LUMEN ET VIRTUS

Modelo de classificação para previsão de óbito por insuficiência cardíaca

https://doi.org/10.56238/levv15n39-075

Daniel Baldini Filipe
Especialista em Data Science e Analytics
Universidade de São Paulo

José Erasmo Silva
Doutor em Administração
Universidade Federal da Bahia

RESUMO
A insuficiência cardíaca é uma síndrome na qual o coração é incapaz de bombear o sangue em níveis ideais por todo o corpo. Nas últimas décadas houve um aumento expressivo de óbitos causados por esta condição. Associa-se este crescimento a mudanças de comportamento e ao envelhecimento, entre outros fatores. Diante do crescente desafio apresentado por este cenário, torna-se essencial desenvolver um modelo de classificação sofisticado que possa prever o óbito e que seja facilmente integrável aos sistemas de saúde existentes. À luz deste desafio, foram criados cinco modelos de classificação para uma amostra de 299 pacientes e seus resultados foram comparados. Esta amostra contém 12 variáveis explicativas além da variável resposta que indica se o paciente foi a óbito durante o tratamento. Os resultados atingidos foram satisfatórios e mostram que, embora mais de um modelo apresente bons resultados, considerando-se aspectos técnicos e analíticos, o modelo logístico binário é o que apresentou o melhor equilíbrio das métricas. Por ser um algoritmo de alto desempenho, fácil interpretação e de baixo custo de treinamento, disponibilização e execução, conclui-se que o modelo logístico binário criado neste estudo pode ser integrado aos sistemas atuais de modo a auxiliar os profissionais da saúde com seu diagnóstico e, com isso, estes podem sugerir alterações de hábitos e/ou tratamento ao paciente para que este tenha uma vida mais longa e saudável.

Palavras-chave: Saúde, Problema no Coração, Machine Learning, Análise de Dados, Balanceamento de Dados.

1 INTRODUÇÃO

Nos anos 1950 tiveram início as pesquisas para criação do que hoje conhecemos como “machine learning”. Em 1957, Frank Rosenblatt criou o primeiro “software” capaz de aprender à medida que recebesse dados ao longo do tempo (Taulli, 2020). Sua ideia tinha por base o neurônio humano e ficou conhecida como o “perceptron” de Rosenblatt. Porém, principalmente pelas limitações computacionais da época, só era possível criar redes neurais com um “perceptron”, o que fez com que este tema ficasse esquecido por algumas décadas.

Nos anos 1980, as ideias de Rosenblat voltaram à tona e isso trouxe o que alguns autores chamam de revolução da Inteligência Artificial [IA]. Com o poder computacional muito superior aos dos anos 1950, o conceito do “perceptron” de Rosenblatt foi expandido dando origem ao que atualmente chamamos de “deep learning”.

O uso do “machine learning” na área da saúde vem crescendo significativamente nos últimos anos. Durante a pandemia da COVID-19, estes algoritmos mostraram-se eficazes e importantes tanto para o desenvolvimento da vacina (Park, 2021) quanto para o gerenciamento de materiais e equipamentos hospitalares tão escassos naquele período.

A adoção de técnicas de “machine learning” para detecção de doenças cardiovasculares vem ganhando espaço, pois se estima que 620 milhões de pessoas vivam sob alguma enfermidade deste grupo e que ele seja responsável por aproximadamente 17,9 milhões de mortes por ano. Segundo estudo publicado pela British Heart Foundation, este número deverá aumentar nos próximos anos devido a fatores como: mudanças de comportamento, crescimento e envelhecimento da população mundial e crescimento da taxa de sobrevivência após ataques cardíacos e acidentes vasculares cerebrais (BHF, 2023; WHO, 2023).

Para que estas técnicas consigam prever e controlar de forma adequada o aparecimento ou a evolução de doenças cardiovasculares é crucial a identificação e o acompanhamento dos fatores de risco tais como hipertensão, colesterol, obesidade entre outros (Uddin et al., 2022). Sinais visíveis tais como o aumento abruto de peso, desenvolvimento de edemas, inchaço do tornozelo também devem ser considerados fatores relevantes e serem acompanhados para que as medidas necessárias sejam tomadas a tempo prevenindo assim a evolução para situações mais graves (Escolar et al., 2021).

Parte do grupo das doenças cardiovasculares, a insuficiência cardíaca é uma síndrome na qual o coração é incapaz de bombear o sangue de forma a suprir todas as necessidades metabólicas dos tecidos com a quantidade adequada de sangue (Rohde et al., 2018). Segundo Ferreira (2020), pode ainda gerar inchaços ou congestões devido ao acúmulo de sangue em uma região ocasionado pela redução do seu fluxo.

De acordo com Arruda et al. (2022), o período entre 1998 e 2019 no Brasil foi marcado por um significativo número de óbitos relacionados à insuficiência cardíaca, especialmente em adultos com

mais de 50 anos. Durante esses anos, o total de mortes atribuídas a esta condição chegou a 567.789. Este dado ressalta a crescente preocupação com a insuficiência cardíaca como uma causa importante de mortalidade na população mais velha do país. A tendência observada ao longo dessas duas décadas sugere a necessidade de uma atenção maior à saúde cardíaca e ao manejo de doenças crônicas, especialmente em idosos, para combater essa causa crescente de mortalidade.

Diante da crescente preocupação com as doenças cardíacas no Brasil, especialmente motivada pelo envelhecimento da população, surge a necessidade de explorar métodos de pesquisa que ofereçam informações sobre os fatores de risco associados a eventos cardíacos. Neste contexto, os modelos de regressão logística binária e multinomial são utilizados por pesquisadores que tenham por interesse avaliar a probabilidade de ocorrência de infarto em um grupo de pessoas baseado em características físicas tais como: peso, cintura abdominal e também em seus hábitos de saúde tais como: hábitos alimentares, prática de exercícios físicos, tabagismo entre outros. Para este estudo o infarto é a variável a ser prevista e o evento pode ou não ocorrer em função das variáveis explicativas providas a ele. Desta forma o infarto trata-se de uma variável qualitativa dicotômica e como o intuito é estimar a probabilidade de ocorrência deste evento a regressão logística binária pode ser utilizada (Fávero e Belfiore, 2017).

Portanto, diante do crescente desafio representado pela insuficiência cardíaca, torna-se essencial desenvolver um modelo de classificação sofisticado, utilizando técnicas avançadas de "machine learning". Este modelo deve ser capaz de analisar e interpretar uma variedade de dados clínicos e comportamentais - semelhantes aos utilizados no exemplo de regressão logística binária de Fávero et al. (2017) - para prever com precisão a probabilidade de óbito causado por insuficiência cardíaca.

Além disso, é crucial que este modelo seja projetado para se integrar de forma eficiente aos sistemas de saúde existentes, permitindo que profissionais da área utilizem suas previsões para tomar decisões clínicas mais informadas. Ao fazer isso, o modelo não só contribuirá para a detecção precoce e o tratamento mais eficaz da insuficiência cardíaca, mas também ajudará a prolongar e melhorar a qualidade de vida dos pacientes. A tendência de aumento das mortes por insuficiência cardíaca, como destacado por Arruda et al. (2022), enfatiza a urgência de tal inovação. Assim, o objetivo deste trabalho é desenvolver um modelo de classificação que não apenas avance o campo do machine learning na saúde, mas que também tenha um impacto significativo e prático no manejo e tratamento de doenças cardiovasculares.

2 MATERIAL E MÉTODOS

A base de dados utilizada para este estudo foi obtida no seguinte endereço: *https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records.* Ela contém os registros médicos de 299 pacientes e possui as seguintes variáveis: Idade (“age”), sexo (“sex”), se o paciente é fumante (“smoking”), se possui diabetes (“diabetes”), anemia (“anaemia”) e pressão alta (“high_blood_pressure”). Além destes dados mais conhecidos, ela traz também: a quantidade de plaquetas (“platelets”) que são as células do sangue responsáveis pela coagulação do sangue; A fração de ejeção (“ejection_fraction”) que é a porcentagem de sangue ejetada pelo coração a cada batimento; A creatinina sérica (“serum_creatinine”) que se trata de um resíduo químico produzido pela morte de células musculares; A creatinofosfoquinase (“creatinine_phosphokinase”) que é uma enzima presente nos tecidos musculares, no coração e no cérebro; O sódio sérico (“serum_sodium”) que traz a quantidade de sódio no sangue do paciente e, por fim, a variável tempo (“time”) que traz a quantidade de dias que o paciente foi acompanhado pelo médico. A última informação é a variável resposta, “death_event”, que determina se o paciente foi a óbito durante o tratamento.

