CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil

Machine learning for predictive analyses in health: an example of an application to predict death in the elderly in São Paulo, Brazil

Machine learning para análisis predictivos en salud: ejemplo de aplicación para la predicción de óbitos en ancianos de São Paulo, Brasil

Hellen Geremias dos Santos [1]
Carla Ferreira do Nascimento [1]
Rafael Izbicki [2]
Yeda Aparecida de Oliveira Duarte [3]
Alexandre Dias Porto Chiavegatto Filho [1]

doi: 10.1590/0102-311X00050818

Resumo

Este estudo objetiva apresentar as etapas relacionadas à utilização de algoritmos de machine learning *para análises preditivas em saúde. Para isso, foi realizada uma aplicação com base em dados de idosos residentes no Município de São Paulo, Brasil, participantes do estudo* Saúde Bem-estar e Envelhecimento *(SABE) (n = 2.808). A variável resposta foi representada pela ocorrência de óbito em até cinco anos após o ingresso do idoso no estudo (n = 423), e os preditores, por 37 variáveis relacionadas ao perfil demográfico, socioeconômico e de saúde do idoso. A aplicação foi organizada de acordo com as seguintes etapas: divisão dos dados em treinamento (70%) e teste (30%), pré-processamento dos preditores, aprendizado e avaliação de modelos. Na etapa de aprendizado, foram utilizados cinco algoritmos para o ajuste de modelos: regressão logística com e sem penalização, redes neurais,* gradient boosted trees *e* random forest*. Os hiperparâmetros dos algoritmos foram otimizados por validação cruzada 10-*fold*, para selecionar aqueles correspondentes aos melhores modelos. Para cada algoritmo, o melhor modelo foi avaliado em dados de teste por meio da área abaixo da curva (AUC) ROC e medidas relacionadas. Todos os modelos apresentaram AUC ROC superior a 0,70. Para os três modelos com maior AUC ROC (redes neurais e regressão logística com penalização de lasso e sem penalização, respectivamente), foram também avaliadas medidas de qualidade da probabilidade predita. Espera-se que, com o aumento da disponibilidade de dados e de capital humano capacitado, seja possível desenvolver modelos preditivos de* machine learning *com potencial para auxiliar profissionais de saúde na tomada de melhores decisões.*

Previsões; Mortalidade; Idoso

Correspondência
H. G. Santos
Faculdade de Saúde Pública, Universidade de São Paulo.
Av. Dr. Arnaldo 715, 1º andar, São Paulo, SP
01246-904, Brasil.
hellengeremias@usp.br

[1] Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, Brasil.
[2] Centro de Ciências Exatas e de Tecnologia, Universidade Federal de São Carlos, São Carlos, Brasil.
[3] Escola de Enfermagem, Universidade de São Paulo, São Paulo, Brasil.

Introdução

A análise preditiva consiste na aplicação de algoritmos para compreender a estrutura dos dados existentes e gerar regras de predição. Esses algoritmos podem ser utilizados em um cenário não supervisionado, em que apenas preditores (covariáveis) estão disponíveis no conjunto de dados, ou em problemas supervisionados, quando, além dos preditores, está disponível também uma resposta de interesse, responsável por guiar a análise [1].

Na área da saúde, modelos preditivos podem ser utilizados para estimar o risco de determinado desfecho ocorrer, dado um conjunto de características socioeconômicas, demográficas, relacionadas ao hábito de vida e às condições de saúde, entre outras. Seus resultados, quando combinados a medidas de saúde pública aplicadas em nível populacional, podem trazer implicações positivas na redução de custos e na efetividade de intervenções, como tratamentos e ações preventivas. Adicionalmente, conhecer o risco de um desfecho ocorrer pode auxiliar gestores responsáveis por formular e avaliar políticas públicas a direcionar intervenções preventivas, considerando a ponderação entre danos e benefícios [2,3]. Historicamente, alguns modelos têm sido desenvolvidos para tentar predizer a ocorrência de desfechos de interesse para a saúde da população. Pesquisadores do *Framingham Heart Study* desenvolveram funções de risco para doença cardiovascular [4,5] que motivaram políticas públicas para o estabelecimento de medidas preventivas direcionadas a indivíduos com maior risco [6]. Da mesma forma, modelos preditivos de diagnóstico e prognóstico de câncer de mama também têm sido relatados na literatura como ferramenta auxiliar na identificação de indivíduos com maior risco, para os quais, estratégias de rastreamento e tratamento profilático representam intervenções com potencial impacto benéfico no prognóstico da doença [7].

Esses modelos, em geral, são derivados do ajuste de modelos lineares, considerados algoritmos mais simples de *machine learning*, como o de regressão logística, para desfechos categóricos, e o de regressão linear, para desfechos contínuos. Na última década, novas abordagens têm sido desenvolvidas para acomodar relações não lineares, solucionar problemas de colinearidade e de alta dimensionalidade dos dados, entre outras particularidades, representando, em alguns casos, generalizações ou extensões dos modelos lineares a fim de torná-los mais flexíveis [8,9].

Nesse contexto, algoritmos de inteligência artificial (*machine learning*) mais flexíveis têm sido utilizados na modelagem preditiva de desfechos de interesse para a saúde, como para predizer o risco de mortalidade [10], de readmissão hospitalar [11] e de desfechos desfavoráveis ao nascimento [12]. Com a disponibilidade crescente de dados relevantes para o desenvolvimento de pesquisas em saúde, esses algoritmos têm o potencial de melhorar a predição do desfecho por capturar relações complexas nos dados, bem como por sua capacidade em lidar com uma grande quantidade de preditores [13,14].

No Brasil, a utilização desses algoritmos em saúde pública ainda é incipiente. Como exemplo, pode-se citar o estudo de Olivera et al. [15] que desenvolveu modelos preditivos de diabetes não diagnosticada a partir de dados de 12.447 adultos entrevistados para o Estudo ELSA (*Estudo Longitudinal da Saúde do Adulto*), utilizando cinco algoritmos de *machine learning* (regressão logística, redes neurais, *naive bayes*, método dos *K* vizinhos mais próximos e *random forest*).

Com o envelhecimento populacional, informações prognósticas relacionadas ao risco de óbito e de outras doenças de prevalência elevada nessa população têm se tornado cada vez mais importantes para médicos, pesquisadores e formuladores de políticas públicas como uma ferramenta para auxiliar em tomadas de decisões referentes ao rastreamento de doenças, ao direcionamento de programas preventivos e à oferta de tratamentos especializados. Dados sobre estado geral de saúde, presença de doenças e limitações funcionais, bem como os de acesso e utilização de serviços de saúde podem contribuir para a estimativa do risco de morte na população idosa. Assim, o presente estudo tem o objetivo de exemplificar e discutir as etapas que compõem uma análise preditiva, utilizando algoritmos de *machine learning* para predizer o risco de óbito em cinco anos em idosos residentes no Município de São Paulo, Brasil, participantes do estudo *Saúde, Bem-estar e Envelhecimento* (SABE).

