas camadas intermediárias; Krishnasree e Rao (2016), que referiram obter acurácia de 91% utilizando método próprio de Regularização Bayesiana; Umamaheswari, Valarmathi e Jasmine (2017), que obtiveram valores de acurácia variando entre 86.49-91.49% utilizando diferentes tipos de classificadores, dentre eles, RF e SVM.

O problema desses ajustes é que em nenhum momento se considera a presença de dependência entre as variáveis explicativas ou se divide a base de dados em grupo de treinamento e teste - metodologia básica da mineração de dados e classificação -, o que pode ter como consequência problemas de *overfitting* (KOUROU et al., 2015). Além disso, tais metodologias distintas em cada uma dessas pesquisas dificultam a comparação com os resultados das médias de acurácias relatadas nesta dissertação. Da literatura aqui abordada, apenas Das et al. (2009) considerou o particionamento da base de dados em treinamento/teste e realizou uma seleção de variáveis, alcançando acurácia de 89.01% com um classificador do tipo RN. Tais observações sobre as metodologias encontradas na literatura, referente a estas bases de dados, trazem à reflexão sobre a necessidade de padronização do delineamento do plano de pesquisa, no qual se faz importante inserir as diversas técnicas consolidadas na Estatística, desde a amostragem, coleta e tratamento dos dados, até a forma de apresentação das análises. Assim, compreende-se que seria útil, no âmbito do desenvolvimento e produção científica, a construção de um protocolo

universal de pesquisa voltado para área da saúde.

A Tabela 35 traz as referências principais utilizadas para comparação neste trabalho. A pesquisa original é a de Detrano et al. (1989) que coletou os dados do banco Cleveland e o utilizou para classificação de cardiopatas. A partir dela, surgem outras pesquisas testando diversos tipos de classificadores e/ou diferenciando ente si na quantidade de variáveis explicativas utilizadas.

Dominic, Gupta e Khare (2015) relatam ter obtido acurácia de 98% utilizando classificador do tipo *Adaboost*, com oito variáveis do banco *Cleveland* (Tabela 35). Contudo, tanto no ajuste de tal modelo quanto na predição dos cardiopatas, utilizaram os dados em sua totalidade, não segregando a amostra em grupo de treinamento e teste. Isto compromete a avaliação da adequação modelo, haja vista que, uma vez ajustado aos dados, o modelo deve classificar bem estes mesmos dados, e testar este modelo em outro grupo amostral mostra-se uma forma mais justa de avaliar sua adequação aos padrões da amostra (YADAV e SHUKLA, 2016).

Tabela 35 – Referências principais.

Autores (Ano)	teste e treinamento	Resposta Binária	# Variáveis	Algorítimo	Acurácia	Banco
Detrano et al. (1989)	Sim	Sim	13	Logistic-derived discriminant function	77.00%	Cleveland
Gennari et al. (1989)	Não	Sim	13	Classit	78.90%	Cleveland
Kahramanli e Allahverdi (2008)	Sim	Sim	13	RN Fuzzy	87.4%	Cleveland
Das et al. (2009)	Sim	Sim	5	RN	89.01%	Cleveland
Tu et al. (2009)	Sim	Sim	13	DT Bagging	78.91% 81.41%	Cleveland
Patel et al. (2015)	Sim	Não	13	J48 J48+Pruning Logistic Tree RandomForest	- 56.76% 55.77% -	Cleveland
Dominic, Gupta e Khare (2015)	Não	Não	8 8 - - 7 - 8	NB DT SVM Logistic MLP Adaboost	66% 86% - - 64% - 98%	Long Beach/ Switzerland Cleveland - - Switzerland - Cleveland
Sayad e Halkarnikar (2014)	Sim	Sim	13	MLP	94%	Cleveland
Krishnasree e Rao (2016)	Sim	Sim	13	Regressão Múltipla Regularização Bayesiana	86% 91%	Cleveland
Umamaheswuari et al. (2017)	Sim	Sim	13	RandomForest GBM LDA SVM Stacking	86.49% 86.49% 85.14% 83.78% 91.89%	Cleveland
Putra et al. (2019)	Sim	Não	-	RandomForest	92.25%	Cleveland

Segundo Dominic, Gupta e Khare (2015), os bancos de dados *Long Beach* e *Switzerland* apresentam discrepância na proporção de indivíduos nas categorias originais da variável resposta (que vai de 0 à 4). Quando dicotomizadas para as categorias de cardiopatas e não cardiopatas, a desproporcionalidade entre as classes continua existindo, agora entre as duas classes (cardiopatas e não cardiopatas), como pode ser visualizado na Tabela 1, na qual para o banco *Long Beach* haviam 51 sujeitos não cardiopatas e 149 cardiopatas, e para o banco *Switzerland*, a quantidade era de 8 não cardiopatas para 115 cardiopatas.

Esse desbalanceamento nas categorias de predição mostrou-se um problema no momento da estimação, principalmente na base *Switzerland*, pois havia pouca informação sobre a classe dos não cardiopatas, neste caso, e a maioria dos modelos ajustados não conseguiram prever corretamente os indivíduos dessa classe, como verifica-se na Tabela 34.

Outro problema encontrado no processo de classificação, também mencionado por Dominic, Gupta e Khare (2015), refere-se ao fato de haver grande quantidade de dados faltantes (NA) nas bases de dados aqui utilizadas, principalmente na *Long Beach* e *Switzerland*. Tal condição acarreta na redução de observações utilizadas na classificação, e, consequentemente, no cálculo das acurácias. Nesta pesquisa, o SVM foi o classificador que sofreu maior impacto, no tocante à exclusão de observações com dados faltantes, e isso pode justificar a alta variabilidade nas médias de acurácias, considerando que menor quantidade de elementos na matriz de confusão implica que, uma mínima mudança na classificação de tais indivíduos resulta em maior impacto no resultado da acurácia (STOCKWELL e PETERSON, 2002).

5.4 PARÂMETROS DE CAMPELLO DE SOUZA NO CONTEXTO DOS MODELOS SELECIONADOS

Nesta seção se propõe apresentar as análises dos modelos definidos pelas técnicas de seleção de atributos, já descritas na metodologia desta dissertação, com ênfase nos parâmetros Campello de Souza, a fim de identificar qual deles mostrou-se mais relevante na classificação de doentes cardíacos.

A Tabela 36 apresenta as frequências com que os critérios de seleção determinaram os parâmetros Campello de Souza como relevantes, considerando a classificação de pessoas com e sem cardiopatia. Dos sete aqui descritos, a PAM foi escolhida pelos critérios em 22 modelos (52.38%), do total de 42, o que mostra a relevância deste parâmetro no contexto da classificação de indivíduos com doença cardíaca. Em seguida observa-se o IPPA, que foi escolhido pelos

critérios em 13 dos 42 modelos, correspondendo à 30.95%.

Tabela 36 – Frequências absolutas e relativas dos parâmetros em ambos os cenários.

PAM		IPPA		RC		IPPARC		HM		α		α_2		N
n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	
22	**52.38**	13	30.95	11	26.19	9	21.43	9	21.43	9	21.43	8	19.05	42

Quando consideram-se os 12 modelos válidos - que aqui foram definidos como modelos que passaram pelo processo de seleção pelo VIF - com maiores médias de acurácias no grupo teste, cenário 2, e obtêm-se as médias das medidas de desempenho (sensibilidade, especificidade e VPP), se observa, pela Tabela 37, que os modelos com acurácias médias mais altas para cada banco foram: modelo F com classificador LR, para *Cleveland*; modelo F com classificador NB, para *Hungarian*; modelo E com classificador LR, para *Long Beach*. No caso da base de dados *Switzerland*, o modelo com acurácia média alta foi o modelo E, com classificador RF, entretanto, o único modelo dos três definidos como válidos neste banco de dados que apresentou média de Especificidade não nula (média de 5.96% e desvio padrão igual à 16.78%), foi o modelo C, com classificador *Adaboost*.

