Universidade de São Paulo
Faculdade de Saúde Pública

Hellen Geremias dos Santos

**Comparação da performance de algoritmos
de *machine learning* para a análise preditiva
em saúde pública e medicina**

São Paulo
2018

HELLEN GEREMIAS DOS SANTOS

**Comparação da performance de algoritmos de
machine learning para a análise preditiva em saúde
pública e medicina**

Tese apresentada ao Programa de Pós-Graduação em Epidemiologia da Faculdade de Saúde Pública da Universidade de São Paulo para obtenção do título de Doutor em Ciências.

Área de concentração: métodos e técnicas de análise em epidemiologia.

Orientador: Prof. Dr. Alexandre Dias Porto Chiavegatto Filho.

São Paulo
2018

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Catálogo da Publicação

Ficha elaborada pelo Sistema de Geração Automática a partir de dados fornecidos pelo(a) autor(a)
Bibliotecária da FSP/USP: Maria do Carmo Alvarez - CRB-8/4359

Geremias dos Santos, Hellen
Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina / Hellen Geremias dos Santos; orientador Alexandre Dias Porto Chiavegatto Filho. -- São Paulo, 2018. 206 p.

Tese (Doutorado) -- Faculdade de Saúde Pública da Universidade de São Paulo, 2018.

1. Modelos de predição. 2. Expectativa de Vida. 3. Mortalidade. 4. Prognóstico. 5. Qualidade de vida. I. Dias Porto Chiavegatto Filho, Alexandre, orient. II. Título.

AGRADECIMENTOS

A Deus por sua providência em dar sentido e meta às minhas escolhas, e por sua ação constante em minha vida como cuidador, torcedor e guia.

Aos meus pais, Gisneide e Helio, e ao meu irmão, Wesley, pelo amor, carinho e apoio incondicional durante minha trajetória de vida – pessoal, acadêmica e profissional. São eles os mediadores de todas as minhas conquistas.

Ao meu namorado, Marcio, pelo amor, amizade e companheirismo.

Às amigas, Danielli e Agatha, pela companhia, carinhosa e alegre, durante minha estada em São Paulo.

À Professora Selma Maffei de Andrade e à amiga Francine pela motivação e participação ímpar na minha formação como epidemiologista e pesquisadora.

Aos amigos pós-graduandos Francimário, Sitso, Shu, Marina, Carla, Ilana, Patrícia, Joana, Alejandra, Elisangela, Luciana e Priscila, por sua companhia, incentivo, gentileza e disposição em me ajudar durante o doutorado.

Ao meu orientador, Alexandre Dias Porto Chiavegatto Filho, pela competência com que conduziu o desenvolvimento da tese e o meu progresso e qualificação como pesquisadora.

Aos autores dos artigos que compuseram a tese pela oportunidade de aprendizado e contribuições valiosas durante a elaboração e revisão dos manuscritos.

À Renilda Shimono e à Vânia dos Santos Silva, pelo acolhimento, atenção e ajuda dispensados durante o doutorado.

Aos professores da Faculdade de Saúde Pública pela excelência e compromisso dedicados à minha formação como docente e pesquisadora.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de estudos, que possibilitou minha dedicação exclusiva ao doutorado.

SANTOS, H. G. **Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina.** 2018. 206f. Tese (Doutorado) – Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, 2018.

RESUMO

Modelos preditivos estimam o risco de eventos ou agravos relacionados à saúde e podem ser utilizados como ferramenta auxiliar em tomadas de decisão por gestores e profissionais de saúde. Algoritmos de *machine learning* (ML), por sua vez, apresentam potencial para identificar relações complexas e não-lineares presentes nos dados, com consequências positivas na performance preditiva desses modelos. A presente pesquisa objetivou aplicar técnicas supervisionadas de ML e comparar sua performance em problemas de classificação e de regressão para prever respostas de interesse para a saúde pública e a medicina. Os resultados e discussão estão organizados em três artigos científicos. O primeiro apresenta um tutorial para o uso de ML em pesquisas de saúde, utilizando como exemplo a predição do risco de óbito em até 5 anos (frequência do desfecho 15%; n=395) para idosos do estudo “Saúde, Bem-estar e Envelhecimento” (n=2.677), segundo variáveis relacionadas ao seu perfil demográfico, socioeconômico e de saúde. Na etapa de aprendizado, cinco algoritmos foram aplicados: regressão logística com e sem penalização, redes neurais, *gradient boosted trees* e *random forest*, cujos hiperparâmetros foram otimizados por validação cruzada (VC) 10-fold. Todos os modelos apresentaram área abaixo da curva (AUC) ROC (*Receiver Operating Characteristic*) maior que 0,70. Para aqueles com maior AUC ROC (redes neurais e regressão logística com e sem penalização) medidas de qualidade da probabilidade predita foram avaliadas e evidenciaram baixa calibração. O segundo artigo objetivou prever o risco de tempo de vida ajustado pela qualidade de vida de até 30 dias (frequência do desfecho 44,7%; n=347) em pacientes com câncer admitidos em Unidade de Terapia Intensiva (UTI) (n=777), mediante características obtidas na admissão do paciente à UTI. Seis algoritmos (regressão logística com e sem penalização, redes neurais, árvore simples, *gradient boosted trees* e *random forest*) foram utilizados em conjunto com VC aninhada para estimar hiperparâmetros e avaliar performance preditiva. Todos os algoritmos, exceto a árvore simples, apresentaram discriminação (AUC ROC > 0,80) e calibração satisfatórias. Para o terceiro artigo, características socioeconômicas e demográficas foram utilizadas para prever a expectativa de vida ao nascer de municípios brasileiros com mais de 10.000 habitantes (n=3.052). Para o ajuste do modelo preditivo, empregou-se VC aninhada e o algoritmo *Super Learner* (SL), e para a avaliação de performance, o erro quadrático médio (EQM). O SL apresentou desempenho satisfatório (EQM=0,17) e seu vetor de valores preditos foi utilizado para a identificação de *overachievers* (municípios com expectativa de vida superior à predita) e *underachievers* (município com expectativa de vida inferior à predita), para os quais características de saúde foram comparadas, revelando melhor desempenho em indicadores de atenção primária para os *overachievers* e em indicadores de atenção secundária para os *underachievers*. Técnicas para a construção e avaliação de modelos preditivos estão em constante evolução e há poucas justificativas teóricas para se preferir um algoritmo em lugar de outro. Na presente tese, não foram observadas diferenças substanciais no desempenho preditivo dos algoritmos aplicados aos problemas de classificação e de regressão analisados. Espera-se que a maior disponibilidade de dados estimule a utilização de algoritmos de ML mais flexíveis em pesquisas de saúde futuras.

Palavras-chave: Expectativa de Vida; Modelos de Predição; Mortalidade; Prognóstico; Qualidade de Vida.

SANTOS, H. G. **Comparison of machine learning algorithms performance in predictive analyzes in public health and medicine**. 2018. 206f. Thesis (Ph. D.) – Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, 2018.

ABSTRACT

Predictive models estimate the risk of health-related events or injuries and can be used as an auxiliary tool in decision-making by public health officials and health care professionals. Machine learning (ML) algorithms have the potential to identify complex and non-linear relationships, with positive implications in the predictive performance of these models. The present research aimed to apply various ML supervised techniques and compare their performance in classification and regression problems to predict outcomes of interest to public health and medicine. Results and discussion are organized into three articles. The first, presents a tutorial for the use of ML in health research, using as an example the prediction of death up to 5 years (outcome frequency=15%; n=395) in elderly participants of the study “Saúde, Bem-estar e Envelhecimento” (n=2,677), using variables related to demographic, socioeconomic and health characteristics. In the learning step, five algorithms were applied: logistic regression with and without regularization, neural networks, gradient boosted trees and random forest, whose hyperparameters were optimized by 10-fold cross-validation (CV). The area under receiver operating characteristic (AUROC) curve was greater than 0.70 for all models. For those with higher AUROC (neural networks and logistic regression with and without regularization), the quality of the predicted probability was evaluated and it showed low calibration. The second article aimed to predict the risk of quality-adjusted life up to 30 days (outcome frequency=44.7%; n=347) in oncologic patients admitted to the Intensive Care Unit (ICU) (n=777), using patients’ characteristics obtained at ICU admission. Six algorithms (logistic regression with and without regularization, neural networks, basic decision trees, gradient boosted trees and random forest) were used with nested CV to estimate hyperparameters values and to evaluate predictive performance. All algorithms, with exception of basic decision trees, presented acceptable discrimination (AUROC > 0.80) and calibration. For the third article, socioeconomic and demographic characteristics were used to predict the life expectancy at birth of Brazilian municipalities with more than 10,000 inhabitants (n=3,052). Nested CV and the Super Learner (SL) algorithm were used to adjust the predictive model, and for evaluating performance, the mean squared error (MSE). The SL showed good performance (MSE=0.17) and its vector of predicted values was used for the identification of underachievers and overachievers (i.e. municipalities showing worse and better outcome than predicted, respectively). Health characteristics were analyzed revealing that overachievers performed better on primary health care indicators, while underachievers fared better on secondary health care indicators. Techniques for constructing and evaluating predictive models are constantly evolving and there is scarce theoretical justification for preferring one algorithm over another. In this thesis no substantial differences were observed in the predictive performance of the algorithms applied to the classification and regression problems analyzed herein. It is expected that increase in data availability will encourage the use of more flexible ML algorithms in future health research.

Keywords: Life Expectancy; Mortality; Predictive Models; Prognostic; Quality of Life.

APRESENTAÇÃO

Esta tese foi desenvolvida para a obtenção do título de Doutor em Ciências pelo Programa de Pós-Graduação em Epidemiologia (PPG-Epi) da Faculdade de Saúde Pública da Universidade de São Paulo, na linha de pesquisa de métodos e técnicas de análise em epidemiologia, e contou com o apoio do Conselho de Aperfeiçoamento de Pessoal de Nível Superior, por meio da concessão de bolsa de doutorado.

A resolução CoPGr 6875, de 06 de agosto de 2014, referente ao regulamento vigente à época do meu ingresso no PPG-Epi, reconhece dois formatos para o trabalho final: tese tradicional e coletânea de artigos científicos. Essa resolução prevê ainda, para a defesa, a redação de pelo menos dois artigos de autoria principal do aluno ou coautoria, submetidos ou publicados; no caso de coautoria é preciso apresentar declaração formal dos autores concordando com a utilização do artigo e aceitando a condição de que este não seja utilizado em outra Tese ou Dissertação. Para o meu trabalho, optei pelo formato de coletânea de artigos e apresento, em anexo, a referida declaração para o Artigo 3 – *Overachieving municipalities in public health: a machine-learning approach*.

O presente trabalho objetivou comparar a performance de algoritmos de *machine learning* para a análise preditiva de problemas de regressão e de classificação frequentemente presentes nas áreas de saúde pública e medicina, e está organizado da seguinte forma: descrição da importância de análises preditivas em saúde, com exemplos correspondentes à aplicação de algoritmos de *machine learning* nesse contexto; em seguida, apresento aspectos metodológicos referentes à utilização desses algoritmos, bem como às características principais daqueles que foram aplicados aos problemas de classificação e de regressão que compõem a tese. A seção de métodos traz detalhes sobre os bancos de dados analisados e os métodos e técnicas utilizados para ajuste e avaliação de performance dos modelos preditivos. Na seção de resultados e discussão, para cada estudo, inicialmente é apresentada uma análise descritiva e, na sequência, os artigos científicos submetidos para publicação, cujos comprovantes encontram-se em anexo. Ao final do trabalho foram ainda destacadas as conclusões e considerações finais referentes ao projeto inicialmente proposto e perspectivas futuras.

Os códigos em R, elaborados para cada artigo da tese, encontram-se em repositório do GitHub (<https://github.com/Hellengeremias/Tese-HellenGeremias>).

Sumário

1 INTRODUÇÃO.....	11
1.1 IDENTIFICAR FATORES DE RISCO OU PREDIZER O RISCO DE UM EVENTO OCORRER?.....	11
1.2 PREDIÇÃO PARA O DIAGNÓSTICO	14
1.3 PREDIÇÃO PARA A PREVENÇÃO.....	16
1.4 PREDIÇÃO PARA O PROGNÓSTICO.....	17
1.5 DESAFIOS E LIMITAÇÕES	20
2 MACHINE LEARNING E ANÁLISE PREDITIVA: CONSIDERAÇÕES METODOLÓGICAS.....	21
2.1 DEFINIÇÕES GERAIS	21
2.2 PRÉ-PROCESSAMENTO	24
2.2.1 Transformação de dados: preditores individuais	24
2.2.2 Transformação de dados: múltiplos preditores.....	25
2.2.3 Exclusão de preditores não informativos	26
2.2.4 Soluções para dados faltantes	26
2.2.5 Organização de preditores qualitativos	28
2.3 APRENDIZADO: FERRAMENTAS PARA O TREINAMENTO DE MODELOS	29
2.3.1 Métricas para avaliação de performance	29
2.3.2 Técnicas de reamostragem	32
2.3.3 Hiperparâmetros de algoritmos de aprendizagem	35
2.3.4 Equilíbrio entre viés e variância	36
2.3.5 Dados desbalanceados	37
2.3.6 Comparação de modelos.....	38
2.4 APRENDIZADO: ALGORITMOS DE MACHINE LEARNING	39
2.4.1 Métodos lineares para regressão.....	39
2.4.1.1 Modelo de regressão linear.....	39

2.4.1.2 Modelos relacionados a técnicas de regularização	41
2.4.2 Alternativa para não-linearidade em regressão	42
2.4.2.1 Multivariate Adaptive Regression Splines	42
2.4.3 Métodos lineares para classificação	45
2.4.3.1 Regressão logística	46
2.4.4 Outras alternativas para não linearidade.....	46
2.4.4.1 Método dos K vizinhos mais próximos	48
2.4.4.2 Support Vector Machines	50
2.4.4.3 Redes Neurais Artificiais	55
2.4.5 Métodos baseados em árvores de decisão	59
2.4.5.1 Árvores simples	59
2.4.5.2 Bagging.....	59
2.4.5.3 Random forest	63
2.4.5.4 Boosting.....	64
2.4.5.5 Cubist.....	66
2.4.6 Super Learner	68
2.5 AVALIAÇÃO DO MODELO FINAL.....	70
3 OBJETIVOS	72
3.1 OBJETIVO GERAL.....	72
3.2 OBJETIVOS ESPECÍFICOS	72
3.2.1 Artigo 1	72
3.2.2 Artigo 2.....	72
3.2.3 Artigo 3.....	72
4 MÉTODOS.....	73
4.1 ARTIGO 1	73

4.1.1 Fonte de dados e delineamento do estudo	73
4.1.2 População de estudo	75
4.1.3 Pareamento dos dados	75
4.1.4 Variáveis de estudo.....	76
4.1.5 Análise dos dados	78
4.1.6 Aspectos éticos	79
4.2 ARTIGO 2	79
4.2.1 Fonte de dados e delineamento do estudo	79
4.2.2 População de estudo	80
4.2.3 Variáveis de estudo.....	81
4.2.4 Processamento e análise dos dados	83
4.2.5 Aspectos éticos	84
4.3 ARTIGO 3	85
4.3.1 Fonte de dados e delineamento do estudo	85
4.3.2 População de estudo	86
4.3.3 Variáveis de estudo.....	87
4.3.4 Análise dos dados	88
5 RESULTADOS E DISCUSSÃO	90
5.1 ARTIGO 1	92
5.1.1 Análise descritiva	92
5.1.2 Machine Learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo.....	98
5.2 ARTIGO 2	121
5.2.1 Machine learning to predict 30-day quality-adjusted survival in critically ill patients with cancer.....	121
5.3 ARTIGO 3	142
5.3.1 Análise descritiva	142
5.3.2 Overachieving municipalities in public health: a machine-learning approach.....	144
6 CONCLUSÕES E CONSIDERAÇÕES FINAIS	174
REFERÊNCIAS BIBLIOGRÁFICAS	182

APÊNDICES	182
Apêndice A – Dicionário de variáveis socioeconômicas e demográficas utilizadas para a predição da expectativa de vida ao nascer de municípios brasileiros.....	188
Apêndice B – Dicionário de características de saúde utilizadas na comparação de municípios overachievers e underachievers.	188
ANEXOS	194
Anexo A – Aprovação da Comissão Nacional de Ética em Pesquisa, Estudo SABE 2000.	196
Anexo B – Aprovação do Comitê de Ética em Pesquisa da Faculdade de Saúde Pública, Estudo SABE 2006.....	196
Anexo C – Aprovação do Comitê de Ética em Pesquisa da Faculdade de Saúde Pública, Estudo SABE 2010.....	197
Anexo D – Aprovação do Comitê de Ética para Análise de Projetos de Pesquisa da Diretoria Clínica do Hospital das Clínicas e da Faculdade de Medicina da Universidade de São Paulo, Estudo QALY.....	198
Anexo E – Aprovação do Comitê de Ética em Pesquisa do Hospital de câncer de Barreto, Estudo QALY.....	199
Anexo F – Comprovante de submissão, Artigo 1	200
Anexo G – Comprovante de submissão, Artigo 2	201
Anexo H – Comprovante de submissão, Artigo 3	202
Anexo I – Declaração de ciência e aceitação de utilização de artigo em caso de coautoria	204
CURRÍCULO LATTES – ALUNA	205
CURRÍCULO LATTES – ORIENTADOR	206

1 INTRODUÇÃO

O desenvolvimento e aplicação de análises preditivas em saúde estão relacionados a duas questões importantes: qual a diferença entre identificar fatores de risco para determinado evento de interesse e a de predizer o risco desse evento, quando o mesmo ainda não é conhecido (diagnóstico) ou não ocorreu (prognóstico)? O risco predito pode contribuir para a tomada de decisão por profissionais e gestores de saúde em casos individuais ou em estudos com enfoque coletivo?

Este capítulo foi desenvolvido para discutir tais questões e apresentar pesquisas realizadas na área da saúde que objetivaram ajustar modelos preditivos utilizando como ferramenta algoritmos de *machine learning* (ML) – desde os métodos convencionais, como a regressão linear e a logística, até os mais flexíveis, como o *random forest* e as redes neurais, cujas características serão apresentadas de forma detalhada no capítulo 2.

1.1 IDENTIFICAR FATORES DE RISCO OU PREDIZER O RISCO DE UM EVENTO OCORRER?

A epidemiologia é a ciência que propõe e avalia métodos utilizados pela saúde pública em ações de prevenção, controle e tratamento de doenças (ROUQUAYROL; GOLDBAUM, 2003). Um dos seus principais objetivos consiste em conhecer os determinantes da relação entre indivíduos que sofreram um dado evento, como doença, agravos relacionados à saúde ou óbito, e a população a qual eles pertencem (ALMEIDA FILHO; ROUQUAYROL, 2006). Para tanto, a observação, sistematização e interpretação de dados referentes a esses eventos em população específica, durante um período de tempo determinado, representam importantes etapas da análise epidemiológica, e contribuem para a descrição de perfis de morbidade e de mortalidade (SCHMIDT; DUNCAN, 2003).

No contexto clínico, a epidemiologia exerce papel semelhante ao descrito para a saúde pública. Nesse cenário, de modo geral, a análise epidemiológica contribui para o desenvolvimento e avaliação de protocolos, úteis no estabelecimento de orientações diagnósticas e terapêuticas, e na identificação de preditores relacionados ao curso clínico de doenças, conhecidos como fatores prognósticos (ALMEIDA FILHO; ROUQUAYROL, 2006).

O conceito de risco e de seus correlatos – fator de risco e grupo de risco –, apresenta estreita relação com a contribuição da epidemiologia tanto para a prática de saúde pública como para a da clínica médica. O risco caracteriza-se pela probabilidade de ocorrência de determinado evento relacionado à saúde em uma população específica, durante um período de tempo determinado, e é estimado pela proporção de indivíduos que sofreram o evento de interesse (ALMEIDA FILHO; ROUQUAYROL, 2006).

O fator de risco, por sua vez, corresponde a uma característica que antecede a ocorrência do evento de interesse e que, quando presente, implica aumento na probabilidade desse evento ocorrer, comparada à observada na ausência ou menor exposição a essa característica (ALMEIDA FILHO; ROUQUAYROL, 2006), bem como na taxa instantânea de mudança (ou velocidade) com que indivíduos desenvolvem tal evento de interesse na população (SZKLO; NIETO, 2014). Por fim, o grupo de risco é representado por indivíduos que apresentam um ou mais fatores de riscos em comum, que torna maior a probabilidade de ocorrência de um evento relacionado à saúde, comparada àquela observada para indivíduos que não apresentam tais fatores ou que estão menos expostos aos mesmos (CESAR, 1998).

Embora o enfoque de risco – identificar fatores de risco individuais e grupos de risco – seja alvo de críticas por priorizar intervenções voltadas à modificação de comportamentos individuais em lugar de medidas de maior impacto populacional, como as de âmbito estrutural (BARATA, 2013; LUIZ; COHN, 2006), este ainda desempenha papel importante na prática da saúde pública, sobretudo no que diz respeito ao controle de doenças e à redução de ameaças específicas à saúde no curto prazo, e pode ser utilizado em conjunto com abordagens populacionais e aquelas voltadas a grupos vulneráveis. Por exemplo, níveis mais elevados de determinada exposição podem antecipar o aumento na frequência de uma dada doença e, portanto, justificar intervenções em populações de risco (SEMENZA; SUK; MANISSERO, 2008).

A identificação de fatores de risco envolve a aplicação de métodos epidemiológicos ao estudo de possíveis associações entre um ou mais fatores suspeitos e o evento de interesse – doença ou outros agravos relacionados à saúde e óbito (ROUQUAYROL; GOLDBAUM, 2003). Essa atividade contribui para a realização de prevenção primária, correspondente à promoção da saúde antes do aparecimento da doença, e de prevenção secundária, caso o fator de risco seja um marcador para uma dada doença. Nesse caso, a prevenção objetiva facilitar a cura por meio de detecção precoce e tratamento em tempo oportuno (ALMEIDA FILHO; ROUQUAYROL, 2006).

Além disso, fatores de risco podem ser utilizados como preditores do evento a que estão associados por meio do ajuste de modelos preditivos, que por sua vez, irão resultar em uma probabilidade estimada de um indivíduo apresentar esse evento. Na área da saúde, ainda que maior ênfase seja dada à identificação de fatores de risco, mediante aplicação de testes de hipóteses, conhecer a probabilidade estimada de determinado desfecho ocorrer pode subsidiar a ponderação entre danos/custos e benefícios: se essa probabilidade for relativamente baixa, o benefício de um tratamento ou de uma ação preventiva também o será e, por outro lado, se for alta, iniciar o tratamento ou ações preventivas pode ser benéfico, mesmo que haja um dano ou custo associado a essas intervenções (STEYERBERG, 2009).

Diante desse cenário, organizações governamentais e de saúde podem usar o resultado de modelos preditivos na elaboração e avaliação de políticas públicas. Como exemplo, programas voltados à prevenção de doenças crônicas, que têm sido o foco de políticas públicas desenvolvidas em diversos países, podem utilizar valores preditos (risco ou probabilidade de um evento ocorrer) para direcionar suas ações àqueles que, de fato, possam se beneficiar das mesmas, a fim de alcançar resultados mais efetivos (OLIVERA et al., 2017).

No contexto da prática clínica, estruturar a informação disponível considerando resultados de modelos preditivos pode auxiliar na "tomada de decisão compartilhada", onde médicos e pacientes participam ativamente da decisão sobre técnicas diagnósticas e intervenções terapêuticas. Adicionalmente, conhecer o risco predito de um paciente para determinado evento pode ser útil na comunicação entre profissionais de saúde e subsequente estabelecimento de condutas visando a melhora do prognóstico (STEYERBERG, 2009).

Um dos principais resultados esperados, decorrentes de tomadas de decisão em saúde, é a redução da morbidade e da mortalidade, o que depende basicamente da performance do modelo preditivo, utilizado para identificar indivíduos com risco elevado para determinado desfecho, e da eficácia das ações de serviços de saúde (CESAR, 1998). Nos últimos anos, têm aumentado o número de publicações referentes a estudos que utilizaram algoritmos de ML na modelagem preditiva de respostas de interesse para a saúde, como mortalidade ou outros desfechos não fatais em populações específicas. Com a crescente disponibilidade de dados relevantes para o desenvolvimento de pesquisas em saúde, esses algoritmos têm o potencial de melhorar a performance de modelos preditivos por capturar relações complexas e não-lineares nos dados, bem como por sua capacidade em lidar com uma grande quantidade de preditores, até mesmo mais preditores que observações (CHEN; ASCH, 2017; OBERMEYER; EMANUEL, 2016).

Para que possamos discutir incertezas presentes na prática clínica e de saúde pública e apresentar exemplos de modelos preditivos com potencial para apoiar tomadas de decisão, a seguir são apresentadas algumas questões-chave, descritas em Schmidt & Duncan (2003): que certeza diagnóstica autoriza informar o indivíduo sobre a doença (por exemplo, diabetes) e iniciar seu manejo clínico específico? Como estimar o risco de complicações (por exemplo, as cardiovasculares), de modo a orientar a decisão de iniciar, ou não, ações preventivas? Respostas para essas questões podem basear-se, entre outras evidências, na compreensão e estimação da probabilidade de ocorrência (risco) de doenças e agravos à saúde.

É importante destacar que, independente da resposta que um modelo objetiva prever, se correspondente à investigação diagnóstica ou ao prognóstico de uma doença ou agravo à saúde já conhecido, os algoritmos e técnicas utilizados para o ajuste desse modelo serão os mesmos, pois para ambos os casos o interesse está no risco predito para a resposta como ferramenta auxiliar na tomada de decisão relacionada à terapia ou à prevenção.

1.2 PREDIÇÃO PARA O DIAGNÓSTICO

O diagnóstico de determinada doença baseia-se na avaliação de um conjunto de dados clínicos, tais como os dados iniciais do paciente, os resultados de seus exames e sua evolução clínica. Nesse cenário, o raciocínio probabilístico objetiva prever o risco de determinada doença ao longo do processo diagnóstico a fim de julgá-lo contra um espectro de probabilidades para decisões clínicas. Cada evento apresenta limites diferentes para o que é considerado probabilidade alta ou baixa, pois essa classificação depende dos custos e benefícios associados, respectivamente, ao diagnóstico incorreto e correto (SCHMIDT; DUNCAN, 2003).

Na prática clínica, predições relacionadas à probabilidade de uma doença sem sinal ou sintoma aparente, podem contribuir para a tomada de decisão sobre condutas subsequentes no diagnóstico (solicitar testes mais custosos ou arriscados) e escolha da terapia a ser instituída. Além disso, o rastreamento de determinada doença em indivíduos assintomáticos possibilita melhor tratamento do que quando a mesma é diagnosticada tardiamente. No entanto, é importante destacar que o rastreamento só é útil se houver melhora no prognóstico, comparado ao observado em situações de não rastreamento. Além disso, a realização de um exame diagnóstico deve considerar o fato de que algumas doenças não são tratáveis ou seu curso natural é muito similar ao observado quando o tratamento é instituído (STEYERBERG, 2009).

Historicamente, diversos modelos baseados em métodos convencionais, como a regressão logística, foram desenvolvidos para prever a ocorrência de doença em indivíduos assintomáticos na população ou de outros desfechos de interesse para a saúde pública. Pesquisadores do *Framingham Heart Study* desenvolveram modelos para prever o risco de doença cardiovascular (D'AGOSTINO et al., 2001; KANNEL; MCGEE; GORDON, 1976), que motivaram a elaboração de políticas públicas para o estabelecimento de medidas preventivas, como abordagens não farmacológicas (dieta e prática de atividade física), bem como tratamento medicamentoso com estatinas, direcionado a indivíduos com risco mais elevado para doenças cardiovasculares (EXPERT PANEL ON DETECTION, EVALUATION, 2001).

Mais recentemente, estudo realizado no Reino Unido objetivou comparar modelos resultantes de 4 algoritmos (*random forest*, regressão logística, *boosting* e redes neurais) com recomendações estabelecidas pelo Colégio Americano de Cardiologia para prever o primeiro evento (doença) cardiovascular em 10 anos, utilizando dados de uma coorte prospectiva de 378.256 pacientes. A performance dos modelos preditivos foi avaliada por meio da área abaixo da curva (AUC) ROC (*receiver operating characteristic*) e medidas de sensibilidade e especificidade. A frequência de eventos cardiovasculares foi de 6,6% e o modelo baseado na aplicação do algoritmo de redes neurais foi o que apresentou melhor performance preditiva, com AUC ROC de 0,764 (IC95%: 0,759;0,769), sensibilidade de 67,5% e especificidade de 70,7%, resultando em 355 predições corretas adicionais de doença cardiovascular quando comparado ao modelo baseado nas recomendações do Colégio Americano de Cardiologia (AUC ROC: 0,728; IC95%: 0,723;0,735) (WENG et al., 2017).

No contexto brasileiro, Olivera et al. (2017) desenvolveram modelos preditivos de diabetes não diagnosticada a partir de dados de 12.447 adultos entrevistados para o Estudo Longitudinal de Saúde do Adulto (ELSA), utilizando 5 algoritmos (regressão logística, redes neurais, *naive bayes*, método dos *K* vizinhos mais próximos e *random forest*). A frequência de diabetes não diagnosticada foi de 11,0% e os modelos com melhor performance foram os decorrentes do ajuste de redes neurais e de regressão logística, sem diferenças relevantes entre os mesmos. De acordo com o modelo final escolhido (regressão logística), entre os 403 indivíduos do conjunto de dados de teste que tinham diabetes não diagnosticada, 274 foram identificados como casos positivos (sensibilidade de 68%). Com base nos parâmetros estimados para o modelo de regressão logística, os autores desenvolveram um escore de risco,

implementado como ferramenta *web* para estimativa do risco de diabetes não diagnosticada em dados futuros.

1.3 PREDIÇÃO PARA A PREVENÇÃO

Tanto no caso da prevenção como no do tratamento, questões relacionadas à custo-efetividade são úteis em decisões coletivas de gestores para a disponibilização de ações preventivas, procedimentos diagnósticos e opções terapêuticas nos serviços de saúde. Ações preventivas podem ocorrer em diferentes momentos da história da doença: antes da instalação dos fatores de risco ou da doença, antes do diagnóstico clínico ou antes da ocorrência de complicações decorrentes da doença, incapacitação ou óbito, e devem apresentar benefícios que excedam seus custos. Da mesma forma, para que determinada terapia tenha impacto benéfico no prognóstico de uma doença, ou seja, repercussão positiva na melhora de seu curso clínico, deve exceder quaisquer efeitos colaterais, danos ou custos econômicos (SCHMIDT; DUNCAN, 2003).

Como exemplo de contribuição da análise preditiva para o estabelecimento de ações preventivas são apresentadas, a seguir, duas pesquisas: uma relacionada a identificação de gestantes com maior risco para desfechos desfavoráveis ao recém-nascido e outra referente à identificação de indivíduos com risco mais elevado para transtorno do estresse pós-traumático (TEPT), após ter vivenciado alguma experiência traumática.

Pan et al. (2017) utilizaram dados de 6.457 nascimentos de Illinois, Estados Unidos, para avaliar a performance de 4 algoritmos (*random forest*, análise discriminante linear, regressão logística penalizada e *naive bayes*) na predição de desfechos desfavoráveis à criança no momento do nascimento (prematuridade, baixo peso, admissão em unidade de terapia intensiva) e morte no primeiro ano de vida, utilizando 17 preditores relacionados a características da gestação. O objetivo final da análise preditiva foi estabelecer um limiar de elegibilidade para a admissão de gestantes em um programa de acompanhamento especializado. Os algoritmos apresentaram performance preditiva semelhante e superaram a estratégia corrente, utilizada pelo programa para identificar gestantes com prioridade de acompanhamento. Ao ordenar o risco predito e analisar gestantes alocadas no decil de maior risco, os autores observaram maior valor preditivo positivo (VPP) para o modelo de regressão logística com penalização. Os resultados desse modelo indicaram 170 gestantes adicionais, em comparação à estratégia utilizada pelo programa como critério de elegibilidade, que poderiam

ter um recém-nascido com algum evento adverso caso não fossem acompanhadas pelo programa.

Kessler et al. (2014) empregaram algoritmos de ML (regressão logística convencional e penalizada, *random forest* e *super learner*) para a modelagem preditiva de TEPT a partir de dados retrospectivos de 47.466 exposições traumáticas, obtidos de inquéritos sobre saúde mental conduzidos pela Organização Mundial da Saúde em 24 países. Foram utilizadas como variáveis preditoras fatores sociodemográficos, tipos de exposições traumáticas, presença anterior do mesmo tipo de exposição traumática e de distúrbios mentais. A prevalência de TEPT foi de 4,0%; na análise de performance do modelo final (*super learner*) observou-se que 95,6% dos casos de TEPT estavam presentes entre os 10% de exposições traumáticas preditas como de maior risco para TEPT. Adicionalmente, as 47.466 exposições traumáticas foram divididas em 20 grupos, a partir da ordenação do risco predito para TEPT e, no grupo de risco mais alto, esse desfecho foi observado em 56,3% das exposições. Os autores destacam o fato de que a identificação de indivíduos expostos a experiências traumáticas com alto risco para TEPT pode ajudar os serviços de saúde a melhorar o direcionamento de ações preventivas e, consequentemente, o seu custo-benefício.

1.4 PREDIÇÃO PARA O PROGNÓSTICO

Além das situações descritas anteriormente, a classificação de indivíduos portadores de determinada doença, como as cardiovasculares, ou de populações específicas, como a de idosos, de acordo com seu risco predito para determinado desfecho pode ser útil para a comunicação entre profissionais de saúde e subsequente estabelecimento de condutas visando a melhora prognóstica (STEYERBERG, 2009). Nesse contexto, o risco de morte é frequentemente analisado como resposta de interesse, por sua relevância para diversas condições agudas e crônicas, bem como para a decisão sobre a instituição de um dado tratamento, como por exemplo, o cirúrgico.

Em especial na área médica, a predição do risco de morte pode ser avaliada a partir da ocorrência de determinado episódio relacionado ao cuidado do paciente (por exemplo, admissão hospitalar ou admissão em Unidade de Terapia Intensiva – UTI), ou em um período de tempo específico (por exemplo, dentro de 30 dias após a internação). Nesses cenários, taxas de mortalidade ajustadas, calculadas como a razão entre a mortalidade observada e a predita, decorrente da soma de probabilidades individuais de mortalidade, podem ser utilizadas na

comparação da qualidade do cuidado entre instituições ou entre unidades de uma instituição que prestam o mesmo tipo de assistência. Do ponto de vista do uso de recursos, modelos preditivos podem ajudar a identificar pacientes de UTI com baixo risco de morte, que poderiam ser tratados em serviços de menor complexidade, ou em tomadas de decisão sobre condutas médicas no final da vida (BARNATO; ANGUS, 2004; KEEGAN; SOARES, 2016).

A maioria dos modelos preditivos de mortalidade hospitalar é desenvolvida com auxílio da regressão logística, tais como versões do *Acute Physiology and Chronic Health Evaluation (APACHE)* (KNAUS et al., 1991) e do *Simplified Acute Physiology Score (SAPS)* (LE GALL; LEMESHOW; SAULNIER, 1993; METNITZ et al., 2005) e o *Mortality Probability Models (MPM)* (LEMESHOW et al., 1993). Há ainda modelos baseados em critérios objetivos, estabelecidos por especialistas, para a seleção e atribuição de pesos aos preditores de mortalidade hospitalar, como o *Sepsis-related Organ Failure Assessment (SOFA)* (VINCENT et al., 1996). Recentemente, Pirracchio et al. (2015) desenvolveram um modelo baseado no algoritmo *super learner*, aplicado a uma coleção de modelos paramétricos e não-paramétricos, para prever mortalidade hospitalar. Os autores compararam o desempenho desse modelo, denominado *Super ICU Learner Algorithm (SICULA)*, com o do SOFA, o da versão original do SAPS-II e o de versões modificadas do SAPS-II e do APACHE-II e observaram melhor discriminação para o SICULA (AUC ROC 0,88), quando comparado aos demais modelos (SOFA: 0,71; SAPS-II: 0,78; SAPS-II modificado: 0,83; APACHE-II modificado: 0,82), na predição de mortalidade hospitalar.

O algoritmo *super learner* também foi utilizado por Rose (2013) para prever o risco de óbito em 5 anos (prevalência de 13%) entre 2.066 residentes em Sonoma, Califórnia, com idade a partir de 54 anos. A performance preditiva do *super learner*, medida por meio do erro quadrático médio (EQM), foi de 0,0904 e o R^2 de 0,201, evidenciando 20,1% de variabilidade explicada pelo modelo resultante do *super learner*.

Em relação à aplicação de outros algoritmos de *machine learning*, Austin et al. (2012) utilizaram árvores de decisão simples, *bagging*, *boosted trees*, *random forest* (métodos *ensemble*) e regressão logística para prever o risco de morte em até 30 dias em pacientes hospitalizados com infarto agudo do miocárdio ou insuficiência cardíaca congestiva. Os autores concluíram que os métodos *ensemble* aumentaram o desempenho preditivo da árvore de decisão simples, porém não apresentaram performance superior ao modelo de regressão logística.