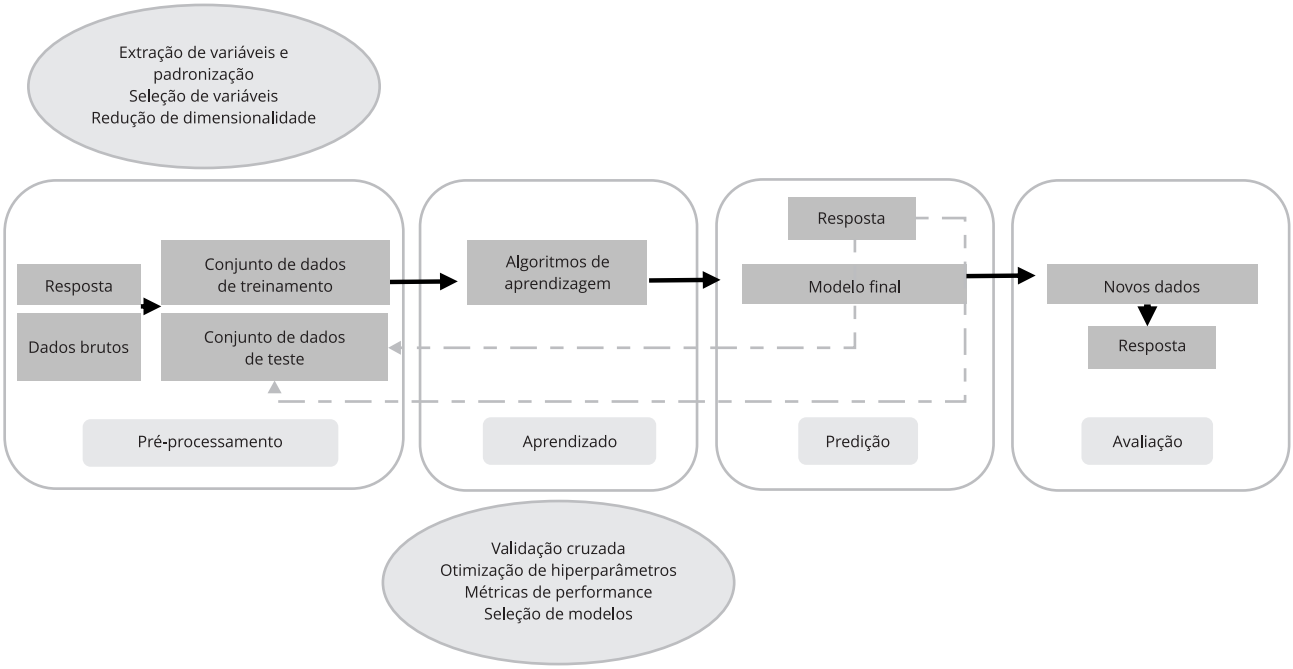
Métodos

Para a presente aplicação, algoritmos de *machine learning* serão utilizados em um cenário supervisionado, no qual, cada observação do conjunto de dados dispõe de um vetor de mensurações para os preditores x_i, $i = 1, 2, ..., n$, bem como para a resposta de interesse, y_i. O objetivo principal consiste em ajustar um modelo que relacione a resposta, Y, aos preditores, X, a fim de predizer esse evento em observações futuras. O tipo de variável resposta a ser predita define dois subgrupos de aprendizado supervisionado: o de regressão, para variáveis quantitativas, e o de classificação, para as do tipo categórica (qualitativa). O ajuste de modelos preditivos, em ambos os casos, pode ser representado pelas seguintes etapas: divisão do conjunto de dados em treinamento e teste, pré-processamento, aprendizado e avaliação de modelos (Figura 1).

A divisão da amostra em dados de treinamento e de teste é realizada para verificar se um modelo apresenta boa performance não apenas em dados utilizados para seu ajuste (treinamento), mas também capacidade de generalização para novas observações (teste). As divisões mais utilizadas são 60:40, 70:30 ou 80:20, dependendo do tamanho do conjunto de dados – em geral, quanto maior o número de observações, maior será a proporção do conjunto inicial utilizada para o treinamento [16]. No presente estudo, foram utilizados 70% dos dados para treinamento dos algoritmos (n = 1.875; 277 óbitos) e 30% para teste da performance preditiva dos modelos ajustados (n = 802; 118 óbitos). O seguinte passo é realizar uma predição acurada da resposta de interesse (no caso, Y: óbito em 5 anos), denotada por $\hat{Y}$, a partir dos valores de um vetor de preditores, X, contendo informações como idade, sexo, presença de hipertensão arterial, entre outras. O aprendizado de um problema de classificação consiste no particionamento do espaço dos preditores em regiões correspondentes às categorias da

Figura 1

Roteiro para aplicação de algoritmos de *machine learning* em análise preditiva.



Fonte: Raschka & Mirjalili [32].

resposta de interesse. A fronteira de decisão entre essas regiões, denominada classificador, representa o modelo preditivo estimado por determinado algoritmo [1]. Portanto, nesse contexto, o objetivo do aprendizado é construir um classificador, *f(X)*, que faça boas predições da resposta de interesse em observações futuras [17]:

$$f(x_{n+1}) \approx y_{n+1}, \ldots f(x_{n+m}) \approx y_{n+m}$$

É importante mencionar que, em problemas de predição, o interesse principal ao estimar o classificador *f* está na acurácia das predições para dados novos, e não na forma exata que a fronteira de decisão estimada assume. A acurácia de $\hat{y}$ como predição para *Y* depende de duas quantidades: erro redutível e irredutível. Em geral, $\hat{f}$ não irá representar uma estimativa perfeita de *f*, e tal imprecisão irá introduzir um erro ao valor predito. Entretanto, esse erro pode ser reduzido por meio da utilização de algoritmos apropriados para estimar *f*. Por outro lado, mesmo que fosse possível obter um classificador perfeito a partir dos preditores *X*, devido ao erro irredutível, esse classificador estaria sujeito a erros. Por exemplo, dois indivíduos com valores idênticos para os preditores mensurados podem apresentar desfechos distintos. Logo, não é possível que o classificador acerte em todos os casos.

Desse modo, os algoritmos de *machine learning* são utilizados com o objetivo de estimar *f* e, portanto, de minimizar o erro redutível. Entre os diversos algoritmos disponíveis, alguns são pouco flexíveis (menos complexos), porém interpretáveis, como é o caso da árvore de decisão e da regressão logística. Por outro lado, há abordagens mais flexíveis (mais complexas), o que torna mais difícil compreender como cada preditor está individualmente associado à resposta de interesse, como, por exemplo, no caso das redes neurais [18].

É importante observar que modelos menos complexos podem apresentar melhor performance que aqueles que são mais flexíveis por estarem menos sujeitos ao sobreajuste e, consequentemente, resultarem em predições mais acuradas para *Y* em novas observações, especialmente para o caso de bancos de dados pequenos. Entretanto, não há um único algoritmo capaz de apresentar boa performance em todas as aplicações, sendo importante comparar alguns algoritmos com características distintas para selecionar aquele que resulte em um modelo com performance preditiva satisfatória para o problema em questão [18].

De modo geral, os algoritmos podem ser agrupados nas seguintes categorias: lineares, não lineares e baseados em árvores de decisão [18]. Para a presente aplicação, foram selecionados para comparação cinco algoritmos: regressão logística e regressão logística penalizada (lineares), redes neurais (não linear), *gradient boosted trees* e *random forest* (baseados em árvores de decisão). O Quadro 1 descreve as principais características desses algoritmos, tais como hiperparâmetros a serem otimizados e o cálculo do *ranking* de importância dos preditores.

Pré-processamento

O pré-processamento é guiado pelos algoritmos que serão utilizados para o ajuste de modelos preditivos. De modo geral, sua aplicação está relacionada às seguintes atividades: (1) transformação de variáveis quantitativas (via padronização ou normalização); (2) redução de dimensionalidade do conjunto de dados (exclusão de preditores altamente correlacionados ou utilização de análise de componentes principais); (3) exclusão de variáveis/observações com dados faltantes ou utilização de técnicas de imputação (média, mediana ou valor mais frequente no caso de preditores numéricos ou categoria adicional, representativa dos indivíduos sem informação, no caso de preditores categóricos); (4) organização de variáveis qualitativas (decomposição das variáveis categóricas em um conjunto de variáveis indicadoras que serão utilizadas como preditores).

Vale destacar que os parâmetros estimados por procedimentos de pré-processamento, como os decorrentes da padronização das variáveis e do cálculo do valor para a imputação de dados faltantes, são obtidos em dados de treinamento e, posteriormente, aplicados aos dados de teste (bem como em novas observações) antes da realização das predições. Tal procedimento é adotado para que a performance dos dados de teste seja fidedigna ao real desempenho do modelo preditivo em dados futuros [16].

Para a presente aplicação, na etapa de pré-processamento, valores faltantes para a variável quantitativa índice de massa corporal (IMC) (n = 387) foram imputados a partir da mediana observada para essa variável no conjunto de treinamento. Para as variáveis categóricas escolaridade e história de

Quadro 1

Principais características dos algoritmos de *machine learning* utilizados neste tutorial e seus pacotes no R.