Um fato que deve ser mencionado é que, nesse contexto, os modelos válidos com maiores médias de acurácias, os quais correspondem aos modelos destacados em negrito na Tabela 37, foram os selecionados pelo critério AIC+VIF, e em todos eles, a PAM está presente.

Tabela 37 – Modelos válidos com maiores médias de acurácia no cenário 2, grupos teste, suas respectivas métricas de desempenho da predição e os parâmetros Campello de Souza.

Bancos	Modelo	# variáveis	Classificador	Acurácia	Sensib.	Especif.	VPP	Parâmetros Campello de Souza
Cleveland	B	41	*Adaboost*	98.58 (1.80)	96.82 (3.93)	99.98 (0.20)	99.97 (0.32)	PAM, IPPA, HM, α
	D	6	LR	81.32 (3.83)	75.76 (7.15)	85.94 (4.53)	81.00 (6.40)	-
	F	**27**	**LR**	**99.20 (1.17)**	**98.23 (2.46)**	**100.00 (0.00)**	**100.00 (0.00)**	**PAM, IPPA, RC**
Hungarian	B	20	NB	83.10 (4.25)	64.71 (10.51)	93.17 (3.95)	84.39 (7.72)	PAM, IPPA, HM
	D	4	LR	80.84 (3.52)	74.60 (7.16)	84.31 (4.83)	72.15 (7.42)	-
	F	**18**	**NB**	**83.56 (4.03)**	**64.16 (10.09)**	**94.17 (3.52)**	**86.20 (6.75)**	**PAM, RC**
Long Beach	B	19	*Adaboost*	79.52 (4.90)	88.00 (4.83)	55.05 (14.21)	85.33 (5.30)	α
	C	5	*Adaboost*	78.77 (4.54)	88.65 (5.60)	50.55 (16.23)	84.19 (5.76)	-
	E	**19**	**LR**	**85.47 (6.14)**	**88.78 (6.16)**	**74.23 (16.85)**	**92.22 (5.13)**	**PAM, IPPA**
Switzerland	B	14	RF	93.12 (3.06)	99.33 (1.35)	0.00 (0.00)	93.38 (3.03)	PAM
	C	4	*Adaboost*	92.52 (3.25)	98.44 (2.18)	5.96 (16.78)	93.74 (3.32)	PAM
	E	**12**	**RF**	**93.51 (3.25)**	**99.74 (0.86)**	**0.00 (0.00)**	**93.33 (3.08)**	**PAM, HM**

Com relação ao desempenho desses modelos em geral, o que melhor estimou e apresentou maiores métricas de desempenho da predição foi o modelo F com classificador LR, no banco *Cleveland*, o qual contém, além de algumas das variáveis nativas da base de dados, a

PAM, o IPPA e o RC, que correspondem aos parâmetros de Campello de Souza com maiores frequências nos modelos selecionados, em ambos os cenários 1 e 2.

Tal modelo apresentou acurácia média de 99.20%, Sensibilidade média de 98.23%, Especificidade média igual à 100.00% e VPP médio também igual à 100.00%. Para este mesmo banco de dados, Das et al. (2009) encontrou valores de 89.01% de acurácia, 80.95% de Sensibilidade e 95.91% de Especificidade; Kahramanli e Allahverdi (2008) relatam ter alcançado 87.4% de acurácia, Sensibilidade de 93% e Especificidade de 78.5%; e o artigo de Sayad e Halkarnikar (2014) apresenta valores de 94% de acurácia, 92% de Sensibilidade e 92.5% de Especificidade. Esses resultados demonstram a competitividade do modelo F com o classificador LR, o qual mostrou ótima predição tanto dos não cardiopatas dentro do grupo de pessoas sem doença cardíaca (Especificidade de 100.00%) quanto dos cardiopatas dentro do grupo de pessoas que se estima terem doença cardíaca (VPP igual à 100.00%).

Considerando os classificadores, a Regressão Logística e o Adaboost foram os métodos com maiores médias de acurácias, cada um presente em 33.33% dos 12 modelos selecionados pelo VIF. Dominic, Gupta e Khare (2015), em estudo semelhante, relatam que os classificadores *Adaboost*, NB e DT obtiveram as maiores acurácias para estes bancos de dados, e que dentre estes, *Adaboost* e DT se sobressaíram, com acurácias variando de 75.00% à 97.65%. Este resultado é um indicativo da complementaridade das variáveis envolvidas no que diz respeito à informação necessária dos indivíduos para predizer as classes cardiopatas e não cardiopatas, considerando que o classificador NB estima as probabilidades condicionais de cada variável do modelo sob o pressuposto de independência entre os atributos.

Quando partimos para a análise dos parâmetros Campello de Souza no contexto desses 12 modelos válidos com maiores médias de acurácias, é notória a relevância da PAM na classificação dos cardiopatas, isto em quase todos os modelos dispostos na Tabela 37. Este resultado concorda com o que já foi discutido sobre a Tabela 36 e sobre as medidas de consistência adaptadas.

Dos 12 modelos da Tabela 37, as variáveis que aparecem com mais frequência são: `V4` (sexo), presente em 11 destes modelos (91.67%); o parâmetro PAM, que aparece em 8 dos 12 modelos (66.67%); `V61` (distância da artéria descendente anterior esquerda) e `V40` (depressão do segmento ST induzida pelo exercício com relação ao repouso), ambas com frequências relativas de 58.33% cada; e `V6` (dor no peito provocada pelo esforço físico), presente em metade dos modelos (50.00%). Estas variáveis mostraram-se relevantes na classificação dos cardiopatas nas bases de dados aqui apresentadas.

Com relação ao tempo de processamento dos modelos da Tabela 37 aplicados a seus respectivos classificadores, observa-se na Tabela 38 que os classificadores LR e RF apresentaram as menores médias do tempo de execução, com valores de 6.80 até 27.50 milissegundos.

Tabela 38 – Tempos de Processamento dos modelos válidos com maiores médias de acurácia no cenário 2, grupos teste.