Fatores prognósticos são também importantes no contexto do estudo de câncer, por exemplo, quando se está interessado em prever o risco de eventos adversos ao iniciar

determinado tratamento quimioterápico. Embora a quimioterapia diminua o risco de recorrência de câncer em estágios iniciais e melhore a sobrevida e os sintomas em estágios mais avançados, há indícios de que esse tratamento, com frequência, seja iniciado tardiamente, o que implica taxas elevadas de mortalidade. Elfiky & Parikh (2017) objetivaram desenvolver e validar um modelo baseado no algoritmo *boosting* para prever risco de morte 30 dias após início de novos regimes quimioterápicos (frequência de morte de 2,1% no conjunto de dados de validação). Registros eletrônicos de saúde de um centro médico especializado de Boston, para o período de 2004-2011, foram utilizados para ajustar o modelo preditivo e dados de 2012-2014 para a validação do modelo ajustado. A frequência de mortalidade em 30 dias no grupo com risco predito mais alto (último decil) foi de 22,6% e entre aqueles com risco predito mais baixo (primeiro decil), nenhum paciente morreu.

Para algumas perguntas de pesquisa, além do risco de morte, outros desfechos não fatais podem ser interessantes como, por exemplo, resultados centrados no paciente (pontuação em questionários sobre qualidade de vida) ou indicadores mais amplos de carga de doença (internação ou readmissão hospitalar), relacionados ao planejamento da assistência (STEYERBERG, 2009).

Nesse sentido, os algoritmos *support vector machine*, *AdaBoost*, regressão logística, *naive bayes* foram aplicados aos dados de registros eletrônicos de saúde de um hospital de Boston para prever hospitalizações no período de um ano em pacientes com doenças cardíacas. A frequência de resultados falsos positivos (FP) ($1 - \text{especificidade}$) e a de verdadeiros positivos (VP) (*sensibilidade*) foram utilizadas para avaliação de performance dos modelos ajustados. O modelo resultante do algoritmo *AdaBoost* foi o que apresentou melhor desempenho. Os resultados evidenciaram que, com uma frequência de FP de 30%, é possível prever com sucesso 82% dos pacientes com doenças cardíacas que serão hospitalizados no ano seguinte. Os autores afirmam que a predição de hospitalizações nesse subgrupo pode contribuir para a implementação de ações preventivas, como agendamento de consultas médicas, rastreamento de possíveis complicações e atendimento por profissionais de saúde a fim de garantir a adesão à terapia medicamentosa. Tais ações são menos onerosas que a hospitalização e, se bem-sucedidas, podem reduzir os custos de cuidados hospitalares (DAI et al., 2015).

Estudo realizado na Califórnia, Estados Unidos, utilizou regressão logística, *random forest* e redes neurais em dados de registros eletrônicos de saúde para prever readmissão 30 dias após a alta hospitalar, e compararam os resultados com o escore LACE, modelo até então utilizado pelo serviço para prever essa resposta. Os autores utilizaram como métrica para

avaliação de performance o VPP, a sensibilidade e a AUC ROC. Assumindo que intervenções devem ser aplicadas a 25% dos pacientes com risco predito mais alto para readmissão hospitalar em 30 dias, o escore LACE apresentou VPP de 20%, sensibilidade de 49% e AUC ROC de 0,72 e o modelo baseado em redes neurais, que apresentou melhor desempenho, VPP de 24%, sensibilidade de 60% e AUC ROC de 0,78 (JAMEI et al., 2017).

Por fim, modelos preditivos podem ser também aplicados em situações cuja unidade de análise seja o território (por exemplo, o município), em vez do indivíduo, situação que também pode contribuir para o planejamento e avaliação de ações de saúde pública. Guo et al. (2017) avaliaram diferentes algoritmos (*support vector machine*, *lasso*, *boosting*, modelo aditivo generalizado, regressão linear e binomial negativa) para prever incidência de dengue em províncias da China para o período de 2011 a 2014. Os autores utilizaram como métrica para avaliação de performance a raiz do erro quadrático médio (*Root Mean Squared Error – RMSE*); o modelo com melhor desempenho foi o derivado do algoritmo *support vector machine* que mostrou-se acurado para prever epidemias nas últimas 12 semanas, bem como o aumento inesperado no número casos (surto) ocorrido em 2014. Além disso, este modelo foi o com melhor desempenho para predição da dinâmica da doença e da ocorrência de surtos em outras áreas da China, não utilizadas no ajuste do modelo preditivo.

1.5 DESAFIOS E LIMITAÇÕES

A utilização de modelos preditivos não é uma prática nova para a saúde pública e a medicina. Da predição do risco para doenças cardiovasculares à estratificação de risco de pacientes em UTIs, é frequente o planejamento de ações de saúde apoiado por valores preditos (MATHENY; OHNO-MACHADO, 2014). Nos últimos anos, o aumento na quantidade e disponibilidade de dados gerados por serviços de saúde, o custo elevado das ações realizadas por esses serviços e a necessidade de qualificação de sua assistência têm incentivado a realização de pesquisas voltadas à predição de desfechos de interesse para a área da saúde. Nesse contexto, algoritmos de ML têm potencial para contribuir na integração de dados, extração de informações e identificação de relações complexas e não-lineares, podendo melhorar a performance de modelos preditivos desenvolvidos para apoiar questões clínicas e de saúde pública (BHARDWAJ; NAMBIAR; DUTTA, 2017; CHEN; ASCH, 2017).

No entanto, há alguns desafios relacionados à realização de análise preditiva em saúde, de modo geral, e à aplicação de algoritmos de ML, em particular. Uma característica inerente

aos dados de saúde refere-se à distribuição desbalanceada das classes de respostas categóricas, ou seja, maior frequência para a classe de indivíduos saudáveis e, por outro lado, uma classe minoritária, representativa de indivíduos doentes ou com resposta desfavorável. Geralmente, a classe minoritária é a mais importante e também a que mais sofre com classificações erradas, pois os algoritmos, para alcançar melhor desempenho (mais acertos), tendem a priorizar a especificidade em vez da sensibilidade, o que pode implicar redução do desempenho de modelos preditivos quando aplicado a novos dados (MENA et al., 2012).

Com relação à variedade de algoritmos disponíveis para utilização em problemas de classificação e de regressão, é importante destacar que, diferente dos modelos de regressão logística ou linear, modelos mais flexíveis, tais como *boosting* e *random forest*, não são interpretáveis ou passíveis de modificação por profissionais com *expertise* relacionada à resposta de interesse. Adicionalmente, as técnicas e métodos utilizados para construir e avaliar os modelos, a definição da resposta de interesse, disponibilidade de variáveis informativas como candidatas a preditores, bem como o objetivo da análise preditiva, podem exercer papel importante na performance de um modelo. Portanto, um balanço entre acurácia e interpretabilidade deve ser considerado na escolha do modelo preditivo final (CARUANA et al., 2015).

O próximo capítulo apresenta conceitos gerais relacionados à aplicação de ML para a realização de análise preditiva, bem como as principais características dos algoritmos utilizados nos estudos desenvolvidos para a tese.

2 MACHINE LEARNING E ANÁLISE PREDITIVA: CONSIDERAÇÕES METODOLÓGICAS

2.1 DEFINIÇÕES GERAIS

A análise preditiva tem como atividade principal estimar o risco de eventos futuros com base em experiências passadas, para orientar a tomada de decisão atual. A operacionalização desse processo envolve um conjunto de ferramentas, ou algoritmos, utilizados para compreender os dados existentes e gerar regras de predição. De modo geral, essas ferramentas

podem ser alocadas em uma das seguintes categorias de aprendizagem: a supervisionada e a não supervisionada (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; KUHN; JOHNSON, 2013).

Em um cenário de aprendizado supervisionado, cada observação i , $i=1, 2, \dots, n$ do conjunto de dados dispõe de um vetor de mensurações para variáveis preditoras (*input* ou variáveis independentes), x_i , bem como de mensuração correspondente à resposta de interesse y_i (*output*, desfecho ou variável dependente), e um modelo que relacione a resposta aos preditores é ajustado com o objetivo de prever essa resposta em observações futuras, para as quais estão disponíveis apenas dados referentes aos preditores (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; JAMES et al., 2014). A distinção entre o tipo de variável resposta resulta em dois subgrupos de aprendizagem supervisionada: o de regressão, para variáveis quantitativas, e o de classificação, para as do tipo categórica (qualitativa) (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). Quanto às características dos preditores, em geral, a maioria dos algoritmos pode ser aplicada independentemente do tipo de variável preditora disponível, contanto que os preditores sejam adequadamente pré-processados antes da análise dos dados (JAMES et al., 2014).

No caso do aprendizado não supervisionado, para toda observação i , $i=1, 2, \dots, n$, há um vetor de mensurações x_i (*input*), correspondente aos preditores disponíveis no conjunto de dados, porém não há uma variável resposta conhecida, y_i , responsável por guiar a análise. Nesse contexto, o objetivo principal do ajuste de modelos consiste em compreender relações entre variáveis ou entre observações, a fim de identificar como estas estão organizadas ou agrupadas (JAMES et al., 2014; RASCHKA, 2017).

Nos trabalhos apresentados nesta tese, métodos supervisionados serão utilizados para a modelagem preditiva de respostas quantitativas e categóricas de interesse para a área da saúde e as seguintes etapas de um sistema de aprendizado de máquina (*machine learning*) serão realizadas: divisão aleatória do conjunto de dados em treinamento e teste, pré-processamento, aprendizado e seleção de modelos em dados de treinamento, predição da resposta de interesse em dados de teste e avaliação dos modelos selecionados (Figura 1).

Figura 1: Roteiro para aplicação de *machine learning* em análise preditiva.

Fonte: RASCHKA (2017).

A primeira etapa, referente à divisão aleatória do conjunto de dados original em treinamento e teste, é realizada a fim de determinar se um algoritmo apresenta boa performance não apenas no conjunto de dados utilizado para o ajuste do modelo preditivo (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 inicial do conjunto de dados – em geral, quanto maior o número de observações, maior será a proporção do conjunto inicial utilizada para o treinamento de modelos (RASCHKA, 2017).

Mais detalhes sobre o ajuste de modelos, bem como sobre a estimativa do erro de predição associado à sua aplicação em novas observações, serão explorados ao longo do texto. No entanto, é importante que fique claro que em todas as etapas do sistema de aprendizado apenas o conjunto de treinamento é explorado, garantindo, ao final desse processo, a obtenção de uma estimativa mais acurada da performance preditiva do modelo ajustado quando aplicado a novas observações.

Em relação à notação utilizada ao longo do texto, letras maiúsculas farão referência a aspectos gerais das variáveis, e as minúsculas a valores observados para as mesmas. A variável resposta será identificada por Y , as preditoras por X e, caso X seja um vetor, seus componentes serão acessados por subscritos, X_j , $j=1,2,...,p$. Matrizes serão representadas por letras maiúsculas em negrito, como a matriz $\mathbf{X}$, referente ao conjunto de n vetores p -dimensionais x_i , $i= 1, 2, ..., n$. Um vetor será destacado em negrito apenas quando for n -dimensional e fizer

referência ao vetor de mensurações correspondente à resposta de interesse, y , ou a todas as observações da j -ésima variável do conjunto de dados, x_j .

O próximo tópico apresenta a etapa de pré-processamento, que explora características do conjunto de dados de treinamento a fim de organizá-lo para o aprendizado de modelos preditivos (KUHN; JOHNSON, 2013; RASCHKA, 2017).

2.2 PRÉ-PROCESSAMENTO

De modo geral, os dados originais, em sua forma bruta, não irão resultar na performance ótima de um algoritmo, sendo necessário considerar características relacionadas à dimensão do conjunto de dados (número de observações e de preditores disponíveis), à variável resposta (categórica ou contínua, balanceada ou desbalanceada, simétrica ou assimétrica) e aos preditores (variável contínua, contagem, variável categórica, variáveis correlacionadas/associadas, escalas diferentes, presença de valores faltantes e de dados esparsos) (KUHN; JOHNSON, 2013; RASCHKA, 2017).

2.2.1 Transformação de dados: preditores individuais

Técnicas de pré-processamento estão, de modo geral, relacionadas à transformação dos dados originais, à seleção de variáveis relevantes, à exclusão ou imputação de dados faltantes e à organização de dados categóricos. A transformação de dados apresenta estreita relação com a melhora do desempenho de muitos algoritmos que requerem preditores medidos na mesma escala. A normalização e a padronização representam as abordagens mais comuns para modificação da escala de uma variável: a primeira refere-se ao redimensionamento das variáveis para uma escala que varie entre 0 e 1, ou seja, dentro de um intervalo delimitado, e a segunda, à centralização das variáveis na média 0 com desvio padrão 1 (KUHN; JOHNSON, 2013; RASCHKA, 2017).

A transformação de dados também pode ser útil na correção de distribuições assimétricas. Primeiramente, é preciso identificar variáveis que apresentam essa característica, por exemplo, a partir da estatística de assimetria ou da razão entre o maior e o menor valor da variável preditora. De acordo com a estatística de assimetria, quanto mais distante de 0, mais assimétrica é a distribuição da variável, já para a razão maior-menor valor, resultados maiores

que 20 são indicativos de distribuição assimétrica. Transformações associadas ao *log*, *raiz quadrada* ou *valor inverso* podem contribuir para a correção de assimetria. Alternativamente, *Box e Cox* propuseram uma família de transformações indexadas por um parâmetro, θ , que pode ser estimado a partir dos dados de treinamento, e cujas transformações anteriormente sugeridas representam casos particulares (KUHN; JOHNSON, 2013).

2.2.2 Transformação de dados: múltiplos preditores

A identificação de preditores altamente correlacionados, e que, portanto, estão medindo a mesma informação subjacente, pode contribuir para que o sistema de aprendizado não resulte em modelos instáveis ou que apresentem erros numéricos. Uma abordagem para lidar com preditores que apresentam essa característica consiste em remover um número mínimo de preditores, a fim de garantir que todas as correlações de pares de variáveis estejam abaixo de um determinado limiar, ou utilizar técnicas para redução de dimensionalidade, a fim de alocar os preditores em um subespaço de menor dimensão (KUHN; JOHNSON, 2013; RASCHKA, 2017).

Nesse último caso, para a maioria das técnicas disponíveis, os novos preditores são funções dos preditores originais. Um exemplo é a análise de componentes principais (ACP), cujo objetivo é explicar a estrutura de variância-covariância de um conjunto de variáveis com base em suas combinações lineares (KUHN; JOHNSON, 2013; RASCHKA, 2017). A aplicação da ACP tem como premissa que parte importante da variabilidade do conjunto de dados pode ser explicada por q componentes principais, com $q < p$, sendo p o número de preditores originais (JOHNSON; WICHERN, 2007).

Uma das principais desvantagens da ACP está relacionada ao seu caráter não supervisionado, que desconsidera a variável resposta ao resumir a variabilidade dos preditores. Como consequência, os componentes principais não irão representar, necessariamente, bons preditores para a resposta de interesse. Como alternativa à ACP, o método dos mínimos quadrados parciais (*Partial Least Squares* - *PLS*) pode ser utilizado a fim de considerar a variável resposta no processo de extração dos componentes principais (KUHN; JOHNSON, 2013).

O *PLS* foi originalmente desenvolvido como alternativa ao método dos mínimos quadrados ordinários quando há multicolinearidade entre os preditores. Esse método deriva combinações lineares dos preditores (componentes principais) que resumem ao máximo a

covariância com a resposta de interesse, ao passo que explicam parte da variabilidade presente nos dados originais. O número de componentes representa um parâmetro de ajuste (ou hiperparâmetro) do *PLS* que pode ser estimado por técnicas de reamostragem, descritas neste capítulo, em tópico subsequente (KUHN; JOHNSON, 2013).

2.2.3 Exclusão de preditores não informativos

Alguns modelos podem ser prejudicados por preditores com distribuição degenerada ou que apresentem variância próxima de 0. Nesses casos, a exclusão desses preditores do conjunto de dados de treinamento antes da modelagem preditiva pode resultar em melhora da performance do modelo e/ou de sua estabilidade numérica. Vale destacar que essa decisão deve ser tomada antes da aplicação de técnicas para redução da dimensionalidade, apresentadas no tópico anterior (KUHN; JOHNSON, 2013).

2.2.4 Soluções para dados faltantes

Há casos em que alguns preditores não apresentam valores para uma dada amostra, seja por erro no processo de coleta dos dados, em que alguns campos tenham sido deixados em branco, porque certas mensurações não são aplicáveis a determinadas observações ou porque o entrevistado optou por não responder à determinada questão. Para que o dado faltante seja adequadamente considerado na etapa de análise, é importante observar seu padrão e frequência para entender porque o valor está ausente e identificar o mecanismo que o gerou (KUHN; JOHNSON, 2013; RASCHKA, 2017).

De modo geral, o dado faltante pode estar relacionado a mecanismos completamente aleatórios, aleatórios porém mensuráveis, ou não aleatórios. No primeiro caso, pode-se considerar que as observações com dados faltantes ocorreram completamente ao acaso e, portanto, não diferem das demais observações (LITTLE; RUBIN, 2002). Para essa condição, a perda de poder da análise que se deseja realizar representa a consequência mais importante do dado faltante (ASSUNÇÃO, 2012).

O mecanismo aleatório e mensurável, por sua vez, é o mais frequente e ocorre quando o dado faltante pode ser explicado por uma ou mais variáveis presentes no conjunto de dados (ASSUNÇÃO, 2012). Nesse caso, a distribuição dos dados faltantes é a mesma que a dos dados

observados dentro de subgrupos, definidos por variáveis correlacionadas àquela que apresenta o dado faltante (ZHANG, 2003). Por exemplo, no contexto clínico, pacientes com febre mais frequentemente tem dados para exame de hemograma que aqueles que não apresentam febre e que, portanto, estão mais propensos a ter dados faltantes para esse exame (GORELICK, 2006).

Por fim, o mecanismo não aleatório ocorre quando a probabilidade de o dado estar ausente depende do valor da própria variável. Além disso, nesse contexto, o dado faltante apresenta algum padrão ou tendência dependente de características que o pesquisador não observa ou controla e esse padrão pode estar relacionado à resposta de interesse, o que caracteriza um *missing* informativo, implicando viés na análise que se deseja realizar devido a diferenças entre observações com e sem o valor para uma dada variável (KUHN; JOHNSON, 2013).

Em sistemas de recomendação, classificações de produtos por clientes pode resultar em *missing* informativo, pois é provável que o cliente avalie, mais frequentemente, produtos que ele tenha ficado muito (ou pouco) satisfeito. Nesse caso, é mais provável que os dados observados estejam polarizados, com poucos valores correspondentes aos pontos centrais de determinada escala de classificação (KUHN; JOHNSON, 2013). Em contextos clínicos, a ausência de um dado (resultado) para determinado teste diagnóstico pode não ser aleatória, mas sim tendenciosa para indivíduos cujo resultado do teste teria indicado uma função normal (por exemplo, desfecho ausente) e, portanto, informativa para a resposta de interesse (VAN DER HEIJDEN et al., 2006).

Uma das abordagens mais diretas para solucionar o problema do valor faltante corresponde à exclusão da variável ou da observação que apresenta essa característica. Entretanto, tal abordagem apresenta desvantagens: muitas observações podem ser removidas, o que reduz a confiabilidade das análises, ou no caso da remoção de variáveis, pode-se perder variáveis informativas para a resposta de interesse. Alternativamente, técnicas de interpolação podem ser utilizadas, como a imputação do dado faltante a partir da média, mediana ou valor mais frequente de um determinado preditor ou de técnicas de imputação múltipla, em que outras variáveis preditoras presentes no conjunto de dados são utilizadas para a predição do valor faltante, por exemplo por meio de modelos de regressão ou do algoritmo *K-Nearest Neighbors* (KNN) (RASCHKA, 2017).

A maioria das técnicas de imputação são aplicáveis a situações em que o mecanismo que gerou o dado faltante é aleatório (mensurável ou não). No caso de o mecanismo não aleatório (e não mensurável) ter sido o responsável por gerar o dado faltante, torna-se necessário

especificar um modelo para a não-resposta, no entanto, métodos para analisar dados faltantes nesse cenário estão além do escopo deste trabalho. Mais informações estão disponíveis em (LITTLE; RUBIN, 2002).

2.2.5 Organização de preditores qualitativos

Em relação à organização de dados qualitativos, a distinção entre variáveis nominais e ordinais representa o primeiro passo. No caso de variáveis ordinais, é possível converter valores do tipo texto em inteiros, e no caso de variáveis nominais, decompor a variável original em um conjunto de variáveis indicadoras (*dummy*), representadas por valores binários (0-1), para cada categoria da variável nominal, de modo que, se a variável for composta por c categorias, então c variáveis indicadoras serão criadas (KUHN; JOHNSON, 2013; RASCHKA, 2017).

No ajuste de modelos preditivos, usualmente, $c - 1$ variáveis indicadoras são utilizadas como preditores, embora essa decisão esteja relacionada ao algoritmo escolhido para a modelagem. Caso o algoritmo inclua um intercepto, adicionar as c variáveis ao modelo irá resultar em problemas numéricos. Por outro lado, se não houver intercepto, utilizar o conjunto completo de variáveis indicadoras pode contribuir para a interpretação do modelo ajustado (KUHN; JOHNSON, 2013; RASCHKA, 2017).

Por fim, vale destacar que parâmetros estimados durante a etapa de pré-processamento, como os decorrentes da padronização de preditores e da redução de dimensionalidade, devem ser obtidos a partir de dados de treinamento e, então aplicados para transformar o conjunto de dados de teste antes da realização das previsões. Caso contrário, a performance resultante dos dados de teste pode ser muito otimista, e não corresponder ao real desempenho do modelo preditivo em dados futuros (RASCHKA, 2017).

O próximo tópico descreve características relacionadas ao treinamento de modelos preditivos, etapa subsequente ao pré-processamento dos dados.

2.3 APRENDIZADO: FERRAMENTAS PARA O TREINAMENTO DE MODELOS

2.3.1 Métricas para avaliação de performance

A avaliação da performance de um algoritmo de ML em um determinado conjunto de dados é realizada por meio da mensuração do quão bem as previsões decorrentes do modelo ajustado reproduzem o valor observado para a resposta de interesse. Portanto, é preciso quantificar o quanto o valor predito ($\hat{y}_i$) para a resposta de uma observação se aproxima de seu valor observado, y_i (JAMES et al., 2014). No contexto de regressão, o erro quadrático médio (EQM) é a medida utilizada com mais frequência, e é dado por:

$$EQM = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

A raiz dessa quantidade resulta em um valor na mesma unidade que os dados originais. Sua interpretação refere-se à distância média entre os valores observados e os preditos pelo modelo. O EQM será pequeno se as respostas preditas pelo modelo forem muito próximas das observadas e será grande se, para algumas observações, a resposta predita e a observada diferirem substancialmente. Na etapa de treinamento de algoritmos de aprendizado, o EQM pode ser utilizado para comparar modelos com diferentes preditores, hiperparâmetros ou modelos decorrentes de algoritmos distintos. Adicionalmente, como o objetivo final da modelagem preditiva é obter previsões acuradas em novos dados, não utilizados para o ajuste do modelo, a performance preditiva do modelo selecionado deve ser avaliada a partir da mensuração de seu EQM em dados de teste (JAMES et al., 2014).

Modelos de classificação, por sua vez, usualmente, resultam em dois tipos de previsão: uma contínua (p_k^*), relacionada à probabilidade de cada uma das classes, k , $k = 1, 2, \dots, K$, da resposta de interesse, e outra categórica (por exemplo, classe 0: desfecho ausente e classe 1: desfecho presente), referente à classe predita para uma dada observação (KUHN; JOHNSON, 2013).

As previsões 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. A partir da previsão contínua, uma nova observação é atribuída à classe em que p_k^* é máxima. Por exemplo, para respostas dicotômicas, a nova observação é atribuída à classe 1 se $p_k^* > 0,5$, caso 0,5 seja o ponto de corte escolhido.

Esse ponto de corte pode ser alterado de acordo com o objetivo do problema de predição sob análise. Na prática, um classificador para resposta dicotômica pode implicar dois erros: atribuir incorretamente um indivíduo que não sofreu o desfecho de interesse à classe correspondente à presença do desfecho (classe 1), ou atribuir incorretamente um indivíduo que sofreu o desfecho de interesse à classe correspondente à ausência dessa resposta (classe 0) (JAMES et al., 2014).

A partir da definição de um ponto de corte para p_k^* , uma matriz de confusão, representada pela tabulação cruzada de classes observadas e previstas para observações do conjunto de dados de teste, é comumente utilizada para visualizar esses erros. A Tabela 1 ilustra essa matriz para quando o desfecho de interesse é dicotômico. As caselas da diagonal principal (a e d) denotam casos em que as classes são corretamente previstas, enquanto que as caselas fora da diagonal (b e c) representam os erros de classificação (KUHN; JOHNSON, 2013).

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)	$c + d$
		$a + c$	$b + d$	

A acurácia ou taxa de erro geral ($(a+d) / (a+b+c+d)$) representa a concordância entre as classes previstas e observadas e representa a métrica mais simples derivada de uma matriz de confusão. Entretanto, há desvantagens em utilizar essa medida, pois a mesma não faz distinção entre o tipo de erro cometido (classificar uma observação como “desfecho presente” quando o mesmo está ausente ou classificá-la como “desfecho ausente” quando este está presente) e não considera a frequência natural (prevalência) de cada classe (JAMES et al., 2014; KUHN; JOHNSON, 2013).

No entanto, muitas vezes, é de interesse do pesquisador determinar qual o erro derivado de um modelo preditivo (classificador), ou seja, conhecer seu desempenho específico, de acordo com as classes da resposta de interesse, que pode ser descrito por medidas de sensibilidade e especificidade. A sensibilidade ($a / (a+c)$) é a proporção de verdadeiros positivos entre todos os indivíduos cuja resposta de interesse foi, de fato, observada, e a especificidade ($d / (b+d)$)

refere-se à proporção de verdadeiros negativos entre aqueles com resposta de interesse (observada) ausente (JAMES et al., 2014; KUHN; JOHNSON, 2013).

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 para avaliar a sensibilidade e especificidade decorrentes de todos os pontos de corte possíveis para p_k^* . Sabe-se que um ponto de corte igual a 0,5 minimiza a taxa de erro geral. No entanto, se o pesquisador estiver mais interessado em prever corretamente a presença do desfecho (classe 1), por exemplo, é possível alterar o ponto de corte de modo a aumentar o número de verdadeiros positivos. Para cada ponto de corte, a frequência de verdadeiros positivos (sensibilidade) e de falsos positivos ($1 -$ especificidade) são plotadas uma versus a outra (JAMES et al., 2014; KUHN; JOHNSON, 2013).

O desempenho geral de um classificador, resumido em todos os pontos de corte possíveis para p_k^* , é dado pela área abaixo da curva (AUC) ROC. Essa medida quantifica o poder discriminatório de um modelo, ou seja, sua habilidade em ordenar os indivíduos do risco mais alto para o mais baixo, separando os com resposta positiva para o desfecho daqueles com resposta negativa e, quanto maior a AUC (mais próxima de 1), melhor o desempenho de um classificador. Assim como o EQM para problemas de regressão, a AUC ROC pode ser útil na comparação de dois ou mais modelos com diferentes preditores, hiperparâmetros ou mesmo classificadores decorrentes de algoritmos completamente diferentes (JAMES et al., 2014; KUHN; JOHNSON, 2013).

É importante destacar que a AUC ROC não avalia a habilidade do modelo em atribuir probabilidades acuradas de um evento ocorrer. Portanto, além dessa medida, aquelas referentes à calibração do risco predito também são interessantes, pois medem a precisão com que a probabilidade predita corresponde à taxa de eventos observada em um conjunto de dados. São exemplos a curva de calibração, a distribuição da probabilidade predita entre as classes da resposta de interesse e o teste de Hosmer e Lemeshow (MEURER; TOLLES, 2017; PENCINA; D'AGOSTINO, 2015).

Para a construção da curva de calibração, o risco predito para a resposta de interesse em dados de teste é ordenado e dividido em grupos, baseados nos percentis da probabilidade estimada ou em valores fixos dessa estimativa: $[0; 0,10]$, $(0,10; 0,20]$, ..., $(0,90; 1]$. Para cada grupo, calcula-se a taxa observada da resposta de interesse (número de indivíduos que sofreram o evento dividido pelo total de observações). A curva de calibração corresponde ao ponto médio

de cada grupo, no eixo x , e à taxa de eventos observada para o respectivo grupo, no eixo y . Uma curva próxima de uma linha de 45° é indicativa de que o modelo preditivo estima probabilidades bem calibradas (KUHN; JOHNSON, 2013).

2.3.2 Técnicas de reamostragem

Em geral, o aprendizado de modelos preditivos é composto por dois objetivos principais: selecionar e avaliar modelos. No primeiro caso, a performance de diferentes modelos é estimada a fim de escolher aquele que apresenta melhor desempenho. Já no segundo, após a escolha de um modelo, busca-se estimar seu erro de predição (erro de generalização) em novas observações (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

Se o pesquisador dispõe de um conjunto de dados grande, a melhor abordagem para a seleção de modelos consiste em subdividir, aleatoriamente, o conjunto de treinamento em treinamento e validação. Dessa forma, o conjunto de dados original fica dividido em três partes: treinamento, validação e teste. Os dados de treinamento serão utilizados para ajustar os modelos e os de validação para selecionar um modelo com base em sua performance preditiva. Por fim, os dados de teste serão utilizados para avaliar o erro de generalização do modelo selecionado (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). A figura 2, a seguir, ilustra a divisão do conjunto de dados original.

Figura 2: Divisão do conjunto de dados original em treinamento, validação e teste para aplicações de *machine learning*.

Fonte: RASCHKA (2017).

Entretanto, há situações em que o pesquisador não dispõe de dados suficientes para dividi-los em três partes, e técnicas de reamostragem podem ser utilizadas para aproximar o conjunto de validação por meio da reutilização de observações do conjunto de treinamento original (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

Entre as técnicas de reamostragem, a validação cruzada *k-fold* representa uma das mais utilizadas em problemas de ML. Essa técnica consiste na divisão aleatória do banco de treinamento original em k partes de tamanhos aproximadamente iguais, em que $k-1$ irão representar dados de treinamento para ajuste do modelo preditivo e a outra parte ficará reservada para a estimativa de sua performance. O processo se repete até que todas as partes tenham participado tanto do treinamento como da validação do modelo, resultando em k estimativas de performance que serão resumidas, usualmente, pelo cálculo da média e do erro padrão (KUHN; JOHNSON, 2013).

O caso em que o número de partições é igual ao número de observações do conjunto de treinamento original, $K = N$, é conhecido como validação cruzada *leave-one-out*. Não há uma regra geral para a escolha de k , embora a divisão dos dados em 5 ou 10 partes seja usual. Conforme k aumenta, a diferença de tamanho do conjunto de treinamento original e dos subconjuntos reamostrados se torna menor, e a medida em que essa diferença diminui, o viés da técnica de *validação cruzada* também se torna menor. Por outro lado, há um aumento no tempo necessário para obter o resultado final da validação cruzada (KUHN; JOHNSON, 2013).

Vale destacar que, técnicas de reamostragem, como a validação cruzada *k-fold*, devem ser aplicadas a todos os passos que compõem o aprendizado de um modelo preditivo, a única exceção está relacionada ao pré-processamento inicial, não supervisionado, que pode ser realizado no conjunto de treinamento original (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

A figura 3, a seguir, resume o processo de validação cruzada *k-fold*, com $k=10$. Os dados de treinamento foram divididos em 10 partes, e em cada iteração, 9 partes foram utilizadas para o treinamento do modelo com diferentes hiperparâmetros, e uma para estimar sua performance preditiva. Ao final do processo, as performances estimadas de cada modelo, E_k , são utilizadas para calcular sua performance média.

Figura 3: Processo de validação cruzada *k-fold*, $k=10$.

Fonte: RASCHKA (2017).

Repetições da técnica de validação cruzada *k-fold* podem ser realizadas a fim de aumentar a precisão das estimativas de performance (KUHN; JOHNSON, 2013). Outra alternativa consiste na utilização de validação cruzada aninhada (*nested cross validation*) (VARMA; SIMON, 2006), 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.

A figura 4, a seguir, resume o processo de validação cruzada aninhada, com $k=10$ para a validação cruzada externa e para a validação cruzada interna. Na validação cruzada externa, os dados originais foram divididos em 10 partes e, a cada iteração, 9 partes foram consideradas para treinamento e uma para teste. Na validação cruzada interna, realizada a cada iteração da validação cruzada externa, o conjunto de treinamento foi também dividido em 10 partes e, para as iterações 1 a 10, 9 partes foram utilizadas para o treinamento de modelos com diferentes hiperparâmetros, e uma para estimar sua performance preditiva.

Figura 4: Processo de validação cruzada aninhada.

Fonte: RASCHKA (2017).

A validação cruzada aninhada é interessante quando o objetivo do pesquisador é selecionar um entre diferentes algoritmos de ML, resultando em uma boa estimativa do resultado esperado, em termos de erro de predição, para a otimização dos hiperparâmetros de um algoritmo, seguido de sua aplicação em novos dados (RASCHKA, 2017; VARMA; SIMON, 2006).

2.3.3 Hiperparâmetros de algoritmos de aprendizagem

Em *machine learning* (ML), a maioria dos algoritmos apresenta um ou mais parâmetros que controlam a complexidade (ou o equilíbrio entre viés e variância) do modelo ajustado. Tais parâmetros, denominados parâmetros de sintonização (*tuning parameter*) ou hiperparâmetros, podem ser diretamente especificados antes do ajuste do modelo preditivo, pois não são diretamente estimados pelos dados de treinamento, ou otimizados por validação cruzada, pois não há uma fórmula analítica disponível para o cálculo do seu valor apropriado (KUHN; JOHNSON, 2013). São exemplos de hiperparâmetros o número k de vizinhos mais próximos, utilizados pelo algoritmo *KNN*, e o número m de árvores de decisão presentes no algoritmo *Boosting*.

Como os hiperparâmetros apresentam relação com a complexidade (flexibilidade) de um modelo preditivo, escolhas inadequadas para o seu valor podem resultar, por exemplo, em sobreajuste (*overfitting*) e performance ruim do modelo em novas observações. Na prática, uma métrica é escolhida para avaliar o erro de predição, como o EQM (em problemas de regressão) e a AUC ROC (em problemas de classificação) e, para um dado algoritmo, essa métrica é avaliada em uma lista de valores candidatos ao hiperparâmetro por validação cruzada, com o objetivo de selecionar aquele que resulte em um modelo que minimiza o erro de predição (KUHN; JOHNSON, 2013).

2.3.4 Equilíbrio entre viés e variância

Determinados valores do hiperparâmetro irão implicar modelos mais simples, ao passo que outros irão resultar em modelos mais complexos. De modo geral, modelos simples apresentam variância baixa e viés alto, ou seja, sua função estimada não irá apresentar modificações substanciais entre diferentes conjuntos de treinamento (apresenta poucos parâmetros estimados – variância baixa), porém pode não ser uma boa aproximação para o padrão presente nesses dados (viés alto). Por outro lado, modelos mais complexos apresentam variância alta e viés baixo: sua função estimada pode ser composta por muitos parâmetros, de modo que pequenas modificações no conjunto de treinamento podem implicar funções estimadas muito diferentes (variância alta). No entanto, para um conjunto de treinamento em particular, tal função se aproxima muito bem do padrão presente nos dados e, dessa forma, apresenta viés reduzido. Tais características estão relacionadas a modelos que se ajustam bem aos dados de treinamento, mas que apresentam desempenho ruim quando aplicados a novas observações, ou seja, apresentam baixa capacidade de generalização (*overfitting*) (IZBICKI; SANTOS, 2018).

A figura 5 ilustra três situações para um problema de classificação com dois preditores (X_1 e X_2) e 200 observações de treinamento (classe 1, $n = 104$ e frequência de 52%, ilustrada em azul; e classe 2, $n = 96$, frequência de 48%, ilustrada em vermelho). Para cada situação, um número k de vizinhos foi determinado, definindo uma vizinhança, onde a proporção (p) de observações da classe 1 (azul) correspondeu à resposta predita, de acordo com a seguinte regra: classe 1 se $p > 0,5$ e classe 2 caso contrário. Portanto, na figura a seguir, de acordo com essa regra, as regiões coloridas indicam a classificação de todas as observações de treinamento como classe 1 (azul) ou classe 2 (vermelha).

Figura 5: Equilíbrio entre viés e variância.

Fonte: figura gerada a partir da adaptação de código em R disponibilizada pelo Professor Tiago Mendonça dos Santos no curso “Big Data: transformando dados em informação” do Insper – Instituto de Ensino e Pesquisa.

Observe que, de acordo com o primeiro modelo ($k=200$) todas as observações foram classificadas como classe 1, ou seja, a classe mais frequente, pois todo o conjunto de dados foi considerado para compor a vizinhança de uma observação, situação que indica viés alto e variância baixa. O segundo modelo, por sua vez, baseou-se na estimativa de k por validação cruzada ($k=9$) e representa, para esse caso, o modelo que equilibra viés e variância. Por fim, para o terceiro modelo, apenas um vizinho compôs a vizinhança de uma observação ($k=1$), de modo que a região considerada para realizar determinada predição ficou extremamente delimitada, caracterizando um modelo com viés baixo, variância alta e, portanto, mais instável a modificações do conjunto de treinamento.