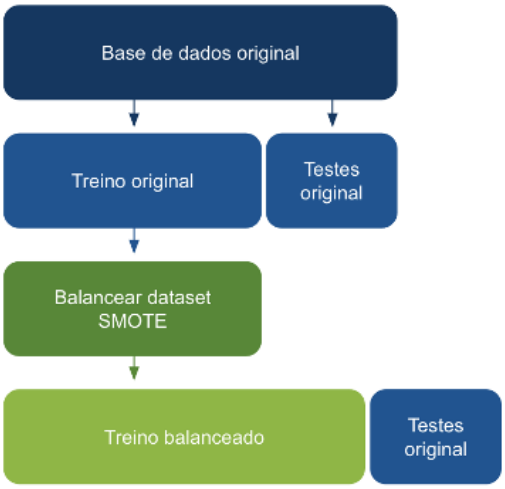
Não continha qualquer valor inexistente nos dados e também não foi necessário nenhum tratamento ou alteração das informações. As variáveis “anaemia”, “diabetes”, “high_blood_pressure”, “sex”, “smoking” e “death_event” foram convertidas do tipo inteiro para o tipo fator e a variável “age” foi convertida para inteiro.

Além de ajustar os modelos com a base de dados original, este estudo também utilizou-se da técnica "Synthetic Minority Oversampling Technique” [SMOTE] para adicionar observações à categoria minoritária na base de treino e ajustar os modelos a partir do novo conjunto de dados.

O SMOTE é considerado o mais proeminente método para contornar a frequente questão de dados desbalanceados, pois ao invés de replicar dados já existentes na base, ele usa interpolação linear dos dados da classe minoritária e técnicas como a do “K nearest neighbors” para criar exemplos sintéticos da classe minoritária (Elreedy et al., 2022) o que pode ajudar na generalização do modelo e também evitar seu “overfitting”.

A base de testes não foi balanceada em nenhum cenário permanecendo, portanto somente observações existentes na base de dados original como exibido na Figura 1.

Figura 1. Fluxo de preparação dos dados para modelos de base balanceada



Fonte: Dados originais da pesquisa

Em busca do modelo mais adequado para previsão da variável resposta foi feita a implementação, previsão e análise da resposta dos seguintes modelos: Regressão logística binária, máquina de vetores e suporte, árvore de decisão, “gradient boosting” e “naive Bayes”.

Para facilitar a reprodução dos resultados, o valor 2024 foi utilizado como semente em todas as etapas deste estudo. Determinar uma semente trata-se de uma técnica utilizada para buscar garantir que os resultados obtidos sejam sempre iguais, pois esta semente assegura que todas as execuções possuam a mesma aleatoriedade.

3 REGRESSÃO LOGÍSTICA

O modelo logístico foi desenvolvido no fim do século XIX, porém tornou-se popular a partir da segunda década do século XX. Esta técnica estatística visa prever a probabilidade de um determinado conjunto de dados pertencer a uma categoria (Cramer, 2002).

Além de amplamente utilizada na área da saúde, ela também mostra-se eficaz para previsões nas áreas financeiras, de seguros, entre outras para definir, por exemplo, se o crédito deve ou não ser concedido, ou ainda qual é o nível de risco de determinado cliente sofrer um acidente (Fávero e Belfiore, 2017).

Neste estudo foi utilizado o procedimento “stepwise”, pois ele mantém somente as variáveis preditoras necessárias para o nível de significância escolhido.

As principais vantagens deste modelo são: requerer baixo desempenho computacional e ser de fácil implementação e interpretação. Suas principais limitações são relacionadas a multicolinearidade e/ou heterocedasticidade nas variáveis explicativas.

4 MÁQUINAS DE VETORES DE SUPORTE

No início dos anos 1990, como alternativa às redes neurais artificiais para tarefas de classificação e regressão não lineares, a técnica de máquinas de vetores de suportes foi desenvolvida e ganhou relevância por possuir maior capacidade de generalização mesmo sendo estimada a partir de uma base de dados sem muitos registros (Gholami e Fakhari, 2017).

Para estimativa deste modelo foi utilizada a padronização das variáveis utilizando-se da técnica do "Z-Score": eq. (1). Esta técnica faz com que as variáveis padronizadas apresentem média 0 e variância 1 utilizando-se da form.(1) (Fávero e Belfiore, 2017).

$$Z = \frac{X - \mu}{\sigma} \qquad (1)$$

em que, Z: é o índice; X: é o valor que precisa ser convertido; μ: é a média de todos os valores de X; σ: é o desvio padrão da amostra.

Além do Z-Score, este modelo também fez uso da validação cruzada, ou "cross-validation". Segundo Guerbai et al. (2022) esta técnica consiste em separar os dados em "n" subconjuntos de mesmo tamanho. O algoritmo então seleciona n-1 partes para estimação e a parte não considerada é utilizada para testes. Esta operação é repetida até que todas as partes tenham sido usadas tanto para estimação quanto para testes. A acurácia do modelo é calculada o final deste processo iterativo.

As principais vantagens das máquinas de vetores de suporte são: baixa influência por ruídos nos dados, ser capaz de classificar problemas não lineares e ser efetivo em grandes volumes de dados. As principais desvantagens estão na dificuldade de interpretação e visualização do hiperplano, ser um algoritmo que requer maior poder computacional para ser executado, em comparação com a regressão logística e a possibilidade de "overfitting*"* devido à escolha de hiperparâmetros (Boswell, 2002).

5 ÁRVORES DE DECISÃO

Trata-se de um algoritmo poderoso e versátil capaz de fazer a previsão de dados categóricos e contínuos. A partir de testes sequenciais ele compara um atributo a um limite adicionando a observação à classe mais frequente. Por ser o componente fundamental das "random forests", é possível dizer que se trata de um dos algoritmos mais importantes atualmente (Géron, 2022; Kotsiantis, 2011).

Quando os dados não possuem muitas variáveis e/ou não são necessárias muitas ramificações na árvore para chegar à previsão, trata-se de um modelo de fácil implementação, interpretação e de rápida execução, caso contrário este modelo torna-se de difícil interpretação, a velocidade de execução é afetada e também pode incorrer em "overfitting" (Sosnovshchenko et al., 2018).

5.1 GRADIENT BOOSTING

Trata-se de uma técnica avançada para modelagem preditiva em dados estruturados, empregando a estratégia de "boosting de gradiente com modelos lineares componente a componente" (glmboost), uma abordagem que se distingue pelo seu foco na construção sequencial e incremental de modelos. Ao contrário do "gradient boosting" tradicional, que frequentemente recorre a árvores de decisão como modelos base para reduzir resíduos e melhorar a acurácia usando de iterações, o glmboost otimiza a função de perda em um contexto de modelos lineares generalizados (GLMs), ajustando-se variável por variável, o que facilita uma maior interpretabilidade dos ajustes internos do modelo. Esta metodologia não apenas permite uma análise detalhada da contribuição de cada variável explicativa, mas também melhora a precisão das predições de maneira mais transparente, contrastando com as limitações de visibilidade dos ajustes internos característicos de algoritmos considerados "caixa-preta" (Hothorn et al. 2023; Saupin, 2022).

Entre as principais vantagens do algoritmo glmboost, destaca-se sua capacidade de oferecer modelos preditivos robustos sem exigir alto desempenho computacional, além de ter implementação intuitiva. Isso é particularmente útil em campos que demandam um alto grau de explicabilidade dos modelos, como em aplicações científicas e de pesquisa. Contudo, é importante reconhecer que, assim como outras técnicas de modelagem, o glmboost não está isento de desafios, como a necessidade de gerenciar multicolinearidade e heterocedasticidade nas variáveis explicativas. No entanto, sua abordagem focada em modelos componente a componente oferece uma via mais transparente para o entendimento e a interpretação dos ajustes do modelo, mitigando algumas das preocupações comuns associadas a métodos de "machine learning" tradicionais (Hothorn et al. 2023; Sosnovshchenko et al., 2018).

5.2 NAIVE BAYES

Desenvolvido com base no teorema criado por Thomas Bayes (1702 – 1761), o “naive Bayes” é um classificador que gera uma tabela de probabilidades para o fenômeno em questão a partir da seguinte equação: eq (2).

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \qquad (2)$$

em que, P: é a probabilidade; P(A|B): é a probabilidade do evento B dado o evento A; P(A): é a probabilidade do evento A; P(B) é a probabilidade do evento B.

Este algoritmo vem sendo utilizado para classificação em diversos cenários, tais como a classificação de e-mails como “SPAM”, recomendação de produtos, diagnósticos na área da saúde e até mesmo para controlar veículos autônomos (Wickramasinghe e Kalutarage, 2020).