Algoritmo (Pacote)	**Hiperparâmetros otimizados**	**Principais características**	***Ranking* de importância dos preditores**
Regressão logística *(glm)*	-	A função que relaciona os preditores à resposta de interesse está restrita a formas lineares.	Valor absoluto da estatística *t* para cada parâmetro do modelo.
Regressão logística penalizada *(glmnet)*	*Alpha:* porcentagem correspondente ao tipo de regularização a ser aplicada: 0: *ridge*; 1: *lasso*; > 0 e < 1: *elastic net*.	Objetiva obter estimadores viesados, porém com variância reduzida, para os parâmetros de um modelo de regressão usual, o que resulta em um modelo preditivo menos complexo e, portanto, menos sujeito ao sobreajuste em novas observações.	Valor absoluto dos coeficientes correspondentes ao modelo selecionado na etapa de otimização dos hiperparâmetros.
	Lambda: parâmetro de regularização. Pode assumir valores no intervalo [0, ∞]. Quanto maior o seu valor, maior a penalização das estimativas dos parâmetros da regressão.		
Redes neurais *(nnet)*	*Size*: refere-se ao número de unidades latentes da camada intermediária (oculta) de uma rede neural.	Modelo em 2 estágios. *Estágio 1:* aplica transformações não lineares (usualmente a função sigmoide) às combinações lineares dos preditores, dando origem às unidades latentes. *Estágio 2:* relaciona as unidades latentes à resposta de interesse.	Utiliza combinações dos valores absolutos dos pesos do modelo de redes neurais (parâmetros das combinações lineares do primeiro estágio e parâmetros associados às variáveis latentes no segundo estágio do modelo).
	Decay: parâmetro de regularização. Pode assumir valores no intervalo [0, ∞]. Quanto maior o seu valor, maior a penalização das estimativas dos parâmetros das redes neurais.		
Gradient boosted trees (xgboost)	*Nrounds*: número de árvores (iterações) presentes no modelo final.	Objetiva combinar predições de um conjunto de classificadores com taxa de erro apenas ligeiramente inferior a de uma classificação aleatória (árvores de decisão com poucas divisões) para construir um comitê, responsável pela predição final.	Tabulação e soma da influência relativa de cada preditor em cada árvore que compõe o modelo final (melhoria empírica decorrente da utilização desse preditor na realização de uma partição da árvore). Cálculo da média por preditor para todas as iterações, para ter uma visão geral de sua contribuição para o modelo final.
	Maxdepth: profundidade da árvore, relacionada ao número de divisões presentes em cada árvore.		
	Eta: parâmetro de regularização relacionado ao controle da contribuição de cada árvore para a função de predição final. Pode assumir valores no intervalo [0, ∞].		
Random forest (randomForest)	*Mtry:* número de preditores selecionados aleatoriamente como candidatos em cada divisão das árvores de decisão que compõem o classificador final.	Objetiva combinar predições de um conjunto de classificadores complexos (árvores de decisão com muitas divisões), aplicados a amostras *bootstrap* do conjunto de treinamento. Diferencial: seleção aleatória de preditores a serem utilizados, a fim de reduzir a correlação entre as árvores que serão agregadas para produzir a predição final.	Em cada árvore, a precisão da predição correspondente às observações que não compuseram a amostra *bootstrap* é calculada. Esse mesmo procedimento é realizado após permutar cada um dos preditores. A diferença entre essas duas medidas de precisão é calculada e, posteriormente, uma média dessa diferença é computada e normalizada para cada preditor.

queda, cujo número de valores faltantes correspondeu a 377 e 181, respectivamente, uma categoria adicional foi criada para representar tais observações. Para os demais preditores, a ausência de informação resultou em eliminação da observação correspondente ao dado faltante. As variáveis quantitativas IMC e idade foram padronizadas a partir de média e desvio padrão observados no conjunto de treinamento.

O próximo tópico descreve características relacionadas ao treinamento de modelos preditivos com foco em problemas de classificação.

Aprendizado: técnicas de reamostragem para otimização de hiperparâmetros

Em *machine learning* há dois tipos de parâmetros a serem estimados: os parâmetros usuais de um algoritmo, como, por exemplo, os pesos de uma regressão logística; e os parâmetros de ajuste, ou hiperparâmetros, relacionados ao controle da flexibilidade de um algoritmo, como a penalização aplicada aos parâmetros da regressão logística para se obter estimadores viesados, porém com variância reduzida ou o número de unidades latentes presentes na camada oculta de uma rede neural [19].

O controle da flexibilidade de um algoritmo é dependente do balanceamento entre viés e variância. O viés está relacionado à correspondência entre o valor predito por um modelo, para uma dada observação, e o valor real observado, e a variância refere-se à sensibilidade das predições à variabilidade das observações de treinamento. De modo geral, um modelo flexível tem variância alta, pois seu resultado será muito diferente para bancos de dados distintos. Já um modelo pouco flexível tem variância baixa, porém pode apresentar viés alto. O quanto de flexibilidade permitir quando se desenvolve um modelo preditivo é o que, em última instância, torna a utilização de algoritmos de *machine learning* uma tarefa desafiadora [8].

Nesse cenário, o aprendizado de modelos preditivos é composto por dois objetivos principais: selecionar e avaliar os modelos. No primeiro caso, para um dado algoritmo que possua hiperparâmetros, a performance de diferentes modelos, baseados em variações dos valores para os hiperparâmetros, é avaliada para selecionar aquele que resulte em melhor desempenho (equilíbrio entre viés-variância). Já no segundo, após a definição do modelo, busca-se estimar seu erro de predição (erro de generalização) em novas observações [1].

Se o pesquisador dispõe de um conjunto de dados grande, a melhor abordagem para ambos os objetivos consiste em dividir aleatoriamente o conjunto de dados original em três partes: treinamento, validação e teste. Em situações em que o conjunto de dados não é grande o suficiente para ser dividido em três partes, técnicas de reamostragem podem ser utilizadas para aproximar o conjunto de validação por meio da reutilização de observações do conjunto de treinamento [1].

Entre as técnicas de reamostragem, a validação cruzada *k-fold* representa uma das mais utilizadas em problemas de *machine learning*. Essa técnica consiste na divisão aleatória do banco de treinamento em k partes de tamanhos iguais, em que $k-1$ irá compor os dados de treinamento para o ajuste de modelos, e a outra parte ficará reservada para a estimativa de sua performance. O processo continua até que todas as partes tenham participado tanto do treinamento como da validação do modelo, resultando em k estimativas de performance. Repetições desse processo podem ser utilizadas para aumentar a precisão dessas estimativas, de modo que, a cada repetição, diferentes partições do conjunto de treinamento são consideradas para compor cada uma das k partes do processo de validação cruzada [19].

Não há uma regra precisa para a escolha de k, embora a divisão dos dados em 5 ou 10 partes seja mais comum. Conforme k aumenta, a diferença de tamanho do conjunto de treinamento original e dos subconjuntos reamostrados se torna menor, e, à medida que essa diferença diminui, o viés da técnica de validação cruzada também se torna menor. Por outro lado, o tempo necessário para obter o resultado final da validação cruzada se torna maior [19]. Para a presente aplicação, utilizou-se validação cruzada com $k = 10$, repetida 10 vezes.

Aprendizado: medidas para avaliação de performance

Uma vez estabelecido o valor de k, é preciso definir uma medida para estimar a performance dos modelos ajustados. Tais medidas são importantes tanto na etapa de seleção quanto na de avaliação dos modelos preditivos. Seu cálculo objetiva mensurar o quanto o valor predito para uma observação

se aproxima de seu valor observado. Quando a resposta de interesse é uma variável categórica, dois tipos de predição podem ser obtidos: uma contínua (p_k^*), que é uma estimativa da probabilidade de a nova observação pertencer a cada uma das classes, k, $k = 1,2,\ldots, k$, da resposta de interesse, e outra categórica (por exemplo, 0: desfecho ausente e 1: desfecho presente), que é uma predição para o valor da resposta de uma nova observação [19,20].

As predições contínuas são especialmente interessantes por possibilitarem a utilização do classificador (modelo ajustado) em diferentes cenários, a partir do estabelecimento de pontos de corte de acordo com o interesse do pesquisador, em termos de sensibilidade (S) e especificidade (E). Ao definir um ponto de corte para p_k^*, é possível obter uma matriz de confusão, representada pela tabulação cruzada de classes observadas e preditas para os dados de teste, conforme ilustrado na Tabela 1, em que *a* e *d* denotam casos com resposta corretamente predita, e *b* e *c* representam erros de classificação. A sensibilidade ($a/(a+c)$) é a proporção de verdadeiros positivos (VP) entre todos os indivíduos cuja resposta de interesse foi observada, e a especificidade ($d/(b+d)$) refere-se à proporção de verdadeiros negativos (VN) entre aqueles com resposta de interesse ausente [20].