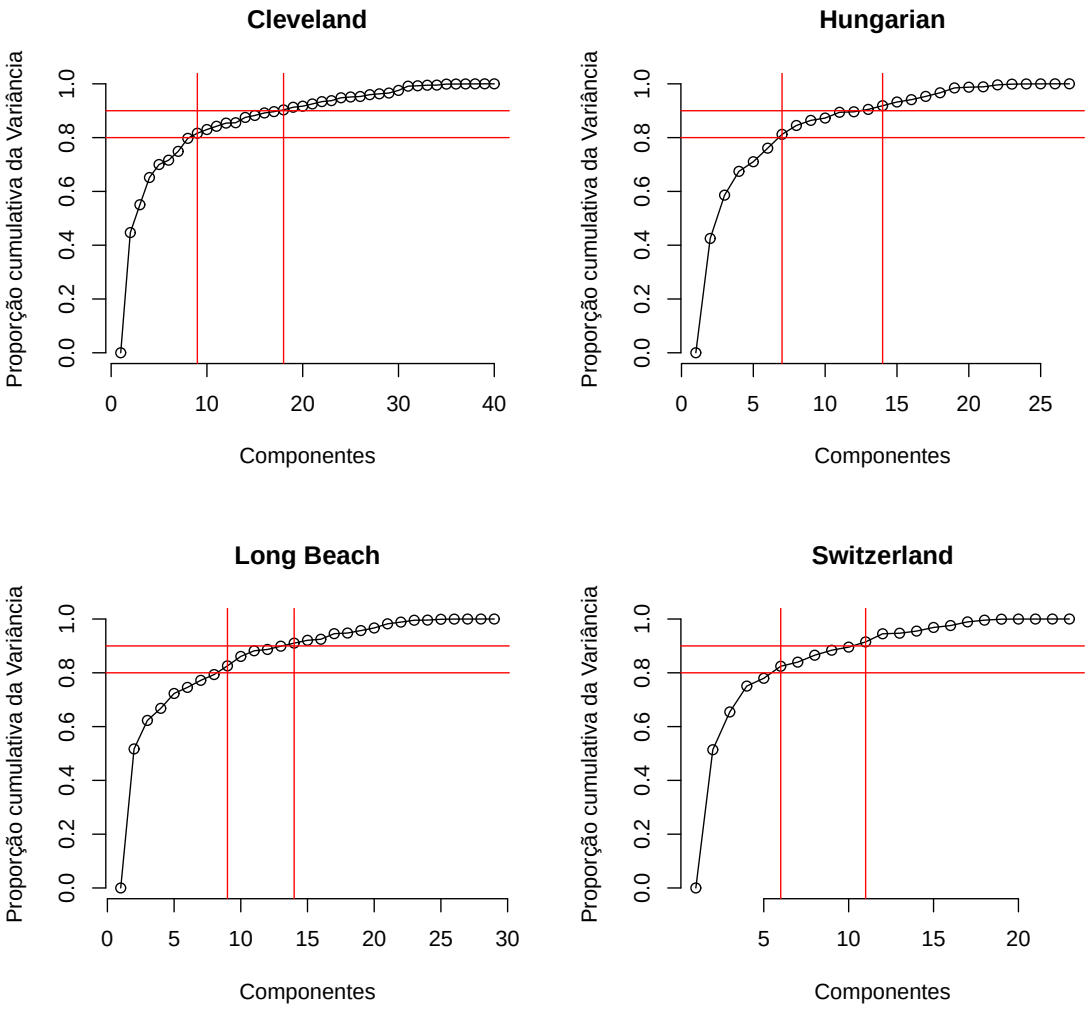
Banco de Dados	Modelo	Classificador	Tempo em milissegundos	
			Média	Desvio Padrão
Cleveland	B	*Adaboost*	275.50	(63.06)
	D	**LR**	**6.80**	**(8.39)**
	F	LR	27.30	(10.53)
Hungarian	B	NB	239.70	(46.85)
	D	**LR**	**12.10**	**(8.80)**
	F	NB	166.90	(35.01)
Long Beach	B	*Adaboost*	214.80	(57.92)
	C	*Adaboost*	126.30	(38.73)
	E	**LR**	**17.50**	**(9.47)**
Switzerland	**B**	**RF**	**27.78**	**(9.09)**
	C	*Adaboost*	75.15	(11.98)
	E	**RF**	**23.64**	**(8.97)**

5.5 COMPONENTES PRINCIPAIS E CLASSIFICAÇÃO

Nesta seção apresenta-se os resultados das classificações realizadas utilizando componentes principais como variáveis explicativas, as quais tiveram o objetivo de reduzir a dimensionalidade da matriz explicativa e aproveitar maior quantidade possível de informação disponível nas variáveis quantitativas dos bancos de dados, considerando o contexto de predição dos cardiopatas.

Para a estimativa das componentes principais foi utilizada a matriz de correlação, tendo em vista que as variáveis quantitativas apresentavam-se em diferentes unidades de medida. Após construir componentes para cada base de dados, os gráficos *Screeplot* foram delineados e apresentados na Figura 17, nas quais foram traçadas retas (em vermelho) que delimitam a quantidade de componentes que possuem entre 80% e 90% de representabilidade da variância total dos dados explicativos. Observa-se que, com no mínimo nove componentes, em todos os bancos, já se obtém pouco mais de 80% de variabilidade total dos dados explicada. Assim, considerou-se 10 Componentes, um número aceitável para aplicação dos classificadores em todos os bancos de dados.

Figura 17 – Screeplots - Componentes Principais versus variabilidade cumulativa explicada.



As Tabelas 39 e 40 dispõem das métricas de desempenho dos modelos na classificação nos Bancos de treinamentos e testes, respectivamente.

Para os grupos treinamento, destacam-se os classificadores SVM, RN e *Adaboost* com acurácias variando entre 92.25% (RN no banco *Hungarian*) e 100% (SVM, em todas as bases de dados e RN no banco *Switzerland*), e desvios padrões variando de 0.00% (SVM, em todas as bases de dados e RN no banco *Switzerland*) à 2.10% (*Adaboost* no banco *Switzerland*).

Se tratando de acurácia, em geral, o classificador RF teve o pior desempenho em todos os quatro bancos no grupo treinamento, enquanto o SVM obteve desempenhos excelentes nos mesmos grupos, com valores de acurácias, sensibilidades, especificidades e VPPs iguais à 100% e desvios padrões nulos. Este resultado, em específico, corrobora com o fato de que o classificador SVM, tem melhor desempenho na classificação quando as variáveis de entrada são de natureza contínua (KOTSIANTIS, ZAHARAKIS e PINTELAS, 2007).

Tabela 39 – Métricas de desempenho da predição dos modelos construídos com 10 Componentes Principais, Bancos treinamento.

Bancos	Classificador	Acurácia	Sensib.	Especif.	VPP
Cleveland	NB	87.77 (1.74)	81.09 (3.03)	94.50 (1.76)	93.77 (1.88)
	RF	83.78 (2.09)	80.29 (3.19)	87.20 (3.29)	86.56 (2.95)
	SVM	**100.00 (0.00)**	**100.00 (0.00)**	**100.00 (0.00)**	**100.00 (0.00)**
	LR	94.67 (1.57)	94.19 (1.87)	95.16 (1.73)	95.20 (1.66)
	RN	**99.40 (0.57)**	**99.51 (0.88)**	**99.27 (0.75)**	**99.29 (0.75)**
	Adaboost	**99.97 (0.14)**	**100.00 (0.00)**	**99.94 (0.28)**	**99.95 (0.26)**
Hungarian	NB	74.73 (1.75)	46.60 (4.52)	91.75 (1.64)	77.47 (3.40)
	RF	72.75 (2.82)	49.58 (4.88)	86.69 (3.40)	69.73 (4.59)
	SVM	**100.00 (0.00)**	**100.00 (0.00)**	**100.00 (0.00)**	**100.00 (0.00)**
	LR	80.92 (2.35)	62.47 (5.05)	92.02 (2.19)	82.82 (3.42)
	RN	**92.25 (1.66)**	**83.78 (3.80)**	**97.37 (1.46)**	**95.18 (2.46)**
	Adaboost	**99.96 (0.15)**	**99.99 (0.14)**	**99.94 (0.23)**	**99.90 (0.38)**
Long Beach	NB	78.27 (3.73)	91.07 (3.78)	33.56 (9.80)	82.63 (2.84)
	RF	72.55 (4.28)	93.10 (3.44)	1.16 (2.57)	76.49 (2.95)
	SVM	**100.00 (0.00)**	**100.00 (0.00)**	**100.00 (0.00)**	**100.00 (0.00)**
	LR	79.26 (3.21)	95.95 (2.30)	21.02 (11.21)	80.87 (2.74)
	RN	**99.51 (0.74)**	**99.87 (0.44)**	**98.25 (2.98)**	**99.50 (0.85)**
	Adaboost	**99.33 (1.07)**	**100.00 (0.00)**	**96.95 (4.98)**	**99.16 (1.31)**
Switzerland	NB	93.49 (3.73)	97.65 (2.43)	52.37 (24.23)	95.31 (2.69)
	RF	90.61 (2.68)	99.22 (1.13)	0.00 (0.00)	91.25 (2.34)
	SVM	**100.00 (0.00)**	**100.00 (0.00)**	**100.00 (0.00)**	**100.00 (0.00)**
	LR	93.96 (3.81)	99.57 (0.89)	40.27 (35.37)	94.20 (3.58)
	RN	**100.00 (0.00)**	**100.00 (0.00)**	**100.00 (0.00)**	**100.00 (0.00)**
	Adaboost	**98.22 (2.10)**	**100.00 (0.00)**	**79.35 (26.89)**	**98.13 (2.20)**

Quando avalia-se o cenário dos grupos de teste, o classificador LR obtém melhores valores de acurácia no banco *Cleveland* (91.03%), *Hungarian* (77.83%) e *Long Beach* (73.63%), com desvios padrões de 3.59%, 4.28% e 6.71%, respectivamente. Já no banco *Switzerland*, os melhores valores de acurácias foram de 89.83% (RF), 89.03% (SVM) e 88.82% (*Adaboost*), com desvios padrões de 5.03%, 5.34% e 5.75%, respectivamente.