Entre as vantagens deste algoritmo estão sua velocidade de processamento, a boa predição para problemas que possuam mais de duas categorias possíveis e sua capacidade de previsão quando as variáveis preditoras são categóricas. Geralmente, por entender que as variáveis preditoras não guardam relações entre si, isso pode ser considerado uma desvantagem já que em problemas reais isso raramente ocorre. Além disso, quando determinada categoria não está na base de dados de treinamento, o algoritmo recai sobre um problema conhecido como "zero-frequency problem" (Bruce et al., 2023).

6 AJUSTE, AVALIAÇÃO E ESCOLHA DO MODELO

Para o ajuste dos modelos, a base de dados original foi dividida em duas partes ficando 80% das observações para treino e 20% para testes.

Para a avaliação de desempenho de modelos de classificação, cinco indicadores derivados da matriz de confusão são comumente utilizados. São eles: Acurácia, sensibilidade, especificidade, precisão e AUC.

A matriz de confusão é uma ferramenta de visualização de dados que exibe quantas previsões foram feitas correta e incorretamente pelo modelo em relação aos dados originais.

A Tabela 1 exibe uma matriz de confusão genérica e nela visualiza-se as seguintes informações: Verdadeiros Negativos [VN] – quantidade de observações com o valor real e previsto iguais a 0. Verdadeiros Positivo [VP] – quantidade de observações com o valor real e previsto iguais a 1. Falsos Positivo [FP] – quantidade de observações em que o valor real é 0 e o previsto foi 1. Falsos Negativos [FN] - quantidade de observações em que o valor real é 1 e o previsto foi 0.

Tabela 1. Matriz de confusão genérica

Predito	Resultados	
	Referência 0	Referência 1
Predito 0	VN	FN
Predito 1	FP	VP

Fonte: Dados originais da pesquisa

É importante dizer que os valores exibidos na matriz de confusão alteram-se dependendo do ponto de corte, ou "cutoff", escolhido pelo pesquisador. Como os modelos de classificação preveem um valor entre 0 e 1, é necessário que o pesquisador determine até qual valor, ou ponto de corte, a previsão deve ser considerada como "não evento", ou 0, para que acima disso seja considerado como "evento", ou 1. Para cada um dos modelos estimados neste estudo, foi calculada a acurácia para todos os de corte entre 0,10 e 0,90. O valor escolhido para cada modelo foi aquele que apresentou a maior acurácia das previsões na base de dados de testes.

A acurácia é a percentagem de acertos do modelo em relação ao total de previsões. Este resultado é obtido com o uso da eq. (3) e retorna um valor entre 0 e 1.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \quad (3)$$

O valor obtido representa o percentual de acertos do modelo em relação ao total de casos. Por exemplo, caso o resultado tenha sido de 0,8456, sabe-se que o modelo previu corretamente 84,56% do total de casos.

A sensibilidade é a proporção de acertos do modelo em relação ao total de casos positivos. O resultado é obtido por meio da eq. 4 e retorna um valor entre 0 e 1.

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \quad (4)$$

Um resultado de 0,7615 indica que o modelo previu corretamente 76,15% dos casos positivos presentes na amostra ou, em outras palavras, o modelo previu como negativo 23,85% dos casos positivos.

A especificidade é definida como a proporção de acertos do modelo em relação ao total de casos negativos. Este resultado é obtido utilizando a eq. (5) e retorna um valor entre 0 e 1.

$$\text{Especificidade} = \frac{VN}{VN + FP} \quad (5)$$

Similar à sensibilidade, se o resultado obtido foi de 0,9150 pode-se dizer que o modelo previu corretamente 91,50% dos casos negativos presentes na amostra ou, em outras palavras, o modelo previu como positivo 8,50% dos casos negativos.

A precisão é caracterizada pela proporção de previsões corretamente positivas dentro daquelas previstas como positivas pelo modelo. O resultado é obtido utilizando a eq. (6) e retorna um valor entre 0 e 1.

$$\text{Precisão} = \frac{VP}{VP + FP} \quad (6)$$

Caso o resultado seja 0,9 pode-se concluir que 90% dos valores previstos como positivos eram de fato positivos ou, em outras palavras, 10% do que o modelo previu como positivo era negativo.

Por fim apresenta-se o AUC, ou a “Area under ROC curve” que, segundo Šimundić (2009), trata-se de uma métrica global do diagnóstico de acurácia. Este índice, que utiliza como parâmetros a sensibilidade e a especificidade obtidas a partir das previsões do modelo, traz um valor entre 0 e 1 no qual qualquer valor inferior a 0,5 demonstra que o modelo criado não é capaz de categorizar em um nível aceitável as observações, podendo então o modelo ser considerado como não utilizável. Ainda

segundo a autora, estas métricas globais existem para comparar dois ou mais diagnósticos para avaliar qual é o mais adequado, o que é um dos desafios deste estudo.

Como suporte para a escolha do modelo mais adequado foi utilizado o método "Analytic Hierarchy Process" [AHP]. Segundo Saaty (1977), por meio do AHP, os fatores importantes para a tomada de decisão são organizados para facilitar a visualização de seus relacionamentos e auxiliar o tomador de decisão a comparar os elementos de forma mais assertiva através da exibição dos dados em mesma magnitude.

Como resultado do processo, o AHP atribui um índice de preferência global a cada uma das alternativas analisadas. A alternativa que contém o maior índice deve ser considerada como a preferida conforme os critérios fornecidos ao método. Porém, ele não deve ser entendido como fator absoluto para a tomada de decisão e sim como um índice adicional a ser levado em consideração.

Este estudo utilizou-se da linguagem R para carregar, ajustar e analisar os dados bem como para gerar as previsões de óbito por insuficiência cardíaca utilizando diferentes modelos de classificação binária. Na Tabela 2 encontram-se os pacotes utilizados.

Tabela 2. Lista de pacotes utilizados neste estudo

Pacote	Descrição	Versão
readr	Ler arquivo csv	2.1.4
jtools	Obter detalhes sobre modelo de regressão	2.2.2
dply_r	Função "if_else"	1.1.4
caret	Criar matriz de confusão	6.0-94
pROC	Visualizar a curva ROC	1.18.5
e1071	Estimar máquina de vetores e suporte	1.7-14
rpart	Estimar árvore de decisão	4.1.21
rpart.plot	Visualizar árvore de decisão	3.1.1
mboost	Estimar modelo "gradient boosting"	2.9-9
naivebayes	Estimar modelo "naive Bayes"	0.9.7
unbalanced	Balancear base de dados	2.1

Fonte: Resultados originais da pesquisa

7 RESULTADOS E DISCUSSÃO

Durante a análise dos dados por meio de suas estatísticas descritivas, foi notada significativa diferença da ordem de grandeza entre algumas variáveis, bem como expressivas diferenças entre seus valores mínimos e máximos, como mostra a Tabela 3.

As diferenças de grandeza foram observadas, por exemplo, entre as variáveis "age" e "platelets". A variável "platelets" pode ser também utilizada como referência de análise da amplitude dos dados.

Na Tabela 4 é possível identificar que a variável dependente "DEATH_EVENT" é desbalanceada, pois contém apenas 32% de registros iguais a 1. Esta distribuição pode prejudicar o

ajuste de alguns modelos, pois este recebe mais informações relacionadas a uma categoria específica o que poderá gerar viés nas previsões.

Tabela 3. Estatísticas básicas das variáveis discretas e contínuas da base de dados

Variáveis	min	q1	mediana	média	q3	max
age	40,0	51,0	60,0	60,829	70,0	95,0
creatinine_ phosphokinase	23,0	116,5	250,0	581,839	582,0	7861,0
ejection_ fraction	14,0	30,0	38,0	38,084	45,0	80,0
platelets	25100,0	212500,0	262000,0	263358,03	303500,0	850000,0
serum_ creatinine	0,5	0,9	1,1	1,394	1,4	9,4
serum_ sodium	113,0	134,0	137,0	136,625	140,0	148,0
time	4,0	73,0	115,0	130,261	203,0	285,0

Fonte: Resultados originais da pesquisa

Tabela 4. Frequência das variáveis categóricas da base de dados.

Variáveis	0	1
anaemia	170	129
diabetes	174	125
high_blood_pressure	194	105
sex	105	194
smoking	203	96
DEATH_EVENT	203	96

Fonte: Resultados originais da pesquisa

Nesta seção são apresentadas as acurácias obtidas em cada um dos modelos e em cada divisão da base de dados. Por tratar-se de cinco modelos distintos executados tanto na base original quanto na base balanceada, é importante que sejam apresentados os dados de cada um individualmente e em seguida que seja feita a análise e contraposição das informações a fim de identificar qual modelo parece ser o mais adequado e quais são suas limitações no escopo deste estudo.