Um balanço entre sensibilidade e especificidade pode ser apropriado quando há diferentes penalidades associadas a cada tipo de erro. Nesse caso, a curva ROC (*receiver operating characteristic*) representa uma ferramenta adequada para avaliar a sensibilidade e a especificidade decorrentes de todos os pontos de corte possíveis para a probabilidade predita, de modo que o desempenho geral de um classificador pode ser avaliado pela área abaixo da curva (AUC) ROC: quanto maior a AUC (mais próxima de 1), melhor a performance do modelo [20]. A AUC ROC pode ser útil na comparação entre dois ou mais modelos com diferentes preditores, diferentes hiperparâmetros ou mesmo classificadores provenientes de algoritmos completamente distintos. Na etapa de seleção de modelos, a AUC ROC também é frequentemente utilizada como métrica para otimização dos hiperparâmetros no processo de validação cruzada *k-fold* [18,19].

Embora seja uma medida adequada para avaliar o poder discriminatório de modelos preditivos, a AUC ROC assume que um resultado falso positivo (FP) é tão ruim quanto um resultado falso negativo (FN) e, além disso, não informa a acurácia da magnitude global do risco predito. Assim, outras medidas, não diretamente relacionadas à performance preditiva, também são interessantes para avaliar um modelo, como aquelas referentes à calibração do risco predito – curva de calibração e distribuição da probabilidade predita –, que avaliam a habilidade do modelo em estimar, de forma acurada, a probabilidade do desfecho para um indivíduo, ou seja, medem a precisão com que a probabilidade predita corresponde à taxa de eventos observada em um conjunto de dados [20,21].

Seleção e avaliação de modelos para o risco de óbito em cinco anos

Para cada algoritmo dessa aplicação, exceto a regressão logística, uma lista de valores candidatos para os hiperparâmetros foi estabelecida, e, na sequência, utilizando validação cruzada *k-fold*, realizou-se a análise da performance preditiva de cada um desses modelos, por meio da AUC ROC, para selecionar aquele com melhor desempenho. Posteriormente, o modelo selecionado foi aplicado aos dados de

Tabela 1

Matriz de confusão para problemas de classificação com resposta dicotômica.

		Resposta observada		
		Presente	**Ausente**	
Resposta predita	Presente	Verdadeiro positivo (*a*)	Falso positivo (*b*)	*a* + *b*
	Ausente	Falso negativo (*c*)	Verdadeiro negativo (*d*)	*c* + *d*
		a + c	b + d	

teste para avaliar seu erro de predição em observações futuras, novamente utilizando a AUC ROC. Como a regressão logística não apresenta hiperparâmetros, essa foi ajustada uma única vez aos dados de treinamento, e, na sequência, o modelo com seus parâmetros ajustados foi avaliado nos dados de teste. Adicionalmente, curvas de calibração e a distribuição da probabilidade predita de óbito em cinco anos para as categorias da resposta de interesse no banco de teste foram avaliadas como medidas de calibração.

Estudo SABE

Como exemplo de aplicação dos algoritmos de machine learning para análise preditiva em saúde, foram analisados dados do estudo SABE, realizado com indivíduos de 60 anos ou mais, residentes no Município de São Paulo. Esse estudo teve início no ano 2000 com 2.143 indivíduos (amostra probabilística de 1.568 idosos distribuídos em 72 setores censitários, acrescida de 575 idosos para compensar a taxa mais elevada de mortalidade da faixa etária de 75 anos ou mais e do sexo masculino) [22].

Da amostra inicial, 1.115 pessoas foram novamente entrevistadas em 2006, e uma nova coorte probabilística de 298 indivíduos com idade entre 60 e 64 anos foi adicionada ao seguimento [23]. Por fim, no ano de 2010, uma nova onda foi conduzida com 978 idosos das coortes anteriores, e, adicionalmente, uma amostra probabilística de 367 pessoas com idade entre 60 a 64 anos passou a compor o estudo [24].

A coleta de dados foi realizada no domicílio dos idosos por meio de um questionário padrão, com perguntas relacionadas às condições de saúde, demográficas e socioeconômicas do idoso. Para a presente análise, foram utilizados dados da coorte de 2000 com acréscimo dos indivíduos que ingressaram no estudo SABE em 2006 e em 2010, resultando em 2.808 idosos. A variável resposta, de natureza binária, correspondeu à ocorrência de óbito em até cinco anos após o ingresso do idoso no estudo (n = 423).

A data do óbito foi obtida pelo pareamento probabilístico dos dados do estudo SABE com microdados do Sistema de Informações sobre Mortalidade, da Secretaria de Saúde do Município de São Paulo, para o período de janeiro de 2000 a setembro de 2016. Esse procedimento foi realizado com auxílio do software LinkPlus (https://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm), utilizando o nome, a data de nascimento e o nome da mãe como variáveis de comparação, o sexo como variável de blocagem e o endereço, quando necessário, para revisão manual.

Foram selecionadas 37 variáveis como potenciais preditores de óbito em idosos, de acordo com a literatura relacionada ao tema [25,26]. Os preditores foram organizados nos seguintes blocos:

a) Demográfico: idade (numérica contínua) e sexo (binária);

b) Socioeconômico: escolaridade em anos de estudo (nenhum, 1-3, 4-7, 8 e mais), percepção de suficiência de renda (binária), viver em lar unipessoal (binária);

c) Estado de saúde e características comportamentais: tabagismo (fuma, já fumou, nunca fumou), quantos dias na semana ingere bebidas alcoólicas (nenhum, menos de 1, 1-3, 4 ou mais), índice de massa corpórea (numérica contínua), número de refeições em um dia (1, 2, 3 ou mais), autoavaliação de saúde (excelente/muito boa/boa e regular/ruim/muito ruim);

d) Morbidades: doenças autorreferidas (hipertensão, diabetes, câncer, doença pulmonar, doença vascular cerebral, doença cardíaca e psiquiátrica), história de queda no último ano (binária) e comprometimento cognitivo (binária), avaliado e classificado de acordo com a versão modificada do *Mini Exame do Estado Mental* [27];

e) Condição de mobilidade (binárias, sim ou não): dificuldade para caminhar vários quarteirões, levantar de uma cadeira, subir vários lances de escada, agachar, ajoelhar ou se curvar, levantar os braços acima dos ombros, empurrar ou puxar objetos pesados, levantar objetos com mais de 5kg ou levantar uma moeda.

f) Estado funcional (binárias, sim ou não): dificuldade para se vestir, tomar banho, se alimentar, deitar/levantar da cama, ir ao banheiro, preparar refeições quentes, cuidar do próprio dinheiro, tomar os próprios remédios ou fazer tarefas domésticas pesadas.

Uma vez que o questionário do estudo SABE passou por modificações ao longo das três ondas, sempre que necessário, as variáveis foram recategorizadas para tornar as informações compatíveis. O estudo SABE obteve aprovação do Comitê de Ética em Pesquisa da Faculdade de Saúde Pública da Universi-

dade de São Paulo (nº dos pareceres: 315/99, 83/06, 2044). Todas as etapas da modelagem preditiva foram realizadas com auxílio do software R. Para a etapa de aprendizado dos modelos, foram utilizadas funções disponíveis no pacote CARET (*Classification and Regression Training*).

Resultados

A Tabela 2 apresenta resultados das etapas de aprendizado (treinamento) e avaliação (teste) dos modelos preditivos. A métrica utilizada para otimização e seleção de modelos durante o aprendizado foi a AUC ROC, portanto, o modelo com maior AUC ROC, para cada algoritmo, foi o escolhido para a avaliação de performance nos dados de teste. Em relação ao *ranking* de importância dos preditores (cinco mais importantes) para os modelos selecionados ao fim da etapa de aprendizado, observa-se que diferentes abordagens produziram *rankings* com preditores similares, porém em sequência diferente.

Por fim, a avaliação de performance dos modelos selecionados evidenciou desempenho satisfatório, com AUC ROC superior a 0,70 para todos os modelos. Os melhores resultados foram observados para os modelos de redes neurais, regressão logística com penalização de lasso (*least absolute shrinkage and selection operator*) e regressão logística simples, respectivamente, sem grandes diferenças entre eles.

Em relação à interpretação das medidas apresentadas para avaliação de performance dos modelos em dados de teste, serão apresentados, a seguir, três cenários, utilizando, como exemplo, as redes neurais. Primeiramente, observa-se que, devido à baixa frequência de óbito na amostra analisada (15%), o ponto de corte para a probabilidade predita de 0,50 resulta em especificidade elevada (0,98), porém em baixa sensibilidade (0,08).