Nesse contexto, de equilíbrio entre viés e variância, há problemas em que modelos menos complexos ainda representarão a melhor alternativa, pois estão menos sujeitos ao sobreajuste, resultando em predições mais acuradas para Y em novas observações. Entretanto, não há um único algoritmo capaz de apresentar boa performance em todas as aplicações, sendo interessante compará-los para escolher o mais adequado ao problema que se está estudando (JAMES et al., 2014).

2.3.5 Dados desbalanceados

Em problemas de classificação, o desbalanceamento ocorre quando uma ou mais classes apresentam proporção muito baixa quando comparada a outras classes da resposta de interesse. Algumas alternativas para reduzir o impacto do desbalanceamento na performance preditiva são representadas por: modificação dos pesos (probabilidade *a priori*) das observações de

treinamento cuja resposta de interesse está presente; otimização do modelo a fim de aumentar a sensibilidade da classe menos frequente; escolha de pontos de corte alternativos a fim de reduzir a taxa de erro da classe menos frequente; e técnicas de reamostragem *post hoc* (*down-sampling*, *up-sampling* e *SMOTE* – *synthetic minority over-sampling technique*) aplicadas a fim de balancear as classes da resposta de interesse (KUHN; JOHNSON, 2013).

2.3.6 Comparação de modelos

Diversos algoritmos de ML têm sido desenvolvidos para solucionar diferentes problemas, sendo essencial comparar, ao menos, algumas dessas ferramentas para selecionar aquela que resulta em um modelo com melhor performance preditiva, pois nenhum algoritmo domina todos os outros, em todos os conjuntos de dados possíveis; ou seja, em um conjunto de dados particular, um determinado algoritmo pode funcionar melhor, mas o mesmo pode não ser o que apresenta desempenho mais satisfatório em outro conjunto de dados (JAMES et al., 2014). Nesse contexto, uma vez que o valor ótimo dos hiperparâmetros tenha sido determinado para diferentes algoritmos, resultando em um conjunto de modelos preditivos candidatos ao modelo final, uma questão permanece: como escolher entre esses modelos?

Em muitos casos, alguns modelos serão equivalentes em termos de performance, logo é possível ponderar os benefícios dos diferentes algoritmos, como complexidade computacional, facilidade de implementar a função de predição estimada e interpretabilidade. A seguir, uma sugestão de comparação de modelos para escolha do modelo final descrita em Kuhn & Johnson (2013) é apresentada:

1. Avaliar, inicialmente, modelos menos interpretáveis e mais flexíveis, como por exemplo, *support vector machine*, *random forest* e *boosting*. Para muitos problemas, estes modelos irão apresentar resultados mais acurados.
2. Posteriormente, investigar modelos mais simples que os do passo anterior, como o *MARS* e as regressões penalizadas (*lasso* e *ridge*).
3. Se os modelos forem equivalentes, utilize o modelo mais simples que aproxime de modo razoável a performance de modelos mais complexos.

Nesse contexto, o próximo tópico apresenta as principais características de algoritmos utilizados na etapa de aprendizado de modelos preditivos, ajustados para os problemas de classificação e regressão que compõem a tese.

2.4 APRENDIZADO: ALGORITMOS DE MACHINE LEARNING

2.4.1 Métodos lineares para regressão

2.4.1.1 Modelo de regressão linear

O modelo de regressão linear representa um bom ponto de partida para a compreensão de algoritmos de ML, pois muitas das ferramentas atuais podem ser vistas como generalizações ou extensões desse modelo (JAMES et al., 2014). Como os trabalhos apresentados nesta tese estão voltados a problemas de predição, a natureza inferencial dos modelos de regressão linear não será explorada. Mais detalhes podem ser encontrados em literaturas específicas sobre o tema, como em Kutner, Nachtsheim e Neter (2004) e em McCullagh & Nelder (1989).

Modelos de regressão linear são utilizados para predizer respostas contínuas e resultam em descrição interpretável de relações entre variáveis preditoras e a resposta de interesse. Nesse contexto, objetiva-se “aprender” os pesos (ou parâmetros) da equação que melhor descreve a relação entre os preditores e a variável resposta para utilizá-los na predição dessa resposta em novas observações (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). Sua representação tem a seguinte forma:

$$f(X) = \beta_0 + \sum_{j=1}^p \beta_j X_j$$

Os parâmetros β_j , $j=1, 2, \dots, p$, são desconhecidos e os preditores X_j podem ser quantitativos, transformações de variáveis quantitativas (*log*, *raiz quadrada* ou *representações polinomiais*), variáveis indicadoras, códigos numéricos de variáveis ordinais ou ainda representações de interação entre variáveis preditoras. Entretanto, independente das características de X_j , o modelo será sempre linear nos parâmetros (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

A partir dos dados de treinamento, o método dos mínimos quadrados ordinários pode ser utilizado para estimar valores de $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ que minimizam a soma de quadrados dos resíduos (SQR), também denominada função perda do modelo de regressão linear:

$$SQR(\beta) = \sum_{i=1}^n (y_i - f(x_i))^2$$

$$= \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2$$

Para minimizar a SQR, considere $\mathbf{X}$, uma matriz de dimensão $n \times (p+1)$ em que cada linha representa uma observação com a respectiva mensuração para os preditores (com 1 na primeira posição), e $\mathbf{y}$ um vetor n -dimensional de mensurações para a resposta de interesse, no conjunto de treinamento. Dessa forma, a SQR pode ser escrita como:

$$SQR(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

Diferenciando com relação a β ,

$$\begin{aligned} \frac{\partial SQR}{\partial \beta} &= -2\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta) \\ &= -2(\mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{X}\beta) \end{aligned}$$

Assumindo que $\mathbf{X}$ tem posto completo e que $\mathbf{X}^T \mathbf{X}$ é positiva definida, pode-se igualar esta derivada a *zero*:

$$\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta) = 0$$

$$\mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{X}\beta = 0$$

para obter a solução única de estimativas para β :

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Assim, os valores preditos no conjunto de dados de treinamento serão dados por:

$$\hat{\mathbf{y}} = \mathbf{X} \hat{\beta} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

De modo que $\hat{y}_i = \hat{f}(x_i)$, e o valor predito de uma nova observação, x^* , é dado por $\hat{f}(x^*) = (1 : x^*)^T \hat{\beta}$.

O EQM, descrito em tópico anterior, representa a média da SQR minimizada para estimar os parâmetros do modelo e pode ser utilizado como medida da performance do modelo de regressão linear:

$$EQM = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$$

Estimativas decorrentes do método dos mínimos quadrados ordinários podem apresentar algumas características indesejáveis: 1) diante de um número elevado de preditores, $\mathbf{X}^T \mathbf{X}$ pode ser não singular, seja porque $\mathbf{X}$ não tem posto completo, como consequência da presença de preditores perfeitamente correlacionados, ou porque o número de preditores é maior que o de observações; 2) o método dos mínimos quadrados ordinários implica estimadores não viesados

para β , porém que podem apresentar variância alta (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

2.4.1.2 Modelos relacionados a técnicas de regularização

Técnicas de regularização podem ser aplicadas a fim de se obter estimadores viesados, porém com variância reduzida, para os parâmetros do modelo de regressão linear. A regularização impõe penalizações à complexidade do modelo preditivo e, dessa forma, pode contribuir para a redução do efeito de sobreajuste e, conseqüentemente, para a melhora da performance preditiva do modelo em novas observações. As técnicas mais utilizadas são o *ridge*, o *lasso* e o *elastic net* (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

A solução da regressão *ridge* é também uma função linear de $\mathbf{y}$, cujas estimativas dos parâmetros resultam da minimização de uma SQR (função perda) com penalização quadrática:

$$\hat{\beta}^{ridge} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}$$

O aumento do valor de λ implica aumento da penalização das estimativas de β . Observe que o intercepto β_0 , referente ao valor médio da resposta quando todas as variáveis independentes assumem valor *zero*, não é adicionado ao termo de penalização, pois o objetivo é penalizar parâmetros associados aos preditores.

Se a matriz de preditores for reparametrizada de forma a apresentar média *zero* para cada preditor, então os parâmetros de interesse podem ser obtidos em duas etapas: 1) β_0 será estimado pela média da resposta de interesse; 2) os demais parâmetros serão estimados por regressão *ridge* sem intercepto (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). Além de média *zero*, cada preditor pode ser dividido por seu respectivo desvio padrão, resultando em uma matriz de preditores padronizados, $\mathbf{Z}$, com p colunas. Assim, a SQR penalizada pode ser escrita na forma matricial como:

$$SQR(\lambda) = (\mathbf{y} - \mathbf{Z}\beta)^T (\mathbf{y} - \mathbf{Z}\beta) + \lambda \beta^T \beta$$

A solução *ridge*, após derivar com relação à β , é representada por:

$$\hat{\beta}^{ridge} = (\mathbf{Z}^T \mathbf{Z} + \lambda \mathbf{I})^{-1} \mathbf{Z}^T \mathbf{y}$$

Embora a regressão *ridge* esteja associada à obtenção de estimadores que apresentam variância reduzida, a mesma não resulta em parâmetros estimados iguais a *zero*. Uma alternativa, quando o interesse da regularização é também realizar uma seleção supervisionada

de variáveis, é representada pelo método *lasso*, em que a penalização quadrática é substituída por uma penalização baseada no valor absoluto do parâmetro:

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin} \left\{ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

Quando λ apresenta valores altos, determinados parâmetros poderão apresentar estimativas iguais a *zero*. Esse tipo de penalização torna a solução não linear em y e, diferente da regressão *ridge*, não apresenta solução analítica para β , ou seja, não há uma forma fechada para expressar o vetor de parâmetros estimados (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

Por fim, o método *elastic net*, representa uma combinação das duas penalizações, a quadrática e a baseada no valor absoluto, contribuindo tanto para a restrição das estimativas dos parâmetros como para a obtenção de soluções esparsas (KUHN; JOHNSON, 2013; RASCHKA, 2017):

$$\hat{\beta}^{Enet} = \underset{\beta}{argmin} \left\{ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda_1 \sum_{j=1}^p \beta_j^2 + \lambda_2 \sum_{j=1}^p |\beta_j| \right\}$$

A técnica de validação cruzada *k-fold* pode ser utilizada para otimizar o valor dos hiperparâmetros (λ , no caso da regressão *ridge* e do *lasso* e λ_1 e λ_2 no caso do *elastic net*) (KUHN; JOHNSON, 2013; RASCHKA, 2017).

2.4.2 Alternativa para não-linearidade em regressão

2.4.2.1 Multivariate Adaptive Regression Splines

As técnicas apresentadas anteriormente são especialmente importantes para solucionar problema de colinearidade e de dimensionalidade presentes no conjunto de dados. Além disso, podem amenizar o efeito do sobreajuste ao reduzir o número de parâmetros estimados e, consequentemente, a complexidade do modelo ajustado. Entretanto, o modelo de regressão linear pode apresentar problemas adicionais, como os relacionados à não linearidade da relação preditor-resposta (JAMES et al., 2014). Caso estejam presentes, estruturas não-lineares podem ser modeladas por meio da adição de novas variáveis à matriz de preditores, decorrentes de transformações das variáveis originais (como as representações polinomiais). Outra alternativa consiste em utilizar algoritmos que, por natureza, modelam relações não lineares entre preditores e respostas contínuas e que não requerem a especificação da forma exata da não

linearidade antes de sua aplicação aos dados de treinamento. A regressão adaptativa multivariada *splines* (*Multivariate Adaptive Regression Splines – MARS*) é um exemplo e será apresentada com mais detalhes no próximo tópico (KUHN; JOHNSON, 2013).

Outros algoritmos aplicáveis a problemas de regressão, porém também utilizados no contexto de classificação, como o método dos K vizinhos mais próximos (*K-Nearest Neighbors – KNN*), o *support vector machine* e as redes neurais artificiais (*artificial neural networks*), serão apresentados após o modelo de regressão logística. Por fim, algoritmos que utilizam árvores de decisão (*bagging*, *random forest*, *boosting* e *cubist*) serão apresentados ao final desta seção.

A principal motivação do método *Multivariate Adaptive Regression Splines (MARS)* está relacionada à utilização de pares refletidos de cada preditor X_j , em vez do preditor original, como termos do modelo preditivo. Esses pares resultam da aplicação de expansões em funções base, lineares por partes, da forma $(x - a)_+$ e $(a - x)_+$, em que o símbolo “+” indica parte positiva, e a representa o ponto de corte para a formação do par:

$$(x_j - a)_+ = \begin{cases} x_j - a, & x_j > a \\ 0, & x_j \leq a \end{cases}$$

E,

$$(a - x_j)_+ = \begin{cases} a - x_j, & x_j < a \\ 0, & x_j \geq a \end{cases}$$

Dessa forma, pode-se obter uma coleção de funções base, representada por:

$$C = \{(X_j - a)_+, (a - X_j)_+\}_{a \in \{x_{1j}, x_{2j}, \dots, x_{nj}\}} \\ j=1, 2, \dots, p.$$

O modelo é construído a partir da seleção de variáveis presentes no conjunto C , e pode ser descrito da seguinte forma:

$$f(X) = \beta_0 + \sum_{m=1}^M \beta_m h_m(X)$$

onde cada $h_m(X)$ é uma função em C , ou o produto de uma ou mais funções desse conjunto. Uma vez que h_m é escolhida, o coeficiente β_m é estimado por um modelo de regressão linear padrão.

O principal diferencial do *MARS* está na definição das funções $h_m(X)$. A construção do modelo, inicia-se por uma função constante, $h_0(X)=1$, e todas as funções do conjunto C são candidatas para as etapas seguintes. Em cada etapa, um novo par de funções é caracterizado por

produtos entre uma função h_m , presente no conjunto de modelos M , e um par de funções do conjunto C , como representado a seguir:

$$\hat{\beta}_{M+1}h_l(X).(X_j - a)_+ + \hat{\beta}_{M+2}h_l(X).(a - X_j)_+, h_l \in M$$

Na primeira etapa do método, como a única função de M é a função constante, o novo par de funções a ser adicionado ao modelo terá a seguinte forma:

$$\beta_1(X_j - a)_+ + \beta_2(a - X_j)_+; \quad a \in \{x_{ij}\}$$

A escolha do par de funções de C está relacionada à combinação preditor-ponto de corte responsável pela maior redução do EQM. Uma vez que o valor de a e dos parâmetros β_1 e β_2 tenham sido estimados, o par de funções correspondente será adicionado ao conjunto M e, na etapa seguinte, além do produto resultante da multiplicação pela função constante, produtos do tipo $h_m(X).(X_j - a)_+$ e $h_m(X).(a - X_j)_+$, $a \in \{x_{ij}\}$, serão considerados como possíveis escolhas para o par de novas funções a serem adicionadas ao modelo. A figura 6 ilustra os três primeiros passos do método, com destaque em vermelho para o produto entre funções de M , à esquerda, e de C , à direita, selecionadas em cada passo (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

Figura 6: Ilustração dos 3 primeiros passos do método *MARS*.

Fonte: Hastie, Tibshirani & Friedman (2009).

Ao final desse processo, o modelo, usualmente, resulta em sobreajuste dos dados de treinamento. Portanto, um procedimento de seleção de variáveis é aplicado a fim de remover, sequencialmente, variáveis relacionadas ao menor aumento no EQM, produzindo assim um modelo mais parcimonioso. O número de variáveis a serem retidas do modelo é considerado um hiperparâmetro, otimizado por validação cruzada *k-fold* (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

2.4.3 Métodos lineares para classificação

Em problemas de aprendizado associados a respostas qualitativas a predição de interesse resulta na classificação da observação em uma das categorias (ou classes) da variável resposta.

Nesse contexto, o modelo preditivo pode apresentar dois tipos de valores preditos: um referente à probabilidade da observação pertencer a cada uma das classes, e outro correspondente ao rótulo da classe predita para essa observação, como por exemplo, aos valores 0 ou 1, representativos, respectivamente, da ausência e presença de determinada condição de interesse em respostas binárias (JAMES et al., 2014; KUHN; JOHNSON, 2013).

De modo geral, o aprendizado de um problema de classificação consiste no particionamento do espaço dos preditores em regiões correspondentes às categorias da resposta de interesse. A fronteira de decisão entre essas regiões, denominada classificador, representa o modelo preditivo estimado por determinado algoritmo de aprendizado (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). Portanto, o objetivo do aprendizado é construir um classificador, $f(X)$, que faça boas previsões da resposta de interesse em observações futuras (IZBICKI; SANTOS, 2018):

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

A fronteira de decisão pode ser pouco ou muito flexível, dependendo do algoritmo utilizado. O caso mais simples, e com resultados satisfatórios para a maioria dos problemas de classificação, está relacionado a estimativas de fronteiras de decisão lineares, como as resultantes do ajuste de um modelo de regressão logística, descrito a seguir.

2.4.3.1 Regressão logística

A regressão logística consiste em um modelo linear para problemas de classificação, cujo objetivo é prever a probabilidade de cada observação pertencer a uma das classes da resposta de interesse. Para que essa probabilidade possa ser estimada, a variável resposta do conjunto de treinamento é modelada por meio da distribuição binomial, que tem como parâmetro, p , a probabilidade de ocorrência de uma classe específica. É importante destacar que a probabilidade estimada de determinado evento deve estar limitada ao intervalo $[0,1]$ e apresentar relação direta com os preditores mensurados para cada observação do conjunto de dados (KUHN; JOHNSON, 2013). Para tanto, utiliza-se a função logística:

$$p(X) = \Pr(Y = k \mid X = x) = \frac{\exp(\beta_0 + \sum_{j=1}^p \beta_j X_j)}{1 + \exp(\beta_0 + \sum_{j=1}^p \beta_j X_j)}$$

que resulta no modelo logístico, responsável por modelar a relação entre a probabilidade de determinada resposta, $p(X) = Pr(Y = k | X = x)$, e o conjunto de preditores, X . É importante observar que em predição não se assume que essa relação atenda a pressupostos do modelo, requeridos quando há propósito inferencial, apenas se espera que ela resulte em um bom classificador (IZBICKI; SANTOS, 2018).

O ajuste do modelo logístico pode ser realizado a partir da aplicação do método da máxima verossimilhança, que estima valores para o vetor de parâmetros, β . Essas estimativas corresponderão àquelas que resultam em uma probabilidade predita para cada observação do conjunto de treinamento, $\hat{p}(x_i)$, mais próxima da verdadeira classe a qual a observação pertence (JAMES et al., 2014). A função de verossimilhança formaliza, matematicamente, o método:

$$L(\beta) = \prod_{i=1}^n \left(p(X)^{y_i} (1 - p(X))^{1-y_i} \right)$$

de modo que as estimativas de β serão escolhidas a fim de maximizar essa função. Por fim, o vetor de parâmetros estimados, $\hat{\beta}$, será utilizado para prever a resposta de interesse em novas observações (JAMES et al., 2014). Mais detalhes sobre o método da máxima verossimilhança e sobre as características inferenciais do modelo logístico podem ser encontrados em Agresti (2013) e em Hosmer, Lemeshow e Sturdivant (2013).

Para uma resposta dicotômica, rotulada com 0s e 1s, se $p(X)$ é a probabilidade associada à presença de determinada resposta, a chance $(p/1-p)$ desse desfecho ocorrer será:

$$\frac{Pr(Y = 1 | X = x)}{Pr(Y = 0 | X = x)} = \frac{p(X)}{1 - p(X)} = \exp(\beta_0 + \sum_{j=1}^p \beta_j X_j)$$

Em inferência estatística, a chance é a medida de escolha para descrever a relação entre a resposta de interesse e os preditores mensurados. Além disso, a transformação *logit*, quando aplicada à essa medida, torna o modelo linear em X :

$$\log \left(\frac{p(X)}{1 - p(X)} \right) = \beta_0 + \sum_{j=1}^p \beta_j X_j$$

A fronteira de decisão linear do modelo de regressão logística estará relacionada à escolha de um ponto de corte para $p(X)$. Por exemplo, se o ponto de corte escolhido for $p(X) = 0,5$, indivíduos com $p(X) > 0,5$ serão classificados como um evento (presença de determinada resposta) e aqueles com $p(X) < 0,5$ como um não evento (ausência de determinada resposta). Logo, $p(X) = 0,5$ irá definir a fronteira de decisão linear para tal modelo, cujo *log* da chance, nesse caso, será *zero* (JAMES et al., 2014). Por fim, vale destacar que as penalidades descritas

no contexto de regressão linear (*ridge*, *lasso* e *elastic net*) podem também ser utilizadas em conjunto com a regressão logística.

2.4.4 Outras alternativas para não linearidade

2.4.4.1 Método dos K vizinhos mais próximos

O método dos K vizinhos mais próximos (*K-Nearest Neighbors – KNN*) representa uma alternativa não paramétrica, tanto para problemas de regressão como para os de classificação, quando a relação entre os preditores e a resposta de interesse requer um modelo mais flexível, como por exemplo nos casos de não linearidade (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; KUHN; JOHNSON, 2013).

A etapa de treinamento do *KNN* consiste em armazenar os preditores e a resposta de interesse, de forma que a predição da resposta para uma nova observação é realizada utilizando diretamente o conjunto de treinamento, e não uma função ajustada a partir desses dados. Assim, dada uma nova observação, x^* , representada por um vetor p -dimensional de mensurações para os preditores de interesse, $x^* = (x_1^*, \dots, x_p^*)^T$, a predição da resposta é realizada considerando as observações individuais do conjunto de treinamento mais próximas (ou similares) à x^* .

Basicamente, esse método apresenta duas características importantes que antecedem a predição da resposta de interesse para a nova observação: a definição do número k de vizinhos mais próximos que irão compor a vizinhança da nova observação; e a escolha da medida de distância responsável por identificar as k observações do conjunto de treinamento mais próximas, ou mais similares, à nova observação (RASCHKA, 2017).

O número k de vizinhos mais próximos é considerado um hiperparâmetro que determina o quão bem o modelo ajustado irá generalizar para dados futuros. A escolha de k apresenta relação direta com o limiar viés-variância: se $K = N$, em que N representa o total de observações do conjunto de treinamento, a resposta predita será sempre a mesma, o que caracteriza um cenário de viés alto e variância baixa. Por outro lado, se $K = 1$, o valor predito estará mais vulnerável a ruídos ou dados discrepantes presentes no conjunto de treinamento, apresentando, desta forma, variância alta, porém menor viés que o caso descrito anteriormente. A otimização de k pode ser realizada por técnicas de reamostragem, como a validação cruzada *k-fold* (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; KUHN; JOHNSON, 2013; LANTZ, 2013).

Como descrito anteriormente, o *KNN* utiliza a proximidade entre a nova observação e as demais amostras do conjunto de treinamento para definir uma vizinhança. O termo proximidade implica medir a distância entre as observações, com base nos preditores disponíveis. Frequentemente, uma medida simples de distância, denominada distância euclidiana, pode ser utilizada para medir proximidade:

$$d(x^*, x_i) = \sqrt{\sum_{j=1}^p (x^*_j - x_{ij})^2}$$

onde x^* e x_i representam vetores p -dimensionais correspondentes às mensurações para os preditores da nova observação e de uma das observações do conjunto de treinamento, respectivamente. Para a utilização dessa métrica é importante que os dados sejam padronizados na etapa de pré-processamento de modo que a contribuição de cada preditor para o cálculo da distância seja igual. Por fim, após a otimização de k , o conjunto de treinamento pode ser utilizado para prever a resposta de interesse em novas observações (RASCHKA, 2017; WU; KUMAR, 2009).

Após a definição do número de vizinhos e da medida de distância a ser utilizada, na abordagem mais simples, em problemas de regressão, a resposta de x^* corresponde à média das respostas observadas em sua vizinhança, $N_k(x)$, definida por k amostras de treinamento com vetor de preditores, x_i , mais similares à x^* , ou seja, por seus k vizinhos mais próximos:

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

Para problemas de classificação, a resposta predita será representada pela classe mais frequente, observada na vizinhança de x^* . Adicionalmente, para cada classe j da resposta de interesse, é possível calcular a probabilidade condicional de a nova observação pertencer a j -ésima classe por meio da fração de pontos em $N_k(x)$ cujo valor da resposta é j :

$$P(Y = j | X = x^*) = \frac{1}{K} \sum_{x_i \in N_k(x)} I(y_i = j)$$

A classe j predita para x^* será representada por aquela que apresentar a maior probabilidade condicional.

2.4.4.2 Support Vector Machines

O *Support Vector Machine* (SVM) representa uma generalização de um algoritmo denominado classificador de margem máxima (*maximal margin classifier*), que acomoda fronteiras não lineares em problemas de classificação. Adicionalmente, o SVM difere dos demais algoritmos utilizados para classificação por não estimar, diretamente, probabilidades, mas sim a classe da resposta de interesse para uma nova observação.

A seguir, as principais características do classificador de margem máxima serão exploradas para, posteriormente, apresentar mais detalhes sobre o algoritmo SVM. Para tanto, considere uma matriz de dados $\mathbf{X}$, de dimensão $n \times p$, correspondente a n observações do conjunto de treinamento em um espaço p -dimensional, uma variável resposta composta pelas classes $\{-1, 1\}$ e uma nova observação representada por um vetor p -dimensional de mensurações para os preditores de interesse, $\mathbf{x}^* = (x_1^*, \dots, x_p^*)^T$.

Nesse cenário, o classificador de margem máxima pode ser utilizado para estimar uma fronteira de decisão linear, que separa perfeitamente as observações do conjunto de treinamento de acordo com a classe a que as mesmas pertencem, ou seja, estimar a equação de um hiperplano cujas regiões de classificação contenham apenas uma das classes da resposta de interesse:

$$\begin{cases} \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} > 0, & y_i = 1 \\ \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} < 0, & y_i = -1 \end{cases}$$

Portanto, de acordo com esse classificador, se $f(\mathbf{x}^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \dots + \beta_p x_p^*$ for positivo, $\mathbf{x}^*$ será classificada na categoria 1 e, se negativo, na categoria -1 (JAMES et al., 2014).

Em geral, se os dados podem ser perfeitamente separados, então um número infinito de hiperplanos candidatos à fronteira de decisão pode ser definido. Nesse caso, o hiperplano que maximiza a margem de separação (menor distância perpendicular) entre observações de treinamento das duas classes, denominado hiperplano de margem máxima, será o escolhido para representar a fronteira de decisão. Por fim, o classificador de margem máxima irá considerar a localização da nova observação com relação à fronteira de decisão, para realizar a classificação. A figura 7 ilustra o classificador de margem máxima para o caso separável quando apenas dois preditores (x_1 e x_2) são utilizados. Observe que, nesse caso, o hiperplano de margem máxima está representado em $\mathbb{R}^2$, por uma reta. Para o caso em que p preditores

são utilizados, a fronteira de decisão será representada por um hiperplano em $\mathbb{R}^p$ (JAMES et al., 2014).

Figura 7: Classificador de margem máxima para o caso separável. A linha contínua central representa o hiperplano de margem máxima e a margem é representada por ambas as distâncias entre o hiperplano e as linhas pontilhadas, definidas por vetores suporte.

Fonte: James et al. (2014).

Embora esse classificador seja um algoritmo simples para problemas de classificação, sua implementação só é possível se existir um hiperplano que separe perfeitamente os dados de treinamento de acordo com a classe a que os mesmos pertencem. No entanto, para a maioria dos problemas reais, não é possível definir esse hiperplano, pois algumas observações de diferentes categorias da variável resposta podem estar sobrepostas (JAMES et al., 2014).

Nesse caso, para fronteiras de decisão lineares, é possível ampliar o conceito de hiperplano separador a fim de considerar hiperplanos de margens suaves, também denominados hiperplanos de vetores suporte, que ao separar observações de classes distintas toleram que algumas estejam localizadas do lado incorreto da margem, ou até mesmo, do hiperplano separador (JAMES et al., 2014). O algoritmo associado aos hiperplanos de vetores suporte é denominado classificador de vetor suporte (*support vector classifier*) e está representado na figura 8, a seguir.

Figura 8: Classificador de margem máxima para o caso não separável. A linha contínua central representa o hiperplano de margem máxima e a margem é representada por ambas as distâncias entre o hiperplano e as linhas pontilhadas, definidas por vetores suporte.

Fonte: James et al. (2014).

A solução para o problema de otimização desse classificador está associada à identificação de $\beta_0, \beta_1, \dots, \beta_p$ que maximize o tamanho da margem, representado por M , entretanto com restrições relacionadas à classificação errada de observações do conjunto de treinamento com relação ao hiperplano ou à margem:

$$\begin{aligned}
 & \underset{\beta_0, \beta_1, \dots, \beta_p}{\text{maximizar}} \quad M, \\
 & \text{sujeito a} \quad \sum_{j=1}^p \beta_j^2 = 1, \\
 & \quad y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}) \geq M(1 - \varepsilon_i), \\
 & \quad \varepsilon_i \geq 0, \sum_{i=1}^n \varepsilon_i \leq C
 \end{aligned}$$

onde ε_i , $i = 1, 2, \dots, n$, são variáveis (*slack variables*) relacionadas à localização de uma observação com relação à margem ou ao hiperplano separador. Se a i -ésima observação estiver no lado correto da margem, então $\varepsilon_i = 0$, caso contrário, $\varepsilon_i > 0$. Adicionalmente, se a observação estiver do lado errado não só da margem, mas também do hiperplano separador, então $\varepsilon_i > 1$ (JAMES et al., 2014).

A constante C representa um hiperparâmetro referente ao custo do número e severidade de violações toleráveis (classificações incorretas) com relação à margem e ao hiperplano. Dessa

forma, C controla o limiar viés-variância do algoritmo de aprendizado: conforme seu valor aumenta, maior a margem e, portanto, mais tolerante a classificações incorretas será o modelo ajustado, situação equivalente a obter um classificador mais viesado, porém com menor variância; por outro lado, a medida que C diminui, a margem torna-se mais estreita, de modo que a mesma será raramente violada, o que conduz a um modelo com viés baixo, porém com variância alta. Como descrito anteriormente, por ser um hiperparâmetro, C pode ser otimizado por validação cruzada *k-fold* (JAMES et al., 2014).

O problema de otimização relacionado ao classificador de vetor suporte utiliza o produto interno entre as observações, em vez das observações propriamente ditas (JAMES et al., 2014). O produto interno de duas observações x_i e $x_{i'}$ é dado por:

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^p x_{ij} x_{i'j}$$

Deste modo, o *classificador de vetor suporte* pode ser representado como:

$$f(x) = \beta_0 + \sum_{i=1}^n \alpha_i \langle x^*, x_i \rangle$$

Logo, ao avaliar $f(x^*)$ é preciso calcular o produto interno entre a nova observação e cada observação do conjunto de treinamento. Entretanto, α_i será diferente de zero apenas para os vetores suporte, ou seja, para as observações do conjunto de treinamento responsáveis pela definição da margem, ou que violam a margem do hiperplano separador. Se S corresponder à coleção de índices dos vetores suporte, o classificador pode ser reescrito da seguinte forma:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \langle x^*, x_i \rangle$$

Agora, considere uma generalização do produto interno da forma:

$$K(x_i, x_{i'})$$

onde K é uma função, denominada *Kernel*, que quantifica a similaridade entre duas observações. Para o classificador de vetor suporte, essa função é conhecida como *Kernel* linear e quantifica a similaridade entre as observações por meio da correlação de Pearson dos preditores padronizados:

$$K(x_i, x_{i'}) = \sum_{j=1}^p x_{ij} x_{i'j}$$

O classificador de vetor suporte é uma abordagem natural quando a fronteira de decisão entre as categorias da resposta de interesse é linear. Entretanto, algumas vezes, a fronteira de decisão não é linear, e determinados artifícios podem ser utilizados para que esse

comportamento seja captado pelo classificador, como por exemplo, aumentar o espaço dos preditores, adicionando variáveis correspondentes a transformações (como a *polinomial*) dos preditores originais. Nesse sentido, o algoritmo *SVM* representa uma extensão do classificador de vetor suporte que resulta da ampliação do espaço de variáveis de uma maneira específica, utilizando outras formas, que não a linear, para a função *Kernel* (JAMES et al., 2014).

Os *Kernels polinomial* e *radial* representam as formas não lineares mais frequentemente utilizadas em problemas de aprendizado que apresentam relações não lineares entre os preditores e a resposta de interesse. Um *Kernel polinomial* de grau d é representado como:

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^p x_{ij} x_{i'j})^d$$

E o *radial* como:

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^p (x_{ij} - x_{i'j})^2)$$

Estas funções conduzem a uma fronteira de decisão muito mais flexível que a linear. Nestes casos, a função, $f(x)$, tem a seguinte forma:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x^*, x_i)$$

Há uma extensão do algoritmo *SVM* para respostas contínuas. É importante recordar que, no modelo de regressão linear, o método dos mínimos quadrados ordinários foi utilizado a fim de identificar estimativas para $\beta_0, \beta_1, \dots, \beta_p$ responsáveis por minimizar uma função perda, representada pela SQR. No caso do *SVM*, o problema de otimização visa identificar estimativas que minimizam uma função perda diferente, relacionada apenas aos resíduos maiores, em valor absoluto, que determinada constante, condição que estende o conceito de margem, do classificador de vetor suporte, para o contexto de regressão (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; JAMES et al., 2014).

Primeiramente, considere o modelo de regressão linear, $f(x) = \beta_0 + x^T \beta$, e a minimização de H para estimar β :

$$H(\beta, \beta_0) = \sum_{i=1}^n V(y_i - f(x_i))$$

A função V corresponde a uma medida de erro “ ε -insensitiva”, que ignora erros de tamanho menor que ε e, dessa forma, define os vetores suporte em regressão:

$$V_H = \begin{cases} 0, & |r| \leq \varepsilon \\ |r| - \varepsilon, & c.c \end{cases}$$

Assim como em problemas de classificação, a função solução do *SVM* para problemas de regressão terá a seguinte forma:

$$f(x) = \beta_0 + \sum_{i=1}^n \alpha_i \langle x^*, x_i \rangle$$

e apenas o subconjunto de valores estimados para α_i referentes aos vetores suporte serão diferentes de zero. Além disso, a solução depende de produtos internos entre os vetores de mensurações dos preditores, em vez das observações propriamente ditas, logo diferentes *Kernels* podem ser utilizados para os produtos internos, como o *polynomial* e o *radial*, anteriormente apresentados (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

Nesses casos, o modelo será da forma $f(x) = \beta_0 + \sum \beta_m h_m(x)$, onde m corresponde ao número de funções base que aproximam a função de regressão, e a solução do *SVM* será representada por:

$$f(x) = \beta_0 + \sum_{i=1}^n \alpha_i K(x^*, x_i)$$

$$\text{com } K(x_i, x_{i'}) = \sum_{m=1}^M h_m(x_i) h_m(x_{i'}).$$

2.4.4.3 Redes Neurais Artificiais

Uma rede neural artificial, em sua forma mais simples, caracteriza-se por um modelo em dois estágios, aplicável à predição tanto de respostas contínuas como de respostas categóricas. Esse modelo é, usualmente, representado por um diagrama de rede, como o da Figura 9, que no caso de problemas de regressão, apresenta $K=1$ com apenas um desfecho (*output*), Y_1 no topo do diagrama. Para problemas de classificação com resposta definida por k classes, K unidades estarão presentes no topo do diagrama, Y_k , $k = 1, \dots, K$, correspondentes a variáveis binárias (0-1) identificadoras de cada classe, em que a k -ésima unidade modela a probabilidade da classe k (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

Figura 9: Diagrama de rede para um modelo de redes neurais com apenas uma camada intermediária de variáveis latentes.

Fonte: Hastie, Tibshirani & Friedman (2009).

No primeiro estágio, a rede neural modela a variável resposta a partir de um conjunto intermediário de m variáveis ou unidades latentes (*hidden units*), resultantes de combinações lineares dos preditores originais (KUHN; JOHNSON, 2013). Seu diferencial, nessa etapa, está na possibilidade de utilizar uma transformação não linear das combinações, denominada função de ativação $\sigma(\cdot)$, resultando na seguinte representação:

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), \quad m = 1, \dots, M$$

onde Z_m são as unidades latentes representadas na camada intermediária (ou camada oculta) do diagrama de rede, que serão as variáveis de entrada (*input*) para o modelo do segundo estágio.

A função sigmoide $\sigma(\alpha_{0m} + \alpha_m^T X) = 1 / (1 + e^{-(\alpha_{0m} + \alpha_m^T X)})$ geralmente é a escolhida para $\sigma(\cdot)$.