8 MODELO LOGÍSTICO BINÁRIO

Após executado o procedimento “stepwise”, identificou-se que somente as variáveis “age”, "ejection_fraction", "serum_creatinine" e "time" são relevantes no intervalo de confiança de 95% como pode ser verificado na Tabela 5. Os coeficientes apresentados na tabela abaixo serão discutidos na próxima seção deste estudo.

Tabela 5. Tabela de resposta do procedimento “stepwise”

Variável	Estimadores	Erro padrão	2.5%	97.5%	z val.	p
(Intercept)	0.8400	1.1465	-1.4071	3.0872	0.7327	0.4637
age	0.0372	0.0163	0.0052	0.0693	2.2754	0.0229
ejection_fraction	-0.0687	0.0162	-0.1004	-0.0369	-4.2381	0.0000
serum_creatinine	0.7171	0.1898	0.3450	1.0891	3.7779	0.0002
time	-0.0205	0.0031	-0.0266	-0.0144	-6.5851	0.0000

Fonte: Resultados originais da pesquisa

Na Tabela 6 é verificado que na base de dados de treino original, que contém 80% das observações, o modelo conseguiu uma acurácia de 82,4%. Na base de treino balanceada é possível observar que a acurácia foi acrescida em 2,1% chegando em 84,6%

Tabela 6. Matriz de confusão de treino do modelo logístico binário

Predito	Resultados			
	Original		Balanceado	
	Referência 0	Referência 1	Referência 0	Referência 1
	----------%----------			
Predito 0	143	15	150	12
Predito 1	27	54	38	124

Fonte: Resultados originais da pesquisa
Nota: Na base Original o ponto de corte foi de 0,55 e na base Balanceada foi de 0,65

A base de dados de testes continha 20% das observações da base de dados original. A predição do modelo estimado a partir da base de dados original teve uma acurácia de 91,7%. Já a predição do modelo estimado a partir da base balanceada teve acurácia de 90%, fato que pode ser verificado na Tabela 7.

Tabela 7. Matriz de confusão de testes do modelo logístico binário

Predito	Resultados			
	Original		Balanceado	
	Referência 0	Referência 1	Referência 0	Referência 1
	----------%----------			
Predito 0	43	2	41	4
Predito 1	3	12	2	13

Fonte: Resultados originais da pesquisa
Nota: Na base Original o ponto de corte foi de 0,55 e na base Balanceada foi de 0,65

9 MÁQUINA DE VETORES E SUPORTE

Mesmo este algoritmo sendo capaz de fazer “cross-validation” dos dados, foram executados dois cenários.

Na avaliação do modelo, adotou-se a técnica de validação cruzada com 5 subconjuntos, assegurando que cada segmento do conjunto de dados seja utilizado para teste uma vez, com os

segmentos restantes empregados no treinamento. Os hiperparâmetros estabelecidos para o modelo incluíram um kernel do tipo “polynomial”, um valor de regularização (C) de 3, e “gamma” definido como 0.1. Além disso, optou-se por não padronizar os dados (“scale” definido como “FALSE”) pois esta estratégia não apresentou qualquer melhora no desempenho do modelo. Com estas configurações, observou-se uma acurácia de 44,8% utilizando os dados em sua forma original, que aumentou para 47,4% após o processo de balanceamento das classes, evidenciando uma melhoria no desempenho do modelo com dados balanceados.

Foi feita uma segunda tentativa neste mesmo cenário na qual o hiperparâmetro “scale” foi alterado para “TRUE”. Com este parâmetro o algoritmo padroniza os valores das variáveis utilizando-se da transformação “Z-Score” e a acurácia verificada para o modelo com os dados originais foi de 91,3% e de 93,2% com os dados balanceados como pode ser verificado com base na Tabela 8.

Tabela 8. Matriz de confusão de máquina de vetores e suporte

Predito	Resultados			
	Original		Balanceado	
	Referência 0	Referência 1	Referência 0	Referência 1
	----------%----------			
Predito 0	201	24	183	17
Predito 1	2	72	9	175

Fonte: Resultados originais da pesquisa
Nota: Na base Original o ponto de corte foi de 0,45 e na base Balanceada foi de 0,57

No segundo cenário a base de dados foi separada em treinamento e testes. Utilizando-se dos mesmos parâmetros da segunda tentativa, na base de treino original o modelo obteve acurácia de 91,9%. Como pode ser verificado por meio da Tabela 9, o modelo com a base de treino balanceada foi de 95,1%.

Tabela 9. Matriz de confusão de treino da máquina de vetores e suporte

Predito	Resultados			
	Original		Balanceado	
	Referência 0	Referência 1	Referência 0	Referência 1
	----------%----------			
Predito 0	156	15	155	9
Predito 1	2	66	7	153

Fonte: Resultados originais da pesquisa
Nota: Na base Original o ponto de corte foi de 0,45 e na base Balanceada foi de 0,57

A base de dados de testes continha 20% das observações da base de dados original. A predição do modelo estimado a partir da base de dados original teve acurácia de 83,3%. Já a predição do modelo estimado a partir da base balanceada teve acurácia de 85% conforme matriz de confusão da Tabela 10.

Tabela 10. Matriz de confusão de testes da máquina de vetores e suporte

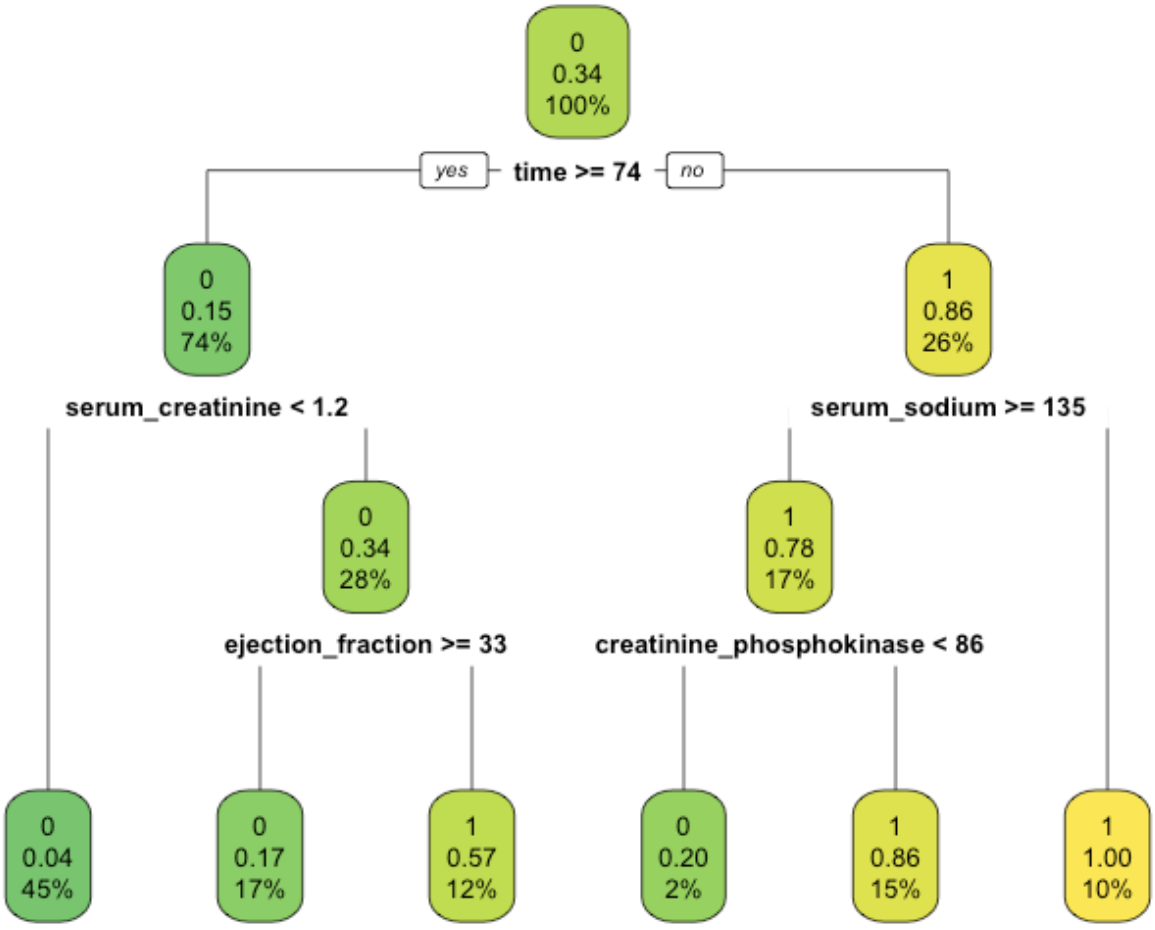
Predito	Resultados			
	Original		Balanceado	
	Referência 0	Referência 1	Referência 0	Referência 1
	----------%----------			
Predito 0	41	6	38	2
Predito 1	4	9	7	13

Fonte: Resultados originais da pesquisa
Nota: Na base Original o ponto de corte foi de 0,45 e na base Balanceada foi de 0,57

10 ÁRVORE DE DECISÃO

Para a execução deste algoritmo foi utilizada a base de dados original com todas as variáveis preditoras. Na Figura 2 é possível identificar que não são necessários muitos galhos para que a árvore chegue ao resultado. Além disso, é possível notar que as variáveis “time”, “serum_creatinine”, “serum_sodium”, “creatinine_phosphokinase” e “ejection_fraction” foram consideradas para chegar ao o resultado.