Para o ponto de corte que maximiza a sensibilidade e a especificidade (p ótimo igual a 0,143), pode-se afirmar que, com uma frequência de falso positivos de 30%, é possível prever, com sucesso, 70% dos idosos que irão morrer dentro de 5 anos, ou seja, dos 118 óbitos presentes nos dados de teste, 83 seriam corretamente preditos.

Por fim, outra alternativa para a escolha de p é assumir que intervenções devem ser aplicadas a 10% dos idosos com risco predito mais alto para óbito em 5 anos ($p \geq 0{,}316$ para as redes neurais). Nesse caso, com uma frequência de falso positivos de 7,02%, seria possível prever, com sucesso, 27,12% dos idosos que irão morrer dentro de 5 anos, ou seja, dos 118 óbitos presentes nos dados de teste, 32 seriam corretamente identificados.

Antes da escolha do modelo final a ser utilizado na prática para predizer óbito em cinco anos em observações futuras, é importante avaliar a consistência da probabilidade predita quando comparada à taxa de eventos observada para a resposta de interesse. Essa análise pode ser realizada por meio da distribuição da probabilidade predita para a ocorrência de óbito em cinco anos, segundo as respostas observadas para o desfecho, e por meio da construção de uma curva de calibração, conforme apresentado na Figura 2, para o modelo de redes neurais e os modelos de regressão logística com e sem penalização.

A distribuição da probabilidade predita de óbito em 5 anos dos três modelos apresentou comportamento semelhante, com distribuição assimétrica, concentrada à esquerda, tanto para aqueles que não morreram em cinco anos (n = 684) como para os que sofreram esse desfecho (n = 118). Em relação à curva de calibração, a probabilidade de óbito em cinco anos para os três modelos apresentou acurácia elevada até o 6º decil, a partir do qual reduziu substancialmente, em especial para o modelo de redes neurais que apresentou probabilidade predita inferior 0,64 e, portanto, para os decis subsequentes, qualidade inferior a dos modelos de regressão logística com e sem penalização.

Por fim, após a avaliação de performance dos modelos selecionados, a regressão logística com penalização de lasso representa uma escolha interessante como modelo final para utilização em dados futuros, por sua menor complexidade quando comparada à regressão logística usual e às redes neurais. Caso seja esse o modelo escolhido, as estimativas de seus parâmetros a serem utilizadas na predição da resposta de interesse em novas observações podem ser definidas por meio do ajuste de um modelo de regressão logística penalizada de lasso ao conjunto de dados completo, utilizando os hiperparâmetros otimizados na etapa de treinamento.

Tabela 2

Aprendizado (treinamento) e avaliação (teste) de modelos preditivos.

Algoritmos	Regressão logística	Regressão logística com penalização	Redes neurais	*Gradient boosted trees*	*Random forest*
TREINAMENTO (resultados para o melhor modelo)					
Hiperparâmetros otimizados	-	*alpha* = 1 *lambda* = 0,003	*size* = 3 *decay* = 2	*nrounds* = 100 *maxdepth* = 1 *eta* = 0,3	*mtry* = 7
AUC ROC Média (dp) da validação cruzada	0,803 *	0,766 (0,07)	0,767 (0,06)	0,765 (0,07)	0,738 (0,05)
Ranking de importância das variáveis					
1	Idade	Dificuldade para tomar banho	Idade	Idade	Idade
2	Consumo de tabaco	Idade	Dificuldade para tomar banho	Dificuldade para tomar banho	*Mini Exame do Estado Mental*
3	Dificuldade para tomar banho	Consumo de tabaco	Consumo de tabaco	Índice de massa corporal	Dificuldade para ir ao banheiro
4	Sexo	Dificuldade para comer	Dificuldade para ir ao banheiro	Sexo	Dificuldade para tomar banho
5	Diabetes mellitus	Sexo	Diabetes mellitus	*Mini Exame do Estado Mental*	Sexo
Teste					
AUC (IC95%)	0,773 (0,732; 0,814)	0,777 (0,735; 0,818)	0,779 (0,738; 0,820)	0,768 (0,724; 0,813)	0,744 (0,699; 0,789)
Pontos de corte para *p* (risco predito)					
p = 0,5					
S (VP)	0,144 (17)	0,130 (15)	0,08 (10)	0,144 (17)	0,110 (13)
1 - E (FP)	0,03 (21)	0,026 (18)	0,020 (13)	0,026 (18)	0,019 (13)
p ótimo **					
S (VP)	0,653 (77)	0,670 (79)	0,703 (83)	0,712 (84)	0,700 (79)
1 - E (FP)	0,251 (172)	0,253 (173)	0,300 (205)	0,304 (208)	0,328 (238)
p 10% com risco mais alto **					
S (VP)					
1 - E (FP)					
Óbitos observados segundo o risco predito					
n (%) entre os 10% com risco mais alto	34 (28,814)	36 (30,508)	32 (27,119)	34 (28,814)	33 (27,966)
n (%) entre os 10% com risco mais baixo ***	-	-	-	-	-

dp: desvio padrão; E: especificidade; FP: falso positivo; IC95%: intervalo de 95% de confiança; S: sensibilidade; VP: verdadeiro positivo.

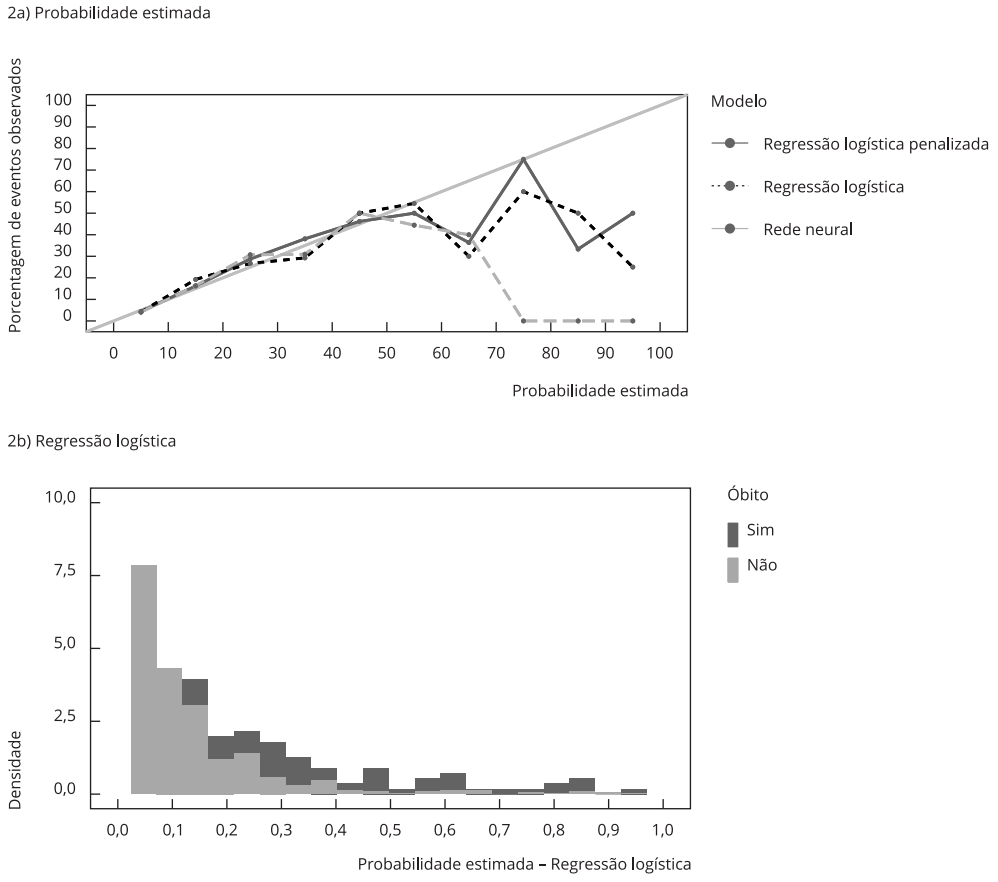
* AUC ROC do ajuste do modelo uma única vez aos dados de treinamento;