Observe que, se $\sigma(\cdot)$ for a função identidade, o resultado será um modelo linear. Logo, pode-se concluir que uma rede neural representa uma generalização não linear de modelos lineares para problemas de regressão e de classificação (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

Uma vez que o número de unidades latentes, Z_m , tenha sido definido, cada uma é relacionada à resposta de interesse, caracterizando o segundo estágio do modelo de redes neurais:

$$\begin{aligned} T_k &= \beta_{0k} + \beta_k^T Z, \quad k = 1, \dots, K \\ f_k(X) &= g_k(T), \quad k = 1, \dots, K \end{aligned}$$

A função $g_k(T)$ é responsável pela transformação final dos vetores T_k . Em problemas de regressão, utiliza-se a função identidade, $g_k(T) = T_k$, e em classificação, uma transformação não linear, denominada *softmax*, é aplicada em T_k a fim de garantir que os valores preditos estejam no intervalo $[0,1]$:

$$g_k(T) = \frac{\exp(T_k)}{\sum_{k=1}^K \exp(T_k)}$$

Os parâmetros desconhecidos, θ , do modelo de redes neurais representados por $\{\alpha_{0m}, \alpha_m; m=1,2,...,M\}$ e $\{\beta_{0k}, \beta_k; k=1,2,...,K\}$, devem ser estimados a fim de ajustar corretamente os dados de treinamento. Para regressão, a SQR é utilizada como função perda:

$$R(\theta) = \sum_{k=1}^K \sum_{i=1}^n (y_{ik} - f_k(x_i))^2$$

E para classificação, pode-se utilizar a função *cross-entropy* (*deviance*):

$$R(\theta) = -\sum_{i=1}^n \sum_{k=1}^K y_{ik} \log f_k(x_i)$$

Minimizar $R(\theta)$ pode representar um problema de otimização numérica desafiador, pois é preciso inicializar os parâmetros, a partir de valores aleatórios, e utilizar técnicas especializadas para sua solução, como o *back-propagation*. Além disso, devido ao número elevado de parâmetros que precisam ser estimados, a solução de $R(\theta)$ tem tendência ao sobreajuste dos dados de treinamento e, portanto, técnicas de regularização devem ser utilizadas (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

Em redes neurais uma abordagem denominada *weight decay* é utilizada como técnica de regularização, resultando em efeito similar ao da regressão *ridge* em modelos lineares. De acordo com essa técnica, uma penalidade é adicionada à função perda, que passa a ser representada por $R(\theta) + \lambda J(\theta)$, em que $J(\theta)$ é dado por:

$$J(\theta) = \sum_{km} \beta_{km}^2 + \sum_{ml} \alpha_{ml}^2$$

e $\lambda \geq 0$ é um hiperparâmetro otimizado por validação cruzada *k-fold*. Em geral, os valores de λ variam entre 0 e 0,1 e, conforme seu valor aumenta, maior a restrição dos pesos dos modelos a valores próximos de *zero* e menor o efeito do sobreajuste (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; KUHN; JOHNSON, 2013).

O modelo em dois estágios descrito anteriormente, com apenas uma camada de unidades latentes, representa a estrutura mais simples de uma rede neural. O número de unidades latentes

em uma camada, por sua vez, depende do número de preditores e de observações presentes no conjunto de treinamento, sendo comum adicionar um número razoavelmente grande de unidades e treinar o modelo com regularização (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; KUHN; JOHNSON, 2013).

Outros modelos, em que mais camadas ocultas são utilizadas como passo intermediário entre os preditores originais e a resposta de interesse, podem ser desenvolvidos, caracterizando as redes neurais artificiais profundas. Nesse contexto, um conjunto de algoritmos, denominado *deep learning*, é utilizado para treinar as redes neurais de forma mais eficiente. De modo geral, o *deep learning* alcança alta flexibilidade e performance por meio da representação dos dados observados como uma hierarquia aninhada de funções matemáticas simples, cada uma associada a uma camada oculta diferente do modelo. Assim, a cada aplicação de uma função diferente, uma nova representação dos dados de entrada (*input*) é obtida e, quanto mais profundas as camadas, mais abstratos são os recursos extraídos dos dados para obter uma dada representação (GOODFELLOW; BENGIO; COURVILLE, 2016).

Por exemplo, em análise de imagem, mapear diretamente um conjunto de *pixels* para a identificação de um objeto representa um problema extremamente desafiador. O *deep learning* analisa esse problema por meio da divisão de um mapeamento único e complexo em uma sequência de mapeamentos aninhados mais simples, cada um descrito por uma camada oculta diferente. Dessa forma, assim como em redes neurais simples, a primeira camada é representada por dados observados e, na sequência, as camadas ocultas extraem características cada vez mais abstratas da imagem: diante dos *pixels*, a primeira camada oculta pode identificar as bordas da imagem comparando o brilho de *pixels* vizinhos. Em seguida, dada a descrição das bordas, a segunda camada oculta pode procurar por cantos e contornos, que são identificáveis por coleções de arestas. Por fim, a partir da descrição dos cantos e contornos, a terceira camada oculta pode detectar partes inteiras de objetos específicos ao identificar coleções específicas de cantos e contornos. Tal processo, de descrição da imagem por partes, ao final, possibilita o reconhecimento do objeto que a mesma contém (GOODFELLOW; BENGIO; COURVILLE, 2016).

Deep learning tem apresentado resultados promissores na área médica, ao ser aplicado em análise de imagens como ferramenta auxiliar no diagnóstico de doenças, como o câncer de pele (ESTEVA et al., 2017) e a retinopatia diabética (GULSHAN et al., 2016). Mais informações sobre esse conjunto de algoritmos podem ser encontrados em GoodFellow, Bengio e Courville (2016).

2.4.5 Métodos baseados em árvores de decisão

2.4.5.1 Árvores simples

Métodos baseados em árvores representam uma alternativa interessante para a construção de modelos preditivos quando a relação entre os preditores e a resposta de interesse é não linear e complexa. As árvores de regressão, para respostas contínuas, e as de classificação, para respostas categóricas, representam as abordagens mais simples (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; JAMES et al., 2014).

Para ambos os casos, o algoritmo pode ser resumido em duas etapas: 1) partição do espaço dos preditores, ou seja, do conjunto de valores possíveis para $X_1, X_2, \dots, X_p$, em J regiões distintas e disjuntas $R_1, R_2, \dots, R_J$; 2) ajuste de um modelo simples (usualmente uma constante) para predição da resposta de interesse em cada região. Como o conjunto de regras utilizado para o particionamento do espaço dos preditores pode ser completamente descrito por uma árvore, esta abordagem é conhecida como árvore de decisão (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; JAMES et al., 2014).

A construção de uma *árvore* de regressão é baseada na minimização da seguinte SQR:

$$\sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

onde $\hat{y}_{R_j}$ é a resposta média das observações de treinamento alocadas na j -ésima região. Entretanto, é computacionalmente inviável considerar todas as partições possíveis do espaço dos preditores em J regiões, e uma abordagem denominada divisão binária recursiva (*recursive binary splitting*) é, frequentemente, utilizada para a construção da árvore: a mesma inicia-se com todas as observações de treinamento pertencentes a uma única região. Na sequência, todos os preditores $X_1, X_2, \dots, X_p$ e pontos de corte, s , possíveis são considerados a fim de escolher a combinação preditor-ponto de corte que resulta na menor SQR, ou seja, para um j e s quaisquer, objetiva-se definir um par de regiões (JAMES et al., 2014):

$$R_1(j, s) = \{X \mid X_j < s\} \quad e \quad R_2(j, s) = \{X \mid X_j \geq s\}$$

E busca-se os valores de j e s que minimizem:

$$\sum_{i: X_i \in R_1(j, s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i: X_i \in R_2(j, s)} (y_i - \hat{y}_{R_2})^2$$

onde $\hat{y}_{R_1}$ e $\hat{y}_{R_2}$ representam, respectivamente, a resposta média das observações de treinamento na $R_1(j, s)$ e na $R_2(j, s)$. Posteriormente, uma ou ambas as regiões resultantes são divididas em mais duas regiões, e esse processo continua até que um critério de parada (usualmente, um número de observações mínimo para compor uma região) seja alcançado (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; JAMES et al., 2014).

Uma vez que as regiões $R_1, R_2, \dots, R_j$ tenham sido criadas, a predição da resposta para uma nova observação corresponderá à resposta média das observações de treinamento da região a qual a nova observação foi alocada. Para atribuir uma nova observação a determinada região, os valores mensurados para seus preditores são utilizados (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; JAMES et al., 2014).

A árvore completa (cheia), T_0 , com J regiões, é, geralmente, muito complexa e, portanto, predispõe ao sobreajuste do conjunto de treinamento e ao aumento do erro de generalização para novas observações. Esse problema pode ser solucionado por meio do controle do tamanho de T_0 , denominado poda, que resulta em uma subárvore T , menos complexa e com performance preditiva potencialmente melhor que a de T_0 em novas observações (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; JAMES et al., 2014).

O critério para podar T_0 é semelhante ao utilizado por técnicas de regularização descritas anteriormente. No presente caso, um hiperparâmetro λ é responsável por controlar o limiar entre o tamanho da árvore e a qualidade de ajuste do modelo, e sua otimização pode ser realizada por validação cruzada k -fold (KUHN; JOHNSON, 2013).

Para cada valor de λ há uma subárvore correspondente, T_λ , para a qual a função perda, descrita a seguir, é mínima:

$$\sum_{m=1}^{|T|} \sum_{i: x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \lambda |T|$$

A quantidade $|T|$ indica o número de nós terminais (*terminal nodes*) presentes na subárvore T , R_m é a região correspondente ao m -ésimo nó terminal, e $\hat{y}_{R_m}$ a resposta predita para observações de R_m . Conforme λ aumenta, menor será a subárvore T_λ responsável por minimizar a função perda e, portanto, mais interpretável será o modelo ajustado (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; JAMES et al., 2014).

A árvore de classificação é muito similar à de regressão, exceto pelo fato de que a predição da resposta para uma nova observação está relacionada à classe mais frequente entre

as observações de treinamento da região à qual a nova observação pertence. Assim, se o objetivo do problema de aprendizado for classificar uma resposta que assume os valores $1, 2, \dots, K$, para a aplicação da divisão binária recursiva e do processo de poda, outros critérios, que não a SQR, serão utilizados, tais como a taxa de erro de classificação, o índice Gini e o *cross-entropy* (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; JAMES et al., 2014; RASCHKA, 2017). A taxa de erro de classificação corresponde à fração de observações de treinamento de uma determinada região que não pertencem à classe mais frequente desta mesma região:

$$E = 1 - \max_k (\hat{p}_{mk})$$

onde $\hat{p}_{mk}$ representa a proporção de observações da m -ésima região que são da k -ésima classe.

O índice Gini representa uma medida da variância total ao longo das K classes da resposta de interesse, e pequenos valores indicam que um determinado nó terminal contém, predominantemente, observações de uma única classe:

$$G = \sum_{k=1}^K \hat{p}_{mk} (1 - \hat{p}_{mk})$$

Da mesma forma, a medida *cross-entropy* apresentará valores pequenos se o m -ésimo nó for puro, ou seja, se R_m apresentar frequência elevada de observações pertencentes a mesma classe, e é definida por:

$$D = - \sum_{k=1}^K \hat{p}_{mk} \log \hat{p}_{mk}$$

Árvores de decisão simples geram um conjunto de condições interpretáveis e fáceis de implementar. Devido à lógica de sua construção, podem ser aplicadas a conjuntos de treinamento que apresentam dados faltantes ou diferentes tipos de preditores (esparços, assimétricos, contínuos, categóricos, entre outros) sem a necessidade de pré-processamento, não requerem a especificação de uma forma para a relação entre os preditores e a resposta, e podem, ainda, conduzir à seleção de variáveis (KUHN; JOHNSON, 2013).

No entanto, os modelos preditivos resultante de árvores de decisão são, em geral, instáveis, ou seja, pequenas alterações nos dados de treinamento podem implicar mudanças na estrutura da árvore ou em suas regras e, conseqüentemente, alterar a interpretação do modelo ajustado. Nesse contexto, diversos métodos foram desenvolvidos a fim de melhorar o desempenho preditivo de árvores de decisão simples por meio da construção de múltiplas árvores, em vez de uma árvore única, a fim de combinar suas predições em uma predição final,

mais acurada. São exemplos os métodos *bagging*, *random forest* e *boosting* (JAMES et al., 2014; RASCHKA, 2017).

2.4.5.2 Bagging

O método *bootstrap* é utilizado em situações em que é difícil calcular diretamente o desvio padrão de determinada quantidade de interesse. Em ML, esse método pode ser empregado para calcular a performance de modelos, representando uma técnica de reamostragem. Além disso, o mesmo pode ser utilizado em conjunto com algoritmos que produzem modelos instáveis, a fim de melhorar sua performance preditiva, procedimento conhecido como agregação *bootstrap* ou *bagging* (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; KUHN; JOHNSON, 2013).

A árvore de decisão é um dos algoritmos que apresenta melhora substancial de sua performance preditiva quando utilizado em conjunto com o método *bootstrap* e, portanto, um dos mais utilizados como algoritmo base para o *bagging*. Nesse caso, considere um conjunto de treinamento $Z = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ com $z_i = (x_i, y_i)$, $i = 1, \dots, n$, onde x_i representa um vetor de valores para os preditores e y_i o valor da resposta de interesse, contínua ou categórica, para a i -ésima observação (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

O *bagging*, inicia-se com a aplicação do método *bootstrap* para obter B conjuntos de dados de tamanho n amostrados com reposição do conjunto de treinamento, em que a unidade amostral é representada pelo par $z_i = (x_i, y_i)$. Posteriormente, para cada conjunto de amostras *bootstrap*, Z_b^* , $b = 1, \dots, B$, o algoritmo de árvore de decisão é aplicado, resultando em B modelos preditivos (árvores de regressão ou de classificação) e, portanto, em $\hat{y}_b^*$ previsões para uma mesma observação (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; JAMES et al., 2014).

É importante destacar que, diferente da árvore simples, na agregação *bootstrap* a árvore cheia deve ser utilizada, ou seja, o processo de poda não é aplicado. Dessa forma, cada árvore individual terá viés baixo, porém variância alta. Se a resposta for contínua, a agregação *bootstrap* implica redução da variância a partir do cálculo da média das previsões $\hat{y}_1^*(x), \hat{y}_2^*(x), \dots, \hat{y}_B^*(x)$ decorrentes das B árvores, que irá representar a predição final:

$$\hat{y}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \hat{y}_b^*(x)$$

Para classificação, considere $\hat{G}_b(x)$ a predição da b -ésima árvore, correspondente a determinada classe da resposta de interesse. Nesse caso, o classificador *bagging* será baseado no voto majoritário:

$$\hat{G}_{bag}(x) = \text{voto majoritário}\{G_b(x)\}_1^B$$

A predição final da resposta de interesse para uma nova observação corresponderá à classe mais indicada pelas B árvores que foram agregadas (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

O método *bagging* apresenta como desvantagem o fato de as B árvores agregadas apresentarem correlação alta, devido à utilização de todos os preditores como candidatos em todas as etapas de divisão das B árvores (KUHN; JOHNSON, 2013). Como alternativa, a fim de reduzir a correlação entre as árvores que serão agregadas, pode-se utilizar o método *random forest*, descrito a seguir.

2.4.5.3 Random forest

O *random forest* é muito semelhante ao *bagging*, porém realiza um sorteio aleatório dos preditores no conjunto de treinamento a fim de reduzir a correlação entre as árvores agregadas. Desse modo, assim como descrito no tópico anterior, o *random forest* inicia-se com a obtenção de B conjuntos de dados de tamanho n por amostragem com reposição do conjunto de treinamento e, posteriormente, para cada conjunto, estima uma árvore de decisão (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; JAMES et al., 2014).

No entanto, ao aplicar a divisão binária recursiva durante o processo de construção da árvore, para cada segmentação do espaço dos preditores, ou seja, em cada nó da árvore, uma amostra aleatória do conjunto de p preditores, sem reposição e de tamanho m , é escolhida, e somente no subgrupo amostrado realiza-se a escolha da combinação preditor-ponto de corte responsável por essa segmentação. Tal processo continua até que algum critério de parada seja alcançado. Observe que, se $m = p$, então o resultado do *random forest* será idêntico ao do *bagging*. Embora seja recomendado $m \approx p/3$ em problemas de regressão e $m \approx \sqrt{p}$ nos de classificação, esse valor representa um hiperparâmetro do *random forest* e pode ser otimizado por validação cruzada *k-fold* (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; JAMES et al., 2014).

A predição da resposta de interesse em novas observações é idêntica à descrita para o *bagging*. Em problemas de regressão, a resposta média é dada por:

$$\hat{y}_{rf}(x) = \frac{1}{B} \sum_{b=1}^B \hat{y}_b^*(x)$$

Onde $\hat{y}_b^*$ é resposta predita pela b -ésima árvore *random forest*. Para classificação, como no caso do *bagging*, considere $\hat{G}_b(x)$ a predição da b -ésima árvore *random forest*, correspondente a determinada classe da resposta de interesse. Nesse caso, o classificador *random forest* será baseado no voto majoritário:

$$\hat{G}_{rf}(x) = \text{voto majoritário}\{G_b(x)\}_1^B$$

Ou seja, a classe predita para uma nova observação será a mais indicada entre as B árvores *random forest* (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

2.4.5.4 Boosting

O método *boosting* foi originalmente desenvolvido para problemas de classificação e, posteriormente, ampliado para os de regressão. Sua ideia principal, que deu origem ao método *AdaBoost*, consiste em combinar a predição de um conjunto de classificadores fracos, cuja taxa de erro é ligeiramente inferior a de uma classificação aleatória, a fim de construir um comitê poderoso, responsável pela predição final (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

A incorporação, pelo método *AdaBoost*, dos conceitos de função perda e modelo aditivo, conduziu a generalizações para problemas de classificação, bem como a sua extensão para problemas de regressão, caracterizando o método *gradient boosting*, cujo objetivo principal, dado uma função perda e um algoritmo base, é identificar um modelo aditivo que minimize a função perda (KUHN; JOHNSON, 2013).

O método *gradient boosting* na sua forma mais simples é, usualmente, inicializado por um modelo associado ao melhor palpite para a resposta de interesse (por exemplo, a média), para o qual os resíduos são calculados. A seguir, um algoritmo base é aplicado a esses resíduos que, nesse caso, representam a resposta de interesse, para ajustar um modelo que minimize determinada função perda, por exemplo, a SQR. Esse modelo é, então, adicionado ao anterior, a fim de atualizar o valor predito, e tal processo continua até que um número pré-especificado de iterações seja alcançado (KUHN; JOHNSON, 2013). Observe que o modelo inicial possui

viés alto, porém variância baixa e, a cada passo, o valor predito é atualizado de modo a diminuir o viés e aumentar a variância do modelo atualizado (IZBICKI; SANTOS, 2018).

O método *gradient boosting* pode ser aplicado a diversos algoritmos, entretanto, a árvore de decisão é frequentemente escolhida como algoritmo base, devido à possibilidade de utilizá-la como um classificador fraco por meio da construção de uma árvore com poucas divisões (profundidade reduzida) (KUHN; JOHNSON, 2013). Nesse cenário, um *gradient boosting* básico irá apresentar dois hiperparâmetros: a profundidade da árvore, relacionada ao número de divisões em cada árvore, e o número de iterações, m , que está diretamente relacionado ao número de árvores do modelo *boosting* final, e ambos podem ser otimizados por validação cruzada *k-fold* (JAMES et al., 2014; KUHN; JOHNSON, 2013).

Conforme descrito no tópico referente às árvores simples, tanto para problemas de regressão como para os de classificação, a árvore de decisão objetiva particionar o espaço dos preditores em R_j regiões disjuntas, para $j = 1, 2, \dots, J$ e prever a resposta de interesse, γ_j , para a região correspondente. Assim, a regra de predição de uma árvore de decisão pode ser representada da seguinte forma:

$$x \in R_j \Rightarrow f(x) = \gamma_j$$

E a árvore de decisão completa pode ser formalmente expressa como:

$$T(x; \theta) = \sum_{j=1}^J \gamma_j I(x \in R_j)$$

onde $\theta = \{R_j, \gamma_j\}_1^J$. As regiões R_j serão definidas a partir da minimização de uma função perda, $L(\cdot)$:

$$\hat{\theta} = \arg \min_{\theta} \sum_{i=1}^N L(y_i, T(x_i, \theta))$$

e, dado $\hat{R}_j$, a predição da j -ésima região, $\hat{\gamma}_j$, corresponderá à resposta média, para árvores de regressão e, à classe mais frequente, para árvores de classificação (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

O método *gradient boosting*, utilizado em conjunto com *árvores de decisão*, objetiva adicionar novas *árvores* à função de predição $\hat{f}$, em cada iteração m , $m=1, 2, \dots, M$, sem ajustar novamente os coeficientes das *árvores* previamente adicionadas, implicando uma aproximação gradual para f (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). O modelo resultante de sua aplicação pode ser representado da seguinte forma:

$$f_M = \sum_{m=1}^M T(x; \theta_m)$$

Tal que $\theta_m = \{R_{jm}, \gamma_{jm}\}_1^{J_m}$.

Seja $f_{m-1}(x)$ o modelo atual, então a cada iteração m , uma *função perda* é otimizada:

$$\hat{\theta}_m = \arg \min_{\theta_m} \sum_{i=1}^N L(y_i, f_{m-1}(x_i) + T(x_i; \theta_m))$$

e a função de predição, atualizada:

$$f_m(x) = f_{m-1}(x) + T(x; \theta_m)$$

Quando o número de iterações especificado é alcançado, a função de predição estimada será:

$$\hat{f}(x) = f_M(x)$$

O *gradient boosting* pode ser regularizado por meio da aplicação de um termo de penalização (λ) à *árvore* que será adicionada à função de predição:

$$f_m(x) = f_{m-1}(x) + \lambda T(x; \theta_m)$$

Esse termo controla a contribuição de cada *árvore* para a função de predição final e também representa um hiperparâmetro do *gradient boosting* (JAMES et al., 2014).

2.4.5.5 Cubist

Assim como os algoritmos *bagging*, *random forest* e *boosting*, o *cubist* utiliza uma combinação de predições resultantes de árvores simples (algoritmo base) para obter o valor predito final. Sua principal diferença, quando comparado a esses algoritmos, está no estimador da resposta em cada árvore: para o *bagging*, o *random forest* e o *boosting* o estimador utilizado é a média, e para o *cubist*, um modelo de regressão linear.

Além disso, o *cubist* é aplicado a um algoritmo base específico, denominado *M5*, que difere de uma árvore de regressão, basicamente em três características: 1) um critério diferente para o particionamento do espaço dos preditores é utilizado; 2) o nó terminal de uma árvore prediz a resposta de interesse a partir de um modelo de regressão linear, conforme comentado anteriormente; 3) a resposta predita para uma nova observação corresponde a uma combinação das predições de diferentes modelos ao longo de um caminho entre o nó terminal e o topo de uma árvore.

A partição do espaço dos preditores inicia-se pelo cálculo do desvio padrão (dp) para a resposta de interesse no conjunto de treinamento (S). Posteriormente, o desvio padrão da resposta de interesse em subconjuntos dos dados de treinamento (S_j) resultantes de diferentes combinações preditor-ponto de corte é calculado como uma medida de erro (QUINLAN, 1992). A combinação que maximize a redução na taxa de erro esperada (TEE) é então escolhida para a partição inicial do conjunto de dados:

$$TEE = dp(S) - \sum_j \frac{n_j}{n} sd(S_j)$$

Essa métrica determina se a variação total da partição atual, ponderada por seu número de observações, é menor que a variação total da partição imediatamente anterior. Após a partição inicial, um modelo linear é ajustado para os subconjuntos resultantes, utilizando como variável explicativa o preditor escolhido para o particionamento. Nas divisões subsequentes, esse processo se repete, entretanto, o ajuste do modelo linear considera como variáveis explicativas não apenas o preditor responsável pelo particionamento atual, mas também todos aqueles que o precedem, responsáveis por partições anteriores. Além disso, após a primeira partição, o $dp(S)$ da TEE é substituído pelo erro associado ao modelo de regressão linear ajustado (KUHN; JOHNSON, 2013).

O processo de formação da *árvore* continua até que não haja mais melhoras na TEE ou observações suficientes para a realização de novas divisões. Uma vez que o conjunto completo de modelos lineares tenha sido formado, cada um é submetido a um procedimento de simplificação: para um dado modelo, uma taxa de erro ajustada (TEA) inicial é calculada,

$$TEA = \frac{n^* + p}{n^* - p} \sum_{i=1}^{n^*} |y_i - \hat{y}_i|,$$

onde n^* é o número de observações utilizadas para o ajuste do modelo e p o número de parâmetros estimados. Na sequência, cada termo do modelo é retirado e uma nova TEA é calculada. Caso haja redução na TEA , o respectivo termo é retirado do modelo. Esse procedimento é aplicado de forma independente para cada modelo linear presente na árvore (KUHN; JOHNSON, 2013). Mais informações sobre o algoritmo *M5*, como técnicas de suavização da resposta predita (basicamente, uma combinação linear simples dos modelos que compõem a árvore) e de poda da árvore (avaliação de uma taxa de erro com e sem uma subárvore, a partir do nó terminal da árvore cheia) estão descritas em Quinlan (1992).

O algoritmo *cubist* representa uma extensão do algoritmo *M5*, que modifica a técnica de suavização da resposta predita e incorpora dois hiperparâmetros: M comitês, que exercem efeito

semelhante ao descrito para as iterações do algoritmo *boosting*; e K vizinhos do conjunto de treinamento mais próximos da observação para a qual se está interessado em prever o desfecho. Ambos podem ser estimados por validação cruzada *k-fold* (KUHN; JOHNSON, 2013).

Os comitês representam sequências de modelos, em que cada modelo é afetado pelo resultado do modelo imediatamente anterior, por meio da aplicação do algoritmo base a versões repetidamente modificadas do conjunto de treinamento com o objetivo de corrigir erros no valor predito. Por exemplo, se a predição do primeiro comitê for baixa para a resposta de uma dada observação, o segundo comitê vai ser ajustado para compensar esse erro predizendo um valor mais alto, e assim sucessivamente. Logo, a variável resposta utilizada pelo m -ésimo comitê será uma resposta ajustada pelo erro do comitê previamente aplicado:

$$y_m^* = y + (y - \hat{y}_{m-1})$$

Ao final desse processo, a estimativa da resposta de interesse corresponderá à média simples das predições resultantes dos M comitês. Adicionalmente, a predição final de uma nova observação pode ser ajustada selecionando as k amostras de treinamento mais similares (vizinhos) a essa observação:

$$\frac{1}{K} \sum_{l=1}^K \omega_l [t_l + (\hat{y} - \hat{t}_l)]$$

onde t_l é a resposta observada para o l -ésimo vizinho, $\hat{t}_l$ a predição cubista para esse vizinho e ω_l um peso calculado com base na distância (por exemplo, a distância de *Manhattan*) entre o l -ésimo vizinho e a nova observação (KUHN; JOHNSON, 2013).

2.4.6 Super Learner

O algoritmo *super learner* está relacionado ao algoritmo *stacking*, introduzido no contexto de redes neurais por David H. Wolpert (1992) e adaptado para o contexto de regressão por Leo Breiman (1996). Trata-se de um algoritmo delineado para encontrar a combinação ótima de uma coleção de algoritmos, a partir da minimização do erro (risco) de validação cruzada associado a cada algoritmo dessa coleção (VAN DER LAAN; POLLEY; HUBBARD, 2007).

Para a apresentação das definições e estratégia de ajuste do *super learner*, considere o seguinte cenário: o pesquisador dispõe de um conjunto de dados $W_i = (X_i, Y_i), i = 1, \dots, n$, tal que Y_i é o desfecho de interesse e X_i um conjunto p -dimensional de preditores, ambos observados para a observação i . O objetivo consiste em estimar a função $\psi_0(X) = E(Y | X)$ por meio de um algoritmo que mapeie W_i em uma função de predição. Essa função pode ser expressa como o minimizador do valor esperado de uma função perda aplicada aos dados observados:

$$\psi_0(X) = \arg \min_{\psi} E[L(W, \psi(X))]$$

tal que L , a função perda, pode ser representada, por exemplo, pela SQR, no caso de problemas de regressão. Nesse contexto, não só um algoritmo, mas sim uma coleção de algoritmos pode ser proposta para estimar ψ_0 , de modo que o algoritmo *super learner* é aplicado a fim de selecionar alguns (ou todos) os algoritmos dessa coleção com base em uma combinação ótima dos valores preditos por esses algoritmos.

Polley & Laan (2010), descrevem o seguinte roteiro para o ajuste do algoritmo *super learner*, aplicado a uma coleção de algoritmos, $\mathcal{L}$, de tamanho M , $m = 1, \dots, M$:

- **Passo 1:** ajustar cada algoritmo presente em $\mathcal{L}$ ao conjunto de dados completo $W_i, i = 1, \dots, n$, para estimar $\hat{\psi}_m(X)$, $m = 1, 2, \dots, M$.
- **Passo 2:** dividir o conjunto de dados $W_i, i = 1, \dots, n$ em amostras de treinamento e validação de acordo com um processo de validação cruzada k -fold, $k = 1, \dots, K$. Seja a k -ésima parte (*fold*) o conjunto de validação, $V(k)$, e as $K-1$ partes restantes o conjunto de treinamento, $T(k)$. A cada iteração k da validação cruzada, ajuste cada algoritmo de $\mathcal{L}$ em $T(k)$, aplique o estimador (modelo) resultante em $V(k)$ e salve as predições correspondentes, representadas por:

$$\hat{\psi}_{m,T(k)}(X_i), \quad X_i \in V(k), \quad k = 1, \dots, K$$

- **Passo 3:** para cada algoritmo em $\mathcal{L}$, agrupe as predições resultantes do processo de validação cruzada de forma a construir uma matriz Z , de dimensão $n \times M$, expressa por:

$$Z = \{\hat{\psi}_{m,T(k)}(X_{V(k)}), \quad k = 1, \dots, K; m = 1, \dots, M$$

Tal que $X_{V(k)} = W_i \in V(k)$ representa o vetor de preditores das amostras presentes no conjunto de validação na k -ésima iteração da validação cruzada.

- **Passo 4:** seja uma família de vetores de pesos, α , correspondentes a combinações de pesos para os algoritmos presentes em $\mathcal{L}$, proposta da seguinte forma:

$$h(z | \alpha) = \sum_{m=1}^M \alpha_m \hat{\psi}_{m,T(k)}(X_{V(k)}), \quad \alpha_m \geq 0 \forall m, \quad \sum_{m=1}^M \alpha_m = 1$$

De todas as α -combinações possíveis, escolha aquela que minimiza o risco de validação cruzada dos M algoritmos, $\sum_{m=1}^M \alpha_m \hat{\psi}_m$, dada por:

$$\hat{\alpha} = \arg \min_{\alpha} \sum_{i=1}^n (Y_i - h(z_i | \alpha))^2$$

- **Passo 5:** combine $\hat{\alpha}$ com $\hat{\psi}_m(X)$, $m=1,2,...,M$ para criar o ajuste final do *super learner*:

$$\hat{\psi}_{SL}(X) = \sum_{m=1}^M \hat{\alpha}_m \hat{\psi}_m(X)$$

Observe que, embora seja possível escolher apenas um algoritmo entre os M presentes em $\mathcal{L}$, usualmente uma média ponderada dessa coleção é escolhida.

É importante destacar que o *super learner* é um algoritmo desenvolvido para a seleção de modelos. Logo, não apresenta uma estimativa adequada para avaliação de performance em dados futuros (risco ou erro de predição). Caso, essa seja uma medida de interesse, é preciso realizar um processo de validação cruzada adicional (*nested cross validation*), aplicado aos passos 1 a 5, a fim de estimar de forma adequada a performance preditiva do *super learner* e dos algoritmos presentes em $\mathcal{L}$. Esse procedimento resulta não só em uma estimativa do risco, mas também de seu erro padrão para cada algoritmo testado (PETERSEN et al., 2015).

2.5 AVALIAÇÃO DO MODELO FINAL

Todo processo de aprendizado apresenta como uma de suas etapas a avaliação. No contexto de *ML*, o processo de aprendizado de um modelo preditivo é avaliado por meio de seu desempenho em um conjunto de dados de teste, que não foi utilizado para o ajuste de modelos preditivos. Portanto, após a aplicação de algoritmos (como os descritos no tópico 2.5) ao conjunto de treinamento para a seleção de um modelo que ajuste satisfatoriamente esses dados, sua performance é avaliada em dados novos (conjunto de teste), que não participaram do ajuste do modelo. Para problemas de regressão, a performance é comumente avaliada por meio do

EQM, e para os de classificação, pela AUC ROC (conforme descrito no tópico 2.3.1). Caso essa performance seja satisfatória, o modelo preditivo poderá, então, ser utilizado para prever a resposta de interesse em dados futuros (RASCHKA, 2017).

3 OBJETIVOS

3.1 OBJETIVO GERAL

Aplicar técnicas supervisionadas de *machine learning* (ML) e comparar sua performance em problemas de classificação e de regressão para prever respostas de interesse para a saúde pública e a medicina.

3.2 OBJETIVOS ESPECÍFICOS

3.2.1 Artigo 1

Apresentar as etapas que compõem a aplicação de ML em análise preditiva, utilizando como exemplo a predição do risco de óbito em 5 anos para idosos residentes no município de São Paulo, participantes do estudo Saúde Bem-Estar e Envelhecimento (SABE).

3.2.2 Artigo 2

Predizer o risco de tempo de vida com qualidade de até 30 dias para pacientes com câncer admitidos em Unidade de Terapia Intensiva (UTI) de dois hospitais públicos do Estado de São Paulo especializados no tratamento de câncer.

3.2.3 Artigo 3

Predizer a expectativa de vida ao nascer de municípios brasileiros e, posteriormente, comparar características de saúde de municípios com expectativa de vida ao nascer superior à esperada (*overachievers*) com as daqueles com expectativa de vida inferior à esperada (*underachievers*), visando identificar práticas de saúde pública bem-sucedidas.

4 MÉTODOS

Esta seção descreve os procedimentos metodológicos utilizados para desenvolver cada artigo que compõe a tese. Portanto, serão apresentados fonte de dados e delineamento, população e variáveis de estudo, bem como aspectos referentes à organização e à análise de dados de cada artigo. É importante destacar que os métodos e as técnicas empregados na análise dos dados estão descritos de forma detalhada no capítulo 2 da tese.

4.1 ARTIGO 1

4.1.1 Fonte de dados e delineamento do estudo

O Artigo 1 utilizou dados de 2.808 idosos que ingressaram no estudo Saúde, Bem-estar e Envelhecimento (SABE) nas coortes de 2000 (2.143 idosos), 2006 (298 idosos) ou 2010 (367 idosos) (Figura 10).

Figura 10: Amostra de idosos do Estudo SABE segundo ano de ingresso no estudo (2000, 2006 ou 2010) e reentrevista em coortes subsequentes.

Iniciado no ano 2000 como um estudo multicêntrico realizado em sete cidades da América Latina e Caribe, o estudo SABE teve como principais objetivos coletar informações sobre condições de vida e saúde de idosos e avaliar diferenciais de gênero e socioeconômicos com relação ao estado de saúde e acesso e utilização de serviços de saúde, a fim de projetar as necessidades sociais e de saúde decorrentes do rápido crescimento da população idosa (LEBRÃO; LAURENTI, 2005).

No Brasil, o estudo foi realizado com idosos residentes na zona urbana do município de São Paulo. Seu projeto inicial fixou número mínimo de 1.500 entrevistas distribuídas segundo estratos definidos por sexo e faixa etária. A amostra final para o ano 2000 foi composta por 2.143 indivíduos (coorte A), sendo 1.568 selecionados por amostragem probabilística e 575 decorrentes da ampliação da amostra dentro da faixa etária de 75 anos ou mais, com o objetivo de compensar a frequência de morte mais elevada nessa população (SILVA, 2003).

Para a seleção dos indivíduos entrevistados, primeiramente realizou-se o sorteio de domicílios mediante amostragem por conglomerados em dois estágios, sob o critério de partilha proporcional ao tamanho. O primeiro estágio resultou em uma amostra de 72 setores censitários, e o segundo, em uma amostra de 90 domicílios para cada setor sorteado. Todos os indivíduos residentes nesses domicílios, considerados elegíveis segundo os objetivos da pesquisa, foram identificados e convidados a participar do estudo. Informações detalhadas sobre o desenho amostral estão disponíveis em Silva (2003).

Com subsídios da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e do Ministério da Saúde, sob coordenação do Departamento de Epidemiologia da Faculdade de Saúde Pública da Universidade de São Paulo, o estudo SABE passou a representar um estudo longitudinal (LEBRÃO; LAURENTI, 2005), com realização de seguimento dos idosos entrevistados no ano 2000, bem como de inclusão de novos participantes, nos anos 2006 e 2010.

Da amostra inicial, 1.115 indivíduos foram novamente entrevistados em 2006 e uma nova coorte probabilística de 298 indivíduos com idade entre 60 e 64 anos foi adicionada ao seguimento (coorte B) (LEBRÃO; DUARTE, 2008). Por fim, no ano de 2010, uma nova onda de entrevistas 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 (coorte C) (CORONA; DUARTE; LEBRÃO, 2014).

No ano 2000, a coleta de dados foi realizada no domicílio dos idosos com auxílio de um questionário padrão composto por 11 seções: informações pessoais (A), avaliação cognitiva (B), estado de saúde (C), estado funcional (D), medicamentos (E), acesso e utilização de

serviços (F), rede de apoio familiar e social (G), história de trabalho e fontes de receita (H), características da moradia (J), antropometria (K) e mobilidade e flexibilidade (L). O questionário do ano 2000, utilizado como base para a organização das variáveis da presente pesquisa, encontra-se disponível em http://www.fsp.usp.br/sabe/Artigos/Questionario_2000.pdf.