Figura 2. Árvore de decisão do modelo



Fonte: Resultados originais da pesquisa

A base de dados de treino que continha 80% dos dados da base original teve a acurácia de 86,2%. Já o modelo com a base de treino balanceada apresentou acurácia de 89,8% conforme dados exibidos na Tabela 11.

Tabela 11. Matriz de confusão de treino da árvore de decisão

Predito	Resultados			
	Original		Balanceado	
	Referência 0	Referência 1	Referência 0	Referência 1
	----------%----------			
Predito 0	153	28	151	22
Predito 1	5	53	11	140

Fonte: Resultados originais da pesquisa
Nota: O ponto de corte foi de 0,60 em ambas as bases

A base de dados de testes continha 20% das observações da base de dados original. A predição do modelo estimado a partir da base de dados original teve acurácia de 83,3%. Já a predição do modelo estimado a partir da base balanceada teve acurácia de 81,7% conforme a Tabela 12.

Tabela 12. Matriz de confusão de testes da árvore de decisão

Predito	Resultados			
	Original		Balanceado	
	Referência 0	Referência 1	Referência 0	Referência 1
	----------%----------			
Predito 0	41	6	39	5
Predito 1	4	9	6	10

Fonte: Resultados originais da pesquisa
Nota: O ponto de corte foi de 0,60 em ambas as bases

11 GRADIENT BOOSTING

Para este algoritmo foram criados dois modelos para avaliar qual traria maior eficácia na predição do óbito. O primeiro modelo foi executado levando em consideração apenas as variáveis que o procedimento “stepwise” do algoritmo de regressão logística apontou como relevantes, ou seja, com as variáveis “age”, "ejection_fraction", "serum_creatinine" e "time". Embora seja possível usar a seleção de variáveis de um modelo para informar outro, é importante notar que as premissas e características de cada modelo são diferentes. Portanto, as variáveis importantes para um modelo de regressão logística podem não ser as mais adequadas para um modelo de “gradient boosting”. O segundo modelo foi criado utilizando-se de todas as variáveis presentes na base de dados, porém este apenas considerou as variáveis apresentadas na Tabela 13 como relevantes.

Tabela 13. Variáveis do modelo e seus respectivos coeficientes.

Variável	Valor
(Intercept)	3.939965e+00
age	1.065228e-02
creatinine_phosphokinase	4.992803e-05
ejection_fraction	-2.332532e-02
serum_creatinine	2.348574e-01
serum_sodium	-2.277677e-02
time	-8.056760e-03

Fonte: Resultados originais da pesquisa

O modelo estimado a partir da base de dados original obteve a maior acurácia sendo ela de 82,8% na divisão de treino. Já o modelo estimado a partir da base de dados balanceada apresentou acurácia, levemente menor do que o modelo não balanceado, sendo ela de 82,1%, como pode ser verificado na Tabela 14.

Tabela 14. Matriz de confusão de treino “gradient boosting”

Predito	Resultados			
	Original		Balanceado	
	Referência 0	Referência 1	Referência 0	Referência 1
	----------%----------			
Predito 0	142	16	151	11
Predito 1	25	56	47	115

Fonte: Resultados originais da pesquisa
Nota: Na base Original o ponto de corte foi de 0,51 e na base Balanceada foi de 0,34

Conforme pode ser verificado na Tabela 15, a previsão do modelo não balanceado apresentou acurácia de 90% enquanto a estimativa do modelo balanceado apresentou um pequeno incremento sendo de 91,6%.

Tabela 15. Matriz de confusão de testes “gradient boosting”

Predito	Resultados			
	Original		Balanceado	
	Referência 0	Referência 1	Referência 0	Referência 1
	----------%----------			
Predito 0	42	3	42	3
Predito 1	3	12	2	13

Fonte: Resultados originais da pesquisa
Nota: Na base Original o ponto de corte foi de 0,51 e na base Balanceada foi de 0,34

12 NAIVE BAYES

Como este modelo desconsidera qualquer correlação entre as variáveis, todas foram utilizadas para sua criação. Na Tabela 16 é verificado que na base de dados de treino original, que contém 80%

das observações, o modelo conseguiu uma acurácia de 26,8%. Na base de treino balanceada é possível verificar que a acurácia foi acrescida em 3,2% chegando em 29,9%.

Tabela 16. Matriz de confusão de treino “naive Bayes”

Predito	Resultados			
	Original		Balanceado	
	Referência 0	Referência 1	Referência 0	Referência 1
	----------%----------			
Predito 0	6	23	7	72
Predito 1	152	58	155	90

Fonte: Resultados originais da pesquisa
Nota: O ponto de corte foi de 0,20 em ambas as bases

A base de dados de testes continha 20% das observações somente da base de dados original. A predição do modelo estimado a partir da base de dados original teve uma acurácia de 26,7% e a partir da base de dados balanceada a acurácia foi de 23,3%, conforme pode ser verificado pela Tabela 17.

Tabela 17. Matriz de confusão de testes “naive Bayes”

Predito	Resultados			
	Original		Balanceado	
	Referência 0	Referência 1	Referência 0	Referência 1
	----------%----------			
Predito 0	4	3	5	6
Predito 1	41	12	40	9

Fonte: Resultados originais da pesquisa
Nota: O ponto de corte foi de 0,20 em ambas as bases

13 RESUMO DAS MÉTRICAS OBTIDAS

Para facilitar a visualização e interpretação dos resultados foi criada a Tabela 18 que além de exibir a acurácia alcançada, também mostra a sensibilidade, a precisão e a AUC de cada modelo em cada uma das bases de dados.

Tabela 18. Resumo das principais métricas obtidas nos modelos
(continua)

Modelo	Resultados										
		Original					Balanceada				
	Divi	PC	Acur	Sens	Preci	AUC	PC	Acur	Sens	Preci	AUC
		----------%----------									
Logístico binário	TR	0,55	82,4	78,3	66,7	88,8	0,65	84,6	91,2	76,5	93,1
	TS		91,7	85,7	80,0	87,8		90,0	76,5	86,7	88,9
Gradient boosting	TR	0,51	82,8	77,7	69,1	79,5	0,34	82,1	91,2	70,9	82,7
	TS		90,0	80,0	80,0	86,6		91,6	81,2	86,6	90,0

Tabela 18. Resumo das principais métricas obtidas nos modelos

(conclusão)											
Modelo	Resultados										
		Original					Balanceada				
	Divi	PC	Acur	Sens	Preci	AUC	PC	Acur	Sens	Preci	AUC
	----------%----------										
Máquina de vetores e suporte	CM	0,45	91,3	75,0	97,3	87,0	0,57	93,2	91,1	95,1	93,2
	TR		92,9	81,5	97,1	90,1		95,1	94,4	95,6	95,1
	TS		83,3	60,0	69,2	75,6		85,0	86,7	65,0	85,6
Árvore de decisão	TR	0,60	86,2	65,4	91,4	81,1	0,60	89,8	86,4	92,7	89,8
	TS		83,3	60,0	69,2	75,6		81,7	66,7	62,5	76,7
Naive Bayes	TR	0,20	26,8	71,6	27,6	37,7	0,20	29,9	55,6	36,7	29,9
	TS		26,7	80,0	22,6	61,1		23,3	60,0	18,4	35,6

Fonte: Resultados originais da pesquisa

Nota: Divisão [Divi]; Treino [TR]; Teste [TS]; Completa [CM]; Ponto de corte [PC]; Acurácia em porcentagem [Acur]; Sensibilidade em porcentagem [Sens]; Precisão em porcentagem [Preci]; AUC em porcentagem

14 ANÁLISE DAS MÉTRICAS

Este estudo considerou em sua avaliação uma das métricas mais populares para os modelos de classificação binária, a acurácia. Porém, uma limitação desta métrica é que ela considera apenas um ponto de corte, portanto o pesquisador deve atentar-se a ele pois sua escolha pode alterar significativamente o resultado do estudo. De acordo com Šimundić (2009), a acurácia precisa ser analisada com cautela, pois ela somente mostra que o modelo é capaz de trazer as classificações corretas, mas conclui dizendo que outras medidas de diagnósticos devem ser ponderadas para que a conclusão seja mais assertiva.