Nota-se que há subestimação na predição quando partimos do grupo treinamento para o teste, e que no grupo treinamento, até mesmo para os bancos ditos problemáticos por conta da grande quantidade de dados faltantes (*Long Beach* e *Switzerland*) alcançou-se especificidades de 100%, com desvio padrão igual a 0.00%. Em contrapartida, quando referem-se aos grupos testes, por mais que nem todos os valores de especificidade sejam nulos, os desvios padrões são relativamente altos quando comparados as suas respectivas médias. Isso reflete novamente a problemática no ajuste desses bancos, relatada outrora por Dominic, Gupta e Khare (2015), e a dificuldade em identificar e classificar corretamente os indivíduos não cardiopatas dentro do

Tabela 40 – Métricas de desempenho da predição dos modelos construídos com 10 Componentes Principais, Bancos teste.

Bancos	Classificador	Acurácia	Sensib.	Especif.	VPP
Cleveland	NB	85.18 (3.86)	79.30 (7.27)	91.12 (5.55)	90.01 (5.84)
	RF	83.78 (4.22)	80.91 (7.37)	86.85 (6.39)	86.04 (6.70)
	SVM	83.84 (3.86)	86.38 (6.27)	81.46 (6.94)	82.41 (5.88)
	LR	**91.03 (3.59)**	**90.81 (5.08)**	**91.30 (5.29)**	**91.33 (4.99)**
	RN	87.79 (3.41)	86.54 (6.19)	89.17 (5.29)	88.86 (5.31)
	Adaboost	82.97 (4.46)	84.58 (6.78)	81.59 (7.28)	82.11 (6.97)
Hungarian	NB	74.32 (4.79)	44.60 (10.01)	90.93 (4.19)	73.23 (10.71)
	RF	74.01 (4.36)	51.14 (9.47)	87.05 (5.34)	69.02 (10.60)
	SVM	65.66 (5.06)	51.83 (10.55)	73.58 (7.01)	52.11 (9.07)
	LR	**77.83 (4.28)**	**58.66 (9.31)**	**88.62 (4.73)**	**74.14 (9.80)**
	RN	70.43 (5.18)	56.83 (10.47)	78.13 (6.71)	59.13 (10.22)
	Adaboost	68.16 (5.06)	58.83 (11.57)	73.70 (6.93)	55.37 (8.63)
Long Beach	NB	71.37 (6.87)	87.29 (6.90)	15.04 (12.63)	78.68 (6.45)
	RF	73.49 (5.46)	93.55 (4.84)	1.75 (4.98)	77.47 (6.02)
	SVM	70.06 (6.60)	87.44 (6.95)	8.85 (10.32)	77.49 (6.60)
	LR	**73.63 (6.71)**	**91.39 (6.71)**	**10.59 (10.37)**	**78.60 (6.24)**
	RN	59.83 (7.22)	69.06 (9.10)	27.33 (16.37)	77.37 (7.30)
	Adaboost	70.91 (6.18)	88.76 (7.18)	7.47 (9.37)	77.57 (6.16)
Switzerland	NB	86.14 (6.72)	95.47 (6.31)	0.00 (0.00)	89.94 (5.20)
	RF	**89.83 (5.03)**	**99.49 (1.53)**	**0.00 (0.00)**	**90.29 (5.18)**
	SVM	89.03 (5.34)	98.60 (2.47)	0.00 (0.00)	90.20 (5.21)
	LR	81.43 (10.15)	89.86 (10.26)	2.66 (12.83)	89.55 (5.67)
	RN	81.17 (6.77)	89.40 (8.32)	11.05 (29.72)	90.18 (5.87)
	Adaboost	88.82 (5.75)	98.38 (3.94)	0.00 (0.00)	90.19 (5.17)

grupo de pessoas sem doença cardíaca.

Para o banco *Cleveland*, o qual possui diversas aplicações na literatura, uma acurácia de 91.03% ($\pm$ 3.59%), sensibilidade de 90.81% ($\pm$ 5.08%) e especificidade igual à 91.30% ($\pm$ 5.29%), equivalem a resultados competitivos com outros tipos de tratamento de dados e classificadores, como por exemplo: 89.01% de acurácia, 80.95% de sensibilidade e 95.91% de especificidade utilizando Redes Neuronais (DAS et al., 2009); 87.4% de acurácia, sensibilidade igual à 93%, e especificidade de 78.5% usando uma Rede Neuronal híbrida proposta por Kahramanli e Allahverdi (2008); 92% de sensibilidade, 92.5% de especificidade e 94% de acurácia, também utilizando Redes Neuronais (SAYAD e HALKARNIKAR, 2014).

Tais resultados podem motivar trabalhos futuros voltados para o tratamento das informações quantitativas das bases de dados aqui discutidas, com o objetivo de avaliar o desempenho de outros classificadores que utilizam técnicas de redução da dimensionalidade da matriz explicativa, como por exemplo, a técnica de Aprendizagem Gráfica da Dispersão

Adaptativa (CHEN et al., 2019).

A Tabela 41 traz os tempos de processamento computacional dos classificadores aplicados às 10 primeiras componentes principais. Dentre os seis classificadores, SVM e LR apresentaram os menores custos computacionais, com médias variando entre 5.40 e 14.10 milissegundos, e desvios padrões de 6.83 à 9.05 milissegundos, isto nos quatro bancos de dados (*Cleveland*, *Hungarian*, *Long Beach* e *Switzerland*). Já com relação ao maior tempo de processamento, o *Adaboost* mostrou maior custo computacional, com médias de tempos variando de 95.60 ($\pm$ 24.01) à 141.50 ($\pm$ 29.76) milissegundos.

Tabela 41 – Médias e desvios padrões dos tempos (em milissegundos) de processamento dos ajustes utilizando 10 Componentes principais.

Banco	NB	RF	SVM	LR	RN	*Adaboost*
Cleveland	68.20 (16.29)	55.20 (9.69)	**9.90 (8.35)**	**6.40 (8.23)**	47.10 (18.38)	**118.20 (13.95)**
Hungarian	87.10 (29.38)	78.40 (11.61)	**14.10 (6.83)**	**6.60 (8.31)**	61.50 (16.04)	**141.50 (29.76)**
Long Beach	42.90 (22.26)	35.40 (7.97)	**7.80 (9.05)**	**7.70 (8.51)**	31.20 (9.35)	**110.20 (14.84)**
Switzerland	27.60 (11.02)	17.10 (7.95)	**6.70 (8.42)**	**5.40 (7.97)**	29.40 (12.54)	**95.60 (24.01)**

5.6 INCLUSÃO DA VARIÁVEL `LOCALITY`

Após unificar as bases de dados, a variável `locality`, que diz respeito aos locais (países ou cidades) nos quais foram coletados os dados, foi adicionada com a finalidade de verificar se ela é significativa na classificação dos cardiopatas destes bancos. A Tabela 42 traz as distribuições dos cardiopatas estratificados pelos locais, a qual foi utilizada para implementação do teste Qui-Quadrado de Pearson.