Além de dados do estudo SABE, foram utilizados microdados referentes aos óbitos de indivíduos com 60 anos ou mais, registrados no Sistema de Informações sobre Mortalidade (SIM). Esses dados foram disponibilizados aos pesquisadores do estudo SABE pela Secretaria de Saúde do Município de São Paulo.

4.1.2 População de estudo

Para a presente pesquisa, a população de estudo foi compreendida por indivíduos com 60 anos ou mais residentes na zona urbana do município de São Paulo, entrevistados no ano de 2000 (n=2.143) para o estudo SABE e, adicionalmente, indivíduos que ingressaram no estudo em 2006 (n=298) e em 2010 (n=367), resultando em uma amostra de 2.808 idosos. Fizeram parte também do estudo óbitos ocorridos nessas coortes, entre janeiro de 2000 a setembro de 2016.

4.1.3 Pareamento dos dados

Os microdados do SIM, referentes aos óbitos de indivíduos com 60 anos ou mais residentes no município de São Paulo, foram relacionados aos dados das coortes A, B e C do estudo SABE por pareamento probabilístico com auxílio do *software LinkPlus* (disponível em: <https://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm>), utilizando o nome do indivíduo, a data de nascimento e o nome da mãe como variáveis de comparação, o sexo como variável de bloqueio e o endereço, quando necessário, para revisão manual.

4.1.4 Variáveis de estudo

A variável resposta de interesse desta pesquisa é de natureza binária e correspondeu à ocorrência de óbito em até 5 anos após o ingresso do idoso no estudo ($n=423$, 15%), identificado por pareamento probabilístico.

Os preditores de interesse foram representados por 37 variáveis (Figura 11), selecionadas de acordo com a literatura relacionada ao tema (SCHONBERG et al., 2009; SUEMOTO et al., 2016; YOURMAN et al., 2012), presentes nas seções A, B, C, D, H, K e L do questionário do ano 2000, utilizado como modelo para o presente estudo. Uma vez que foram realizadas alterações nesse questionário entre as três coortes do estudo SABE, sempre que necessário, as variáveis foram recategorizadas para tornar as informações compatíveis. Embora as variáveis da seção F, referentes a características de acesso e utilização de serviços, sejam de importância epidemiológica para a predição de óbito em idosos, as mesmas não foram selecionadas em decorrência de mudanças substanciais na sua forma de pergunta entre os questionários das três coortes.

O comprometimento cognitivo foi avaliado e classificado de acordo com a versão modificada do Mini Exame do Estado Mental (ICAZA; ALBALA, 1999). As variáveis sobre dificuldade de mobilidade tinham como opções de resposta: “sim, apresenta dificuldade”, “não apresenta dificuldade”, “não pode realizar a atividade”, “pode, mas não faz”. Para esta pesquisa optou-se por agrupar as respostas “sim, apresenta dificuldade” e “não pode realizar a atividade” em uma única categoria (sim), bem como as respostas “não tem dificuldade” e “pode, mas não faz” (não). Por fim, variáveis correspondentes à dificuldade em realizar atividades instrumentais, presentes no bloco “Estado funcional”, apresentavam como opção de resposta: “sim”, “não”, “não consegue” e “não costuma fazer” e, para esta pesquisa, foram dicotomizadas em “sim” (agrupamento das categorias “sim” e “não consegue”) e “não” (agrupamento das categorias “não” e “não costuma fazer”).

Figura 11: Preditores de óbito em 5 anos em idosos participantes do estudo SABE.

Demográficas e	<ul style="list-style-type: none"> • Idade em anos (contínua) • Sexo (masculino ou feminino) • Escolaridade em anos de estudo (nenhum, 1-3, 4-7, ≥ 8, sem informação) • Percepção de renda (suficiente ou insuficiente) • Vive em lar unipessoal (sim ou não)
Estado de saúde e características	<ul style="list-style-type: none"> • Tabagismo (fuma, já fumou, nunca fumou) • Consumo de bebida alcoólica em dias na semana (nenhum, <1, 1-3, ≥ 4) • Índice de massa corpórea (contínua) • Nº de refeições em um dia (1, 2, ≥ 3) • Auto avaliação de saúde (excelente/muito boa/boa ou regular/ruim/muito ruim)
Morbidades	<ul style="list-style-type: none"> • Doenças autorreferidas (sim ou não): hipertensão, diabetes, câncer, doença pulmonar, doença vascular cerebral, doença cardíaca e doença psiquiátrica • História de queda no último ano (sim, não, sem informação) • Comprometimento cognitivo (<13 ou ≥ 13)
Mobilidade	<p>Dificuldade para (sim ou não):</p> <ul style="list-style-type: none"> • 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
Estado funcional	<p>Dificuldade para (sim ou não):</p> <p>- Atividades básicas</p> <ul style="list-style-type: none"> • Se vestir • Tomar banho • Se alimentar • Deitar/levantar da cama • Ir ao banheiro <p>- Atividades instrumentais</p> <ul style="list-style-type: none"> • Preparar refeições quentes • Cuidar do próprio dinheiro

4.1.5 Análise dos dados

A análise dos dados seguiu as etapas descritas na Figura 1 do capítulo 2 da tese. Primeiramente, optou-se por eliminar 131 observações com valor faltante para variáveis categóricas, exceto para “escolaridade” e “história de queda no último ano”, em que o número de dados faltantes foi de, respectivamente, 377 e 181. Para essas variáveis, uma categoria adicional foi criada a fim de alocar as observações com valores ausentes. Portanto, o banco de dados utilizado na análise preditiva foi composto por 2.677 observações, das quais 395 (15%) corresponderam a óbitos.

Esse conjunto de dados foi aleatoriamente dividido em dois subconjuntos: banco de treinamento, onde foram alocadas 70% das observações para o aprendizado dos modelos preditivos, e banco de teste, representado pelas observações restantes (30% dos dados originais), a partir do qual realizou-se a avaliação dos modelos selecionados na etapa de aprendizado.

Na etapa de pré-processamento dos dados, para a variável índice de massa corporal (IMC), em que 387 indivíduos não tinham informação, realizou-se o cálculo da mediana nos dados de treinamento, e este valor foi imputado para todas as observações sem informação para IMC, tanto no banco de treinamento como no de teste. Em relação à organização das variáveis categóricas, aquelas com mais de duas categorias foram substituídas por suas respectivas variáveis indicadoras (escolaridade, tabagismo, dias da semana que consome bebida alcoólica, número de refeições por dia e história de queda). Por fim, as variáveis quantitativas, idade e IMC, foram padronizadas utilizando sua respectiva média e desvio padrão.

A análise descritiva dos preditores selecionados para este estudo foi realizada mediante distribuição de frequências absoluta e relativa para as variáveis categóricas e, média e desvio padrão para as variáveis contínuas.

Para a análise preditiva, cinco algoritmos foram selecionados: regressão logística, regressão logística penalizada, redes neurais, *gradient boosted trees* e *random forest*. As características principais dos algoritmos selecionados estão descritas no capítulo 2 da tese, bem como na figura 2 deste artigo. Para cada algoritmo, exceto a regressão logística, hiperparâmetros foram otimizados por meio do estabelecimento de uma lista de possíveis valores e da utilização de validação cruzada com $k=10$, repetida 10 vezes.

Tanto na etapa de treinamento, realizada para a seleção de modelos, como na de teste, em que os modelos selecionados foram avaliados, utilizou-se a área abaixo da curva (AUC)

ROC como medida de performance preditiva. Além dessa medida, a calibração da probabilidade predita foi avaliada por curvas de calibração e da distribuição do risco predito de óbito em até 5 anos segundo as categorias da resposta de interesse nos dados de teste, para a qual é esperada distribuição assimétrica com concentração em valores baixos do risco predito para o grupo de indivíduos que não morreu e, por outro lado, para aqueles que morreram, distribuição assimétrica com concentração em valores altos do risco predito, evidenciando acurácia na magnitude no risco predito para os respectivos grupos.

4.1.6 Aspectos éticos

Por se tratar de pesquisa envolvendo seres humanos, o estudo SABE seguiu as normas referentes à Resolução 196/96 do Conselho Nacional de Saúde – Diretrizes e Normas Regulamentadoras de Pesquisas Envolvendo Seres Humanos. Antes do início de cada entrevista, um Termo de Consentimento Livre e Esclarecido foi lido para o entrevistado e seu consentimento obtido por meio de assinatura. Cada entrevistado recebeu uma versão impressa desse termo, contendo detalhes de identificação e contato dos coordenadores da pesquisa.

O estudo do ano de 2000 foi aprovado pela Comissão Nacional de Ética em Pesquisa (nº do parecer 315/99 – Anexo A). Adicionalmente, as ondas de 2006 e 2010 do estudo SABE obtiveram aprovação do Comitê de Ética em Pesquisa da Faculdade de Saúde Pública de acordo com os pareceres de 83/06 (Anexo B) e 2044/10 (Anexo C).

4.2 ARTIGO 2

4.2.1 Fonte de dados e delineamento do estudo

O presente estudo utilizou dados de um estudo de coorte prospectiva de pacientes com câncer admitidos em UTI intitulado “QALY – Modelo para predizer sobrevida ajustada para a qualidade em pacientes com câncer admitidos em unidade de terapia intensiva” – estudo QALY (CAVALCANTI, 2009), realizado em dois hospitais públicos especializados no tratamento de câncer: Instituto do Câncer do Estado de São Paulo (ICESP) Octávio Frias de Oliveira, localizado no município de São Paulo (SP), e Fundação Pio XII – Hospital do Câncer de Barretos, com sede no município de Barretos (SP).

O estudo QALY teve como objetivos principais desenvolver um modelo para prever sobrevida ajustada para qualidade de vida (QALY) em pacientes com câncer gravemente enfermos, para os quais se considerou a possibilidade de admissão em UTI, e desenvolver um modelo para prever QALY em pacientes com câncer internados em UTI há cinco dias. Como objetivos secundários o estudo se propôs a avaliar a qualidade de vida relacionada à saúde antes do início da doença aguda que motivou a internação na UTI, e no seguimento de 15 dias e 3, 6, 12, 18 e 24 meses.

4.2.2 População de estudo

Foram utilizados como critério de inclusão do paciente no estudo QALY ter idade igual ou superior a 18 anos, ser portador de neoplasia maligna comprovada ou provável e ter sido admitido em UTI. Entre março de 2010 e agosto de 2011, 808 pacientes foram considerados elegíveis para o estudo. Destes, 15 pacientes foram excluídos: 12 por terem o diagnóstico de neoplasia maligna descartado após admissão em UTI, 2 por não aceitarem participar do estudo e um por estar internado há mais de 24 horas na UTI. Portanto, a coorte desse estudo foi composta por 793 pacientes. O seguimento dos pacientes foi realizado por contato telefônico em 15 dias e aos 3, 6, 12, 18 e 24 meses para verificação do *status* vital e preenchimento da escala de qualidade de vida EQ-5D-3L. Em agosto de 2013 o seguimento foi finalizado, com a ocorrência de quatro perdas: uma aos 3 meses e três aos 24 meses de seguimento.

Para a presente pesquisa, foram considerados apenas os pacientes com seguimento completo até os 24 meses (n=789). Adicionalmente, foram excluídos pacientes que apresentaram algum dado ausente para os preditores selecionados para estudo, resultando em amostra final de 777 pacientes (Figura 12).

Figura 12 – Coorte prospectiva de pacientes com câncer admitidos em UTIs de dois hospitais públicos do Estado de São Paulo, especializados no tratamento de câncer, 2010-2011.

4.2.3 Variáveis de estudo

A variável resposta de interesse desta pesquisa resultou da categorização do QALY em menor ou igual a 30 dias ($n=347$, 48%) e maior que 30 dias. O QALY, que por sua vez, foi calculado de acordo com os seguintes passos:

- Foram considerados aos seguintes intervalos de tempo de vida: da admissão até o terceiro mês e entre 3 e 6 meses; 6 e 12 meses; 12 e 18 meses; 18 e 24 meses de seguimento.
- A qualidade de vida relacionada à saúde foi avaliada por meio da aplicação da escala de qualidade de vida EQ-5D-3L no início de cada intervalo de tempo. O escore resultante dessa escala, denominado índice de utilidade, foi calculado de acordo com os coeficientes derivados para o EQ-5D-3L de amostra do Reino Unido, com metodologia *Time Trade-Off* (DOLAN, 1997; SZENDE; OPPE; DEVLIN, 2007).
- O índice de utilidade obtido no início de cada intervalo de tempo foi utilizado para sua respectiva ponderação, exceto para o intervalo entre a admissão e o seguimento aos 3 meses, para o qual o escore EQ-5D-3L obtido aos 15 dias representou a ponderação aplicada. Este processo resultou em uma medida de QALY para cada intervalo de tempo de vida.
- Por fim, para cada paciente, um valor total de QALY aos 24 meses foi obtido por meio da soma dos QALYs correspondentes a cada intervalo de tempo de vida.

Em relação às variáveis preditoras, 27 características coletadas na admissão do paciente à UTI foram selecionadas:

- **Características demográficas:** sexo (feminino ou masculino) e idade (variável contínua);
- **Morbidades:** insuficiência renal crônica (sim ou não), insuficiência pulmonar crônica (sim ou não), diabetes (sim ou não), insuficiência cardíaca crônica, segundo a classificação da *New York Heart Association- NYHA* (THE CRITERIA COMMITTEE OF THE NEW YORK HEART, 1994) (não – classe I: ausência de fadiga, dispneia ou angina (sintomas) durante atividades habituais, ou classe II: sintomas desencadeados por atividades habituais; sim – Classe III: sintomas desencadeados por nível moderado de atividade física, ou seja, por atividades menos intensas que as habituais, ou classe IV: sintomas em repouso ou em níveis mínimos de atividade);
- **Estado de saúde e características comportamentais:** índice de massa corporal (variável contínua), alcoolismo (sim ou não), tabagismo (prévio, atual ou nunca fumou), uso de esteroides (sim ou não);
- Capacidade funcional, avaliada por meio do *Eastern Cooperative Oncology Group* (ECOG) (OKEN et al., 1982);

- Presença de *delirium*, avaliada utilizando a versão em Língua Portuguesa do questionário *Confusion Assessment Method for Intensive Care Units* (CAM-ICU) (GUSMAO-FLORES et al., 2011);
- Tipo de admissão à UTI (médica, cirurgia eletiva ou cirurgia de emergência);
- **História de saúde relacionada ao câncer:** sítio primário do câncer (cérebro e sistema nervoso central, cavidade oral ou faríngea, trato digestivo, sistema respiratório, mama, sistema genital, trato urinário, hematológico – linfoma, leucemia ou mieloma, outros/sítio indeterminado), status do câncer (controlado ou em remissão por mais de 30 dias; ativo, com diagnóstico recente, referente aos pacientes com diagnóstico inferior a 4 meses e que ainda não haviam recebido o primeiro tratamento ou estavam em tratamento há menos de 30 dias; e câncer com recaída ou progressão, correspondente aos pacientes com sinais de recidiva ou progressão da neoplasia), extensão do câncer (limitada, se neoplasia sólida restrita ao órgão acometido; localmente avançada, se invasão de órgãos ou estruturas adjacentes ao órgão acometido pelo tumor ou envolvimento de linfonodo regional; metástases à distância, se evidência de disseminação da neoplasia à distância; e leucemias), cirurgia prévia (não, sim – curativa adjuvante, sim – paliativa), quimioterapia prévia (não, sim – curativa adjuvante, sim – paliativa) e radioterapia prévia (não, sim – curativa adjuvante, sim – paliativa);
- **Complicações atuais relacionadas ao câncer:** efeito de massa intracraniana (sim ou não), obstrução ou compressão das vias aéreas (sim ou não), síndrome da compressão da medula espinal (sim ou não), toxicidade significativa devido à quimioterapia (sim ou não) e sangramento maior (sim ou não).
- Infecção nosocomial (sim ou não), infecção respiratória (sim ou não) e ventilação mecânica na admissão à UTI (sim ou não).

4.2.4 Processamento e análise dos dados

Esta pesquisa, assim como a anterior, seguiu as etapas descritas na Figura 1 do capítulo 2. Na etapa de pré-processamento dos dados, variáveis categóricas com mais de 2 categorias foram substituídas por suas respectivas variáveis indicadoras (tabagismo, capacidade funcional ECOG, tipo de admissão, sítio primário do câncer, status do câncer, extensão do câncer, cirurgia

prévia, quimioterapia prévia e radioterapia prévia) e as variáveis quantitativas, idade e IMC, foram padronizadas utilizando sua respectiva média e desvio padrão.

A análise descritiva dos preditores selecionados para este estudo foi realizada mediante distribuição de frequências absoluta e relativa para as variáveis categóricas, e média e desvio padrão para as variáveis contínuas.

Para a análise preditiva, na etapa de aprendizado dos modelos, utilizou-se validação cruzada aninhada, devido ao menor tamanho amostral ($n=777$), que pode implicar variações na performance dos algoritmos, dependendo da amostra que compuser os bancos de treinamento e de teste, e também devido ao interesse em selecionar um entre os diferentes algoritmos aplicados. Conforme descrito no capítulo 2, a validação cruzada aninhada é composta por um processo de validação cruzada externo (*outer loop*) e um interno (*inner loop*). Para a presente pesquisa, os 777 pacientes foram divididos em treinamento e teste utilizando-se a validação cruzada *leave-one-out* (*outer loop*) e, adicionalmente, aplicou-se validação cruzada em 10 partes no banco de treinamento (*inner loop*) a cada iteração do *leave-one-out*.

Os seguintes algoritmos foram selecionados para aplicação: regressão logística, regressão logística penalizada, redes neurais, árvore de classificação simples, *gradient boosted trees* e *random forest*. As características principais dos algoritmos selecionados estão descritas no capítulo 2 da tese. Para cada algoritmo, exceto a regressão logística, hiperparâmetros foram otimizados por meio do estabelecimento de uma lista de possíveis valores e da utilização de validação cruzada com $k=10$, no *loop* interno da validação cruzada aninhada.

A performance preditiva dos algoritmos foi avaliada segundo o poder discriminatório, medido pela AUC ROC, e a calibração da probabilidade predita, realiza por curvas de calibração e pelo teste de Hosmer e Lemeshow. Adicionalmente, medidas de sensibilidade, especificidade, frequência de casos verdadeiros positivos e falsos positivos foram verificadas para diferentes pontos de corte da probabilidade predita por cada algoritmo. Os pontos de corte foram escolhidos considerando que intervenções, como por exemplo, indicação de cuidados paliativos fora da UTI, poderiam ser aplicadas a 30%, 20% ou 10% dos pacientes com risco predito mais alto para QALY menor ou igual a 30 dias.

4.2.5 Aspectos éticos

Por se tratar de pesquisa envolvendo seres humanos, o estudo QALY seguiu todas as normas referentes à Resolução 196/96 do Conselho Nacional de Saúde – Diretrizes e Normas

Regulamentadoras de Pesquisas Envolvendo Seres Humanos. A participação do paciente no estudo só foi possível se o mesmo, ou no caso de incapacitação, seu representante legalmente constituído, tenha aceitado participar do estudo após esclarecimentos sobre o protocolo do estudo e assinatura do Termo de Consentimento Livre e Esclarecido.

O estudo QALY obteve aprovação do Comitê de Ética para Análise de Projetos de Pesquisa da Diretoria Clínica do Hospital das Clínicas e da Faculdade de Medicina da Universidade de São Paulo (parecer nº 0161/10 – Anexo D) e do Comitê de Ética em Pesquisa do Hospital de câncer de Barretos (parecer nº 337/2010 – Anexo E).

4.3 ARTIGO 3

4.3.1 Fonte de dados e delineamento do estudo

O estudo desenvolvido para o Artigo 3 tem delineamento transversal e está organizado em duas etapas: a primeira correspondente à predição da expectativa de vida ao nascer de municípios brasileiros, e a segunda, à comparação de características de saúde entre dois grupos, um correspondente aos municípios que superaram a expectativa de vida ao nascer predita (*overachievers*) e o outro aos que apresentaram expectativa de vida ao nascer inferior à predita (*underachievers*).

A qualificação de municípios em *overachievers* e *underachievers* foi realizada para cada tercil da expectativa de vida observada, a fim de evitar que apenas municípios com expectativa de vida mais alta (municípios mais ricos) e com expectativa de vida mais baixa (municípios mais pobres) fossem comparados. Portanto, para cada tercil, a expectativa de vida predita foi ordenada e foram selecionados 100 municípios *overachievers* e 100 municípios *underachievers* e, posteriormente, a comparação de características de saúde foi realizada por tercil (Figura 13).

Figura 13: Etapas para realização do estudo de predição da expectativa de vida de municípios brasileiros e comparação de características de saúde de *overachievers* e *underachievers*.

Bancos de dados distintos foram utilizados em cada etapa do estudo e, para a composição desses bancos, foram selecionados dados agregados de pesquisas realizadas pelo Instituto Brasileiro de Geografia e Estatística (IBGE), disponíveis em um banco de tabelas estatísticas, acessado por meio do Sistema IBGE de recuperação automática (SIDRA) (<https://sidra.ibge.gov.br/home/pnadcm>), dados do Atlas do Desenvolvimento Humano no Brasil (<http://www.atlasbrasil.org.br/2013/pt/>) e do Departamento de Informática do Sistema Único de Saúde (DATASUS) (<http://datasus.saude.gov.br/informacoes-de-saude/tabnet>). Todos os dados são referentes ao ano de 2010, quando foi realizado o último Censo Demográfico brasileiro.

4.3.2 População de estudo

A unidade de análise foi representada por municípios brasileiros com mais de 10.000 habitantes, correspondente a 3.052 dos 5.565 municípios brasileiros. Esse critério de seleção foi adotado a fim de reduzir o efeito de variações aleatórias anuais nas variáveis selecionadas para estudo.

4.3.3 Variáveis de estudo

A seleção de dados nas fontes de informação supracitadas, a tabulação e cálculo das variáveis finais, utilizadas como preditores, e a organização dos bancos foram realizadas de forma independente por dois pesquisadores. Além dessa estratégia, um terceiro pesquisador foi responsável por comparar os bancos de dados para investigar e corrigir possíveis inconsistências.

Na primeira etapa do estudo, a resposta de interesse foi representada pela expectativa de vida ao nascer (variável quantitativa), obtida do Atlas do Desenvolvimento Humano no Brasil. Em relação aos preditores, foram utilizadas 35 variáveis quantitativas, correspondentes a características demográficas e socioeconômicas dos municípios, mais uma variável qualitativa referente aos estados em que os municípios estão localizados.

Os preditores quantitativos foram construídos com base em dados disponíveis no SIDRA, acessado em novembro de 2016. Adicionalmente, para os preditores taxa de trabalho infantil, nascimentos *per capita* e percentual de cobertura do programa Bolsa Família utilizou-se dados disponíveis no DATASUS. Um dicionário de variáveis foi elaborado (Apêndice A) com detalhes sobre o número das tabelas selecionadas na busca feita no SIDRA, bem como do cálculo para a definição dos preditores.

Na segunda etapa do estudo, a resposta de interesse foi representada pela qualificação do município em *overachiever* ou *underachiever*, e as variáveis explicativas por 11 características de saúde, correspondentes a indicadores de atenção primária (cobertura das equipes da Estratégia Saúde da Família, cobertura vacinal, cobertura das equipes de Saúde Bucal, número de equipes da Estratégia Saúde da Família/ 10.000 habitantes) e de atenção secundária à saúde (número de leitos/ 10.000 habitantes, número de mamografias realizadas/ 100 mulheres, número de equipamentos de ultrassom/ 10.000 nascidos vivos, número de equipamentos de Raio X/ 10.000 habitantes, número de equipamentos para manutenção da vida/ 10.000 habitantes, percentual de baixo peso ao nascer, percentual de parto cesárea). Os dados utilizados para o cálculo desses indicadores foram obtidos do DATASUS, em maio de 2017. O dicionário contendo detalhes dos cálculos realizados encontra-se no Apêndice B.

4.3.4 Análise dos dados

A primeira etapa do estudo seguiu o roteiro descrito na Figura 1 do capítulo 2 da tese. Para o pré-processamento, todas as variáveis quantitativas foram padronizadas, por meio da subtração de sua média ($\bar{X}_j$) e divisão pelo desvio padrão (S_j):

$$X_j^* = \frac{X_j - \bar{X}_j}{S_j}$$

A variável correspondente ao estado de localização do município estava representada por 26 categorias (o Distrito Federal foi agrupado com o estado de Goiás). Para a análise preditiva, essa variável foi transformada em 26 variáveis indicadoras, uma para cada estado. A variável indicadora correspondente ao estado de São Paulo não foi adicionada à análise, em decorrência de problemas numéricos para algoritmos com intercepto caso todas as variáveis indicadoras fossem adicionadas ao modelo, como é o caso do modelo de regressão linear. Portanto, foram utilizados 60 preditores: 35 variáveis quantitativas e 25 variáveis indicadoras.

Na análise descritiva, realizou-se síntese numérica das variáveis quantitativas, segundo medidas resumo e de dispersão, e avaliou-se a correlação entre essas variáveis, bem como sua correlação com a resposta de interesse. As variáveis indicadoras correspondentes aos estados brasileiros foram analisadas por medidas de frequência.

Para a análise preditiva, utilizou-se validação cruzada aninhada em conjunto com o algoritmo *super learner*, que foi aplicado a uma coleção de algoritmos composta por: regressão linear, ridge, lasso, *elastic net*, *partial least squares*, redes neurais, *support vector machines* linear, polinomial e radial, árvore simples, *random forest*, *gradient boosted trees* e *cubist*. As características principais dos algoritmos da coleção, bem como do *super learner*, estão descritas no capítulo 2 da tese.

A validação cruzada externa (*outer loop*) foi realizada em 10 partes, resultando na divisão do conjunto de dados original em treinamento e teste, de forma iterativa. A validação cruzada interna (*inner loop*) está representada na Figura 14 de acordo com os itens *a*, *b*, *c*, *d* e *e*. O item *a* corresponde à divisão do banco de treinamento inicial em 10 partes ($K=1, 2, \dots, 10$), para treinar cada algoritmo da coleção (item *b*), deixando reservada a k -ésima parte para validação (destacada em cinza). Adicionalmente, a cada iteração da validação cruzada interna, para cada algoritmo da coleção, exceto a regressão linear, hiperparâmetros foram otimizados mediante definição de uma lista de possíveis valores e utilização de validação cruzada 10-*fold*.

Ao final da validação cruzada interna, as previsões obtidas para o banco de treinamento inicial (item c , $Z_{k,m}$, $k = 1, \dots, 10$ e $m = 1, \dots, 16$, k : número de iterações da validação cruzada interna; m : número de algoritmos da coleção) foram relacionadas à resposta de interesse (Y), com o objetivo de estimar o vetor de pesos (α), correspondente à combinação ótima das previsões (item d).

O vetor de pesos estimado ($\hat{\alpha}$), em conjunto com os modelos resultantes da aplicação dos algoritmos da coleção ao banco de treinamento inicial ($\hat{Y}_1$, correspondente à regressão linear, $\hat{Y}_2$, à regressão ridge, e assim sucessivamente, até o último algoritmo da coleção, $\hat{Y}_{16}$, referente ao *cubist*) caracterizou o modelo ajustado para o *super learner* (item e).

Por fim, esse modelo, bem como os decorrentes de cada algoritmo da coleção, foram aplicados aos dados de teste para a obtenção do valor predito da expectativa de vida ao nascer. Esse processo foi realizado 10 vezes, uma para cada iteração da validação cruzada externa, modificando, a cada iteração, o banco de treinamento inicial e os dados de teste.

Figura 14: Representação da validação cruzada interna, realizada para o ajuste de um modelo preditivo com base no algoritmo *super learner*.

Fonte: Polley, Rose & van der Laan (2011).

Ao final do processo iterativo da validação cruzada externa, cada município apresentava um valor predito para a expectativa de vida ao nascer. Esses resultados foram utilizados para avaliar a performance preditiva do *super learner* e dos algoritmos da coleção previamente apresentada, por meio do cálculo do erro quadrático médio (EQM), e também para classificar os municípios em *overachievers* ou *underachievers*, segundo o vetor de valores preditos pelo *super learner*.

Na segunda etapa do trabalho, realizou-se análise descritiva bem como testes de hipótese para comparação das 11 características de saúde entre os municípios *overachievers* e *underachievers*. Essas análises foram realizadas separadamente para cada tercil da expectativa de vida ao nascer observada. Devido à não normalidade da distribuição das características de saúde selecionadas (verificada pelo teste de *Shapiro-Wilk* e gráfico quantil-quantil da distribuição normal), optou-se por utilizar o teste não paramétrico de *Mann-Whitney* para comparar cada característica de saúde entre os grupos de municípios pré-definidos.

5 RESULTADOS E DISCUSSÃO

5.1 ARTIGO 1

5.1.1 Análise descritiva

Esta seção tem a finalidade de analisar, interpretar e discutir os Resultados a fim de evidenciar o alcance dos objetivos propostos e a contribuição ao conhecimento, resultante dos trabalhos desenvolvidos.

5.1 ARTIGO 1

5.1.1 Análise descritiva

A amostra deste estudo foi composta por 2.808 idosos, ingressantes no estudo SABE em 2000, 2006 ou 2010. A frequência de óbito em 5 anos foi de 15,0% (n=423). Em relação às características socioeconômicas e demográficas, a idade média foi de 70,7 anos (desvio padrão de 8,8), mais da metade dos entrevistados era mulher (59,4%), referiu até 7 anos de estudo (72,9%), considerou sua renda insuficiente para as despesas diárias (64,5%) e não residia sozinho (85,5%). A variável referente à escolaridade foi a que apresentou percentual mais elevado para valores faltantes (13,4%) (Tabela 2).

Comparado com a amostra geral, o grupo dos idosos que morreu em até 5 anos após ingresso no estudo SABE apresentou média de idade superior (76,7 anos; desvio padrão de 9,1) e frequência mais elevada de indivíduos do sexo masculino (54,4%), com escolaridade mais baixa (83,7% tinham até 7 anos de estudo) e com percepção de renda insuficiente (68,3%) (Tabela 2). Para o grupo dos idosos que não morreu em 5 anos, a média de idade foi de 70 anos (desvio padrão de 8,3).

Tabela 2: Características socioeconômicas e demográficas de idosos ingressantes no estudo SABE em 2000, 2006 ou 2010.

Variáveis	Total		Óbito em 5 anos			
	n (2808)	%	Sim		Não	
			n (423)	%	n (2385)	%
Sexo						
Masculino	1139	40,6	230	54,4	909	38,1
Feminino	1669	59,4	193	45,6	1476	61,9
Escolaridade (em anos)						
Nenhuma	575	20,5	131	31,0	444	18,6
1 a 3	653	23,3	102	24,1	551	23,1
4 a 7	816	29,1	121	28,6	695	29,1
≥ 8	387	13,8	40	9,5	347	14,5
Sem informação (<i>missing</i>)	377	13,4	29	6,9	348	14,6
Percepção de renda						
Suficiente	955	34,0	119	28,1	836	35,1
Insuficiente	1811	64,5	289	68,3	1522	63,8
Sem informação (<i>missing</i>)	42	1,5	15	3,5	27	1,1
Vive em lar unipessoal						
Sim	407	14,5	67	15,8	340	14,2
Não	2401	85,5	356	84,2	2045	85,7

Quanto às características comportamentais e referentes ao estado de saúde, aproximadamente metade dos idosos avaliou sua saúde como regular ou ruim/muito ruim (54,2%), a prevalência de consumo de álcool e tabaco foi de, respectivamente, 30,3% e 14,7% e a maioria referiu realizar 3 ou mais refeições por dia (72,3%). Observou-se frequência mais elevada de fumo atual (18,9%) e autopercepção de saúde ruim/muito ruim (16,1%) entre os idosos que morreram em até 5 anos após ingresso no estudo SABE (Tabela 3).

O índice de massa corporal médio da amostra geral foi de 26,9 (desvio padrão de 5,1); 387 (10,2%) observações apresentavam valor faltante para essa variável. Segundo a ocorrência de óbito em 5 anos, o IMC foi de, respectivamente 25,6 (desvio padrão de 5,1; 120 – 28,4% de valores faltantes) para os que morreram e de 27,0 (desvio padrão de 5,1; 267 – 11,2% de valores faltantes) para os que não morreram.

Tabela 3: Estado de saúde e características comportamentais de idosos ingressantes no estudo SABE em 2000, 2006 ou 2010.

Variáveis	Total		Óbito em 5 anos			
	n (2808)	%	Sim		Não	
			n (423)	%	n (2385)	%
Tabagismo						
Fuma atualmente	413	14,7	80	18,9	333	14,0
Já fumou, não fuma mais	953	33,9	170	40,2	783	32,8
Nunca fumou	1441	51,3	172	40,7	1269	53,2
Sem informação (<i>missing</i>)	1	0,0	1	0,2	-	-
Consumo de bebida alcóolica (em dias)						
Não consome	1958	69,7	308	72,8	1650	69,2
< 1 dia	318	11,3	43	10,2	275	11,5
1 a 3	336	12,0	37	8,7	299	12,5
≥ 4	196	7,0	35	8,3	161	6,8
Nº de refeições por dia						
1	107	3,8	18	4,3	89	3,7
2	667	23,8	105	24,8	562	23,6
≥ 3	2029	72,3	300	70,9	1729	72,5
Sem informação (<i>missing</i>)	5	0,2	-	-	5	0,2
Auto avaliação de saúde						
Muito boa	262	9,3	33	7,8	229	9,6
Boa	1016	36,2	123	29,1	893	37,4
Regular	1275	45,4	198	46,8	1077	45,2
Ruim/muito ruim	248	8,8	68	16,1	180	7,5
Sem informação (<i>missing</i>)	7	0,2	1	0,2	6	0,3

As morbidades referidas que apresentaram prevalência mais alta foram hipertensão arterial (54,3%), doença cardíaca (18,8%) e diabetes (17,9%). A frequência de ocorrência de queda nos últimos 12 meses foi de 30,6% e 6,4% dos idosos apresentaram informação faltante para essa variável; Todas as morbidades apresentaram frequência mais elevada entre os idosos que morreram em 5 anos, exceto problema nervoso psiquiátrico (Tabela 4).

Tabela 4: Morbidades autorreferidas de idosos ingressantes no estudo SABE em 2000, 2006 ou 2010.

Variáveis	Total		Óbito em 5 anos			
	n (2808)	%	Sim		Não	
	n (423)	%	n (2385)	%		
Hipertensão Arterial						
Sim	1525	54,3	257	60,8	1268	53,2
Não	1267	45,1	165	39,0	1102	46,2
Sem informação (<i>missing</i>)	16	0,6	1	0,2	15	0,6
Diabetes Mellitus						
Sim	502	17,9	97	22,9	405	17,0
Não	2285	81,4	324	76,6	1961	82,2
Sem informação (<i>missing</i>)	21	0,7	2	0,5	19	0,8
Câncer						
Sim	118	4,2	22	5,2	96	4,0
Não	2684	95,6	400	94,6	2284	94,8
Sem informação (<i>missing</i>)	6	0,2	1	0,2	5	0,2
Doença Pulmonar						
Sim	279	9,9	56	13,2	223	9,4
Não	2520	89,7	364	86,1	2156	90,4
Sem informação (<i>missing</i>)	9	0,3	3	0,7	6	0,3
Acidente Vascular Cerebral						
Sim	205	7,3	61	14,4	144	6,0
Não	2598	92,5	361	85,3	2237	93,8
Sem informação (<i>missing</i>)	5	0,2	1	0,2	4	0,2
Doença cardíaca						
Sim	529	18,8	123	29,1	406	17,0
Não	2261	80,5	300	70,9	1961	82,2
Sem informação (<i>missing</i>)	18	0,6	-	-	18	0,8
Problema Nervoso Psiquiátrico						
Sim	402	14,3	50	11,8	352	14,8
Não	2402	85,5	371	87,7	2031	85,2
Sem informação (<i>missing</i>)	4	0,1	2	0,5	2	0,1
Queda nos últimos 12 meses						
Sim	859	30,6	167	39,5	692	29,0
Não	1768	63,0	248	58,6	1520	63,7
Sem informação (<i>missing</i>)	181	6,4	8	1,9	173	7,3
Mini Exame do Estado Mental (pontos)						
< 13	410	14,6	134	31,7	276	11,6
≥ 13	2398	85,4	289	68,3	2109	88,4

Em relação à dificuldade para realizar atividades ligadas à mobilidade, a maioria das perguntas apresentou frequência de resposta positiva em torno de 30%. Dificuldade para subir vários lances de escada (57,4%) e para se curvar, ajoelhar ou agachar (48,0%) foram as mais frequentes. No grupo de idosos que morreram em 5 anos, a frequência de dificuldade para realizar atividades relacionadas à mobilidade foi aproximadamente 10% maior que a da amostra geral (Tabela 5).

Tabela 5: Condição de mobilidade de idosos ingressantes no estudo SABE em 2000, 2006 ou 2010.