Por estar relacionado à área da saúde e tendo como referência tanto o ditado popular que diz que é melhor prevenir do que remediar quanto o conhecimento científico que reconhece o valor de diagnósticos precoces, a segunda métrica escolhida foi a sensibilidade. Quanto maior seu valor, menor a quantidade de falsos negativos, ou seja, menor a quantidade de pessoas que deveriam ser acompanhadas por um profissional de saúde, mas o modelo não foi capaz de prever.

O terceiro índice utilizado foi o AUC, pois como dito anteriormente, trata-se de uma métrica que serve para comparar diferentes modelos e, por este motivo, torna-se uma peça fundamental para análise dos resultados produzidos por este estudo.

Outro aspecto que foi ponderado foi o nível de ganho de desempenho das previsões obtidas a partir dos modelos estimados após o balanceamento dos dados. Embora a técnica utilizada nesta etapa tenha fundamentos teóricos sólidos, as observações geradas são sintéticas. Por tratar-se de análise de fatores de saúde foi definido que, caso as métricas dos modelos estimados a partir da base de dados balanceada fossem pouco superiores às métricas obtidas a partir dos modelos estimados a partir da base original, por questões de confiabilidade, a preferência seria dada ao modelo estimado a partir dos dados originais por este não conter dados sintéticos.

Após definidas as métricas e os aspectos que seriam utilizados para a avaliação, foram analisados os resultados obtidos na base de dados de testes de cada um dos modelos.

O modelo “naive Bayes”, por obter resultados que mostraram a menor acurácia, precisão e AUC, o mostrou-se o menos adequado para a previsão da variável resposta.

As métricas da árvore de decisão não tiveram grande variação na base de testes tanto com o modelo estimado a partir da base original quanto com o modelo gerado a partir da base balanceada. A acurácia alcançada foi de 83,3% na base original e de 81,7% na base balanceada. Desta forma pode-se concluir que, embora o balanceamento tenha contribuído com a generalização deste algoritmo, este modelo não foi capaz de atingir a melhor acurácia entre os modelos criados por este estudo.

Como o algoritmo de máquina de vetores e suporte possui a habilidade de “cross-validation”, pode causar estranhamento o fato de ter sido estimado tanto com a base de dados completa quanto com a divisão entre treino e testes. Porém, é sabido que em modelos SVM pode ocorrer “overfitting” devido à escolha de hiperparâmetros. A comparação entre os resultados obtidos com as duas estratégias pode ser benéfica para o melhor entendimento dos resultados e análise de possível “overfitting”.

A acurácia conseguida com a base de dados balanceada completa foi de 93,2% com sensibilidade de 91,1%, precisão de 95,1% e AUC de 93,2%. Em números absolutos este é o melhor modelo criado neste estudo. Já o mesmo modelo com a base não balanceada obteve acurácia de 91,3% com sensibilidade de 75%, precisão de 97,3% e AUC de 87%.

Ao implementar a técnica da divisão entre treino e testes utilizando os mesmos hiperparâmetros dos modelos estimados anteriormente, a capacidade de predição do modelo é bastante afetada e as métricas de acurácia e sensibilidade aproximam-se àquelas conseguidas na árvore de decisão.

Com base nestes dados é razoável supor que os modelos de base completa estão, de alguma forma, incorrendo em “overfitting” já que, ainda que utilizando os mesmos parâmetros, quando o modelo gerado recebe dados desconhecidos a acurácia é 9,6%, a sensibilidade 21,5% e a precisão 27,8% inferiores ao conseguido quando os dados já são conhecidos. Esta diferença pode significar que os modelos SVM aqui propostos, embora traga resultados que possam ser considerados satisfatórios, não se mostrou suficientemente genérico para ser considerado quando utilizado com dados desconhecidos pelo modelo.

Já o modelo “gradient boosting” mostrou-se capaz de prever 90% dos casos corretamente na base original, sendo a segunda maior acurácia e também segundo maior AUC por não considerarmos o modelo SVM de base completa. Na base balanceada o modelo previu 91,6% dos casos corretamente. Estes modelos também apresentaram alta sensibilidade e precisão chegando a resultados satisfatórios e muito próximos ao conseguido no modelo logístico binário.

O modelo logístico binário teve seu melhor desempenho quando estimado com a base de dados original com acurácia de 91,7%, sensibilidade de 85,7%, precisão de 80% e AUC de 87,8%. Quando

se compara os resultados das previsões nas bases de dados percebe-se que na base original a acurácia foi 1,7% e a sensibilidade 9,2% maiores em comparação com a base balanceada, porém, a precisão é 6,7% menor.

Importante pontuar que o modelo logístico binário estimado com a base de dados original apresenta acurácia e sensibilidade superiores na base de testes em comparação àquela atingida na base de treino. Isto pode indicar que o modelo está satisfatoriamente genérico para prever adequadamente a categoria de uma observação que possua dados não presentes na amostra no momento em que foi estimado.

Reforçando essa perspectiva, a aplicação do método AHP para análise dos resultados consolida a robustez do modelo logístico binário na base de dados original, pois este, ainda que muito próximo ao "gradient boosting" na base balanceada, recebe o maior índice de preferência global. Vale relembrar que uma das principais premissas deste estudo foi dar preferência aos modelos estimados a partir dos dados originais e, portanto, quando comparamos as duas melhores opções utilizando somente estes, o índice de preferência global do modelo logístico binário original está confortavelmente distante do "gradient boosting" estimado a partir da base original como pode ser visto na Tabela 19.

Tabela 19. Índice de preferência global obtida por meio do método AHP

Modelo	Índice de Preferência Global
Logístico binário original	12,03%
"Gradient boosting" balanceada	11,99%
"Gradient boosting" original	11,70%
Logístico binário Balanceada	11,70%
Máquina de vetores e suporte balanceada	11,39%
Árvore de decisão balanceada	10,34%
Máquina de vetores e suporte original	10,30%
Árvore de decisão original	10,30%
"Naive Bayes" original	5,83%
"Naive Bayes" balanceada	4,44%

Fonte: Resultados originais da pesquisa

O fato do modelo logístico binário mostrar-se como o mais adequado para este tipo de previsão tem suporte na literatura quando Fávero e Belfiore (2017) debatem sobre o estudo da possibilidade de infarto como variável a ser prevista e as características físicas e de comportamento como variáveis preditoras. O objetivo então é estimar a probabilidade de ocorrência deste evento dicotômico por meio da regressão logística binária.

O AUC de 87,8% atingido com o modelo logístico binário pode ser considerado ótimo, próximo ao excelente, para o diagnóstico de acurácia (Šimundić, 2009). Entender se a acurácia de 91,7% é adequada para a área da saúde trata-se de algo mais complexo, pois devido às complexidades biológicas e demográficas, não é possível determinar um valor, ou uma faixa de valores, que satisfaça a todos os casos. Porém, deixando os números e refletindo um pouco se pode chegar a uma conclusão:

quanto mais urgente for a solução para o fenômeno estudado, menor poderá ser a acurácia aceita inicialmente.

Um exemplo disto é o que foi visto durante o desenvolvimento das vacinas para a COVID-19. No caso da CoronaVac, na fase 3 dos testes, o imunizante apresentou eficácia global de 62,3% o que inicialmente pode ser considerado baixa eficácia, porém dada a dimensão dos impactos gerados pela doença na saúde mental da população e na economia, o imunizante foi aprovado pelos órgãos competentes (Instituto Butantan, 2022).

Por outro lado, o objeto deste estudo, embora relevante, não apresenta um impacto tão significativo quanto a referida pandemia. Portanto, faz-se necessária a comparação com estudos que também tiveram como objeto a insuficiência cardíaca.