Tabela 42 – Tabela de contingência do banco de dados unificado - Cardiopatas versus local de coleta dos dados.

Y	*Cleveland*	*Hungarian*	*Long Beach*	*Switzerland*	Total
Não Cardiopatas	157	188	51	8	404
Cardiopatas	125	106	149	115	495
Total	282	294	200	123	899

O teste Qui-Quadrado de Pearson obteve estatística $\chi^2 = 160.03$, com três graus de liberdade e p-valor menor que 0.001. Desta forma, ao nível de 1% de significância, conclui-se que os países/cidades incluídos como variáveis na análise mostraram-se atributo importante quando comparados os grupos de cardiopatas e não cardiopatas.

Comparados os locais, dois a dois, apenas não há diferença significativa entre a proporção de cardiopatas de *Cleveland* e *Hungarian*, com estatística $\chi^2 = 3.7629$, para grau de liberdade um e p-valor igual a 0.0524. Em seu estudo, Dominic, Gupta e Khare (2015) relatam observar distribuição balanceada entre os bancos *Cleveland* e *Hungarian* quando subdivididos em indivíduos cardiopatas e não cardipatas.

A relevância da variável `locality` pode ser observada também ao aplicar os métodos de seleção de variáveis mencionados na metodologia desta dissertação, dentre os quais, foi escolhida tanto pelo InfoGain quanto pelo AIC. Este resultado serve como motivador de implementação dos algoritmos de classificação utilizando um banco de dados unificado.

6 CONSIDERAÇÕES FINAIS

As estatísticas apresentadas pela Organização Mundial da Saúde são preocupantes e motivam pesquisas direcionadas às cardiopatias, considerando o fato de que identificar doenças precocemente aumenta a sobrevida dos indivíduos que as portam. Partindo dessa premissa, métodos diagnósticos efetivos são imprescindíveis para tomada de decisão.

Neste trabalho foram utilizados alguns dos parâmetros, descritos por Campello de Souza (2010), como variáveis auxiliares na classificação de pessoas com e sem cardiopatias. Dentre eles destacou-se a PAM, tanto pelas medidas de consistências adaptadas (banco *Cleveland* e *Hungarian*) quanto pela frequência de seleção dentre os modelos válidos. Ainda sobre estes parâmetros, observou-se que, como variáveis auxiliares na classificação, os mesmos não são autossuficientes para tomada de decisão.

O desbalanceamento evidente entre a quantidade de indivíduos cardiopatas e não cardiopatas das amostras *Long Beach* e *Switzerland* comprometeu resultados das suas respectivas médias de acurácia e, principalmente da Especificidade, em ambos os bancos de dados, com destaque para o banco *Switzerland*, que apresentou algumas médias de Especificidade iguais à 0.00. No geral, considerando as quatro bases de dados, as acurácias dos 12 modelos válidos com maiores médias de acurácia no cenário 2, grupos teste, variaram entre 78.77% e 99.20%.

O modelo que apresentou maior acurácia média no grupo teste foi o modelo F, com classificador LR, aplicado no banco *Cleveland*, cenário 2. Para tal modelo, as demais métricas de desempenho foram: Sensibilidade média de 98.23%, Especificidade média igual à 100.00% e VPP médio também igual à 100.00%.

Considerando os classificadores, a Regressão Logística e o Adaboost foram os métodos com maiores médias de acurácias dentre os modelos selecionados pelo VIF. E quanto ao tempo de processamento do ajuste dos modelos, os classificadores LR e RF se mostraram menos custosos computacionalmente.

As variáveis que se mostraram mais relevantes na classificação dos cardiopatas das bases de dados aqui apresentados foram: `V4` (sexo), o parâmetro PAM, `V61` (distância da artéria descendente anterior esquerda), `V40` (depressão do segmento ST induzida pelo exercício com relação ao repouso), e `V6` (dor no peito provocada pelo esforço físico).

Foi constatada na literatura referenciada que a maioria das pesquisas, as quais envolvem este diretório de dados e o problema da classificação de cardiopatas, consideram 13 variáveis explicativas do banco reduzido disponibilizado pelo Repositório de Aprendizado de

Máquina da UC Irvine, sem que houvesse um processo de seleção de variáveis que considerasse a possível dependência entre os atributos explicativos. Além disso, em algumas das pesquisas não houve subdivisão das bases de dados em grupos de treinamento e teste, o que, juntamente com a dependência entre as variáveis explicativas, resulta em problema de *overfitting*. Assim, compreende-se que, no âmbito da área da saúde, é importante criar e implementar um protocolo de análise de dados, que viabilize fundamentar os resultados das pesquisas e permita a comparatividade entre elas de maneira mais justa.

No contexto das variáveis contínuas, a utilização de 10 Componentes Principais promoveu classificação razoável no banco *Cleveland*, entretanto nos demais bancos não apresenta melhora expressiva nas acurácias, nem nas outras métricas de desempenho (Sensibilidade, Especificidade e VPP). Mas técnicas como a ACP que reduzem a dimensionalidade da matriz explicativa sem perder informação significativa não devem ser desconsideradas, haja vista que o uso das Componentes como variáveis explicativas na classificação resultou em acurácias médias competitivas, chegando até à 91.03% de acurácia média.

Por fim, a ideia de unificar os bancos e incluir uma variável que categorize o local de coleta (`locality`) merece ser considerada, haja vista que o teste Qui-Quadrado resultou em rejeição da hipótese de independência entre a variável resposta Y e a `locality`.

Fica como sugestão para trabalhos futuros realizar a análise no contexto de classificação na base de dados unificada, considerando a variável que categoriza os locais de coleta, a fim de compreender melhor os fatores em comum que influenciam na classificação de cardiopatas. É interessante também, reproduzir a metodologia aqui descrita utilizando as bases de dados reduzidas disponíveis no Repositório de Aprendizado de Máquina da UC Irvine, as quais contém as 13 variáveis explicativas comumente abordadas na literatura.

Outra problemática que pode ser explorada é a presença de dados faltantes nos bancos, principalmente no *Long Beach* e *Switzerland*. Trabalhar com técnicas de "imputação" de valores no lugar dos NA's podem ser de muita valia na predição da variável resposta aqui explorada, considerando a presença da probabilidade de erro tanto na predição em si, quanto na "imputação" dos novos dados.

Em vista do desempenho competitivo das componentes principais na classificação dos cardiopatas, quando comparado com os da literatura, fica como sugestão aprofundar a pesquisa se utilizando de técnicas emergentes da área de classificação com base na redução de dimensionalidade da matriz explicativa.

De um ponto de vista mais amplo, a partir das dificuldade de comparação entre os

resultados das literaturas aqui abordadas, fica como sugestão de trabalho futuro, a construção de um protocolo de pesquisas na área de Saúde que norteie os cientistas de tal área, promovendo o uso das técnicas consolidadas na Estatística e permitindo reprodutibilidade dos resultados, bem como a comparatividade entre os estudos.

REFERÊNCIAS

AHA, DW.. **Heart Disease Dataset - Long Beach.** V.A. Medical Center, Long Beach, CA, 1988. Available from: https://archive.ics.uci.edu/ml/machine-learning-databases/he art-disease/long-beach-va.data

AKAIKE, H.. **A new look at the statistical model identification.** IEEE Transactions on Automatic Control 19 (6): 716-723, doi:10.1109/TAC.1974.1100705, MR 0423716, 1974.