Variáveis	Total		Óbito em 5 anos			
	n (2808)	%	Sim		Não	
	n (423)	%	n (2385)	%		
Dificuldade para caminhar várias ruas (quadras)						
Sim	871	31,0	201	47,5	670	28,1
Não	1933	68,8	221	52,2	1712	71,8
Sem informação (<i>missing</i>)	4	0,1	1	0,2	3	0,1
Dificuldade de se levantar após ficar sentado por longo período						
Sim	977	34,8	176	41,6	801	33,6
Não	1828	65,1	246	58,2	1582	66,3
Sem informação (<i>missing</i>)	3	0,1	1	0,2	2	0,1
Dificuldade para subir vários lances de escada sem descansar						
Sim	1613	57,4	284	67,1	1329	55,7
Não	1186	42,2	136	32,2	1050	44,0
Sem informação (<i>missing</i>)	9	0,3	3	0,7	6	0,3
Dificuldade em se curvar, ajoelhar ou agachar						
Sim	1348	48,0	246	58,2	1102	46,2
Não	1456	51,9	176	41,6	1280	53,7
Sem informação (<i>missing</i>)	4	0,1	1	0,2	3	0,1
Dificuldade para estender os braços acima dos ombros						
Sim	425	15,1	90	21,3	335	14,0
Não	2380	84,8	332	78,5	2048	85,9
Sem informação (<i>missing</i>)	3	0,1	1	0,2	2	0,1
Dificuldade para empurrar/ puxar grandes objetos (ex. poltrona)						
Sim	755	26,9	178	42,1	577	24,2
Não	2050	73,0	244	57,7	1806	75,7
Sem informação (<i>missing</i>)	3	0,1	1	0,2	2	0,1
Dificuldade em levantar pesos maiores que 5kg (ex. sacola de compras pesada)						
Sim	868	30,9	179	42,3	689	28,9
Não	1937	69,0	243	57,4	1694	71,0
Sem informação (<i>missing</i>)	3	0,1	1	0,2	2	0,1
Dificuldade em levantar uma moeda de uma mesa						
Sim	176	6,3	61	14,1	115	4,8
Não	2622	93,4	357	84,4	225	95,0
Sem informação (<i>missing</i>)	10	0,4	5	1,2	5	0,2

Na amostra geral, a maioria dos idosos referiu não ter dificuldade para realizar atividades básicas ou instrumentais da vida diária. Entre as atividades básicas, dificuldade para se vestir foi a mais frequente (14,7%) e entre as atividades instrumentais, dificuldade para fazer tarefas domésticas mais pesadas (21,7%). Segundo os grupos de ocorrência de óbito em 5 anos, a dificuldade para realizar atividades básicas e instrumentais foi aproximadamente 3 e 2 vezes mais frequente que a observada na amostra geral, respectivamente (Tabela 6).

Tabela 6: Estado funcional de idosos ingressantes no estudo SABE em 2000, 2006 ou 2010.

Variáveis	Total		Óbito em 5 anos			
	n (2808)	%	Sim		Não	
	n (423)	%	n (2385)	%		
<i>Atividades básicas</i>						
Dificuldade para se vestir						
Sim	413	14,7	130	30,7	283	11,9
Não	2394	84,3	293	69,3	2101	88,1
Sem informação (<i>missing</i>)	1	0,0	-	-	1	0,0
Dificuldade para tomar banho						
Sim	222	7,9	100	23,6	122	5,1
Não	2585	92,1	323	76,4	2262	94,8
Sem informação (<i>missing</i>)	1	0,0	-	-	1	0,0
Dificuldade para comer (ex. cortar a comida, encher o copo)						
Sim	123	4,4	63	14,9	60	2,5
Não	2684	95,6	360	85,1	2324	97,4
Sem informação (<i>missing</i>)	1	0,0	-	-	1	0,0
Dificuldade para deitar/levantar da cama						
Sim	168	6,0	76	18,0	92	3,9
Não	2638	93,9	347	82,0	2291	96,1
Sem informação (<i>missing</i>)	2	0,1	-	-	2	0,1
Dificuldade para ir ao banheiro						
Sim	165	5,9	73	17,3	92	3,9
Não	2641	94,1	349	82,5	2291	96,1
Sem informação (<i>missing</i>)	2	0,1	1	0,2	1	0,0
<i>Atividades instrumentais</i>						
Dificuldade em preparar uma refeição quente						
Sim	199	7,1	80	18,9	119	5,0
Não	2607	92,8	343	81,1	2264	94,9
Sem informação (<i>missing</i>)	2	0,1	-	-	2	0,1
Dificuldade para cuidar do próprio dinheiro						
Sim	337	12,0	111	26,2	226	9,5
Não	2469	87,9	311	73,5	2158	90,5
Sem informação (<i>missing</i>)	2	0,1	1	0,2	1	0,0
Dificuldade para tomar seus remédios						
Sim	414	14,7	129	30,5	285	11,9
Não	2392	85,2	294	69,5	2098	88,0
Sem informação (<i>missing</i>)	2	0,1	-	-	2	0,1
Dificuldade para fazer tarefas domésticas mais leves (ex. arrumar a cama, tirar o pó dos móveis)						
Sim	227	8,1	76	18,0	151	6,3
Não	2579	91,8	347	82,0	2232	93,6
Sem informação (<i>missing</i>)	2	0,1	-	-	2	0,1
Dificuldade para fazer tarefas domésticas mais pesadas (ex. lavar roupa, limpar o banheiro)						
Sim	608	21,7	129	30,5	479	20,1
Não	2198	78,3	294	69,5	1904	79,8
Sem informação (<i>missing</i>)	2	0,1	-	-	2	0,1

5.1.2 Machine Learning para análises preditivas em saúde: exemplo de aplicação para prever óbito em idosos de São Paulo

Folha de rosto

Machine Learning para análises preditivas em saúde: exemplo de aplicação para prever
óbito em idosos de São Paulo

Hellen Geremias dos Santos^{1*}, Carla Ferreira do Nascimento¹, Rafael Izbicki², Yeda
Aparecida de Oliveira Duarte³, Alexandre Dias Porto Chiavegatto Filho¹

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

***Autor correspondente:**

Hellen Geremias dos Santos

Departamento de Epidemiologia, Faculdade de Saúde Pública, Universidade de São Paulo

Avenida Doutor Arnaldo, 715, 1º andar, CEP: 01246-904

Telefone: 55 (11)3061-7737

e-mail: hellengeremias@usp.br

Resumo

Este estudo objetiva apresentar as etapas relacionadas à utilização de algoritmos de machine learning para análises preditivas em saúde. Para isso, realizou-se um tutorial com base em dados de idosos residentes no município de São Paulo, participantes do estudo Saúde Bem-Estar e Envelhecimento (SABE) (n=2.808). A variável resposta foi representada pela ocorrência de óbito em até 5 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. O tutorial foi organizado 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.

Introdução

A redução de custos e a efetividade de intervenções em saúde estão diretamente relacionadas à identificação de indivíduos com maior risco de sofrer determinado desfecho¹. Nesse contexto, a análise preditiva pode subsidiar a ponderação entre danos e benefícios, com o objetivo de auxiliar gestores responsáveis por formular e avaliar políticas públicas a direcionar intervenções preventivas^{2,3}.

Historicamente, alguns modelos têm sido desenvolvidos para tentar prever 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 risco⁶. Modelos preditivos de diagnóstico e prognóstico de câncer também têm sido relatados na literatura, e mostram que o rastreamento mais intenso e o tratamento profilático representam intervenções promissoras em caso de risco elevado para essa doença⁷.

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 prever o risco de mortalidade¹⁰, de readmissão hospitalar¹¹ e de desfechos desfavoráveis ao nascimento¹². 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. (2017)¹ que desenvolveu modelos preditivos de diabetes não diagnosticada a partir de dados de 12.447 adultos entrevistados para o Estudo ELSA, utilizando 5 algoritmos de *machine learning* (regressão logística, redes neurais, *naive bayes*, método dos *K* vizinhos mais próximos e *random forest*). Os melhores resultados de

performance foram observados para o modelo de redes neurais e de regressão logística, sem diferenças relevantes.

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 decisão 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 prever o risco de óbito em 5 anos em idosos residentes no município de São Paulo, participantes do estudo Saúde Bem-Estar e Envelhecimento (SABE).

Métodos

De modo geral, 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 análise¹⁵.

No aprendizado supervisionado, que será o tema deste tutorial, cada observação do conjunto de dados dispõe de um vetor de mensurações para os preditores x_i , $i = 1, 2, \dots, n$, 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 prever 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 treinamento¹⁶.

O seguinte problema de classificação será o foco deste tutorial: fazer 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 outros. O aprendizado de um problema de classificação consiste no particionamento do espaço dos preditores em regiões correspondentes às categorias da resposta de interesse. A fronteira de decisão entre essas regiões, denominada classificador, representa o modelo preditivo estimado por determinado algoritmo¹⁵. Portanto, nesse contexto, o objetivo do aprendizado é construir um classificador, $f(X)$, que faça boas predições da resposta de interesse em observações futuras¹⁸:

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

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 irreduzível. 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 irreduzível, esse classificador está 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 neurais¹⁷.

É importante observar que modelos menos complexos podem apresentar melhor performance que aqueles que são mais flexíveis por estarem menos sujeitos ao sobreajuste e, conseqüentemente, 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 questão¹⁷.

De modo geral, os algoritmos podem ser agrupados nas seguintes categorias: lineares, não lineares e baseados em árvores de decisão. Para o presente tutorial, foram selecionados para comparação 5 algoritmos: regressão logística e regressão logística penalizada (lineares), redes neurais (não linear), *gradient boosted trees* e *random forest* (baseados em árvores de decisão). A Tabela 1 descreve as principais características desses algoritmos.

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:

- Transformação de variáveis quantitativas (via padronização ou normalização);
- Redução de dimensionalidade do conjunto de dados (exclusão de preditores altamente correlacionados ou utilização de análise de componentes principais);
- 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);
- 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 previsões. Tal procedimento é adotado para que a performance dos dados de teste seja fidedigna ao real desempenho do modelo preditivo em dados futuros¹⁶. O próximo tópico descreve características relacionadas ao treinamento de modelos preditivos no contexto de 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 neural¹⁹.

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 (mais complexo) tem variância alta, pois seu resultado será muito diferente para bancos de dados distintos, porém menor viés. Já um modelo pouco flexível (mais simples) 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 desafiadora⁸.

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 observações¹⁵.

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 treinamento¹⁵.

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 $k-1$ irão 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 cruzada¹⁹.

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 a medida em 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 maior¹⁹. Para este tutorial, utilizou-se validação cruzada com $k=10$, repetida 10 vezes.

Aprendizado: medidas para avaliação de performance

Uma vez definido o valor de k , é preciso escolher uma medida para estimar a performance dos modelos que serão 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 , $k = 1, 2, \dots, K$, 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 2, 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 ausente²⁰.

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 especificidade decorrentes de todos os pontos de corte possíveis para p_k^* , 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 modelo²⁰. 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*^{17,19}.

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 5 anos

Para cada algoritmo aplicado neste tutorial, exceto a regressão logística, uma lista de valores candidatos para os hiperparâmetros foi estabelecida e, na sequência, utilizando validação cruzada *k-fold*, realizou-se 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 teste para avaliar seu erro de predição em observações futuras, novamente utilizando a AUC ROC. Como a regressão logística não apresenta hiperparâmetros, a mesma 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 5 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* 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 seguimento²³. 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 estudo²⁴.

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 5 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* (disponível em: <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 bloqueio 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:

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

- 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), auto avaliação de saúde (excelente/muito boa/boa e regular/ruim/muito ruim);
- Morbidades: doenças autorreferidas (hipertensão, diabetes, câncer, doenças 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(ICAZA; ALBALA, 1999);
- 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.
- 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 leves ou 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 (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

Os resultados deste tutorial estão organizados de acordo com as etapas descritas na Figura 1. Foram utilizados 70% dos dados para treinamento dos algoritmos (n=1875; 277 óbitos) e 30% para teste da performance preditiva dos modelos ajustados (n=802; 118 óbitos).

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

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.

A Tabela 3 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 final 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 convencional, respectivamente, sem grandes diferenças entre os mesmos.

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 30,0%, é possível prever com sucesso 70,0% 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 prever óbito em 5 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 5 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 5 anos ($n=684$) como para os que sofreram esse desfecho ($n=118$). Em relação à curva de calibração, a probabilidade de óbito em 5 anos para os três modelos apresentou acurácia elevada até o 6º grupo, decorrente do particionamento da probabilidade predita em 10 grupos de acordo com pontos de corte fixos, 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 grupos subsequentes, qualidade inferior à dos modelos de regressão logística com e sem penalização.

Por fim, após a avaliação de performance dos modelos selecionados, observa-se que 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.

Discussão e conclusões

O presente tutorial 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.

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 preditores, e às técnicas e 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 a definição dos bancos de treinamento e teste. Nesse último caso, como a definição dos bancos é realizada de modo aleatório, dependendo da amostra que compor 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 sensibilidade²⁹. 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 técnicas de reamostragem *post hoc* (*up e down sampling* e *SMOTE – Synthetic Minority Over-sampling Technique*), aplicadas para balancear as classes da resposta de interesse¹⁹.

Algoritmos de *machine learning* buscam resumir o efeito de preditores individuais por meio de métricas específicas, denominadas importância das variáveis. Embora essa medida não tenha interpretação causal ou inferencial, a mesma 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 o compõem, conforme observado na presente análise⁸.

Ainda que os modelos tenham sido ajustados a partir de dados relevantes, e que apresentem boa performance preditiva, os mesmos podem gerar previsões 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 flexíveis¹⁹. 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. 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 saúde³⁰.

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 o mesmo 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 análise¹⁴. 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 contribuir para estratégias de prevenção ou controle²⁹.

Os resultados deste tutorial 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 para aplicação de técnicas de pareamento, 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.

Referências

1. Olivera AR, Roesler V, Iochpe C, et al. Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes - ELSA-Brasil: accuracy study. *Sao Paulo Med J*. 2017;135(3):234-246.
2. Pepe MS. Evaluating technologies for classification and prediction in medicine. *Stat Med*. 2005;24(24):3687-3696.
3. Steyerberg EW. *Clinical Prediction Models*. Vol 53. New York, NY: Springer New York; 2009.
4. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: The Framingham study. *Am J Cardiol*. 1976;38(1):46-51.
5. D'Agostino RB, 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(2):180-187.
6. Expert Panel on Detection, Evaluation and T of HBC in A. 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 J Am Med Assoc*. 2001;285(19):2486-2497.

7. Gail MH. Twenty-five years of breast cancer risk models and their applications. *J Natl Cancer Inst.* 2015;107(5):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(23):1805-1814.
9. Mullainathan S, Spiess J. Machine Learning: An Applied Econometric Approach. *J Econ Perspect.* 2017;31(2—Spring):87-106.
10. Rose S. Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol.* 2013;177(5):443-452.
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(7):1-14.
12. Pan I, Nolan LB, Brown RR, et al. Machine learning for social services: A study of prenatal case management in Illinois. *Am J Public Health.* 2017;107(6):938-944.
13. Obermeyer Z, Emanuel EJ. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med.* 2016;375(13):1216-1219.
14. Chen JH, Asch SM. Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. *N Engl J Med.* 2017;376(26):2507-2509.
15. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York: Springer New York; 2008.
16. Raschka S. *Python Machine Learning.*; 2014.
17. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning.* Vol 103. New York, NY: Springer New York; 2014.
18. Izbicki R, dos Santos TM. *Machine Learning Sob a Ótica Estatística - Uma Abordagem Preditivista Para a Estatística Com Exemplos Em R.*; 2018. <https://rizbicki.files.wordpress.com/2016/09/main2.pdf>.
19. Kuhn M, Johnson K. *Applied Predictive Modeling.* (Springer, ed.). New York, NY: Springer New York; 2013.
20. Meurer WJ, Tolles J. Logistic Regression Diagnostics. *JAMA.* 2017;317(10):1068.
21. Pencina MJ, D'Agostino RB. Evaluating Discrimination of Risk Prediction Models. *Jama.* 2015;314(10):1063.
22. Lebrão ML, Duarte YA de O. *O Projeto SABE No Município de São Paulo: Uma Abordagem Inicial.* 1st ed. Brasília: Organização Pan-Americana de Saúde; 2003.
23. Lebrão ML, Duarte YA de O. *Desafios de um estudo longitudinal : o Projeto SABE.*

- Saúde Coletiva. 2008;5(24):166-167.
24. Corona LP, Duarte YA de O, Lebrão ML. Prevalence of anemia and associated factors in older adults: Evidence from the SABE Study. *Rev Saude Publica*. 2014;48(5):723-731.
 25. Yourman LC, Lee SJ, Schonberg MA, Widera EW, Smith AK. Prognostic Indices for Older Adults: A Systematic Review Lindsey. *JAMA*. 2012;307(2):182.
 26. Suemoto CK, Ueda P, Beltrán-Sánchez H, 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*. 2016;0(0):1-7.
 27. Icaza MG, Albala C. Minimental State Examination (MMSE) del estudio de dementia en Chile: análisis estatístico. 1999:1-18.
 28. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7(1):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.
 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. *Proc 21th ACM SIGKDD Int Conf Knowl Discov Data Min - KDD '15*. 2015:1721-1730.31.
 31. Kessler RC, Rose S, Koenen KC, 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(3):265-274.

Figura 1: Roteiro para aplicação de algoritmos de *machine learning* em análise preditiva.

Figura 2: Histogramas para as probabilidades estimadas e curva de calibração de dados de teste para regressão logística, regressão logística penalizada e redes neurais.

Dados correspondentes ao gráfico de calibração apresentado na figura 2.

	Óbitos	Percentual (%) de óbitos para cada grupo derivado de pontos de corte fixos, estabelecidos para a probabilidade predita									
Observados	Esperados	5,0	15,0	25,0	35,0	45,0	55,0	65,0	75,0	85,0	95,0
	Regressão Logística Penalizada	4,4	16,2	28,7	38,1	46,1	50,0	36,4	75,0	33,0	50,0
	Regressão logística	4,2	19,2	26,7	29,3	50,0	54,5	30,0	60,0	50,0	25,0
	Redes Neurais	4,1	16,3	30,7	31,0	50,0	44,4	40,0	0,0	0,0	0,0

Tabela 1: Principais características dos algoritmos de machine learning utilizados neste tutorial e seus pacotes no R.

Algoritmo (Pacote)	Hiperparâmetros otimizados	Principais características	Importância dos preditores
<i>Regressão logística (glm)</i>	-	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.
<i>Regressão logística penalizada (glmnet)</i>	<p><i>Alpha</i>: porcentagem correspondente ao tipo de regularização a ser aplicada: 0: <i>ridge</i>; 1: <i>lasso</i>; >0 e <1: <i>elastic net</i>.</p> <p><i>Lambda</i>: parâmetro de regularização. Pode assumir valores no intervalo $[0, \infty]$. Quanto maior o seu valor, maior a penalização das estimativas dos parâmetros da regressão.</p>	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.
<i>Redes neurais (nnet)</i>	<p><i>Size</i>: refere-se ao número de unidades latentes da camada intermediária (oculta) de uma rede neural.</p> <p><i>Decay</i>: parâmetro de regularização. Pode assumir valores no intervalo $[0, \infty]$. Quanto maior o seu valor, maior a penalização das estimativas dos parâmetros das redes neurais.</p>	<p>Modelo em 2 estágios.</p> <p><i>Estágio 1</i>: aplica transformações não lineares (usualmente a função sigmoide) às combinações lineares dos preditores, dando origem às unidades latentes.</p> <p><i>Estágio 2</i>: relaciona as unidades latentes à resposta de interesse.</p>	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).
<i>Gradient boosted trees (xgboost)</i>	<p><i>Nrounds</i>: número de árvores (iterações) presentes no modelo final.</p> <p><i>Maxdepth</i>: profundidade da árvore, relacionada ao número de divisões presentes em cada árvore.</p> <p><i>Eta</i>: parâmetro de regularização relacionado ao controle da contribuição de cada árvore para a função de predição final. Pode assumir valores no intervalo $[0, \infty]$.</p>	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.
<i>Random Forest (randomForest)</i>	<i>Mtry</i> : 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 <i>bootstrap</i> 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 <i>bootstrap</i> é calculada. Esse mesmo procedimento é realizado após permutar cada um dos preditores. A diferença entre estas duas medidas de precisão é calculada e, posteriormente, uma média dessa diferença é computada e normalizada, para cada preditor.

Tabela 2: Matriz de confusão para problemas de classificação com resposta dicotômica.

		RESPOSTA OBSERVADA		
		Presente	Ausente	
RESPOSTA PREDITA	Presente	Verdadeiro positivo (<i>a</i>)	Falso positivo (<i>b</i>)	<i>a + b</i>
	Ausente	Falso negativo (<i>c</i>)	Verdadeiro negativo (<i>d</i>)	<i>c + d</i>
		<i>a + c</i>	<i>b + d</i>	

Tabela 3: Aprendizado (treinamento) e avaliação (teste) de modelos preditivos.

Algoritmos	Regressão Logística	Regressão Logística com Penalização	Redes Neurais	<i>Gradient boosted trees</i>	<i>Random Forest</i>
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 Corpórea	Dificuldade para ir ao banheiro
4	Sexo	Dificuldade para comer	Dificuldade para ir ao banheiro	Sexo	Dificuldade para tomar banho
5	Diabetes <i>Mellitus</i>	Sexo	Diabetes <i>Mellitus</i>	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 <i>p</i> (risco predito)					
<i>p</i>= 0,5					
<i>S</i> (VP)	0,144 (17)	0,130 (15)	0,08 (10)	0,144 (17)	0,110 (13)
<i>1- E</i> (FP)	0,03 (21)	0,026 (18)	0,020 (13)	0,026 (18)	0,019 (13)
<i>p</i> ótimo**					
<i>S</i> (VP)	0,653 (77)	0,670 (79)	0,703 (83)	0,712 (84)	0,700 (79)
<i>1- E</i> (FP)	0,251 (172)	0,253 (173)	0,300 (205)	0,304 (208)	0,328 (238)
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	-	-	-	-	-

*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.

5.2 ARTIGO 2

5.2.1 Machine learning to predict 30-day quality-adjusted survival in critically ill patients with cancer

Title page

Machine learning to predict 30-day quality-adjusted survival in critically ill patients with cancer

Hellen Geremias dos Santos, MSc^{1*}; Fernando Godinho Zampieri, MD, PhD^{2,3}; Karina Normilio-Silva, MSc^{2,4}; Gisela Tunes da Silva, PhD⁵; Antonio Carlos Pedroso de Lima, PhD⁵; Alexandre Biasi Cavalcanti, MD, PhD^{2,4}; Alexandre Dias P. Chiavegatto Filho, PhD¹

¹Department of Epidemiology, School of Public Health, University of São Paulo, São Paulo, Brazil.

²Research Institute, Heart Hospital (Hospital do Coração – Hcor), São Paulo, São Paulo, Brazil.

³Intensive Care Unit, Hospital Alemão Oswaldo Cruz, São Paulo, São Paulo, Brazil.

⁴Cancer Institute of the State of São Paulo (Instituto do Câncer do Estado de São Paulo – ICESP), São Paulo, São Paulo, Brazil.

⁵Department of Statistics, Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil.

*Correspondence and reprints:

Hellen Geremias dos Santos

Department of Epidemiology, School of Public Health, University of São Paulo

Avenida Doutor Arnaldo, 715, 1º andar, CEP: 01246-904

Telephone number: 55 (11)3061-7737

e-mail: hellengeremias@usp.br

Abstract

Purpose: To develop and compare predictive performance from different machine learning algorithms to estimate quality-adjusted life year (QALY) equal or less than 30 days.

Methods: Six machine learning algorithms (logistic and penalized logistic regression, artificial neural networks, basic decision trees, random forests and gradient boosted trees) were applied with nested cross-validation loops to predict 30-day QALY for patients admitted in a prospective cohort study conduct in Intensive Care Units (ICUs) from two public Brazilian hospitals specialized in cancer care. The model's predictors were 27 characteristics collected at ICU admission. Discrimination was evaluated using the area under the receiver operating characteristic (AUROC) curve. Sensitivity, 1-specificity, true positive and false positive cases were measured for different estimated probability cutoff points (30%, 20% and 10%). Calibration was evaluated through Hosmer-Lemeshow test and calibration plot.

Results: Except for basic decision trees, the adjusted predictive models were close to equivalent, presenting good results for discrimination (AUROC curves more than 0.80) and similar values for sensitivity, 1-specificity, true positive and false positive cases for the cutoff points analyzed. Artificial neural networks presented the best concordance between observed and predicted risk among subjects with lower predicted probability of 30-day QALY, while gradient boosted trees and random forests showed the best concordance among subjects with a higher predicted probability.

Conclusions: Except for basic decision trees, predictive models derived from different machine learning algorithms presented good discrimination and calibration to stratify QALY risk at 30 days in critically ill oncologic patients admitted to the ICU.

Key words: quality of life; prognosis; machine learning; clinical decision-making; intensive care unit; critically ill patients.

Introduction

Intensive Care Unit (ICU) admission decision is complex and must ideally consider not only survival chance, but also future quality of life. Physicians and nurses may be able to make good predictions on mortality, but are frequently not able to accurately predict quality of life during ICU stay or after ICU discharge [1].

The ability to identify oncologic patients likely to benefit from initiation and continuity of critical care is even more cumbersome due to high illness severity and serious baseline conditions [2]. One-year survival of oncologic patients admitted to the ICU is low [3] and quality of life declines continuously. Models that predict quality-adjusted life year (QALY) could be useful to better inform patients, families and healthcare providers on what should be expected after an eventual discharge from the ICU, and to guide decisions regarding end-of-life care [4].

Machine learning algorithms have been used in the context of medical decision to develop predictive models based on clinical data [5, 6]. Changes are frequently applied in order to improve techniques for constructing and evaluating these models, however, little is known about the advantages for preferring one learning algorithm over another [7]. We aimed at developing and comparing predictive performance from different machine learning algorithms to estimate the risk of QALY equal or less than 30 days using baseline data from patients with cancer admitted in ICU. The time scale for QALY is years. However, in our study, it is more sensible to assess quality-adjusted survival in days, although we will still use the well-known abbreviation QALY throughout the text.

Materials and Methods

Study design and participants

We analyzed data from a prospective cohort study of two public Brazilian hospitals specialized in cancer care (Dr. Octávio Frias de Oliveira Cancer Institute – ICESP and Pio XII Foundation – Barretos Cancer Hospital). Details of the protocol have been published previously [2]. The study was approved by the research ethics committees of both institutions (project number 0161/2010 at ICESP and number 337/2010 at Pio XII Foundation). Written informed consent was obtained from the patients or their surrogates.

From March 2010 to August 2010, patients older than 18 years with a confirmed or probable malignant neoplasm admitted to the ICU for less than 24 hours were included for eligibility. For ICESP ICU, during the collection period, we randomly selected either the night (7 p.m. to 7 a.m.) or the day shift (7 a.m. to 7 p.m.) and enrolled all patients consecutively. Such strategy were adopted because the number of daily admissions to the ICU exceeded the staff capacity to include all participants. For Barretos Cancer Hospital, the participants were consecutively selected.

We assessed vital status and health-related quality of life utility 15 days and 3, 6, 12, and 18 months after ICU admission. Additionally, health-related quality of life utility was assessed before ICU admission and vital status at 24 months. Health-related quality of life utility was evaluated by the EQ-5D-3L descriptive system on five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) [8]; when the utility values were negative, we used the zero value. For those patients with missing values for health-related quality of life utility at 15 days or 3 months (n=237) we used either the previous utility value or the mean between the previous and the next utility value for imputation. The follow up period finished in August 2013.

QALYs were calculated for each follow-up interval (from admission to 3 months, 3-6, 6-12, 12-18 and 18-24 months) by multiplying the correspondent survival time by the EQ-5D-3L utility value. Except for the first interval, for which we considered the EQ-5D-3L utility value from 15-day assessment, we used the utility values obtained at the beginning of the follow-up interval. Moreover, QALYs of each follow-up interval were added to calculate the total QALY, corresponding to the overall 24-month follow-up.

Outcome and predictors

Our outcome was the categorical variable derived from the dichotomization of total QALY into equal or less than 30 days (30-day QALY) and more than 30 days. We also conducted additional sensitivity analyses using different cutoffs of total QALY (0-day and 60-day).

We used 27 characteristics collected at ICU admission as predictors: demographic characteristics (age and sex); body mass index (BMI); five comorbidities (chronic renal failure, chronic pulmonary failure, chronic heart failure and diabetes); three health risk factors (alcoholism, steroid use and smoking); Eastern Cooperative Oncology Group (ECOG)

performance status [9]; delirium, assessed with the Confusion Assessment Method for the ICU questionnaire [10]; type of admission (medical, elective surgery or emergency surgery); nosocomial infection; respiratory infection; invasive mechanical ventilation at baseline; health history related to cancer disease (primary cancer site, cancer status, cancer extension and 3 previous cancer treatments – surgery, chemotherapy and radiotherapy); and current cancer-related complications (the presence of any of: intracranial mass effect, obstruction or compression of the airways, spinal cord compression syndrome, significant toxicity due to chemotherapy and major bleeding).

Statistical Analysis

We tested the performance of six machine learning algorithms to predict 30-day QALY at ICU admission: logistic and penalized logistic regression, artificial neural networks, basic decision trees, random forests and gradient boosted trees. Continuous variables (age and BMI) were standardized to avoid oversized effects due to difference in scales.

We applied nested cross-validation loops, composed by an outer and an inner validation process, to estimate the expected error of different machine learning algorithm [11, 12]. For our analyses, in the outer validation process, patients were randomly divided into a training and a test set using leave one out cross validation. Hyperparameters of the machine learning algorithms were then optimized on the training set by grid search and 10-fold cross-validation (inner validation process). Predictive performance was evaluated by calculating discrimination and calibration for the test data. For discrimination we used area under the receiver operating characteristic (AUROC) curve with 95% confidence intervals [13, 14]. Sensitivity (S), 1 - specificity (E), true positive (TP) and false positive (FP) cases were measured for different estimated probability (p) cutoff points. In practice, individual-level predicted probabilities could be aggregated into group-based risk levels to help prioritize patients for receiving an intervention. For instance, to choose the cutoff points for p we tested the results if the hypothetical decisions regarding care outside ICU should be applied to 30%, 20% or 10% of the patients with the highest predicted risk for 30-day QALY. Calibration of the estimated probability was evaluated through Hosmer-Lemeshow test [15] and calibration plots [16].

The performance results presented in this study refer to new data that was not used for model development and selection. Analyses were performed with R software and the model training was carried out through *caret* (Classification and Regression Training) package.

Results

Patient Characteristics

We assessed 808 patients admitted to the two participating ICUs for eligibility. From those, 15 patients were excluded: in 12 patients cancer was ultimately excluded, two did not provide consent to participate and one was in the ICU for more than 24 hours at the time of inclusion. Therefore, the final study sample consisted of 793 patients. Additionally, for this study, we excluded participants who were lost to follow-up (four cases) or with missing data for one of the 27 selected predictors (12 patients), yielding a final analytic sample of 777 patients (eFig. 1).

During a follow-up of 2 years, the median survival time and the QALY for the 777 patients were 195 and 70 days, respectively; 515 (66.3%) participants died and 347 (44.7%) had 30-day QALY. Baseline patient characteristics are shown in Table 1. Most of patients were fully active at baseline according to ECOG performance status and were admitted at the ICU to an elective surgery. Patients with 30-day QALY had more comorbidities, poorer ECOG performance status, more delirium, were more frequently admitted due to medical-reasons and had more nosocomial and respiratory infection and mechanical ventilation use at baseline.

Details about health history regarding cancer and current cancer-related complications are shown in Table 2. Approximately half of patients had been recently diagnosed with an active cancer (51.0%) and curative adjuvant surgery (53.2%) was the most applied previous treatment. The following characteristics regarding cancer were most prevalent in patients with 30-day QALY: recidive/progressive active cancer (51.9%), distant metastasis (40.9%), previous palliative treatments – surgery (9.8%), chemotherapy (19.6%) and radiotherapy (5.8%) and all the current cancer-related complications.

Predictive Models

For each iteration, the tuning hyperparameter combination associated with the biggest AUROC curve was chosen and applied to the entire training set to derive the adjusted predictive models (one to each algorithm), which were then evaluated in the test data. Except for basic decision trees, the differences between the models are not large, with all presenting good results for discrimination in the test data (Table 3).

The estimated probability cutoff point was chosen for three scenarios: considering the 30%, 20% or 10% of the patients with the highest predicted 30-day QALY risk. Overall, the algorithms showed similar true and false positive frequencies. As an example, for a cutoff point related to 10% of the patients with highest predicted risk for the artificial neural networks model ($p \geq 0.83$), with a false positive rate of only 2% it is possible predict with success 20% (68 from 347 cases) of the patients that will have QALY less or equal to 30 days. On the other hand, for the 30% of the patients with the highest predicted risk for the same model ($p \geq 0.65$), with a false positive rate of 11% it is possible to predict with success 53% (184 from 347 cases) of the patients that will have up to 30-day QALY. The same interpretations could be derived for the other adjusted predictive models based on the results presented in Table 3.

Fig. 1 shows the calibration curve for estimated probabilities derived from the five equivalent predictive models. Even though the models presented acceptable calibration, artificial neural networks demonstrated the best concordance between observed event rate and predicted probabilities among subjects with a lower predicted probability of 30-day QALY. On the other hand, gradient boosted trees and random forests resulted in predictions with the best concordance among subjects with a higher predicted probability of 30-day QALY.

The results of the Hosmer-Lemeshow test presented in Table 4 confirms these results: the calibration seemed to be better for the models derived from artificial neural networks, gradient boosted trees and random forest algorithms.

We performed the same analyses regarding discrimination and calibration using two alternative scenarios: 0-day and 60-days QALY. Both predictive models showed acceptable discrimination (eTable 1 and eTable 3 for 0-day and 60-day QALY models, respectively); calibration was acceptable for 60-day QALY (eTable 2) but poorer for 0-day QALY models (eTable 4).

Variable importance according to each machine learning algorithm showed that different approaches produced rankings with similar predictors. Generally, type of admission, age, ECOG performance status, invasive mechanical ventilation at baseline, BMI and previous chemotherapy treatment were the most important predictors (eTable 5).

Discussion

In the current study, we focused on predicting 30-day QALY based on a set of 27 characteristics that can be easily measured in a clinical setting rather than on studying the

relationship between specific characteristics and this outcome. While the latter is of interest in medicine and clinical epidemiology, the former is also relevant as it can contribute to inform patients and physicians about patient prognosis and to help make better decisions regarding interventions[17].

To the best of our knowledge, this is the first time that machine learning algorithms are applied to analyze QALY. Our main findings were that five from the six popular machine learning algorithms presented similar high predictive performance, achieving AUROC curves greater than 0.80 and tree based ensemble methods (gradient boosted trees and random forest) offered substantially greater predictive performance compared to basic decision trees. Regarding the calibration plot, no method had a clear superior calibration; nonetheless, according to Hosmer-Lemeshow test, artificial neural networks, gradient boosted trees and random forest were the models with better calibration.

Our ability to identify critically ill oncologic patients likely to benefit most from ICU admission and stay is still limited, especially due to the difficulty in accurately predicting what should be expected after an eventual ICU admission and/or discharge [18, 19]. Hospital mortality, as well as mortality at ICU or within 30 to 90 days after ICU admission, are the most common responses evaluated in prognostic studies [6, 20, 21], but they may not be enough to provide conclusions regarding benefits and costs of critical care provision for oncologic patients. Ideally, the prognostic evaluation should consider patient-centered outcome measures, such as quality of life or post-ICU burden [19].

Oeyen et al. [22] aimed to assess long-term outcomes and quality of life in critically ill oncologic patients after ICU discharge. The authors observed a decline in quality of life at 3 months, but an improvement at 1 year, although it remained under baseline quality of life. Even though this study discuss functional outcomes related to critical care, the authors did not integrate quality of life with survival data and did not aim to develop a predictive model for the QALY outcome. More recently, Normilio-Silva et al. (2) assessed long-term survival, health-related quality of life, and QALY of oncologic patients admitted to ICUs and found an association of some individual baseline characteristics with QALY, such as type of admission, previous cancer treatments and ECOG performance status. Despite the fact that these findings may help healthcare providers to understand prognostic determinants, it did not result in a predictive model based on multiple predictors.

We applied popular machine learning with simultaneously multiple baseline characteristics to predict 30-day QALY. Even though our results showed that conventional

logistic regression had similar performance to more flexible machine learning algorithms, it is not possible to establish that this will occur in other scenarios. Austin et al. [17] found similar results in a study that examined the predictive ability of machine learning algorithms to predict 30-day mortality risk in patients hospitalized with cardiovascular diseases. Bagging, random forest and boosted regression trees increased the predictive performance of basic regression trees, but did not result in significantly better predictive performance compared to logistic regression.

On the other hand, Cooper et al. [23] applied machine learning algorithms to predict dire outcomes in patients with community-acquired pneumonia and found better predictive performance for artificial neural networks. Therefore, different algorithms should be applied in the development of predictive models in order to identify the one that presents better performance, since even a small improvement can be translated into important differences in health care costs and patient benefits [23].