** Ponto de corte que maximiza a sensibilidade e a especificidade. Varia de acordo com o modelo preditivo;

*** Nenhum óbito observado entre os 10% de indivíduos com menor risco para o desfecho.

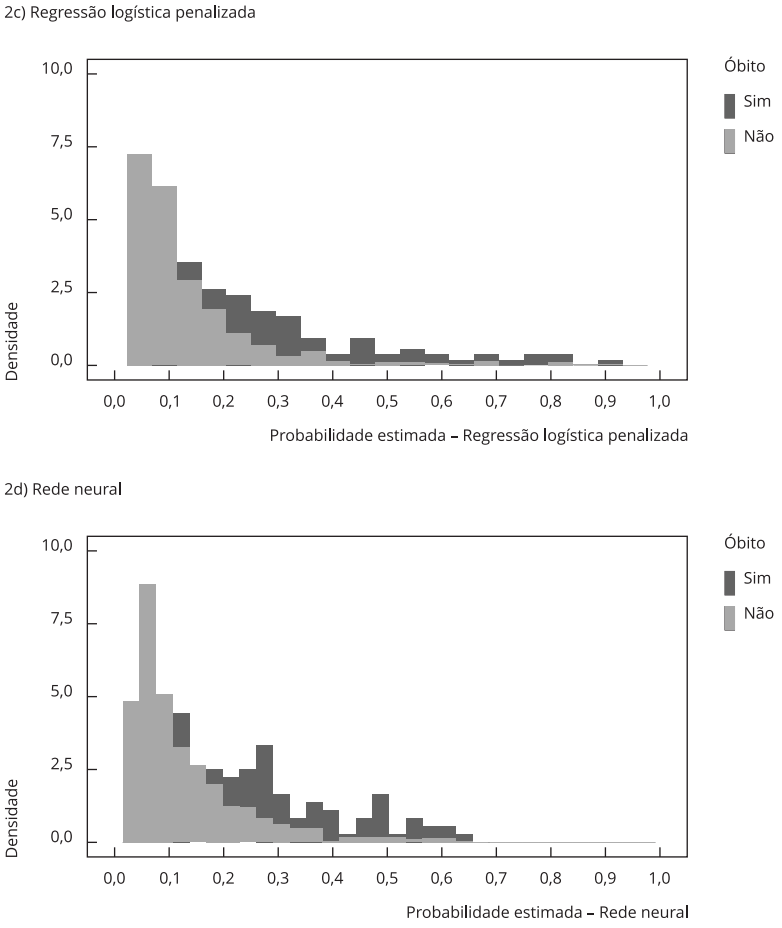
Figura 2

Histogramas para as probabilidades estimadas e curva de calibração de dados de teste para regressão logística, regressão logística penalizada e redes neurais.



(continua)

Figura 2 (continuação)



Discussão e conclusões

A presente aplicação utilizou dados do estudo SABE para exemplificar as etapas envolvidas na utilização de algoritmos de *machine learning* para análises preditivas em saúde. Sabe-se que a performance de um modelo preditivo é dependente do conhecimento relacionado ao problema que se deseja estudar e da disponibilidade de variáveis informativas e com poder discriminatório em relação à resposta de interesse. Portanto, de acordo com literatura relacionada à predição de morte em idosos, foram selecionados preditores sociodemográficos, comportamentais e referentes a doenças, mobilidade e limitações funcionais. Características sobre acesso e utilização de serviços de saúde não foram adicionadas por incompatibilidade das variáveis nas três coortes do estudo. De modo geral, todos os algoritmos avaliados alcançaram AUC ROC superior a 0,70 nos dados de teste, e os melhores resultados foram observados para os modelos de redes neurais, regressão logística penalizada de lasso e regressão logística simples, respectivamente. Resultados semelhantes foram observados por Olivera et al. [15] no ajuste de modelos preditivos de diabetes não diagnosticada em que os melhores resultados de performance foram observados para o modelo de redes neurais e de regressão logística, sem diferenças relevantes.

Diferenças na performance de modelos preditivos podem ser atribuídas às características dos dados, como a definição da resposta de interesse e a disponibilidade de variáveis candidatas a predi-

tores, e às técnicas e aos métodos utilizados para construir e avaliar os modelos, como o refinamento de possíveis valores para os hiperparâmetros dos algoritmos mais flexíveis, na etapa de treinamento, e à definição dos bancos de treinamento e teste. Nesse último caso, como a definição dos bancos é realizada de modo aleatório, dependendo da amostra que compuser os respectivos conjuntos de dados, pode haver variações no resultado de performance, em especial para pequenos conjuntos de dados. Uma possível solução, voltada à melhoria da precisão das estimativas de performance, está relacionada à utilização de validação cruzada aninhada (*nested cross validation*) [28], em que um processo de validação cruzada interno é realizado com o objetivo de estimar os hiperparâmetros e selecionar modelos, e um processo de validação cruzada externo é realizado para avaliar o erro de predição dos modelos selecionados.

Outra característica que pode influenciar o desempenho geral de modelos preditivos em saúde é o fato de a resposta de interesse apresentar distribuição desbalanceada entre suas classes, ou seja, maior frequência para a classe de indivíduos com desfecho ausente e, por outro lado, uma classe minoritária, representativa de indivíduos com resposta positiva que, geralmente, é a classe mais importante, mas também a que sofre com mais classificações erradas. Tal situação ocorre porque os algoritmos, para alcançar melhor desempenho (mais acertos), tendem a priorizar especificidade em vez de sensibilidade [29]. Algumas alternativas para reduzir o impacto do desbalanceamento na performance preditiva são representadas pela escolha de pontos de corte alternativos para a probabilidade predita, para reduzir a taxa de erro da classe menos frequente; pela modificação dos pesos (probabilidade *a priori*) das observações de treinamento cuja resposta de interesse está presente; e por aplicação de técnicas de reamostragem post hoc (*up* e *down sampling* e SMOTE – *Synthetic Minority Over-sampling Technique*), utilizadas para balancear as classes da resposta de interesse [19].

É possível resumir o efeito de preditores individuais por meio de métricas específicas, denominadas importância das variáveis. Embora essa medida não tenha interpretação causal ou inferencial, essa reflete, de forma ordenada, quais variáveis contribuíram mais para a performance do modelo. Uma vez que cada algoritmo ajusta o modelo de modo diferente, espera-se que os rankings correspondentes também apresentem diferenças não só na ordenação, mas também nas variáveis que os compõem, conforme observado na presente análise [8].

Ainda que os modelos tenham sido ajustados a partir de dados relevantes, e que apresentem boa performance preditiva, eles podem gerar predições imprecisas. Entre outras características, esse cenário pode estar relacionado à disponibilidade de um número reduzido de observações, sobretudo de observações com desfecho presente, ou de preditores para o treinamento e, mais frequentemente, ao sobreajuste do modelo para os dados existentes, no caso de algoritmos mais flexíveis [19]. Esse é o caso do estudo atual, em que os algoritmos que costumam apresentar melhor desempenho em grandes bancos de dados, como *random forest* e *gradient boosted trees*, apresentaram as piores performances preditivas. Vale destacar também que modelos mais flexíveis e acurados não são diretamente interpretáveis, ou seja, não é possível representar explicitamente como cada preditor está individualmente associado à resposta de interesse. Logo, em alguns problemas da área da saúde, um balanço entre acurácia e interpretabilidade pode ser desejável na escolha do modelo preditivo final, com o objetivo de facilitar a sua adoção por profissionais de saúde [30].

Em relação à qualidade do risco predito, mesmo que um modelo não apresente boa calibração, ainda pode ser útil em aplicações específicas, como para determinado ponto de corte da probabilidade predita ou para cenários em que a sensibilidade ou a especificidade seja mais relevante. Por exemplo, em casos de utilização de determinado modelo para uma atividade de triagem, em que o objetivo é discriminar indivíduos com risco muito baixo para determinado desfecho. Embora esse modelo possa não apresentar calibração perfeita, ou seja, possa resultar em estimativas incorretas do risco para indivíduos não classificados como de risco muito baixo, o modelo ainda será apropriado para o objetivo a que foi proposto se ele apresentar poder discriminatório aceitável (AUC ROC satisfatória) [20]. Além disso, se for do interesse do pesquisador, é possível recalibrar a probabilidade predita, por exemplo, por meio de um modelo adicional que corrija o padrão observado em sua distribuição.

É importante destacar que mesmo um modelo preditivo com bom poder discriminatório e bem calibrado pode não se traduzir em melhores cuidados à saúde, pois uma predição acurada não diz o que deve ser feito para modificar o desfecho sob análise [14]. Além disso, modelos preditivos de óbito, bem como de doenças crônicas podem basear-se não só em fatores de risco modificáveis, mas também

em características biológicas não modificáveis, como idade e sexo, que, embora contribuam para a performance preditiva do modelo, podem não ser relevantes em estratégias de prevenção ou controle [29].