ANTAL, M.; SZABÓ, L. Z.. **Some remarks on a set of information theory features used for on-line signature verification.** In: 2017 5th International Symposium on Digital Forensic and Security (ISDFS). IEEE, p. 1-5., 2017.

BALDI, P. et al.. **Assessing the accuracy of prediction algorithms for classification: an overview.** Bioinformatics, v. 16, n. 5, p. 412-424, 2000.

BELLMAN, R. E.. **Dynamic Programming.** Republished by Courier Dover Publications. ISBN 978-0-486-42809-3, 2003.

BISHOP, C. M.. **Pattern recognition and machine learning.** Springer, 2006.

BORLAND, L.; PLASTINO, A. R.; TSALLIS, C.. **Information gain within nonextensive thermostatistics.** Journal of Mathematical Physics, v. 39, n. 12, p. 6490-6501, 1998.

BREIMAN, L.. **Random forests.** Machine learning, v. 45, n. 1, p. 5-32, 2001.

CAMPELLO DE SOUZA, F. M..**O apoio ao diagnóstico médico: o que se pode fazer com um tensiômetro e um relógio.** Segunda Edição Revisada e Ampliada, Recife-PE, 2010.

CARVALHO, A. C. C.; SOUSA, J. M. A.. **Cardiopatia isquêmica.** Revista Brasileira de Hipertensão, v. 8, n. 3, p. 297-305, 2001.

CATTELL, R. B.. **The scree test for the number of factors.** Multivariate behavioral research,

v. 1, n. 2, p. 245-276, 1966.

CHAKRABORTY, S. et al.. **Rapid identification of oil-contaminated soils using visible near-infrared diffuse reflectance spectroscopy.** Journal of Environmental Quality, v. 39, n. 4, p. 1378-1387, 2010.

CHAMBERS, J. M.; HASTIE, T. J.. **Statistical Models in S.** Wadsworth & Brooks/Cole, 1992.

CHANG, C.; LIN, C.. **LIBSVM : a library for support vector machines.** ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.

CHARNOV, E. L.; BERRIGAN, D.. **Dimensionless numbers and the assembly rules for life histories.** Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, v. 332, n. 1262, p. 41-48, 1991.

CHEN, P. et al.. **Adaptive sparse graph learning based dimensionality reduction for classification.** Applied Soft Computing, Volume 82, September, 2019.

D'AGOSTINI, G.. **A multidimensional unfolding method based on Bayes' theorem.** Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, v. 362, n. 2-3, p. 487-498, 1995.

DA CUNHA, S. B.; CARVAJAL, S.. **Estatistica Basica - a Arte de Trabalhar com Dados.** Elsevier Brasil, 2009.

DAS, R.; TURKOGLU, I.; SENGUR, A.. **Effective diagnosis of heart disease through neural networks ensembles.** Expert Systems with Applications, v. 36, n. 4, p. 7675-7680, 2009.

DAVISON, A. C.; HINKLEY, D. V.. **Bootstrap methods and their application.** Cambridge University Press, 1997.

DETRANO, R. et al. **International application of a new probability algorithm for the diagnosis of coronary artery disease.** The American Journal of Cardiology, v. 64, n. 5, p. 304-310,

1989.

DETRANO, R. **Heart Disease Dataset - Cleveland.** Cleveland Clinic Foundation, Cleveland, 1988. Available from: https://archive.ics.uci.edu/ml/machine-learning-databas es/heart-disease/cleveland.data

DOBSON, A. J.. **An Introduction to Generalized Linear Models.** London: Chapman and Hall, 1990.

DOMINIC, V.; GUPTA, D.; KHARE, S.. **An Effective Performance Analysis of Machine Learning Techniques for Cardiovascular Disease.** Applied Medical Informatics, Original Research Vol. 36, No. 1 /2015, pp: 23-32.

DRAPER N.R.; SMITH H.. **Applied regression analysis**. Wiley, New York, 1981.

DUA, D; GRAFF, C.. **UCI Machine Learning Repository** [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2019.

ETAATI, L.. **Introduction to R. In: Machine Learning with Microsoft Technologies.** Apress, Berkeley, CA, 2019. p. 15-26.

FERRÉ, L.. **Selection of components in principal component analysis: a comparison of methods.** Computational Statistics & Data Analysis, v. 19, n. 6, p. 669-682, 1995.

FIRTH, D.. **Generalized linear models and Jeffreys priors: an iterative weighted least-squares approach.** In: Computational statistics. Physica, Heidelberg, 1992. p. 553-557.

FOX, J.. **Applied regression analysis, linear models, and related methods.** Sage Publications, Inc, 1997.

FOX, J.; WEISBERG, S.. **An R Companion to Applied Regression.** Third Edition, Sage Publications, 2018.

FRANZ, M. R.; CHIN, M. C.; WANG, D.; STERN, R.; SCHEINMAN, M. M.. **Monitoring of radiofrequency ablation effect by simultaneous monophasic action potential recording** (Abstr.). Pacing Clin Electrophysiol, v. 14, p. 703, 1991.

FREUND, Y.; SCHAPIRE, R. E.. **Experiments with a new boosting algorithm.** In Proceedings of the Thirteenth International Conference on Machine Learning, 1996. pg. 148-156.

FREUND, Y.; SCHAPIRE, R. E.. **A decision-theoretic generalization of online learning and an application to boosting.** Journal of Computer and System Sciences 55. 1997, pg. 119–139.

FREUND, Y.; SCHAPIRE, R.; ABE, N.. **A short introduction to boosting.** Journal-Japanese Society For Artificial Intelligence, v. 14, n. 771-780, p. 1612, 1999.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R.. **The elements of statistical learning.** New York: Springer Series in Statistics, 2001.

GLAROS, A. G.; KLINE, R. B.. **Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model.** Journal of Clinical Psychology, v. 44, n. 6, p. 1013-1023, 1988.

GÜNTHER, F.; FRITSCH, S.. **neuralnet: Training of neural networks.** The R journal, v. 2, n. 1, p. 30-38, 2010.

GUO, L. et al.. **Robust prediction of fault-proneness by random forests.** In: 15th International Symposium on Software Reliability Engineering. IEEE, 2004. p. 417-428.

GUYTON, A. C.; HALL, J. E.. **Tratado de fisiologia médica.** Elsevier Brasil, 2006.

HAIR, J. F. et al.. **Análise multivariada de dados.** Bookman Editora, 2009.

HAND, D. J. **Data Mining.** Encyclopedia of Environmetrics , v. 2, John Wiley & Sons, Ltd., 2006.

HOLLANDER M.; WOLFE D. A.. **Nonparametric Statistical Methods.** John Wiley & Sons, New York, 1999.

HOPPENSTEADT, F. C.; PESKIN, C. S.. **Mathematics in medicine and the life sciences.** Springer Science & Business Media, 2013.

HOSMER JR, D. W.; LEMESHOW, S.; STURDIVANT, R. X.. **Applied logistic regression.** John Wiley & Sons, 2013.

HUSSAIN, M. et al.. **A comparison of SVM kernel functions for breast cancer detection.** In: 2011 Eighth International Conference Computer Graphics, Imaging and Visualization. IEEE, 2011. p. 145-150.

IHME - Institute for Health Metrics and Evaluation. **Global Burden of Disease Collaborative Network.** Global Burden of Disease Study 2017 Results. Seattle, United States, 2017. Disponível em: http://ghdx.healthdata.org/gbd-results-tool.

JANOSI, A.. **Heart Disease Dataset - Hungarian.** Hungarian Institute of Cardiology, Budapest, 1988. Available from: https://archive.ics.uci.edu/ml/machine-learning-databases/ heart-disease/hungarian.data

JAN, M. Y.; HSIU, H.; HSU, T. L.; WANG, Y. Y. L.; WANG, W. K.. **The importance of pulsatile microcirculation in relation to hypertension.** IEEE Engineering in Medicine and Biology Magazine, v. 19, n. 3, p. 106-111, 2000.