Em estudo muito similar a este, Liang e Guo (2023) conseguiram acurácia de 78% e AUC de 71,9%, e complementam dizendo que este número demonstra uma certa superioridade quando comparados com outros estudos recentemente publicados. Já Mahmud et. al (2023) traz um meta-modelo para previsão de insuficiência cardíaca que, utilizando-se da combinação de quatro algoritmos de "machine learning", obteve acurácia de 87%.

Quando se compara os valores aqui alcançados, acurácia de 91,7% e AUC de 87,8%, com outros estudos recentes sobre o mesmo tema é possível concluir que este estudo atingiu bons índices no que se refere às métricas extraídas, sendo, portanto factível que seja considerado como um bom ponto de partida para estudos mais avançados.

15 ANÁLISE DO MODELO LOGÍSTICO ESTIMADO

Como dito nas primeiras páginas deste estudo, um dos fatores positivos deste modelo é sua facilidade de interpretação. Faz-se então necessária a análise das variáveis independentes selecionadas com intervalo de confiança de 95% e seus respectivos coeficientes, presentes na Tabela 20, para ser identificado o impacto de cada um na probabilidade de ocorrência do evento "DEATH_EVENT".

Tabela 20. Resumo das principais métricas obtidas nos modelos

Variável	Estimadores	Erro padrão
(Intercept)	0.8400	1.1465
age	0.0372	0.0163
ejection_fraction	-0.0687	0.0162
serum_creatinine	0.7171	0.1898
time	-0.0205	0.0031

Fonte: Resultados originais da pesquisa

Portanto, percebe-se que "age" ou idade, "ejection_fraction" ou fração de ejeção, "serum_creatinine" ou creatinina sérica e "time" ou tempo de acompanhamento do paciente em dias,

são as variáveis relevantes para o modelo. A variável “smoking”, apenas seria relevante para intervalo de confiança inferior a 88%, por isso não foi considerada no modelo.

Bosch et. al (2019) confirmam que o envelhecimento eleva a chance de insuficiência cardíaca quando mostram que ela ocorre em apenas 0,04% entre pessoas de 18 a 44 anos, porém este número eleva-se para 20,9% para pessoas maiores de 85 anos. Portanto, o fato da variável "age" ser relevante encontra suporte na literatura.

Lam e Solomon (2021) sugerem uma nova classificação insuficiência cardíaca com base na fração de ejeção. Portanto, ainda que esteja sendo objeto de estudos para aprimoramento dos valores, o estudo mostra que a fração de ejeção tem relação direta com a insuficiência cardíaca.

A creatinina sérica é produzida pela morte de células musculares do corpo que deve ser filtrada pelos rins, porém a elevação em seus níveis podem sugerir doenças no coração, ou nos rins, que podem levar à insuficiência cardíaca (BHF, 2023). Desta forma, a literatura também suporta a relação entre a creatinina e a insuficiência cardíaca.

Por último tem-se a quantidade de dias que o paciente está sendo acompanhado por um profissional de saúde. Parece justo concluir que quanto maior este número menor seria o risco de morte por insuficiência cardíaca, pois o paciente estaria recebendo o tratamento adequado.

Portanto, todas as variáveis selecionadas pelo modelo como relevantes tem estudos que confirmam sua relação com a insuficiência cardíaca.

Antes que sejam analisados os coeficientes, é importante lembrar que os modelos logísticos retornam resultados entre 0 e 1, no qual 0 representa a ausência de chance do evento, 1 representa a certeza de ocorrência do evento. Imaginando que o resultado tenha sido 0,56, deve-se entender que a probabilidade de ocorrência do evento é de 56%, segundo o modelo. Caso este valor esteja acima do ponto de corte definido pelo pesquisador, o resultado é então definido como evento, caso esteja abaixo é definido como não-evento.

Para o modelo logístico deste estudo, o resultado é dado a partir da Equação 7 utilizando 0,55 como o ponto de corte.

$$P = \frac{1}{1 + \exp(-(0{,}84 + 0{,}0372 * \text{age} + - 0{,}0687 * \text{ejection\_fraction} + 0{,}7171 * \text{serum\_creatinine} + - 0{,}0205 * \text{time}))} \quad (7)$$

em que, P: é a probabilidade de ocorrência do evento; exp: é a exponencial do negativo do valor obtido pela expressão entre parênteses.

A partir dela conclui-se que para cada ano vivido, o modelo atribui um coeficiente positivo de 0,0372, ou seja, para o modelo a cada ano adicional a possibilidade de morte por insuficiência cardíaca tem um pequeno acréscimo. Para cada unidade adicional de fração de ejeção, que possui o coeficiente negativo de 0,0687, o modelo diminui a possibilidade de morte por insuficiência cardíaca. Seguindo

esta lógica para as demais variáveis pode-se concluir que o parâmetro que mais eleva o risco de óbito é a creatinina sérica, mantendo-se as demais variáveis constantes. Já a quantidade de dias sendo acompanhado é o que menos diminui o risco, segundo o modelo.

Ao ser aplicada a Equação 7 às observações da base de testes, obtém-se acurácia de 91,7%, sensibilidade de 85,7%, precisão de 80% e AUC de 87,8% utilizando o ponto de corte de 0,55.

17 CONSIDERAÇÕES FINAIS

Acredita-se que este estudo alcançou ao objetivo proposto, pois desenvolveu um modelo logístico binário capaz de classificar corretamente 91,7% das observações, resultado superior ao de estudos similares. Admite-se também que seja possível sua incorporação aos sistemas de saúde tanto no que se refere a sua capacidade de classificação quanto em relação aos custos associados a seu treinamento, armazenamento, disponibilização para acesso e execução das previsões a fim de que seu resultado seja adicionado ao histórico clínico do paciente.

Com relação aos aspectos técnicos apresentados, ressalta-se a importância de testes de diferentes modelos para contextos diversos, pois no caso deste estudo, embora modelos bastante sofisticados tenham sido implementados foi a regressão logística que gerou os melhores resultados. Em conjunto com a necessidade de testes de diferentes modelos está a necessidade de escolha da métrica para comparação dos resultados já que esta é uma importante referência para a escolha do modelo adequado. Por último cita-se a vantagem de verificar as estatísticas descritivas dos dados para ser identificada a necessidade de seu balanceamento e, caso haja, recomenda-se a comparação dos resultados obtidos com os dados originais e balanceados a fim de identificar se houve ganho relevante na generalização do modelo com a introdução dos dados sintéticos.

Concentrando-se em aspectos analíticos, é importante que o pesquisador tenha conhecimento das características de cada modelo como, por exemplo, saber que determinado modelo possui tendência ao “overfitting”, para analisar as métricas obtidas a luz desta informação. Outro ponto importante para classificações binárias é a atenção à escolha do ponto de corte, já que este parâmetro é central para a quantificação dos resultados atingidos. Por fim, o pesquisador deve ter clareza de quais métricas são mais relevantes para que, ao comparar os resultados obtidos, seja capaz de optar por aquele que apresente o melhor equilíbrio entre as métricas levando em consideração o objetivo do estudo.

Para pesquisas futuras, recomenda-se a ampliação da faixa etária dos pacientes, a adição de amostra de dados dos sistemas locais de saúde, a diversificação de etnias, a consideração de variáveis de comportamento como, por exemplo, se pratica exercícios físicos, além de informações gerais tais como o aumento abrupto de peso, o desenvolvimento de edemas, inchaço do tornozelo entre outros.

AGRADECIMENTOS

Nesta seção onde devo agradecer a quem contribuiu de maneira relevante a este trabalho, gostaria de deixar meu agradecimento ao meu objeto deste estudo. Agradeço, do fundo do peito, ao meu coração, pois sem ele, literalmente, nada seria possível. Também ao Professor Erasmo por toda dedicação e comprometimento, pois certamente, isto elevou de forma relevante a qualidade do estudo apresentado.

REFERÊNCIAS

Arruda, V. L.; Machado, L. M. G.; Lima, J. C.; Silva, P. R. S.. 2022. Tendência da mortalidade por insuficiência cardíaca no Brasil: 1998 a 2019. Disponível em: <https://www.scielosp.org/article/rbepid/2022.v25/E220021/pt/>. Acesso em: 10 out. 2023.

Bosch, L.; Assmann, P.; Grauw, W.J.C.; Schalk, B.W.M.; Biermans, M.C.J.. 2019. Heart failure in primary care: prevalence related to age and comorbidity. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6683237/>. Acesso em: 29 fev. 2024.

Boswell, D. 2002. Introduction to Support Vector Machine. Thesis. California Institute of Technology, Pasadena, California, United States of America.

British Heart Foundation [BHF]. 2023. Global Heart & Circulatory Diseases Factsheet. Disponível em: <https://www.bhf.org.uk/-/media/files/for-professionals/research/heart-statistics/bhf-cvd-statistics-global-factsheet.pdf>. Acesso em: 10 out. 2023.