Os resultados dessa aplicação estão restritos à referida população, que tem como principal limitação o fato de ser uma amostra relativamente pequena. Espera-se que, com o aumento da disponibilidade e qualidade de dados de óbitos, seja possível melhorar, consideravelmente, a performance desses algoritmos. No entanto, ainda que a amostra seja pequena, atende à suposição de ausência de viés de seleção, ou seja, seus dados são independentes e identicamente distribuídos em relação àqueles que serão observados no futuro e seguem a mesma distribuição da população de onde a amostra foi retirada.

Modelos preditivos baseados em algoritmos de *machine learning* têm sido aplicados na área da saúde com potencial para auxiliar profissionais de saúde na tomada de melhores decisões, como mostram estudos recentes [12,31]. Entretanto, seu sucesso depende da disponibilidade de dados para a etapa de aprendizado de modelos preditivos e de capital humano para entender e desenvolver esses modelos de forma rigorosa e transparente.

Colaboradores

H. G. Santos, C. F. Nascimento e A. D. P. Chiavegatto Filho participaram da concepção do estudo, organização e análise dos dados, interpretação dos resultados, redação e revisão crítica e aprovação final do manuscrito. R. Izbicki contribuiu com a interpretação dos resultados, redação e revisão crítica e aprovação da versão final do manuscrito. Y. A. O. Duarte contribuiu com revisão crítica e aprovação da versão final do manuscrito.

Informações adicionais

ORCID: Hellen Geremias dos Santos (0000-0001-7039-8585); Carla Ferreira do Nascimento (0000-0002-0054-277X); Rafael Izbicki (0000-0003-0379-9690); Yeda Aparecida de Oliveira Duarte (0000-0003-3933-2179); Alexandre Dias Porto Chiavegatto Filho (0000-0003-3251-9600).

Agradecimentos

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) por bolsa de estudo de doutorado. À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (processo 17/09369-8) e à Fundação Lemann (Harvard Brazil Research Fund) pelo financiamento do estudo.

Referências

1. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2008.
2. Pepe MS. Evaluating technologies for classification and prediction in medicine. Stat Med 2005; 24:3687-96.
3. Steyerberg EW. Clinical prediction models: a practical approach to development, validation and updating. New York: Springer; 2009.
4. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. Am J Cardiol 1976; 38:46-51.
5. D'Agostino RB Sr, Grundy S, Sullivan LM, Wilson P. Validation of the Framingham coronary heart disease prediction score: results of a multiple ethnic groups investigation. JAMA 2001; 286:180-7.
6. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive summary of the third report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). JAMA 2001; 285:2486-97.
7. Gail MH. Twenty-five years of breast cancer risk models and their applications. J Natl Cancer Inst 2015; 107:6-11.
8. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J 2017; 38:1805-14.
9. Mullainathan S, Spiess J. Machine learning: an applied econometric approach. J Econ Perspect 2017; 31:87-106.

10. Rose S. Mortality risk score prediction in an elderly population using machine learning. Am J Epidemiol 2013; 177:443-52.
11. Jamei M, Nisnevich A, Wetchler E, Sudat S, Liu E. Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. PLoS One 2017; 12:e0181173.
12. Pan I, Nolan LB, Brown RR, Khan R, van der Boor P, Harris DG, et al. Machine learning for social services: a study of prenatal case management in Illinois. Am J Public Health 2017; 107:938-44.
13. Obermeyer Z, Emanuel EJ. Predicting the future: big data, machine learning, and clinical medicine. N Engl J Med 2016; 375:1216-9.
14. Chen JH, Asch SM. Machine Learning and prediction in medicine: beyond the peak of inflated expectations. N Engl J Med 2017; 376:2507-9.
15. Olivera AR, Roesler V, Iochpe C, Schimidt MI, Vigo A, Barreto SM, et al. Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes - ELSA-Brasil: accuracy study. São Paulo Med J 2017; 135:234-46.
16. Raschka S. Python machine learning. Birmingham: Packt Publishing; 2015.
17. Izbicki R, Santos TM. *Machine Learning* sob a ótica estatística: uma abordagem preditivista para a estatística com exemplos em R, 2018 [versão em desenvolvimento]. http://www.rizbicki.ufscar.br/sml.pdf (acessado em 25/Jun/2019).
18. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. New York: Springer; 2014.
19. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013.
20. Meurer WJ, Tolles J. Logistic regression diagnostics: understanding how well a model predicts outcomes. JAMA 2017; 317:1068-9.
21. Pencina MJ, D'Agostino Sr. RB. Evaluating discrimination of risk prediction models: the C statistic. JAMA 2015; 314:1063-4.
22. Lebrão ML, Duarte YAO, organizadores. O Projeto SABE no município de São Paulo: uma abordagem inicial. Brasília: Organização Pan -Americana da Saúde; 2013.
23. Lebrão ML, Duarte YAO. Desafios de um estudo longitudinal: o Projeto SABE. Saúde Colet (Barueri, Impr.) 2008; 5:166-7.
24. Corona LP, Duarte YA de O, Lebrão ML. Prevalence of anemia and associated factors in older adults: evidence from the SABE Study. Rev Saúde Pública 2014; 48:723-31.
25. Yourman LC, Lee SJ, Schonberg MA, Widera EW, Smith AK. Prognostic indices for older adults: a systematic review. JAMA 2012; 307:182-92.
26. Suemoto CK, Ueda P, Beltrán-Sánchez H, Lebrão ML, Duarte YA, Wong R, et al. Development and validation of a 10-year mortality prediction model: meta-analysis of individual participant data from five cohorts of older adults in developed and developing countries. J Gerontol A Biol Sci Med Sci 2017; 72:410-6.
27. Icaza MG, Albala C. Minimental State Examination (MMSE) del estudio de dementia en Chile: análisis estatístico. Washington DC: Pan American Health Organization; 1999.
28. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 2006; 7:91.
29. Mena LJ, Orozco EE, Felix VG, Ostos R, Melgarejo J, Maestre GE. Machine learning approach to extract diagnostic and prognostic thresholds: application in prognosis of cardiovascular mortality. Comput Math Methods Med 2012; 2012:750151.
30. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: Proceeding of the 21th ACM SigkDD International Conference on Knowlwdge Discovery and Date Mining. Sidney: Association for Computing Machinery; 2015; p. 1721-30.
31. Kessler RC, Rose S, Koenen KC, Karam EG, Stang PE, Stein DJ, et al. How well can post-traumatic stress disorder be predicted from pre-trauma risk factors? an exploratory study in the WHO World Mental Health Surveys. World Psychiatry 2014; 13:265-74.
32. Raschka S. Mirjalili V. Python machine learning: machine learning and deep learning with python, scikit-learn, and tensorflow. 2nd Ed. Birmingham: Packt Publishing; 2017.

Abstract

This study aims to present the stages related to the use of machine learning algorithms for predictive analyses in health. An application was performed in a database of elderly residents in the city of São Paulo, Brazil, who participated in the Health, Well-Being, and Aging Study *(SABE) (n = 2,808). The outcome variable was the occurrence of death within five years of the elder's entry into the study (n = 423), and the predictors were 37 variables related to the elder's demographic, socioeconomic, and health profile. The application was organized according to the following stages: division of data in training (70%) and testing (30%), pre-processing of the predictors, learning, and assessment of the models. The learning stage used 5 algorithms to adjust the models: logistic regression with and without penalization, neural networks, gradient boosted trees, and random forest. The algorithms' hyperparameters were optimized by 10-fold cross-validation to select those corresponding to the best models. For each algorithm, the best model was assessed in test data via area under the ROC curve (AUC) and related measures. All the models presented AUC ROC greater than 0.70. For the three models with the highest AUC ROC (neural networks and logistic regression with LASSO penalization and without penalization, respectively), quality measures of the predicted probability were also assessed. The expectation is that with the increased availability of data and trained human capital, it will be possible to develop predictive machine learning models with the potential to help health professionals make the best decisions.*