JOLLIFFE, I. T.; CADIMA, J.. **Principal component analysis: a review and recent developments.** Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, v. 374, n. 2065, p. 20150202, 2016.

KAHRAMANLI, H.; ALLAHVERDI, N.. **Design of a hybrid system for the diabetes and heart diseases.** Expert Systems with Applications, v. 35, n. 1-2, p. 82-89, 2008.

KAUFMAN, L.; ROUSSEEUW, P. J.. **Finding groups in data: an introduction to cluster**

analysis. John Wiley & Sons, 2009.

KOGAN, B. J.. **Introduction to computational cardiology: mathematical modeling and computer simulation.** Springer Science & Business Media, 2009.

KOTSIANTIS, S. B.; ZAHARAKIS, I.; PINTELAS, P.. **Supervised machine learning: A review of classification techniques.** Emerging artificial intelligence applications in computer engineering, v. 160, p. 3-24, 2007.

KRISHNASREE, K.; RAO, MR N.. **Diagnosis of heart disease using neural networ-ks-comparative study of Bayesian regularization with multiple regression model.** Journal of Theoretical and Applied Information Technology, v. 88, n. 3, p. 638-643, 2016.

KUMARI, R.; SRIVASTAVA, S. Kr.. **Machine learning: A review on binary classification.** International Journal of Computer Applications, v. 160, n. 7, 2017.

KOUROU, K. et al. **Machine learning applications in cancer prognosis and prediction.** Computational and Structural Biotechnology Journal, v. 13, p. 8-17, 2015.

LEE, L. L.; BERGER, T.; AVICZER, E.. **Reliable on-line human signature verification systems.** IEEE Transactions on Pattern Analysis & Machine Intelligence, n. 6, p. 643-647, 1996.

MAGLOGIANNIS, I. G.. **Emerging artificial intelligence applications in computer engineering: real word ai systems with applications in ehealth, hci, information retrieval and pervasive technologies.** IOS Press, Amsterdam, Netherlands, 2007.

MAXIMIANO, J.. **Um olhar cronobiológico sobre o coração e a depressão: entre a biologia e a ritmicidade do diálogo tónico-emocional.** Psilogos: Revista do Serviço de Psiquiatria do Hospital Fernando Fonseca, p. 54-62, 2008.

MCQUEEN, D. M.; PESKIN, C. S.. **Heart simulation by an immersed boundary method with formal second-order accuracy and reduced numerical viscosity.** In: Mechanics for a New Mellennium. Springer, Dordrecht, 2001. p. 429-444.

MENARD, S.. **Applied Logistic Regression Analysis: Sage University Series.** Thousand Oaks, 1995.

MEYER, D. et al.. **e1071**: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, 2019.

MONTGOMERY D.C.; PECK E.A.. **Introduction to linear regression analysis**. Wiley, New York, 1982.

MORAES, R. M.; MACHADO, L. S.. **Assessment systems for training based on virtual reality: A comparison study.** SBC Journal on 3D Interactive Systems, v. 3, n. 1, p. 9-16, 2012.

MORAN, D.; EPSTEIN, Y.; KEREN, G.; LAOR, A., SHEREZ, J.; SHAPIRO, Y.. **Calculation of mean arterial pressure during exercise as a function of heart rate.** Applied Human Science, v. 14, n. 6, p. 293-295, 1995.

MORETTIN, P. A.; BUSSAB, W. O.. **Estatística básica.** Editora Saraiva, 2017.

NETER, J. et al.. **Applied linear statistical models.** Chicago: Irwin, 1996.

O'BRIEN, R. M.. **A caution regarding rules of thumb for variance inflation factors.** Quality & Quantity, v. 41, n. 5, p. 673-690, 2007.

PATEL, J.; TEJALUPADHYAY, D.; PATEL, S.. **Heart disease prediction using machine learning and data mining technique.** Heart Disease, v. 7, n. 1, p. 129-137, 2015.

POLANCZYK, C. A. et al.. **Fatores de risco cardiovascular no Brasil: os próximos 50 anos.** Arq Bras Cardiol, v. 84, n. 3, p. 199-201, 2005.

PUTRA, Y. P.; KHRISNE, D. C.; SUYADNYA, I. M. A.. **Expert System for Early Diagnosis of Heart Disease Using the Random Forest Method.** Journal of Electrical, Electronics and Informatics, v. 3, n. 1, p. 15-18, 2019.

R Core Development Team. **R: A Language and Environment for Statistical Computing.** R Foundation for Statistical Computing, Vienna, 2004.

RAHMAN, M. et al.. **A False Alarm Reduction Method for a Gas Sensor Based Electronic Nose.** Sensors, v. 17, n. 9, p. 2089, 2017.

REGO, L. C.; CAMPELLO DE SOUZA, F. M.. **Improved estimation of left ventricular hypertrophy.** IEEE Engineering in Medicine and Biology Magazine, v. 21, n. 1, p. 66-73, 2002.

RIPLEY, B. D.; HJORT, N. L.. **Pattern recognition and neural networks.** Cambridge University Press, 1996.

RISH, I. et al. **An empirical study of the naive Bayes classifier.** In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence. 2001. p. 41-46.

RITCHIE, H.; ROSER, M.. **Causes of Death - Our World in Data.** Publicado em Fevereiro de 2018 e atualizado em Abril de 2019. Disponível em: https://ourworldindata.org /causes-of-death. Acesso em 30 de Maio de 2019.

RONAN, C. A.. **História ilustrada da ciência da Universidade de Cambridge.** Jorge Zahar Editor Ltda, 1994.

ROSA, J. L. G.. **Fundamentos da inteligência artificial.** Editora LTC, 212 p., Rio de Janeiro, 2011.

ROYSTON P.. **Remark AS R94: A remark on Algorithm AS 181: The W test for normality.** Applied Statistics, 44, 547–551. doi: 10.2307/2986146, 1995.

SAFAVIAN, S. R.; LANDGREBE, D.. **A survey of decision tree classifier methodology.** IEEE Transactions on Systems, Man, and Cybernetics, v. 21, n. 3, p. 660-674, 1991.

SAYAD, A. T.; HALKARNIKAR, P. P.. **Diagnosis of heart disease using neural network.**

International Journal of Advances in Science Engineering and Technology, ISSN: 2321-9009 Volume- 2, Issue-3, July 2014.

SHANNON, C. E.. **A mathematical theory of communication.** Bell system technical journal, v. 27, n. 3, p. 379-423, 1948.

SHI, M.; RENTON, M.. **Modelling mortality of a stored grain insect pest with fumigation: Probit, logistic or Cauchy model?**. Mathematical Biosciences, v. 243, n. 2, p. 137-146, 2013.

SHAWE-TAYLOR, J. et al. **Kernel methods for pattern analysis.** Cambridge University Press, 2004.

SMITS, G. F.; JORDAAN, E. M.. **Improved SVM regression using mixtures of kernels.** In: Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290). IEEE, 2002. p. 2785-2790.

STEINBRUNN, W.. **Heart Disease Dataset - Switzerland.** University Hospital, Zurich, Switzerland, 1988. Available from: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/switzerland.data

STEINER, M. T. A. et al.. **Data Mining como Suporte à Tomada de Decisões-uma Aplicação no Diagnóstico Médico.** XXXVI Simpósio Brasileiro de Pesquisa Operacional, "O impacto da pesquisa operacional nas novas tendências multidisciplinares", v. 23, p. 96-107, 2004.