British Heart Foundation [BHF]. 2023. The heart-kidney link. Disponível em: <https://www.bhf.org.uk/informationsupport/heart-matters-magazine/medical/kidney-heart-link>. Acesso em: 29 fev. 2023.

Bruce, P. C.; Stephens, M. L.; Shmueli, G.; Anandamurthy, M.; Patel, N. R.. 2023, Machine Learning for Business Analytics. In: Bruce, P. C.; Stephens, M. L.; Shmueli, G.; Anandamurthy, M.; Patel, N. R.. Concepts, Techniques, And Applications with JMP PRO. 2 ed. O'Reilly, Hoboken, New Jersey, United States. Disponível em: <https://learning.oreilly.com/library/view/machine-learning-for/9781119903833/c08.xhtml#head-2-51>. Acesso em: 05 jan. 2024

Cramer, J.S. 2002. The Origins of Logistic Regression. Thesis. University of Amsterdam, Amsterdam, The Nederland.

Elreedy, D.: Atiya, A.F.: Kamalov, F.: 2022, A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Disponível em: <https://doi.org/10.1007/s10994-022-06296-4>. Acesso em: 21 fev. 2024.

Escolar, V.: Lozano, A.: Larburu, N.: Kerexeta, J.: Álvarez, R.: Echebarria, A.:Azcona, A.. 2021, Prediction of heart failure decompensations using artificial intelligence and machine learning techniques. Basurto University Hospital, Donostia, San Sebastián, Espanha. Disponível em: <https://www.rccardiologia.com/files/rcc_22_29_4_431-440.pdf>. Acesso em: 17 jan. 2024.

Fávero, L. P. L.; Belfiore, P.. Manual de Análise de Dados: Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata®. 1 ed. Rio de Janeiro: Elsevier, 2017. p. 153-611.

Ferreira, D. 2020. Insuficiência cardíaca. Hospital da Luz, Lisboa, Lisboa, Portugal. Disponível em: <https://www.hospitaldaluz.pt/pt/dicionario-de-saude/insuficiencia-cardiaca>. Acesso em: 16 fev. 2024.

Géron, A. 2022. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow 3e: Concepts, Tools, and Techniques to Build Intelligent Systems. Disponível em: < https://www.oreilly.com/library/view/hands-on-machine-learning/9781098125967/ch06.html>. Acesso em: 5 mar. 2024.

Gholami, R.; Fakhari, N.. 2017. Support Vector Machine: Principles, Parameters, and Applications, 515-535. In: Gholami, R.; Fakhari, N. Handbook of Neural Computation. Academic Press.

Guerbai, Y.; Chibani, Y.; Meraihi, Y.. 2022. Techniques for Selecting the Optimal Parameters of One-Class Support Vector Machine Classifier for Reduced Samples. Disponível em: <https://www.igi-global.com/article/techniques-for-selecting-the-optimal-parameters-of-one-

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., & Hofner, B. (2023). glmboost [Documentação do software]. Disponível em:<https://www.rdocumentation.org/packages/mboost/versions/2.9-9/topics/glmboost>.
Acesso em: 14 mar. 2024

Kotsiantis, S.B.. 2012. Decision trees: a recent overview. Disponível em: < https://link.springer.com/article/10.1007/s10462-011-9272-4>. Acesso em: 05 mar. 2024.

Instituto Butantan. 2022. CoronaVac provou sua eficácia contra Covid-19 no estudo clínico mais criterioso, feito com profissionais de saúde durante pico de casos. Disponível em: <https://butantan.gov.br/noticias/coronavac-provou-sua-eficacia-contra-covid-19-no-estudo-clinico-mais-criterioso-feito-com-profissionais-de-saude-durante-pico-de-casos>. Acesso em: 13 fev. 2024.

Lam, C.; Solomon, S.. 2021. Classification of Heart Failure According to Ejection Fraction. Disponível em: <https://doi.org/10.1016/j.jacc.2021.04.070>. Acesso em: 29 fev. 2024.

Liang, Y.; Guo, C.. 2023. Heart failure disease prediction and stratification with temporal electronic health records data using patient representation. Disponível em: <https://doi.org/10.1016/j.bbe.2022.12.008>. Acesso em: 29 fev. 2024.

Mahmud, I.; Kabir, M.M.; Mridha, M.F.; Alfarhood, S.; Safran. M.; Che, D.. 2023. Cardiac Failure Forecasting Based on Clinical Data Using a Lightweight Machine Learning Metamodel. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10417090/>. Acesso em: 29 fev. 2024.

Park, T. 2021. Behind Covid-19 vaccine development, MIT Schwarzman College of Computing, 2021. Disponível em: <https://news.mit.edu/2021/behind-covid-19-vaccine-development-0518>. Acesso em: 10 out. 2023.

Rohde, L.E.P.: Montera, M.W.: Bocchi, E.A.: Clausell, N.O.: Albuquerque, D.C.: Rassi, S.: Colafranceschi, A.S.: Freitas Junior, A.F.: Ferraz, A.S.: Biolo, A.: Barretto, A.C.P.: Ribeiro, A.L.P.: Polanczyk, C.A.: Gualandro, D.M.: Almeida, D.R.: Silva, E.R.R.: Figueiredo, E.L.: Mesquita, E.T.: Marcondes-Braga, F.G.: Cruz, F.D.: Ramires, F.J.A.: Atik, F.A.: Bacal, F.: Souza, G.E.C.: Almeida Junior, G.L.G.: Ribeiro, G.C.A.: Villacorta Junior, H.: Vieira, J.L.: Souza Neto, J.D.: Rossi Neto, J.M.: Figueiredo Neto, J.A.: Moura, L.A.Z.: Goldraich, L.A.: Beck-da-Silva, L.: Danzmann, L.C.: Canesin, M.F.: Bittencourt, M.I.: Garcia, M.I.: Bonatto, M.G.: Simões, M.V.: Moreira, M.C.V.: Silva, M.M.F.: Oliveira Junior, M.T.: Silvestre, O.M.: Schwartzmann, P.V.: Bestetti, R.B.: Rocha, R.M.: Simões, R.: Pereira, S.B.: Mangini, S.: Alves, S.M.M.: Ferreira, S.M.A.: Issa, V.S.: Barzilai, V.S.: Martins, W.A.: 2028, Diretriz Brasileira de Insuficiência Cardíaca Crônica e Aguda Arquivos Brasileiros De Cardiologia. Disponível em: <https://doi.org/10.5935/abc.20180190>. Acesso em: 16 fev. 2024.

Saaty, T.L.. 1977. How to make a decision: The analytic hierarchy process. European Journal of Operational Research, 48(1), 9-26. Disponível em: <https://doi.org/10.1016/0377-2217(90)90057-I>. Acesso em: 25 mar. 2024.

Saupin, G.. Practical Gradient Boosting: A deep dive into Gradient Boosting in Python. 1 ed. Paris: AFNIL. 2022. p.17.

Šimundić, A. M.. 2009, Measures of Diagnostic Accuracy: Basic Definitions, 2009. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4975285/>. Acesso em: 11 fev. 2024.

Sosnovshchenko, A: Baiev, O.. 2018, Machine Learning with Swift. In: Sosnovshchenko, A: Baiev, O..: Artificial Intelligence for iOS. 1 ed. Packt Publishing, England, Birmingham. Disponível em: <https://learning.oreilly.com/library/view/machine-learning-with/9781787121515/697c4c5f-1109-4058-8938-d01482389ce3.xhtml>. Acesso em: 05 jan. 2024.

Taulli, T.; Introdução à Inteligência Artificial: Uma abordagem não técnica. 1 ed. São Paulo: Novatec, 2020. p. 26.

Uddin, K.M.M.: Ripa, R.: Yeasmin, N.: Biswas, N.: Dey, S.K.. 2022, Machine learning-based approach to the diagnosis of cardiovascular vascular disease using a combined dataset. Dhaka International University; Dhaka, Bangladesh. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666521223000145>. Acesso em: 15 jan. 2024.

World Health Organization [WHO]. 2023. Cardiovascular diseases. Disponível em: <https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1>. Acesso em: 10 out. 2023.

Wickramasinghe, I.: Kalutarage, H.. 2020, Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Disponível em: <https://doi.org/10.1007/s00500-020-05297-6>. Acesso em: 20 feb. 2024.

ANEXO

Projeto R Completo e Excel de apoio disponíveis para download em https://github.com/danielbaldini/TCC_USP_2024_HeartFailureAnalysis