Regarding the outcome variable, we tested two alternative cutoff points for QALY, i.e. 0-day and 60-day QALY. Overall, both predictive models showed acceptable discrimination, but calibration was poorer for 0-day QALY. Nevertheless, even a model with poor calibration could be useful in some clinical settings, considering it discriminate well at some predicted probability cutoff point [13].

Strengths of our study include standardized assessment of predictor and outcome variables using validated scales in a cohort of patients with cancer admitted to two ICUs. Our sampling methods are robust, considering patients were consecutively or randomly selected from patients admitted at ICU and our follow-up was long, as recommended to assess QALY [24]. In addition, EQ-5D-3L utility index and vital status follow-up data were obtained for all but four patients.

Our study also has some limitations. First, the study was conducted at two centers specialized in cancer care in Brazil. For this reason, the patients with cancer in our cohort may have higher 30-day QALY frequency and different risk characteristics than those patients who may have not sought specialized hospital care. Moreover, systematic differences in QALY values for patients with equal characteristics can occur and, consequently, limit the external validity of the adjusted models. Nonetheless, our conclusions are not intended to be generalized to other populations or outcomes, although the methods presented in this study could help researchers to compare machine learning algorithms for classification problems in alternative scenarios.

Another limitation was the small size of the sample. The availability of more individuals and informative predictors could result in an even better predictive performance. For example, predictors could be refined through inclusion of less readily available data, such as gene expression biomarkers [25], as well as parameters to early detection of respiratory failure and sepsis, which are important indications for critical care provision [19].

Our results add new knowledge for the incipient machine learning in healthcare literature. The comparison between different methods revealed quite similar results, thus demonstrating the feasibility of predicting 30-day QALY through easily obtained clinical data, even on a small sample. Assessing 30-day QALY at ICU admission may bring important insights by providing a tool for identification of patients in greatest need of critical care or to suggest the possibility of starting palliative care outside ICUs. Furthermore, knowing oncologic patients at high risk for QALY up to 30 days may optimize research recruitment and may be useful to stratify participants in clinical trials.

Conclusions

Except for basic decision trees, predictive models derived from different machine learning algorithms presented good discrimination and calibration to stratify QALY risk at 30 days for critically ill oncologic patients admitted to the ICU. Future studies are now needed to determine if the availability of 30-day QALY risk could help physicians' decisions regarding ICU admission, stay and discharge.

References

1. Detsky ME, Harhay MO, Bayard DF, Delman AM, Buehler AE, et al (2017) Discriminative accuracy of physician and nurse predictions for survival and functional outcomes 6 months after an icu admission. *JAMA - J Am Med Assoc* 317:2187–2195. doi: 10.1001/jama.2017.4078
2. Normilio-Silva K, Figueiredo AC, Lima ACP, Silva GT, Silva AN, Levites ADD, et al (2016) Long-term survival, quality of life, and quality-adjusted survival in critically ill patients with cancer. *Crit Care Med* 44:1327–1337. doi: 10.1097/CCM.0000000000001648
3. Staudinger T, Stoiser B, Müllner M, Locker GJ, Laczika K, Knapp S, et al (2000) Outcome and prognostic factors in critically ill cancer patients admitted to the intensive care unit. *Crit Care Med* 28:1322–8.
4. Barnato AE, Angus DC (2004) Value and role of intensive care unit outcome prediction models in end-of-life decision making. *Crit Care Clin* 20:345–362. doi: 10.1016/j.ccc.2004.03.002
5. Austin PC, Tu J V., Ho JE, Levy D, Lee DS (2013) Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *J Clin Epidemiol* 66:398–407. doi: 10.1016/j.jclinepi.2012.11.008
6. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ (2015) Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): A population-based study. *Lancet Respir Med* 3:42–52. doi: 10.1016/S2213-2600(14)70239-5
7. Matheny ME, Ohno-Machado L (2014) Generation of Knowledge for Clinical Decision Support. In: Greenes RA (ed) *Clinical Decision Support: Clinical Decision Support The Road to Broad Adoption*, 2nd ed. Elsevier Inc., United States, pp 309–337
8. Szende A, Oppe M, Devlin N (2007) *EQ-5D Value Sets: Inventory, Comparative Review and User Guide*. Springer.
9. Oken MM, Creech RH, Tormey DC, Horton J, Davis TE, McFadden ET, et al (1982) Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol* 5:649–55.
10. Gusmao-Flores D, Salluh JIF, Dal-Pizzol F, Ritter C, Tomasi CD, Lima MASD, et al

- (2011) The validity and reliability of the Portuguese versions of three tools used to diagnose delirium in critically ill patients. *Clin (São Paulo, Brazil)* 66:1917–22. doi: 10.1590/S1807-59322011001100011
11. Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7:91. doi: 10.1186/1471-2105-7-91
 12. Raschka S (2017) *Python Machine Learning*, 2nd ed. Packt Publishing Ltd, Birmingham.
 13. Meurer WJ, Tolles J (2017) Logistic Regression Diagnostics. *JAMA* 317:1068–9. doi: 10.1001/jama.2016.20441
 14. Pencina MJ, D'Agostino RB (2015) Evaluating Discrimination of Risk Prediction Models. *Jama* 314:1063–4 . doi: 10.1001/jama.2015.11082
 15. Hosmer DW, Lemeshow S, Sturdivant RX (2013) *Applied logistic regression*. John Wiley & Sons, Inc., Barradas.
 16. Kuhn M, Johnson K (2013) *Applied Predictive Modeling*. New York, Springer.
 17. Austin PC, Lee DS, Steyerberg EW, Tu J V. (2012) Regression trees for predicting mortality in patients with cardiovascular disease: What improvement is achieved by using ensemble-based methods? *Biometrical J* 54:657–673. doi: 10.1002/bimj.201100251
 18. Biskup E, Cai F, Vetter M, Marsch S (2017) Oncological patients in the intensive care unit: Prognosis, decision-making, therapies and end-of-life care. *Swiss Med Wkly* 147:1–9. doi: 10.4414/smw.2017.14481
 19. Azoulay E, Schellongowski P, Darmon M, Bauer PR, Benoit D, Depuydt P, et al (2017) The Intensive Care Medicine research agenda on critically ill oncology and hematology patients. *Intensive Care Med* 43:1366–1382. doi: 10.1007/s00134-017-4884-z
 20. Torres VBL, Vassalo J, Silva UVA, Caruso P, Torelly AP, Silva E, et al (2016) Outcomes in Critically Ill Patients with Cancer-Related Complications. *PLoS One* 11:e0164537. doi: 10.1371/journal.pone.0164537
 21. Soares M, Silva UVA, Teles JMM, Silva E, Caruso P, Lobo SMA, et al (2010) Validation of four prognostic scores in patients with cancer admitted to Brazilian intensive care units: results from a prospective multicenter study. *Intensive Care Med* 36:1188–1195. doi: 10.1007/s00134-010-1807-7
 22. Oeyen SG, Benoit DD, Annemans L, Depuydt PO, Van Belle SJ, Troisi RI, et al (2013) Long-term outcomes and quality of life in critically ill patients with hematological or solid malignancies: a single center study. *Intensive Care Med* 39:889–898. doi:

10.1007/s00134-012-2791-x

23. Cooper GF, Abraham V, Aliferis CF, Aronis JM, Buchanan BG, Caruana R, et al (2005) Predicting dire outcomes of patients with community acquired pneumonia. *J Biomed Inform* 38:347–366. doi: 10.1016/j.jbi.2005.02.005
24. Ferguson ND, Scales DC, Pinto R, Wilcox ME, Cook DJ, Guyatt GH, et al (2013) Integrating Mortality and Morbidity Outcomes. *Am J Respir Crit Care Med* 187:256–261. doi: 10.1164/rccm.201206-1057OC
25. Starmans MHW, Fung G, Steck H, Wouters BG, Lambin P (2011) A simple but highly effective approach to evaluate the prognostic performance of gene expression signatures. *PLoS One* 6:e28320. doi: 10.1371/journal.pone.0028320

Fig. 1 Calibration curve of observed and predicted quality-adjusted life year equal or less than 30 days for the equivalent predictive models. Dashed black line: logistic regression. Dotted black line: penalized logistic regression. Dashed light grey line: artificial neural networks. Solid dark grey line: gradient boosted trees. Solid light grey line: random forest.

Table 1 Baseline patient characteristics.

Variables	Overall		QALY ≤ 30 days		QALY > 30 days	
Age, mean (sd)	62	(14.2)	62.9	(14.2)	60.6	(14.2)
Female sex, n (%)	331	(42.6)	145	(41.8)	186	(43.3)
BMI, mean (sd)	24.3	(5.3)	23.2	(5.2)	25.1	(5.3)
<i>Comorbidities</i>						
Chronic renal failure, n (%)	19	(2.4)	8	(2.3)	11	(2.6)
Chronic pulmonary failure, n (%)	18	(2.3)	11	(3.2)	7	(1.6)
Chronic heart failure class III or IV New York Heart Association, n (%)	31	(4.0)	17	(4.9)	14	(3.3)
Diabetes, n (%)	119	(15.3)	56	(16.1)	63	(14.7)
<i>Health risk factors</i>						
Alcoholism, n (%)	92	(11.8)	39	(11.2)	53	(12.3)
Use of steroids, n (%)	30	(3.9)	15	(4.3)	15	(3.5)
Smoking, n (%)						
Never	364	(46.8)	168	(48.8)	196	(45.6)
Former	302	(38.9)	135	(38.9)	167	(38.8)
Current	111	(14.3)	44	(12.7)	67	(15.6)
<i>Other characteristics</i>						
ECOG performance status, n (%)						
Fully active	314	(40.4)	94	(27.1)	220	(51.2)
Restricted for physically strenuous activity	209	(26.9)	88	(25.4)	121	(28.1)
Ambulatory and capable of all self-care but unable to carry out work activities	114	(14.7)	60	(17.3)	54	(12.6)
Capable of only limited self-care, confined to bed/chair > 50% time	96	(12.4)	67	(19.3)	29	(6.7)
Completely disabled, cannot carry on any self-care, totally confined to bed/chair	44	(5.7)	38	(11.0)	6	(1.4)
Delirium, n (%)	122	(15.7)	75	(21.6)	47	(10.9)
Type of admission, n (%)						
Medical	343	(44.1)	233	(67.1)	110	(25.6)
Elective surgery	395	(50.8)	89	(25.6)	306	(71.2)
Emergency surgery	39	(5.0)	25	(7.2)	14	(3.3)
Nosocomial infection, n (%)	73	(9.4)	52	(15.0)	21	(4.9)
Respiratory infection, n (%)	109	(14.0)	82	(23.6)	27	(6.3)
Invasive mechanical ventilation at baseline, n (%)	161	(20.7)	88	(25.4)	73	(17.0)

Table 2 Health history related to cancer disease and current cancer-related complications.

Variables	Overall		QALY ≤ 30 days		QALY > 30 days	
<i>Health history related to cancer disease</i>						
Primary cancer site, n (%)						
Brain and Central Nervous System	35	(4.5)	18	(5.2)	17	(4.0)
Oral cavity and pharynx	55	(7.1)	19	(5.5)	36	(8.4)
Digestive tract	273	(35.1)	113	(32.6)	160	(37.2)
Respiratory system	55	(7.1)	38	(11.0)	17	(4.0)
Breast	49	(6.3)	22	(6.3)	27	(6.3)
Genital system	98	(12.6)	42	(12.1)	56	(13.0)
Urinary tract	81	(10.4)	30	(8.6)	51	(11.9)
Hematological (lymphomas/leukemia/myeloma)	62	(8.0)	32	(9.2)	30	(7.0)
Other/undetermined site	69	(8.9)	33	(9.5)	36	(8.4)
Cancer status, n (%)						
Controlled/remission	48	(6.2)	19	(5.5)	29	(6.7)
Active – recent diagnosis	396	(51.0)	148	(42.7)	248	(57.7)
Active – recidive/progressive	333	(42.9)	180	(51.9)	153	(35.6)
Cancer extension, n (%)						
Limited	254	(32.7)	81	(23.3)	173	(40.2)
Locally advanced	279	(35.9)	117	(33.7)	162	(37.7)
Distance metastasis	232	(29.9)	142	(40.9)	90	(20.9)
Leukemia	12	(1.5)	7	(2.0)	5	(1.2)
<i>Previous cancer treatments</i>						
Surgery, n (%)						
No	308	(39.6)	164	(47.3)	144	(33.5)
Yes – curative adjuvant	413	(53.2)	149	(42.9)	264	(61.4)
Yes – palliative	56	(7.2)	34	(9.8)	22	(5.1)
Chemotherapy, n (%)						
No	480	(61.8)	180	(51.9)	300	(69.8)
Yes – curative adjuvant	211	(27.2)	99	(28.5)	112	(26.0)
Yes – palliative	86	(11.1)	34	(19.6)	18	(4.2)
Radiotherapy, n (%)						
No	608	(78.2)	253	(72.9)	355	(82.6)
Yes – curative adjuvant	144	(18.5)	74	(21.3)	70	(16.3)
Yes – palliative	25	(3.2)	20	(5.8)	5	(1.2)
<i>Current cancer-related complications</i>						
Intracranial mass effect, n (%)	30	(3.9)	15	(4.3)	15	(3.5)
Obstruction or compression of the airways, n (%)	30	(3.9)	16	(4.6)	14	(3.3)
Spinal cord compression syndrome, n (%)	21	(2.7)	13	(3.7)	8	(1.9)
Significant toxicity due to chemotherapy, n (%)	42	(5.4)	28	(8.1)	14	(3.3)
Major bleeding, n (%)	53	(6.8)	22	(6.3)	31	(7.2)

Table 3 Performance of the machine learning algorithms on the test data.

Algorithm	AUC ROC (CI95%)	30% highest risk			20% highest risk			10% highest risk		
		TP (S)	FP (I - E)	p	TP (S)	FP (I - E)	p	TP (S)	FP (I - E)	p
Logistic regression	0.81 (0.78;0.84)	186 (0.54)	47 (0.11)	0.67	127 (0.37)	28 (0.07)	0.79	67 (0.19)	11 (0.03)	0.89
Penalized logistic regression	0.81 (0.78;0.84)	174 (0.50)	59 (0.14)	0.60	124 (0.36)	31 (0.07)	0.70	66 (0.19)	12 (0.03)	0.79
Artificial neural networks	0.82 (0.79;0.85)	184 (0.53)	49 (0.11)	0.65	122 (0.35)	33 (0.08)	0.75	68 (0.20)	10 (0.02)	0.83
Basic decision trees	0.61 (0.57;0.65)	115 (0.33)	118 (0.27)	0.68	83 (0.24)	72 (0.17)	0.68	59 (0.17)	19 (0.04)	0.68
Gradient boosted trees	0.82 (0.79;0.85)	181 (0.52)	52 (0.12)	0.63	130 (0.38)	25 (0.06)	0.72	67 (0.19)	11 (0.03)	0.83
Random forest	0.82 (0.79;0.85)	182 (0.52)	51 (0.12)	0.60	127 (0.37)	28 (0.07)	0.67	64 (0.18)	14 (0.03)	0.74

Table 4 Hosmer-Lemeshow test evaluation for differences between the observed and predicted outcome on test data derived from the adjusted 30-day quality-adjusted life year models.

Algorithm	χ^2	p value
Logistic regression	23.51	0.003
Penalized logistic regression	17.88	0.022
Artificial neural networks	13.08	0.109
Basic decision trees	763.75	0.000
Gradient boosted trees	7.25	0.510
Random forest	13.25	0.104

Supplemental Material **eFig. 1** Flow diagram of participants eligible to Quality-adjusted life year study.

eTable 1 Performance of the machine learning algorithms on the test data to predict 0-day quality-adjusted life year.

Algorithm	AUROC curve (CI95%)	30% highest risk			20% highest risk			10% highest risk		
		TP (S)	FP (I - E)	p	TP (S)	FP (I - E)	p	TP (S)	FP (I - E)	p
Logistic regression	0.75 (0.70;0.80)	78 (0.63)	155 (0.24)	0.17	57 (0.46)	98 (0.15)	0.25	36 (0.29)	42 (0.06)	0.42
Penalized logistic regression	0.70 (0.65;0.77)	70 (0.56)	163 (0.25)	0.15	70 (0.56)	85 (0.13)	0.15	44 (0.35)	34 (0.05)	0.25
Artificial neural networks	0.75 (0.70;0.80)	73 (0.59)	160 (0.25)	0.18	55 (0.44)	100 (0.15)	0.25	37 (0.30)	41 (0.06)	0.37
Basic decision trees	0.83 (0.78;0.88)	113 (0.91)	120 (0.18)	0.10	79 (0.64)	76 (0.12)	0.13	35 (0.28)	43 (0.07)	0.19
Gradient boosted trees	0.74 (0.69;0.79)	67 (0.54)	166 (0.25)	0.15	61 (0.49)	94 (0.14)	0.17	31 (0.25)	47 (0.07)	0.29
Random forest	0.76 (0.72;0.81)	78 (0.63)	155 (0.24)	0.20	60 (0.48)	95 (0.15)	0.28	43 (0.35)	35 (0.05)	0.42

eTable 2 Hosmer-Lemeshow test evaluation for differences between the observed and predicted outcome on test data derived from the adjusted 0-day quality-adjusted life year models.

Algorithm	χ^2	p value
Logistic regression	39.79	<0.001
Penalized logistic regression	72.81	<0.001
Artificial neural networks	19.69	0.012
Basic decision trees	485.43	<0.001
Gradient boosted trees	250.57	<0.001
Random forest	21.14	0.007

eTable 3 Performance of the machine learning algorithms on the test data to predict 60-day quality-adjusted life year.

Algorithm	AUROC curve (CI95%)	30% highest risk			20% highest risk			10% highest risk		
		TP (S)	FP (I - E)	p	TP (S)	FP (I - E)	p	TP (S)	FP (I - E)	p
Logistic regression	0.79 (0.76;0.83)	188 (0.49)	45 (0.11)	0.71	134 (0.35)	21 (0.05)	0.81	68 (0.18)	10 (0.03)	0.91
Penalized logistic regression	0.80 (0.77;0.83)	191 (0.50)	42 (0.11)	0.67	136 (0.36)	19 (0.05)	0.75	68 (0.18)	10 (0.03)	0.85
Artificial neural networks	0.80 (0.76;0.82)	194 (0.51)	39 (0.10)	0.70	138 (0.36)	17 (0.04)	0.78	71 (0.19)	7 (0.02)	0.86
Basic decision trees	0.70 (0.62;0.71)	177 (0.47)	56 (0.14)	0.79	99 (0.26)	56 (0.14)	0.79	41 (0.11)	37 (0.09)	0.79
Gradient boosted trees	0.79 (0.76;0.82)	191 (0.50)	42 (0.11)	0.67	130 (0.34)	25 (0.06)	0.75	66 (0.17)	12 (0.03)	0.83
Random forest	0.82 (0.79;0.85)	195 (0.51)	38 (0.10)	0.64	128 (0.34)	27 (0.07)	0.71	67 (0.18)	11 (0.03)	0.78

eTable 4 Hosmer-Lemeshow test evaluation for differences between the observed and predicted outcome on test data derived from the adjusted 60-day quality-adjusted life year models.

Algorithm	χ^2	p value
Logistic regression	23.55	0.003
Penalized logistic regression	7.72	0.461
Artificial neural networks	13.16	0.106
Basic decision trees	403.42	0.000
Gradient boosted trees	8.80	0.359
Random forest	25.33	0.001

eTable 5 Top 5 most important predictors in the training process of machine learning algorithms according to variable importance evaluation function available on *caret* package.

Ranking	1°	2°	3°	4°	5°
Model					
Logistic regression	type of admission	age	ECOG performance status	invasive mechanical ventilation at baseline	BMI
Penalized logistic regression	ECOG performance status	type of admission	previous cancer treatments (chemotherapy)	invasive mechanical ventilation at baseline	respiratory infection
Artificial neural networks	type of admission	ECOG performance status	BMI	invasive mechanical ventilation at baseline	previous cancer treatments (chemotherapy)
Basic decision trees	type of admission	ECOG performance status	respiratory infection	previous cancer treatments (chemotherapy)	Cancer extension
Gradient boosted trees	type of admission	age	BMI	ECOG performance status	invasive mechanical ventilation at baseline
Random forests	BMI	type of admission	age	respiratory infection	ECOG performance status

5.3 ARTIGO 3

5.3.1 Análise descritiva

A amostra deste estudo foi composta por 3.052 municípios brasileiros com mais de 10.000 habitantes (Figura 15). Os estados destacados no mapa foram utilizados como preditores da expectativa de vida (variáveis indicadoras). Municípios que aparecem em branco são aqueles com menos de 10.000 habitantes e que, portanto, não foram selecionados para análise.

Figura 15: Distribuição dos municípios brasileiros com mais de 10.000 habitantes, segundo o estado.

O estado com maior número de municípios foi o de São Paulo ($n = 366$), seguido de Minas Gerais ($n = 362$) e Bahia ($n = 348$). Por outro lado, Roraima ($n = 8$), Amapá ($n = 9$) e Acre ($n = 15$) representaram os estados com menor número de municípios (Figura 16).

Figura 16: Distribuição de frequências dos municípios brasileiros com mais de 10.000 habitantes, segundo o estado em que estão localizados.

A Tabela 7 apresenta medidas resumo e de variabilidade correspondentes às 35 características demográficas e socioeconômicas dos municípios utilizadas como preditores, bem como sua correlação com a expectativa de vida. As variáveis população residente, densidade populacional e PIB per capita apresentaram distribuição assimétrica com concentração à esquerda (mediana menor que a média). Os preditores: percentual de crianças em situação domiciliar de baixa renda ($r = -0,84$), taxa de alfabetização ($r = 0,84$), percentual de domicílios com automóvel ($r = 0,82$), percentual de domicílios com computador ($r = 0,82$), renda mediana ($r = 0,81$), razão de dependência ($r = -0,72$), percentual de brancos ($r = 0,70$) e percentual de domicílios com geladeira ($r = 0,70$) apresentaram correlação com a expectativa de vida igual ou maior que $|0,70|$.

Tabela 7: Análise descritiva de variáveis demográficas e socioeconômicas de municípios brasileiros com mais de 10.000 habitantes.

(continua)

Variável	Média	Desvio padrão	Coefficiente de variação (%)	Mínimo	P ₁₀	P ₂₅	Mediana	P ₇₅	P ₉₀	Máximo	Correlação com Expectativa de Vida (r)
População Residente	58269,8	271948,1	466,7	10004,0	11109,2	14148,8	21498,0	39797,3	92794,0	11253503,0	0,12
Densidade domiciliar	3,5	0,5	14,3	2,7	3,0	3,2	3,4	3,7	4,0	6,9	-0,59
Densidade populacional	177,0	764,8	432,1	0,2	6,1	16,1	36,1	87,3	221,6	13024,6	0,16
Idade mediana	27,8	3,8	13,7	15,9	22,8	25,6	27,8	31,0	32,6	38,5	0,67
Nascimentos per capita	15,0	3,3	22,0	6,0	11,3	12,7	14,5	16,8	19,1	35,2	-0,39
Razão de dependência	60,0	9,8	16,3	38,5	48,8	52,4	58,5	66,3	72,8	102,5	-0,72
PIB per capita	12910,6	15861,7	122,9	2258	3783,5	4901,8	9271,8	15762,3	24864,8	312257,7	0,40
Índice Gini	0,52	0,1	19,2	0,3	0,5	0,5	0,5	0,6	0,6	0,8	-0,38
Renda mediana	577,6	288,6	50,0	158,5	268,1	323,0	531,2	778,2	957,1	2210,7	0,81
Percentual de crianças em situação domiciliar de baixa renda	60,7	23,0	37,9	2,4	28,1	40,3	64,5	82,6	87,4	95,5	-0,84
Taxa de alfabetização	85,2	9,3	10,9	59,7	72	77,2	87,9	93,3	95,7	99,1	0,84
Taxa de desemprego	4,2	1,9	45,2	0,4	2,1	2,9	3,9	5,2	6,6	17,0	-0,01
Taxa de trabalho infantil	11,9	6,1	51,3	0,8	5,5	7,3	10,6	15,0	19,8	63,2	-0,24
Percentual de aposentados	15,2	3,9	25,7	3,6	10,0	12,7	15,4	17,9	20,1	29,8	0,10
Percentual de brancos	43,6	22,9	52,5	4,3	17,8	24,5	37,9	62,6	78,7	96,1	0,70
Percentual de casados	34,3	8,1	23,6	6,6	22,5	29,0	35,1	40,6	43,9	53,4	0,53
Percentual de idosos	11,0	2,9	26,4	2,6	7,2	9,1	11,1	13,1	14,7	20,4	0,21
Percentual de mulheres	49,9	1,4	2,8	38,7	48,1	49,1	50,0	50,8	51,4	54,2	0,17
Percentual de evangélicos	18,3	9,2	50,3	2,0	7,3	11,0	17,3	24,5	30,9	77,3	0,35
Percentual de pessoas com alguma deficiência	24,2	4,1	16,9	4,1	10,2	19,1	21,3	24,1	26,9	38,8	-0,35
Percentual de domicílios com eletricidade	97,0	5,5	5,7	54,2	91,6	97,0	99,1	99,8	99,9	100,0	0,37

(conclusão)

Variável	Média	Desvio padrão	Coefficiente de variação (%)	Mínimo	P₁₀	P₂₅	Mediana	P₇₅	P₉₀	Máximo	Correlação (r) com Expectativa de Vida
Percentual de domicílios com automóvel	29,8	19,8	66,4	0,0	7,7	12,1	24,9	47,6	58,5	78,8	0,82
Percentual de domicílios com computador	23,0	15,5	67,4	0,6	6,0	9,1	19,4	35,3	46,0	72,7	0,82
Percentual de domicílios com geladeira	87,5	12,3	14,1	28,0	70,0	80,5	91,6	97,6	99,0	100,0	0,70
Proporção de residentes em áreas arborizadas	0,5	0,2	40,0	0,0	0,1	0,3	0,4	0,7	0,8	1,0	0,46
Proporção de residentes em áreas pavimentadas	0,5	0,2	40,0	0,0	0,2	0,3	0,5	0,7	0,8	1,0	0,56
Percentual de cobertura do Programa Bolsa Família	72,5	19,6	27,0	0	44,2	60,5	77,0	86,7	95,0	100,0	-0,23
Percentual de residentes com ensino médio completo	17,6	6,6	37,5	1,9	9,7	12,5	16,6	21,9	26,8	47,5	0,60
Percentual de residentes com ensino superior completo	6,0	3,9	65,0	0,4	2,2	3,1	5,1	8,0	10,9	33,8	0,69
Percentual de residentes que trabalham em outro município	10,7	10,6	99,1	0,0	2,4	4,0	7,0	13,1	24,3	65,0	0,16
Percentual de residentes com carga horária de trabalho ≥ 45 horas	28,1	9,3	33,1	3,5	15,7	22,2	28,0	33,9	40,1	63,6	0,42
Percentual de residentes em favela	0,2	0,38	190,0	0,0	0,0	0,0	0,0	0,2	0,5	7,4	0,16
Percentual de residentes em zona urbana	68,7	22,0	32,0	4,2	36,9	51,4	71,9	88,2	95,9	100,0	0,61
Percentual de residentes no município há menos de 10 anos	5,6	4,6	82,1	0,1	1,8	2,9	4,4	6,9	10,5	45,8	0,02
Percentual de estrangeiros	0,1	0,2	200,0	0,0	0,0	0,0	0,0	0,1	0,2	5,7	0,22

A figura 17 apresenta os preditores com correlação maior ou igual a $|0,80|$. A variável correspondente ao percentual de crianças em situação domiciliar de baixa renda foi a que apresentou correlação mais alta com os demais preditores. Essa variável esteve positivamente correlacionada com razão de dependência ($r = 0,82$) e inversamente correlacionada com o percentual de domicílios com automóvel ($r = -0,94$) ou computador ($r = -0,93$), renda mediana ($r = -0,92$), taxa de alfabetização ($r = -0,87$) e proporção de brancos ($r = -0,81$).

Figura 17: Variáveis demográficas e socioeconômicas com correlação maior ou igual a $|0,80|$.

Para a segunda etapa do estudo, características de saúde foram utilizadas para comparar municípios considerados *underachievers* e *overachievers*. A Tabela 8 apresenta uma síntese numérica para essas variáveis, evidenciando alta variabilidade, medida pelo coeficiente de variação, em especial para as variáveis relacionadas à atenção secundária (número de mamografias realizadas a cada 100 mulheres, número de equipamentos de ultrassom a cada 10.000 nascidos vivos e número de leitos, equipamentos de raio X e de manutenção da vida/ 10.000 habitantes) que, adicionalmente, apresentaram distribuição assimétrica com concentração à esquerda (mediana menor que a média).

Em relação às variáveis relacionadas à atenção primária à saúde, observou-se que 25% dos municípios brasileiros com mais de 10.000 habitantes apresentaram cobertura referente às equipes da Estratégia Saúde da Família maior ou igual a 100% ($P_{75} = 100\%$), assinalando para distribuição assimétrica com concentração à direita. A variável correspondente à cobertura das equipes de saúde bucal apresentou comportamento bimodal, com até 10% dos municípios sem cobertura ($P_{10} = 0,0\%$) e, por outro lado, com 25% dos municípios apresentando cobertura de 90% ou mais ($P_{75} = 93,7\%$).

Tabela 8: Análise descritiva de características de saúde de municípios brasileiros com mais de 10.000 habitantes.

Variável	Média	Desvio padrão	Coefficiente de variação (%)	Mínimo	P₁₀	P₂₅	Mediana	P₇₅	P₉₀	Máximo
Cobertura vacinal	78,2	10,1	12,9	30,1	67,8	72,0	77,0	83,0	89,4	182,4
Cobertura das equipes da Estratégia Saúde da Família	72,3	27,6	38,2	0,0	29,8	52,5	79,2	100,0	100,0	100,0
Cobertura das equipes de Saúde Bucal	54,9	35,8	65,2	0,0	0,0	23,5	56,2	93,7	100,0	100,0
Número de equipes da Estratégia Saúde da Família/ 10.000 habitantes	2,4	1,2	50,0	0,0	0,7	1,5	2,5	3,5	4,0	5,8
Percentual de baixo peso ao nacer	7,7	2,2	28,6	1,2	4,9	6,2	7,7	9,0	10,4	18,1
Percentual de parto cesárea	48,4	17,9	37,0	3,3	24,9	34,4	48,3	61,9	71,7	97,0
Número de mamografias realizadas/ 100 mulheres	8,7	10,0	114,9	0,0	0,0	0,2	7,0	13,0	19,0	246,0
Número de equipamentos de ultrassom/ 10.000 nascidos vivos	42,7	42,8	100,2	0,0	0,0	14,0	34,6	58,9	90,9	555,6
Número de leitos/ 10.000 habitantes	17,6	15,5	88,1	0,0	0,0	9,3	15,3	22,5	31,9	245,4
Número de equipamentos de Raio X/ 10.000 habitantes	1,0	0,8	80,0	0,0	0,0	0,4	0,8	1,3	1,9	7,3
Número de equipamentos para manutenção da vida/ 10.000 habitantes	2,6	2,2	84,6	0,0	0,0	1,0	2,2	3,8	5,6	16,1

5.3.2 Overachieving municipalities in public health: a machine-learning approach

Title page

Overachieving municipalities in public health: a machine-learning approach

Type of manuscript: Brief Report.

Alexandre Dias Porto Chiavegatto Filho*

1 - School of Public Health of the University of Sao Paulo, Department of Epidemiology, 715 Av Dr. Arnaldo, Sao Paulo, SP, Brazil 01246-904.

2 – Harvard T. H. Chan School of Public Health, Department of Social and Behavioral Sciences, 677 Huntington Ave, Boston, MA 02115.

Hellen Geremias dos Santos

School of Public Health of the University of Sao Paulo, Department of Epidemiology, 715 Av Dr. Arnaldo, Sao Paulo, SP, Brazil 01246-904.

Carla Ferreira do Nascimento

School of Public Health of the University of Sao Paulo, Department of Epidemiology, 715 Av Dr. Arnaldo, Sao Paulo, SP, Brazil 01246-904.

Kaio Henrique Correa Massa

School of Public Health of the University of Sao Paulo, Department of Epidemiology, 715 Av Dr. Arnaldo, Sao Paulo, SP, Brazil 01246-904.

Ichiro Kawachi

Harvard T. H. Chan School of Public Health, Department of Social and Behavioral Sciences, 677 Huntington Ave, Boston, MA 02115.

* Corresponding Author: Alexandre D. P. Chiavegatto Filho, PhD, School of Public Health, University of Sao Paulo, Department of Epidemiology. 715 Av Dr. Arnaldo, Sao Paulo, SP, Brazil 01246-904. Telephone: 11-3961-7914. Email: alexdiasporto@usp.br.

Abstract

Background: Identifying successful public health ideas and practices is a difficult challenge due to the presence of complex baseline characteristics that can affect health outcomes. We propose the use of machine learning algorithms to predict life expectancy at birth, and then compare health-related characteristics of the under and overachievers (i.e. municipalities that have a worse and better outcome than predicted, respectively).

Methods: Our outcome was life expectancy at birth for Brazilian municipalities and we used as predictors 60 local characteristics that are not directly controlled by public health officials (e.g. socioeconomic factors).

Results: Overachievers presented better results regarding primary health care, such as higher coverage of the massive multidisciplinary program Family Health Strategy. On the other hand, underachievers performed more Caesareans and mammographies, and had more life-support health equipment.

Conclusions: The findings suggest that analyzing the predicted value of a health outcome may bring insights about good public health practices.

Introduction

Machine learning algorithms have been successfully employed for complex health-related problems such as disease diagnosis,¹ mortality risk prediction,² and evaluating adverse birth risks,³ but its potential in public health remains under-explored.

One of the most difficult challenges in public health is to be able to identify successful ideas and practices. Favorable health situations at the local level are not necessarily a consequence of good public health management, as there are a large number of characteristics with potentially complex interactions that can affect local health outcomes (e.g. municipalities with very high per capita income usually have high life expectancy, independently of the quality of their public health management).⁴ We propose a machine learning framework to identify successful public health practices at the local level. We start by training algorithms to predict the value of a municipality's life expectancy at birth using as predictors only local characteristics that are not directly controlled by local public health managers (e.g. socioeconomic factors). Given a high predictive performance of the algorithms, we will then identify the municipalities that have a better outcome than predicted ("overachievers"), and compare their health-related characteristics with the ones that have a worse outcome than predicted ("underachievers").

Methods

We analyzed official public data from municipalities of Brazil, a country with both a recent history of reliable health data,⁵ and the presence of large socioeconomic disparities that can affect health outcomes.^{6,7} For our analyses, we included only the municipalities with more than 10,000 residents to decrease the influence of random annual variations in health outcomes ($n = 3,052$ from a total of 5,565 municipalities). All the data collected is from 2010, the year of the last Census, and the unit of analysis was each municipality. Our public health outcome of interest was life expectancy at birth for each municipality, calculated previously by the United Nations' Development Programme (UNDP).⁸ We used as predictors 60 local characteristics that are not directly controlled by public health officials and are publicly available at the municipal level, such as per capita income, proportion of illiterate residents, proportion of households that have computers, proportion of women, unemployment rate, proportion of

married residents, proportion of white residents, the State that the municipality is situated, among others (see eAppendix 1).

We tested the performance of a total of 16 widely available machine learning algorithms, and then added an ensemble algorithm that creates an optimal weighted average of the individual models (SuperLearner), to predict life expectancy at birth. The algorithms included artificial neural networks, random forests, gradient boosted trees, least squares and penalized linear regression (ridge and lasso), support vector machines, among others (see eAppendix 2). Continuous variables were normalized to avoid oversized effects due to difference in scales, and the categorical variable (the State where the municipality is situated) was transformed into indicator variables.

We then applied nested 10-fold cross-validation, composed by an outer and an inner validation process, to select the hyperparameters and weights of the different machine learning algorithm that composed the SuperLearner, and to test its performance.^{9,10} The use of nested cross-validation performance allows the SuperLearner to be a good estimate of the expected error when tested on new (unseen) data.¹⁰ The predictive performance of the algorithms was measured by the mean squared error (MSE). All the analyzes were performed with R software and model training with *caret* and *SuperLearner* package.¹¹

Health characteristics

Municipalities were separated into tertiles according to life expectancy at birth in order to compare municipalities with similar overall socioeconomic characteristics, where the first tertile is the one with lower life expectancy. Within each group we ranked the under and overachievers based on their performance in life expectancy at birth (i.e. lower or higher than their predicted values). For each tertile, we selected the top 100 overachievers and top 100 underachievers and compared the health characteristics of the two groups using Mann-Whitney U tests, due to non-normality of the distributions. Health characteristics analyzed included vaccination coverage, hospital beds per 10,000 residents, coverage of Estratégia Saúde da Família (Brazil's main primary health program),¹² ultrasound machines per 10,000 live births, among others (see eAppendix 3).

Results

Life expectancy at birth for the 3,052 municipalities ranged from 65.5 in Joaquim Nabuco, Pernambuco, to 78.6 in Brusque, Santa Catarina (total range of 13.1 years). The best performing single algorithm, random forests, presented a cross-validated root mean squared error of 0.175 (Figure 1). For random forests, residing in Minas Gerais State, the illiteracy rate and the proportion of households with an automobile were the most important variables for improving overall predictive performance (see eAppendix 4).