Forecasting; Mortality; Aged

Resumen

El objetivo de este estudio fue presentar las etapas relacionadas con la utilización de algoritmos de machine learning *para análisis predictivos en salud. Para tal fin, se realizó una aplicación en base a datos de ancianos residentes en el Municipio de São Paulo, Brasil, participantes en el estudio* Salud Bienestar y Envejecimiento *(SABE) (n = 2.808). La variable respuesta se representó mediante la ocurrencia de óbito en hasta 5 años tras la inclusión del anciano en el estudio (n = 423), y los predictores fueron representados por 37 variables relacionadas con el perfil demográfico, socioeconómico y de salud del anciano. El aplicación se organizó según las siguientes etapas: división de los datos en formación (70%) y test (30%), preprocesamiento de los predictores, aprendizaje y evaluación de modelos. En la etapa de aprendizaje, se utilizaron cinco algoritmos para el ajuste de modelos: regresión logística con y sin penalización, redes neuronales, gradient boosted trees y random forest. Los hiperparámetros de los algoritmos se optimizaron mediante una validación cruzada 10-*fold*, para seleccionar aquellos correspondientes a los mejores modelos. Para cada algoritmo, el mejor modelo se evaluó con datos de la prueba del área debajo de la curva (AUC) ROC y medidas relacionadas. Todos los modelos presentaron AUC ROC superior a 0,70. Para los tres modelos con mayor AUC ROC (redes neuronales y regresión logística con penalización de Lasso y sin penalización, respectivamente) también se evaluaron medidas de calidad de la probabilidad pronosticada. Se espera que, con el aumento de la disponibilidad de datos y de capital humano capacitado, sea posible desarrollar modelos predictivos de* machine learning *con potencial para ayudar a profesionales de salud en la toma de mejores decisiones.*

Predicción; Mortalidad; Anciano

Recebido em 13/Mar/2018
Versão final reapresentada em 14/Fev/2019
Aprovado em 20/Mai/2019

Santos HG, Nascimento CF, Izbicki R, Duarte YAO, Chiavegatto Filho ADP. *Machine learning* para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. Cad Saúde Pública 2019; 35(7):e00050818.

doi: 10.1590/0102-311XER050818

A revista foi informada sobre erros na Tabela 2. As correções seguem abaixo:

Onde se lê:

Tabela 2

Aprendizado (treinamento) e avaliação (teste) de modelos preditivos.

Algoritmos	**Regressão logística**	**Regressão logística com penalização**	**Redes neurais**	***Gradient boosted trees***	***Random Forest***
TREINAMENTO (resultados para o melhor modelo)					
Hiperparâmetros otimizados	-	*alpha* = 1 *lambda* = 0,003	*size* = 3 *decay* = 2	*nrounds* = 100 *maxdepth* = 1 *eta* = 0,3	*mtry* = 7
AUC ROC Média (dp) da validação cruzada	0,803 *	0,766 (0,07)	0,767 (0,06)	0,765 (0,07)	0,738 (0,05)
Ranking de importância das variáveis					
1	Idade	Dificuldade para tomar banho	Idade	Idade	Idade
2	Consumo de tabaco	Idade	Dificuldade para tomar banho	Dificuldade para tomar banho	*Mini Exame do Estado Mental*
3	Dificuldade para tomar banho	Consumo de tabaco	Consumo de tabaco	Índice de massa corporal	Dificuldade para ir ao banheiro
4	Sexo	Dificuldade para comer	Dificuldade para ir ao banheiro	Sexo	Dificuldade para tomar banho
5	Diabetes mellitus	Sexo	Diabetes mellitus	*Mini Exame do Estado Mental*	Sexo
Teste					
AUC (IC95%)	0,773 (0,732; 0,814)	0,777 (0,735; 0,818)	0,779 (0,738; 0,820)	0,768 (0,724; 0,813)	0,744 (0,699; 0,789)
Pontos de corte para *p* (risco predito)					
p = 0,5					
S (VP)	0,144 (17)	0,130 (15)	0,08 (10)	0,144 (17)	0,110 (13)
1 - E (FP)	0,03 (21)	0,026 (18)	0,020 (13)	0,026 (18)	0,019 (13)
p ótimo **					
S (VP)	0,653 (77)	0,670 (79)	0,703 (83)	0,712 (84)	0,700 (79)
1 - E (FP)	0,251 (172)	0,253 (173)	0,300 (205)	0,304 (208)	0,328 (238)

(continua)

Tabela 2 (continuação)

Algoritmos	Regressão logística	Regressão logística com penalização	Redes neurais	*Gradient boosted trees*	*Random Forest*
p 10% com risco mais alto **					
S (VP)					
1 - E (FP)					
Óbitos observados segundo o risco predito					
n (%) entre os 10% com risco mais alto	34 (28,814)	36 (30,508)	32 (27,119)	34 (28,814)	33 (27,966)
n (%) entre os 10% com risco mais baixo ***	-	-	-	-	-

dp: desvio padrão; E: especificidade; FP: falso positivo; IC95%: intervalo de 95% de confiança; S: sensibilidade; VP: verdadeiro positivo.
* AUC ROC do ajuste do modelo uma única vez aos dados de treinamento;
** Ponto de corte que maximiza a sensibilidade e a especificidade. Varia de acordo com o modelo preditivo;
*** Nenhum óbito observado entre os 10% de indivíduos com menor risco para o desfecho.

Leia-se:

Tabela 2

Aprendizado (treinamento) e avaliação (teste) de modelos preditivos.

Algoritmos	Regressão logística	Regressão logística com penalização	Redes neurais	*Gradient boosted trees*	*Random Forest*
TREINAMENTO (resultados para o melhor modelo)					
Hiperparâmetros otimizados	-	*alpha* = 1 *lambda* = 0,003	*size* = 3 *decay* = 2	*nrounds* = 100 *maxdepth* = 1 *eta* = 0,3	*mtry* = 7
AUC ROC Média (dp) da validação cruzada	0,803 *	0,766 (0,07)	0,767 (0,06)	0,765 (0,07)	0,738 (0,05)
Ranking de importância das variáveis					
1	Idade	Dificuldade para tomar banho	Idade	Idade	Idade
2	Consumo de tabaco	Idade	Dificuldade para tomar banho	Dificuldade para tomar banho	*Mini Exame do Estado Mental*
3	Dificuldade para tomar banho	Consumo de tabaco	Consumo de tabaco	Índice de massa corporal	Dificuldade para ir ao banheiro
4	Sexo	Dificuldade para comer	Dificuldade para ir ao banheiro	Sexo	Dificuldade para tomar banho
5	Diabetes mellitus	Sexo	Diabetes mellitus	*Mini Exame do Estado Mental*	Sexo
Teste					
AUC (IC95%)	0,773 (0,732; 0,814)	0,777 (0,735; 0,818)	0,779 (0,738; 0,820)	0,768 (0,724; 0,813)	0,744 (0,699; 0,789)

(continua)

Tabela 2 (continuação)

Algoritmos	Regressão logística	Regressão logística com penalização	Redes neurais	*Gradient boosted trees*	*Random Forest*
Pontos de corte para *p* (risco predito)					
p = 0,5					
S (VP)	0,144 (17)	0,130 (15)	0,08 (10)	0,144 (17)	0,110 (13)
1 - E (FP)	0,03 (21)	0,026 (18)	0,020 (13)	0,026 (18)	0,019 (13)
p ótimo **					
S (VP)	0,653 (77)	0,670 (79)	0,703 (83)	0,712 (84)	0,700 (79)
1 - E (FP)	0,251 (172)	0,253 (173)	0,300 (205)	0,304 (208)	0,328 (238)
p 10% com risco mais alto					
S (VP)	0,288 (34)	0,305 (36)	0,271 (32)	0,288 (34)	0,280 (33)
1 - E (FP)	0,067 (46)	0,064 (44)	0,070 (48)	0,066 (45)	0,066 (45)
Óbitos observados segundo o risco predito					
n (%) entre os 10% com risco mais alto	34 (28,814)	36 (30,508)	32 (27,119)	34 (28,814)	33 (27,966)
n (%) entre os 10% com risco mais baixo ***	-	-	-	-	-

dp: desvio padrão; E: especificidade; FP: falso positivo; IC95%: intervalo de 95% de confiança; S: sensibilidade; VP: verdadeiro positivo.

* AUC ROC do ajuste do modelo uma única vez aos dados de treinamento;

** Ponto de corte que maximiza a sensibilidade e especificidade. Varia de acordo com o modelo preditivo;

*** Nenhum óbito observado entre os 10% de indivíduos com menor risco para o desfecho.