STOCKWELL, D. R. B.; PETERSON, A. T.. **Effects of sample size on accuracy of species distribution models.** Ecological Modelling, v. 148, n. 1, p. 1-13, 2002.

TU, M. C.; SHIN, D.; SHIN, D.. **Effective diagnosis of heart disease through bagging approach.** In: 2009 2nd International Conference on Biomedical Engineering and Informatics. IEEE, 2009. p. 1-4.

UMAMAHESWARI, K.; VALARMATHI, A.; JASMINE, J.. **Effective diagnosis of heart disease through stacking approach.** Advances in Natural and Applied Sciences, v. 11, n. 9, p.

323-329, 2017.

VAN DER AALST, W.. **Data science in action.** In: Process Mining. Springer, Berlin, Heidelberg, 2016. p. 3-23.

VENABLES, W. N.; RIPLEY, B. D.. **Modern Applied Statistics with S.** Fourth edition. Springer, 2002.

VIEIRA, C. E. C.; SOUZA, R. C.; RIBEIRO, C. C.. **Um estudo comparativo entre três geradores de números aleatórios.** PUC, RioInf.MCC16, 2004.

WANDERLEY, A. L.. **Sobre a dinâmica do Sistema Cardiovascular.** Dissertação de mestrado, Universidade Federal de Pernambuco. CTG. Engenharia Elétrica, 2005.

WIKIPÉDIA, a enciclopédia livre. **Esfigmomanômetro.** Flórida: Wikimedia Foundation, 2018. Disponível em: https://pt.wikipedia.org/wiki/Esfigmoman%C3%B4metro# /media/-File:Sphygmomanometer.jpg. Acesso em: 30 de Maio. 2019.

WHO - World Health Organization. **The top 10 causes of death**. 24 de Maio de 2018. Disponível em: http://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Acesso em: 24 set 2018.

YADAV, S.; SHUKLA, S.. **Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification.** In: 2016 IEEE 6th International Conference on Advanced Computing (IACC). IEEE, 2016. p. 78-83.

ZERBA, K. E.; COLLINS, J. P.. **Spatial heterogeneity and individual variation in diet of an aquatic top predator.** Ecology, v. 73, n. 1, p. 268-279, 1992.

ZHOU, S.; ZHANG, S.; KARYPIS, G.. **Advanced Data Mining and Applications: 8th International Conference, ADMA 2012, Nanjing, China, December 15-18, 2012, Proceedings.** Springer Science & Business Media, 2012.

ZYGMUNT Z. et al.. **'Rcpp' Implementation of 'FSelector' Entropy-Based Feature Selection Algorithms with a Sparse Matrix Support.** Version 0.3.1, UTC, 2019.

ANEXO A – DESCRIÇÃO DAS VARIÁVEIS DOS BANCOS DE DADOS

V1 id: número de identificação do paciente

V2 ccf: número do seguro social

V3 age: idade em anos

V4 sex: sexo (1 = masculino; 0 = feminino)

V5 painloc: localização da dor torácica (1 = subesternal; 0 = caso contrário)

V6 painexer: 1 = provocado pelo esforço; 0 = caso contrário

V7 relrest: 1 = aliviado após o repouso; 0 = caso contrário

V8 pncaden: soma de 5, 6 e 7

V9 cp: tipo de dor torácica

- Valor 1: angina típica
- Valor 2: angina atípica
- Valor 3: dor não anginosa
- Valor 4: assintomáticos

V10 trestbps: pressão arterial sistólica em repouso (em mmHg na admissão no hospital)

V11 htn: 1 = história de hipertensão; 0 = nenhuma história

V12 chol: soro colestório em mg / dl

V13 smoke: acredito que seja 1 = sim; 0 = não (é ou não é fumante)

V14 cigs: número de cigarros por dia

V15 years: número de anos como fumante

V16 fbs: glicemia de jejum> 120 mg / dl (1 = verdadeiro; 0 = falso)

V17 dm: 1 = história de diabetes; 0 = nenhuma história

V18 famhist: história familiar de doença arterial coronária (1 = sim; 0 = nenhum)

V19 restecg: resultados electrocardiográficas em repouso

- Valor 0: normal,
- Valor 1: ter ST-T onda de anormalidade (inversões das ondas T e / ou supradesnivelamento do segmento ST ou depressão> 0,05 mV)
- Valor 2: mostrando hipertrofia ventricular esquerda provável ou definitiva pelo critério de Estes

V20 ekgmo: mês de leitura do ECG durante o exercício

V21 ekias: dia da leitura do ECG durante o exercício

V22 ekgyr: ano de leitura de ECG de exercício

V23 dig: digitálicos utilizados durante o exercício ECG (1 = sim; 0 = não)

V24 prop: beta-bloqueador usado durante o exercício ECG (1 = sim; 0 = não)

V25 nitr: nitratos usados durante o exercício ECG (1 = sim; 0 = não)

V26 pro: bloqueador de canal de cálcio usado durante o exercício ECG (1 = sim; 0 = não)

V27 diuretic: diurético usado durante o exercício ECG (1 = sim; 0 = não)

V28 proto: protocolo de exercícios

1 = Bruce

2 = Kottus

3 = McHenry

4 = rápido Balke

5 = Balke

6 = Noughton

7 = bicicleta 150 kpa min / min

8 = bicicleta 125 kpa min / min

9 = bicicleta 100 kpa min / min

10 = bicicleta 75 kpa min / min

11 = bicicleta 50 kpa min / min

12 = braço ergômetro

V29 thaldur: duração do teste de exercício em minutos

V30 thaltime: tempo em que Foi observada depressão da medida de ST

V31 met: mets atingidos

V32 thalach: frequência cardíaca máxima atingida

V33 thalrest: freqüência cardíaca de repouso

V34 tpeakbps: pico de pressão arterial de exercício (primeira de 2 partes)

V35 tpeakbpd: pico de pressão arterial de exercício (segunda de 2 partes)

V36 dummy

V37 trestbpd: pressão arterial diastólica em repouso

V38 exang: angina induzida por exercício (1 = sim; 0 = não)

V39 xhypo: (1 = sim; 0 = não)

V40 oldpeak: depressão do segmento ST induzida pelo exercício em relação ao repouso

V41 slope: inclinação do segmento ST do exercício de pico

- Valor 1: elevação

- Valor 2: flat

- Valor 3 : descida

V42 rldv5: altura em repouso

V43 rldv5e: altura no pico do exercício

V44 ca: número de grandes vasos (0-3) colorido por flourosopia

V45 restckm

V46 exerckm

V47 restef: fração de ejeção raidonuclidea (repouso)

V48 restwm: movimento da parede (repouso), anormalidade de movimento

0 = nenhum

1 = leve ou moderado

2 = moderado ou grave

3 = acinesia ou dyskmem

V49 exeref: fração de ejeção raidonuclidea (exercício)

V50 exerwm: movimento da parede (exercício)

V51 thal: 3 = normal; 6 = defeito fixo; 7 = defeito reversível

V52 thalsev

V53 thalpul

V54 earlobe: lóbulo da orelha

V55 cmo: mês do cateterismo cardíaco

V56 cday: dia do cateterismo cardíaco

V57 cyr: ano de cateterismo cardíaco

V58 num: diagnóstico de cardiopatia (estado angiográfico)

- Valor 0: <50% de estreitamento do diâmetro

- Valor 1:> 50% de estreitamento do diâmetro

(em qualquer vaso principal: atributos V59 a V68 são medidas dos vasos)

V59 lmt

V60 ladprox

V61 laddist

V62 diag

V63 cxmain

V64 ramus

V65 om1

V66 om2

V67 rcaprox

V68 rcadist

V69 lvx1

V70 lvx2

V71 lvx3

V72 lvx4

V73 lvf

V74 cathef

V75 junk

V76 nome: sobrenome do paciente