The original algorithms were then combined, with weights defined by 10-fold cross validation, to create the SuperLearner, which presented the best predictive performance (MSE: 0.168). The ensemble was used to make the final predictions on life expectancy at birth for each of the 3,052 municipalities and to identify the under and overachievers (eAppendix 5). The cross-validated values of predicted and actual life expectancies for each of the municipalities is presented in Figure 2. Overall, the algorithm performed well on the entire distribution, i.e. when it predicted a high life expectancy the municipality presented a high life expectancy, and vice-versa. The map of the spatial distribution of the under and overachievers according to tertiles of life expectancy is presented in eAppendix 6. It is possible to observe an interestingly broad distribution of under and overachievers throughout Brazil without indications of clustering.

Table 1 presents the results for the differences in health-related characteristics between the under and overachievers. There were considerable differences between the two groups mostly for the second and third tertiles. Overall, overachievers presented better results regarding investments in primary health care, i.e. care usually provided by generalist physicians, dentists and nurses, such as higher coverage of the massive multidisciplinary program *Estratégia Saúde da Família* (Family Health Strategy) as well as basic oral health coverage, and vaccination coverage. On the other hand, underachievers performed more Caesareans and mammographies, and had more life-support health equipment, which are usually provided by medical specialists (secondary health care).

Discussion

Our results show that it is possible to predict with good accuracy the life expectancy of Brazilian municipalities using solely socioeconomic and demographic characteristics. We then analyzed health characteristics of the underachievers and overachievers, i.e. municipalities that

performed worse and better than predicted given their socioeconomic characteristics, respectively, and found that overachievers performed better on primary health indicators, while underachievers fared better on secondary care indicators.

The results confirm the importance of primary healthcare investments identified in previous analyses that applied traditional epidemiological methods,^{13,14} highlighting the necessity of allocating resources to areas with higher marginal benefits, especially on a developing country with scarce resources like Brazil. Contrary to what could be expected, underachievers did not perform worse in all healthcare indicators, but in fact showed higher investments in technologies such as mammographies and Caesareans performed, as well as the availability of life-support equipment.

The results also bring new knowledge for the incipient machine learning in healthcare literature. First, tree-based algorithms (random forests and gradient boosted trees) achieved the highest performance of the individual algorithms. Many recent articles resort to only testing one of the two for prediction problems, instead of comparing the performance of a broader group of algorithms, which our results show is an acceptable strategy. Second, an ensemble of machine learning algorithms (SuperLearner) presented the highest predictive performance, including a 34.9% gain in performance in comparison with standard decision trees.

Our study introduced a new strategy for identifying successful healthcare initiatives, which brings a new approach to the outcome score literature.¹⁵ The same strategy can be applied for other healthcare problems, such as identifying hospitals with a lower than expected mortality rate (after predicting the rate using the characteristics of the patients), identifying companies with more occupational accidents than expected (after predicting this number using the characteristics of industry and workforce), among many other applications.

The study has a few limitations. First, the selection of different algorithms could lead to different results. However, we performed the same analyses using the highest performing individual algorithm (random forests) and found similar results (see eAppendix 7). Second, there could be other relevant predictive variables not accounted for in our analyses. We aimed at including every socioeconomic characteristic that could plausibly affect life expectancy and was available in Brazil, but other unaccounted variables could have an effect on prediction. However, the effect of each individual variable on improving predictive performance was limited. We tested the results by excluding the first, second and third single variable that most improved predictive performance and the difference found was small (eAppendix 8).

Recent advances in machine learning have the potential to improve how healthcare is provided and evaluated.^{16,17} As healthcare costs continue to increase, new pressures arise to cut access or availability of care, which could be more properly resolved by improving the efficiency of healthcare investments and management. Our study shows that assessing the predicted value of a health outcome may bring new insights about good public health practices.

References

- 1 – Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402-2410.
- 2 – Motwani M, Dey D, Berman DS, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *Eur Heart J*. 2017;38:500-507.
- 3 – Pan I, Nolan LB, Brown RR, et al. Machine learning for social services: A study of prenatal case management in Illinois. *Am J Public Health*. 2017;107:938-944.
- 4 – Berman LF, Kawachi I. *Social epidemiology*. New York: Oxford University Press; 2014.
- 5 – Queiroz BL, Freire FHMA, Gonzaga MR, Lima EEC. Completeness of death-count coverage and adult mortality (45q15) for Brazilian states from 1980 to 2010. *Rev Bras Epidemiol*. 2017; Suppl 01:21-33.
- 6 – Victora C. Socioeconomic inequalities in Health: Reflections on the academic production from Brazil. *Int J Equity Health*. 2016;15:164.
- 7 – Landmann-Szwarcwald C, Macinko J. A panorama of health inequalities in Brazil. *Int J Equity Health*. 2016;15:174.
- 8 – United Nations' Development Programme. Human Development Atlas in Brazil. 2013. <http://atlasbrasil.org.br/2013/en>. Accessed February 10, 2018.
- 9 – Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan M. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): A population-based study. *Lancet Respir Med*. 2015;3:42–52.
- 10 – Kennedy C. Guide to SuperLearner. <https://cran.r-project.org/web/packages/SuperLearner/vignettes/Guide-to-SuperLearner.html>. Accessed June 25, 2018.
- 11 – Kuhn M, Johnson K. *Applied predictive modeling*. New York: Springer, 2013.

- 12 – Rasella D, Harhay MO, Pamponet ML, Aquino R, Barreto ML. Impact of primary health care on mortality from heart and cerebrovascular diseases in Brazil: a nationwide analysis of longitudinal data. *BMJ*. 2014;349:g4014.
- 13 – Starfield B, Shi L, Macinko J. Contribution of primary care to health systems and health. *Milbank Q*. 2005;83:457-502.
- 14 – Starfield B, Shi L, Grover A, Macinko J. The effects of specialist supply on populations' health: assessing the evidence. *Health Aff*. 2005;W5:97-107.
- 15 – Wyss R, Glynn RJ, Gagne JJ. A review of disease risk scores and their application in pharmacoepidemiology. *Curr Epidemiol Rep*. 2016;3:277-284.
- 16 – Obermeyer Z, Emanuel EJ. Predicting the future - Big Data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375:1216-1219.
- 17 – Obermeyer Z, Lee TH. Lost in thought - The limits of the human mind and the future of medicine. *N Engl J Med* 2017;377:1209-1211.

Figure 1: Cross-validated performance of the machine learning algorithms according to MSE, RMSE and R^2 .

Figure 2 – Cross-validated results for predicted and actual life expectancy at birth for the 3,052 municipalities.

Table 1- Differences in health-related characteristics between under and overachievers, according to tertiles of life expectancy at birth.

Healthcare variables	1st Tertile				2nd Tertile				3rd Tertile			
	Over	Under	Difference*	CI (95%)	Over	Under	Difference*	CI (95%)	Over	Under	Difference*	CI (95%)
Caesarean deliveries (%)	30.27	33.56	-3.59	-7.17;-0.04	37,08	59,68	-21.34	-25.23;-17.25	57,03	62,90	-6.41	-10.60;-2.35
Family Health Strategy teams per 10,000 residents	3.4	3.2	0,10	-0.20;0.40	3,00	2,20	1,10	0.70;1.40	1,95	1,40	0,40	0.10;0.80
Hospital beds per 10,000 residents	12.9	14.75	-1.60	-4.60;0.60	15,00	16,40	-2.02	-5.50;0.70	16,00	15,45	0,80	-1.90;3.70
Life support equipment per 10,000 residents	0.8	1.3	-0.50	-0.90;-0.10	1,45	3,65	-2.10	-2.60;-1.60	3,00	3,25	-0;00	-0.60;0.50
Low birth weight (%)	7.11	6.64	0,51	-0.03;1.06	6,96	8,20	-0.97	-1.68;-0.36	8,40	8,70	-0.49	-1.08;0.11
Mammographies per 100 women	2,00	3,00	-0.00	-1.00;0.00	3,00	13,00	-9.00	-11.00;-8.00	11,50	14,00	-4.00	-6.00;-2.00
Oral health coverage	69.74	84.72	-0.00	-7.34;0.00	81,23	36,78	28,86	19.64;39.86	38,88	21,87	11,00	0.00;21.30
Primary health coverage for poor residents (last year)	100,00	94.81	0,00	-0.00;0.00	85,51	76,46	14,18	5.31;19.70	63,43	54,77	8,90	0.00;18.42
Ultrasound machines per 10,000 live births	22.1	33.95	-5.60	-14.40;0.00	27,60	43,65	-14.50	-25.30;-0.00	39,15	38,30	0,50	-8.00;11.60
Vaccination coverage (%)	81.23	78.6	2,08	-0.99;5.09	77,48	74,97	2,82	0.53;5.26	75,88	73,52	2,16	0.31;3.95
X-ray machines per 10,000 residents	0.4	0.5	-0.00	-0.20;-0.00	0,65	1,00	-0.50	-0.70;-0.30	1,00	1,10	-0.10	-0.30;0.10

* For Mann-Whitney U tests, the difference in this case is the median of the difference between a sample from the underachievers and a sample from the overachievers, not the difference of the median between the two groups.

Supplemental material

eAppendix 1

We used as predictors of life expectancy at birth 60 publicly available variables that are not directly controlled by health managers. The list, the formula and the source are presented below.

Variable	Formula	Source
Automobile ownership	Number of households with an automobile / Total households	2010 Census
Bolsa Familia coverage	Number of families receiving Bolsa Familia / Total families	Health Ministry, 2010
Child labor	Number of residents from 10 to 15 years old that work / Residents from 10 to 15	2010 Census
College education	Number of residents over 25 that completed college / Residents over 25	2010 Census
Commuting	Number of residents > 10 years old that work outside of municipality / Total residents	2010 Census
Computer ownership	Number of households with a computer / Total households	2010 Census
Dependency ratio	Number of residents < 15 + > 60 years old / Residents between 15 and 59 years old	2010 Census
Disability rate	Number of residents that are disabled / Total residents	2010 Census
Elderly	Number of residents > 59 years old / Total residents	2010 Census
Electricity	Number of households with electricity / Total households	2010 Census
Evangelicals	Number of evangelical residents / Total residents	2010 Census
Favela (slums) residents	Number of residents living in favelas / Total residents	2010 Census
Foreigners	Number of residents that were not born in Brazil / Total residents	2010 Census
Fridge	Number of households with fridges / Total households	2010 Census
Gini Coefficient	Calculated and adjusted by the Instituto de Pesquisas Econômicas e Aplicadas (IPEA)	IPEA, 2010
Green spaces	Number of residents living in arborized areas / Total residents	2010 Census
Highschool completion	Number of residents that completed high school (over 25 years old) / Residents over 25	2010 Census
Household density	Number of residents / Total households	2010 Census
Illiteracy rate	Number of illiterate residents (over 10 years old) / Residents over 10 years old	2010 Census
Married	Number of married residents / Total residents	2010 Census
Median age	Median age of residents	2010 Census
Median income	Median income of residents > 10 years old	2010 Census
Migrants	Number of residents living for less than 10 years at the municipality / Total residents	2010 Census
Municipal density	Number of residents / Total km ² of municipality	2010 Census
Overworking	Number of residents that work 45+ hours a week / Total residents	2010 Census
Paved street	Number of residents living in paved streets / Total residents	2010 Census
Per capita births	Number of births / Total residents	Health Ministry, 2010
Per capita GDP	GDP of the municipality / Total residents	IBGE, 2010
Poor children	Number of poor children (under 15 years old) / Residents under 15	2010 Census
Residents	Total residents	2010 Census
Retired residents	Number of retired residents / Total residents	2010 Census
State of residence *	Dummy variable for each State	2010 Census
Unemployment	Number of residents > 15 years old unemployed / Residents > 15 years old	2010 Census
Urban area	Number of residents living in urban areas / Total residents	2010 Census
Whites	Number of white residents / Total residents	2010 Census
Women	Number of women / Total residents	2010 Census

* 25 dummy variables were included, representing each state. The most populous, São Paulo, was the reference group and the Federal District was included within Goiás.

eAppendix 2

We tested the predictive performance of 16 popular machine learning algorithms to predict life expectancy of Brazilian municipalities using only socioeconomic characteristics. All were trained and had their hyperparameters tuned using the packages *caret* and *SuperLearner* from R (Kuhn & Johnson, 2013). The list of algorithms and their original libraries are presented below.

Algorithm	Original method/library
Artificial neural networks	nnet
Cubist trees	cubist
Elastic Net	enet/elasticnet
Gradient boosted trees	xgbTree/xgboost
K-nearest neighbors	knn
Lasso regression	lasso/elasticnet
Linear regression	lm
Multivariate adaptive regression spline	earth
Partial least squares	pls
Random forest	rf
Regression tree with complexity tuning	rpart
Regression tree with depth tuning	rpart2
Ridge regression	ridge/elasticnet
SuperLearner	SuperLearner
Support vector machines with linear kernel	svmLinear/kernlab
Support vector machines with polynomial kernel	svmPoly/kernlab
Support vector machines with radial kernel	svmRadial/kernlab

eAppendix 3

After predicting life expectancy at birth for each of the 3,052 municipalities using only socioeconomic characteristics, we separated them into tertiles of life expectancy at birth. We then identified the Top 100 underachievers and the Top 100 overachievers (i.e. the 100 municipalities that had lowest life expectancy than predicted, and the 100 that had highest than predicted) for each tertile, and compared differences in healthcare characteristics, listed below.

Variable	Formula	Source
Caesarean deliveries	Live births by caesarean delivery / Total live births	Health Ministry, 2010
Family Health Strategy teams per 10,000 residents	Number of Family Health Strategy teams / Total residents	Health Ministry, 2010
Hospital beds per 10,000 residents	Total hospital beds / Total residents	Health Ministry, 2010
Life support equipment per 10,000 residents	Total life support equipments / Total residents	Health Ministry, 2010
Low birth weight (%)	Number of live births < 2,500g / Total live births	Health Ministry, 2010
Mammographies per 100 women	Mammographies performed on women aged 50 to 69 / Women aged 50 to 69	Health Ministry, 2010
Oral Health Strategy coverage	Oral Health teams from the Family Health Strategy / Total residents	Health Ministry, 2010
Primary health coverage for poor residents	Families on Bolsa Família with primary care follow-up (last year) / Families on Bolsa Família	Health Ministry, 2010
Ultrasound machines per 10,000 live births	Ultrasound machines / Total live births	Health Ministry, 2010
Vaccination coverage	Residents on par with the national immunization list / Target population for the vaccine	Health Ministry, 2010
X-ray machines per 10,000 residents	X-ray machines / Total residents	Health Ministry, 2010

eAppendix 4

Variable importance for predictive models should not be mistaken for causation or even association as applied to inference problems. For random forests, variable importance is calculated by the linear combination of the usage of each variable in the rule conditions and the model. We assessed variable importance with the `varImp()` function of the `caret` package (Kuhn & Johnson, 2013).

eAppendix 5

Top 100 under and overachievers for the **first** tertile of life expectancy at birth.

Underachievers					Overachievers				
Municipality	State	Observed value	SuperLearner predicted value	Difference	Municipality	State	Observed value	SuperLearner predicted value	Difference
Joaquim Nabuco Saúde	PE	65.55	69.72	-4.17	Boa Viagem	CE	71.09	70.08	1.01
Terra Nova	BA	66.12	70.25	-4.13	Arataca	BA	70.46	69.45	1.01
Várzea Nova	BA	67.02	71.01	-3.99	Guajeru	BA	71.20	70.18	1.02
Neópolis	BA	66.12	70.04	-3.92	Major Isidoro Garraão do Norte	AL	70.31	69.29	1.02
Custódia	SE	67.13	71.02	-3.89		PA	70.75	69.72	1.03
Monteiro	PE	67.33	71.20	-3.87	Buriti Bravo	MA	70.10	69.06	1.04
Iaçu	PB	67.51	71.37	-3.86	Potengi	CE	70.54	69.50	1.04
Itapetim	BA	66.81	70.59	-3.78	Buíque	PE	69.75	68.71	1.04
Una	PE	67.25	70.91	-3.66	Santa Maria da Boa Vista	PE	71.39	70.35	1.04
Ribeirão	BA	67.61	71.26	-3.65	Anajatuba	MA	70.69	69.64	1.05
Juripiranga	PE	67.68	71.29	-3.61	Santana do Ipanema	AL	71.17	70.09	1.08
Mirangaba	PB	65.64	69.20	-3.56	Guimarães	MA	71.31	70.22	1.09
Bonito	BA	66.12	69.67	-3.55	Ouricuri	PE	71.40	70.30	1.10
Japoatã	BA	66.12	69.65	-3.53	Pilar	AL	71.26	70.15	1.11
Mari	SE	67.01	70.53	-3.52	Araripe	CE	70.54	69.42	1.12
Tobias Barreto	PB	66.50	69.95	-3.45	Araci	BA	69.79	68.67	1.12
Lagoa Grande	SE	67.13	70.48	-3.35	Pedras de Fogo	PB	71.23	70.11	1.12
Ribeira do Pombal	PE	67.28	70.60	-3.32	Joselândia	MA	69.59	68.46	1.13
Jurema	BA	67.25	70.45	-3.20	Exu	PE	71.47	70.34	1.13
Fátima	PE	65.87	69.02	-3.15	Anajás	PA	71.43	70.30	1.13
Utinga	BA	66.16	69.30	-3.14	Santa Luzia	BA	71.07	69.93	1.14
Valente	BA	67.38	70.50	-3.12	Atalaia	AL	70.13	68.99	1.14
Jijoca de Jericoacoara	BA	68.75	71.85	-3.10	Brejo	MA	70.45	69.31	1.14
Dona Inês	CE	68.35	71.45	-3.10	Porto	PI	70.47	69.32	1.15
Paulo Ramos	PB	66.41	69.49	-3.08	Banzaê	BA	71.15	70.00	1.15
Acarape	MA	65.64	68.70	-3.06	Ingá	PB	71.07	69.92	1.15
	CE	67.56	70.60	-3.04	Serra do Ramalho	BA	71.51	70.36	1.15
Olivença	AL	65.63	68.65	-3.02	Machados	PE	70.99	69.83	1.16
Quijingue	BA	66.36	69.33	-2.97	Bom Jardim	MA	69.98	68.79	1.19
Murici	AL	66.11	69.08	-2.97	Nova Olinda do Maranhão	MA	70.32	69.12	1.20
Itagi	BA	66.62	69.56	-2.94	Viçosa	AL	70.46	69.26	1.20
Paripiranga	BA	68.02	70.96	-2.94	Lagoa de Itaenga	PE	71.44	70.22	1.22
Princesa Isabel	PB	68.19	71.12	-2.93	União dos Palmares	AL	70.83	69.61	1.22
João Dourado	BA	67.52	70.44	-2.92	Humberto de Campos	MA	70.55	69.33	1.22
Condado	PE	67.79	70.71	-2.92	Jaqueira	PE	70.07	68.83	1.24
Conde	BA	67.31	70.22	-2.91	Ipecaetá	BA	71.53	70.28	1.25

Umburanas	BA	67.31	70.22	-2.91	Crisópolis	BA	70.87	69.59	1.28
Penalva	MA	66.57	69.48	-2.91	Dário Meira	BA	71.27	69.99	1.28
Igarapé do Meio	MA	66.21	69.10	-2.89	Pariconha	AL	70.23	68.94	1.29
Arame	MA	67.06	69.93	-2.87	Simões	PI	71.24	69.93	1.31
São Gabriel	BA	67.21	70.07	-2.86	Capitão de Campos	PI	71.35	70.04	1.31
Duque Bacelar Tapauá	MA	66.56	69.41	-2.85	Baixa Grande Barra do Corda	BA	71.49	70.18	1.31
	AM	66.61	69.43	-2.82		MA	70.75	69.44	1.31
Governador Archer Jitaúna	MA	66.21	69.03	-2.82	Chapadinha	MA	71.21	69.88	1.33
	BA	67.25	70.03	-2.78	Itapiúna	CE	71.49	70.14	1.35
Tacaratu	PE	67.63	70.41	-2.78	Loreto	MA	71.27	69.92	1.35
Peixoto de Azevedo	MT	70.65	73.41	-2.76	Augusto Corrêa	PA	71.30	69.92	1.38
Luzilândia	PI	67.00	69.76	-2.76	Santa Bárbara Caldeirão Grande	BA	71.50	70.12	1.38
Montanhas	RN	67.02	69.76	-2.74		BA	70.46	69.07	1.39
São João do Carú Caém	MA	66.01	68.75	-2.74	Parnarama Olho d'Água das Cunhãs	MA	70.50	69.11	1.39
	BA	67.08	69.80	-2.72		MA	70.24	68.84	1.40
Bom Conselho Cedro	PE	67.22	69.90	-2.68	Araruna	PB	70.45	69.02	1.43
	PE	68.00	70.68	-2.68	Sítio do Mato Conceição do LagoAçu	BA	71.46	70.02	1.44
Afonso Bezerra	RN	67.76	70.44	-2.68		MA	69.26	67.82	1.44
Salinas da Margarida Jeremoabo	BA	68.74	71.41	-2.67	Guaratinga Presidente Vargas	BA	70.95	69.51	1.44
	BA	66.36	69.03	-2.67		MA	71.10	69.64	1.46
Sapé	PB	67.64	70.31	-2.67	Quiterianópolis	CE	71.46	70.00	1.46
Vitorino Freire Barra da Estiva Canto do Buriti Santa Inês	MA	66.26	68.92	-2.66	Mulungu do Morro Urbano Santos Santana do Mundaú Santana do Maranhão	BA	70.18	68.71	1.47
	BA	67.99	70.63	-2.64		MA	71.21	69.72	1.49
	PI	68.41	71.01	-2.60		AL	70.13	68.64	1.49
	BA	67.25	69.83	-2.58		MA	70.45	68.94	1.51
Senador Alexandre Costa Itaíba	MA	66.32	68.89	-2.57	Brejões	BA	71.49	69.98	1.51
	PE	66.25	68.81	-2.56	Ipixuna	AM	71.32	69.78	1.54
Inácio Martins Manaíra	PR	70.91	73.46	-2.55	São Bernardo	MA	70.70	69.15	1.55
	PB	66.85	69.40	-2.55	Paratinga	BA	71.46	69.91	1.55
Camamu	BA	67.61	70.14	-2.53	Anagé Poço Redondo Várzea da Roça	BA	71.27	69.70	1.57
Ouroilândia	BA	67.03	69.54	-2.51		SE	70.58	68.99	1.59
Poção	PE	65.59	68.09	-2.50		BA	71.50	69.88	1.62
Tucano	BA	67.69	70.19	-2.50	Pilar Santa Quitéria do Maranhão Salgado de São Félix Presidente Jânio Quadros	PB	71.52	69.89	1.63
Jucati	PE	65.87	68.36	-2.49		MA	70.61	68.97	1.64
Craíbas	AL	66.22	68.67	-2.45		PB	71.20	69.55	1.65
Saboeiro	CE	67.87	70.31	-2.44		BA	71.29	69.63	1.66

Sítio do Quinto Teotônio Vilela	BA	66.69	69.13	-2.44	Matões do Norte	MA	71.10	69.44	1.66
Touros	AL	67.01	69.42	-2.41	São João	PE	71.09	69.42	1.67
Cajueiro	RN	67.95	70.36	-2.41	Anapurus	MA	70.85	69.16	1.69
Cortês	AL	66.45	68.86	-2.41	Primavera	PA	71.30	69.61	1.69
Itaquitinga	PE	67.42	69.83	-2.41	Jutaí	AM	70.93	69.23	1.70
Pedro do Rosário Itaquiraí	PE	67.87	70.27	-2.40	Morro do Chapéu	BA	70.98	69.27	1.71
Wanderlândia	MA	66.77	69.17	-2.40	Igreja Nova	AL	71.23	69.46	1.77
Ipauimirim	MS	71.30	73.70	-2.40	Dois Riachos	AL	70.74	68.95	1.79
Regeneração	TO	69.89	72.26	-2.37	Quebrangulo Vargem Grande	AL	71.52	69.70	1.82
São Luís do Curu	PE	65.89	68.22	-2.33	Itiúba	MA	71.10	69.28	1.82
Sapeaçu	AL	66.09	68.41	-2.32	Gonçalves Dias	BA	71.47	69.65	1.82
Costa Marques	PR	71.05	73.37	-2.32	Pimenteiras Jequiá da Praia	MA	70.46	68.61	1.85
Buritizeiro	AL	67.85	70.17	-2.32	Jenipapo dos Vieiras	PI	71.43	69.53	1.90
Avelino Lopes Manari	AM	66.61	68.91	-2.30	Cantanhede	AL	71.32	69.40	1.92
Águas Belas	AL	67.01	69.30	-2.29	Maragogi	MA	70.96	68.95	2.01
Mata Grande	MA	66.81	69.10	-2.29	Traipu	MA	71.45	69.42	2.03
Carnaíba	PI	68.30	70.57	-2.27	Caracol	AL	70.96	68.92	2.04
Curimatá	BA	69.07	71.33	-2.26	Alto Alegre do Pindaré	AL	70.92	68.84	2.08
Ribeira do Amparo Varjota	PE	67.87	70.13	-2.26	Nina Rodrigues Barra do Choça	PI	71.23	69.13	2.10
Ibateguara	MS	71.30	73.54	-2.24	Peritoró	MA	71.44	69.33	2.11
Groaíras	BA	68.08	70.32	-2.24	Girau do Ponciano	MA	71.45	69.22	2.23
Umbaúba	PE	68.62	70.84	-2.22	Iati	BA	71.21	68.87	2.34
Cantagalo	RN	67.95	70.16	-2.21	Igaci Joaquim Gomes	MA	71.44	69.03	2.41
Jaicós	AL	66.06	68.26	-2.20	Turilândia	AL	70.72	68.31	2.41
	MA	67.33	69.53	-2.20		PE	71.09	68.46	2.63
	BA	66.66	68.85	-2.19		AL	71.45	68.79	2.66
	AM	68.13	70.32	-2.19		AL	70.82	67.72	3.10
	CE	68.72	70.91	-2.19		MA	71.45	68.21	3.24

Top 100 under and overachievers for the **second** tertile of life expectancy at birth.

Underachievers					Overachievers				
Municipality	State	Observed value	SuperLearner predicted value	Difference	Municipality	State	Observed value	SuperLearner predicted value	Difference
Cunha Porã	SC	73.24	76.32	-3.08	Varzelândia Nova	MG	73.37	71.78	1.59
São Lourenço do Oeste	SC	73.43	76.41	-2.98	Timboteua	PA	72.43	70.84	1.59
João Pinheiro	MG	72.25	75.15	-2.90	Cotriguaçu	MT	74.51	72.91	1.60
Paraopeba	MG	72.55	75.16	-2.61	Pedra Branca	CE	71.73	70.12	1.61
Bady Bassitt	SP	73.31	75.90	-2.59	Axixá	MA	71.54	69.93	1.61
Itaperuçu	PR	71.72	74.30	-2.58	Maraú	BA	73.08	71.47	1.61
São Manuel	SP	73.29	75.87	-2.58	Abaetetuba	PA	72.89	71.27	1.62
Araruna	PR	72.08	74.64	-2.56	Acrelândia	AC	73.45	71.83	1.62
Rio Piracicaba	MG	72.63	75.16	-2.53	Vera Cruz	BA	74.02	72.40	1.62
Ponte Serrada	SC	72.40	74.88	-2.48	Candeias do Jamari	RO	74.11	72.49	1.62
Lontras	SC	74.20	76.51	-2.31	Ubatã	BA	72.29	70.66	1.63
Rio dos Cedros	SC	74.61	76.89	-2.28	Amélia Rodrigues	BA	73.59	71.95	1.64
Correia Pinto	SC	73.21	75.44	-2.23	Massapê	CE	71.70	70.05	1.65
Sinimbu	RS	72.50	74.73	-2.23	Brasil Novo	PA	73.55	71.90	1.65
Planalto	PR	72.16	74.32	-2.16	Tavares	PB	71.59	69.93	1.66
Jandaia do Sul	PR	73.04	75.19	-2.15	Alhandra	PB	71.69	70.03	1.66
Bom Jesus dos Perdões	SP	73.10	75.24	-2.14	Orobó	PE	72.19	70.52	1.67
Tocantins	MG	72.99	75.08	-2.09	Queimadas	PB	72.73	71.06	1.67
General Salgado	SP	73.50	75.55	-2.05	Santana do Cariri	CE	71.72	70.04	1.68
BiritibaMirim	SP	72.70	74.75	-2.05	Umarizal	RN	72.49	70.81	1.68
Morungaba	SP	73.10	75.14	-2.04	Alto Paraíso	RO	73.22	71.54	1.68
Palmital	SP	73.83	75.84	-2.01	Itatim	BA	71.93	70.24	1.69
Vespasiano	MG	73.65	75.65	-2.00	Miraíma	CE	71.70	70.00	1.70
Cajamar	SP	73.59	75.58	-1.99	Macaparana	PE	71.70	70.00	1.70
Anastácio	MS	72.33	74.31	-1.98	Boninal	BA	72.35	70.64	1.71
Pederneiras	SP	73.74	75.72	-1.98	Chapada Gaúcha	MG	74.22	72.50	1.72
Carmo da Cachoeira	MG	73.03	74.98	-1.95	Tamboril	CE	71.71	69.97	1.74
Bela Vista de Minas	MG	72.72	74.66	-1.94	Jacareacanga	PA	73.00	71.24	1.76
Itapeva	SP	73.18	75.11	-1.93	Santo Antônio	RN	71.94	70.16	1.78
Candói	PR	71.69	73.61	-1.92	Canavieiras	BA	72.74	70.96	1.78
Joaquim Távora	PR	73.04	74.94	-1.90	Campinópolis	MT	73.17	71.39	1.78
Matipó	MG	71.80	73.68	-1.88	Tartarugalzinho	AP	72.63	70.84	1.79
Pedralva	MG	73.39	75.27	-1.88	Maués	AM	72.97	71.17	1.80
Morro da Fumaça	SC	74.51	76.36	-1.85	Condeúba	BA	72.32	70.52	1.80
Entre Rios de Minas	MG	73.14	74.98	-1.84	João Lisboa	MA	72.07	70.26	1.81
Cesário Lange	SP	73.23	75.07	-1.84	Nossa Senhora do Livramento	MT	74.20	72.38	1.82

Naviraí	MS	73.15	74.99	-1.84	Salvaterra	PA	72.60	70.76	1.84
Sete Quedas	MS	71.70	73.54	-1.84	Barbalha	CE	74.02	72.18	1.84
Venâncio Aires	RS	74.09	75.91	-1.82	Itaparica	BA	74.56	72.68	1.88
Lauro Muller	SC	74.32	76.13	-1.81	Nova Cruz	RN	72.28	70.38	1.90
Senhora dos Remédios	MG	71.56	73.34	-1.78	Cruz do Espírito Santo	PB	71.66	69.74	1.92
Teixeiras	MG	73.26	75.03	-1.77	Matias Olímpio	PI	71.74	69.81	1.93
Santa Margarida	MG	72.07	73.81	-1.74	Orós	CE	72.58	70.64	1.94
São José do Norte	RS	72.52	74.24	-1.72	Macaúbas	BA	72.43	70.48	1.95
Tijucas do Sul	PR	72.54	74.26	-1.72	Milagres	BA	72.32	70.37	1.95
Brotas	SP	73.90	75.62	-1.72	Autazes	AM	72.91	70.96	1.95
São Luís de Montes Belos	GO	73.55	75.27	-1.72	Santa Cruz da Baixa Verde	PE	72.14	70.18	1.96
Arroio Grande	RS	73.43	75.15	-1.72	Afogados da Ingazeira	PE	73.39	71.38	2.01
Wenceslau Braz	PR	72.42	74.13	-1.71	Maragogipe	BA	72.58	70.55	2.03
Cacequi	RS	73.39	75.10	-1.71	Bonito de Santa Fé	PB	72.18	70.15	2.03
São Joaquim	SC	74.01	75.71	-1.70	Amargosa	BA	73.40	71.36	2.04
Camaquã	RS	74.14	75.80	-1.66	Sairé	PE	72.05	70.01	2.04
Catalão	GO	74.12	75.78	-1.66	Granja	CE	71.68	69.64	2.04
Corbélia	PR	73.44	75.08	-1.64	GuajaráMirim	RO	74.39	72.32	2.07
Juquitiba	SP	72.48	74.12	-1.64	Igrapiúna	BA	72.32	70.22	2.10
Aparecida do Taboado	MS	73.23	74.87	-1.64	São Bento do Una	PE	72.34	70.23	2.11
São Miguel Arcanjo	SP	72.95	74.57	-1.62	Mundo Novo	BA	72.43	70.31	2.12
Capanema	PR	73.04	74.64	-1.60	São Sebastião de Lagoa de Roça	PB	72.67	70.52	2.15
Barros Cassal	RS	72.71	74.28	-1.57	São Sebastião do Maranhão	MG	73.94	71.78	2.16
Mantena	MG	73.16	74.71	-1.55	Taipu	RN	71.68	69.51	2.17
Guaíra	SP	74.46	76.01	-1.55	Mata Roma	MA	71.56	69.38	2.18
São Bento do Sapucaí	SP	73.69	75.24	-1.55	Iaciara	GO	74.53	72.34	2.19
Nova Alvorada do Sul	MS	73.52	75.07	-1.55	Itacaré	BA	73.42	71.23	2.19
Nova Xavantina	MT	73.00	74.55	-1.55	Cacimba de Dentro	PB	71.59	69.35	2.24
Guararema	SP	74.03	75.57	-1.54	Triunfo	PE	74.00	71.75	2.25
Três Rios	RJ	73.03	74.56	-1.53	São José do Campestre	RN	72.34	70.09	2.25
Pinhalzinho	SP	73.49	75.02	-1.53	Aliança	PE	72.93	70.67	2.26
Colorado	PR	73.25	74.77	-1.52	Conceição do Almeida	BA	73.31	71.02	2.29
Itamogi	MG	73.61	75.13	-1.52	Santo Estêvão	BA	73.47	71.18	2.29
Sertãoópolis	PR	73.24	74.75	-1.51	Macajuba	BA	71.81	69.50	2.31
Centenário do Sul	PR	72.48	73.99	-1.51	Carira	SE	72.45	70.12	2.33
Mar de Espanha	MG	74.51	76.02	-1.51	Paranhos	MS	73.65	71.31	2.34
Bom Sucesso	MG	73.49	75.00	-1.51	Coremas	PB	72.65	70.31	2.34

Valparaíso de Goiás	GO	73.91	75.41	-1.50	Iguaí	BA	72.18	69.84	2.34
São Pedro da Aldeia	RJ	73.03	74.51	-1.48	Curionópolis	PA	73.53	71.18	2.35
Alto Araguaia	MT	73.13	74.61	-1.48	Poções	BA	72.85	70.49	2.36
Faxinal	PR	72.75	74.23	-1.48	Piranhas	AL	72.14	69.75	2.39
Crixás	GO	73.31	74.79	-1.48	Remígio	PB	72.83	70.40	2.43
Mococa	SP	74.63	76.10	-1.47	Altinho	PE	72.54	70.10	2.44
Taquarituba	SP	73.65	75.12	-1.47	Nova Canaã	BA	72.35	69.89	2.46
Goiás	GO	73.27	74.73	-1.46	Turiação	MA	71.54	69.07	2.47
Arraial do Cabo	RJ	73.31	74.77	-1.46	Ibirataia Sena	BA	72.33	69.84	2.49
Jacinto Machado	SC	74.06	75.52	-1.46	Madureira	AC	73.59	71.04	2.55
Valentim Gentil	SP	73.62	75.08	-1.46	Chã Grande Cândido Sales	PE	73.18	70.53	2.65
Santo Cristo	RS	74.20	75.66	-1.46		BA	72.90	70.24	2.66
Quedas do Iguaçu	PR	72.65	74.10	-1.45	Cocos	BA	73.28	70.62	2.66
Aimorés	MG	73.58	75.01	-1.43	Brejo do Cruz	PB	72.85	70.15	2.70
Maceió	AL	72.94	74.37	-1.43	Panelas Alagoa Grande	PE	71.68	68.96	2.72
Porto Ferreira	SP	74.63	76.06	-1.43		PB	72.83	70.01	2.82
Miranda Santa Cruz das Palmeiras	MS	71.91	73.34	-1.43	Alagoinha	PE	73.04	70.21	2.83
Cachoeira de Minas	SP	73.87	75.29	-1.42	Coribe	BA	73.44	70.54	2.90
	MG	74.25	75.67	-1.42	Alagoinha Pedro Alexandre	PB	72.72	69.81	2.91
Regente Feijó	SP	74.10	75.52	-1.42		BA	71.68	68.77	2.91
São Pedro do Sul	RS	74.45	75.85	-1.40	Irará	BA	73.53	70.59	2.94
Céu Azul	PR	73.69	75.08	-1.39	Rio Formoso	PE	73.56	70.60	2.96
Itirapina	SP	73.77	75.15	-1.38	Pão de Açúcar Matriz de Camaragibe	AL	72.58	69.57	3.01
Rosana	SP	74.10	75.48	-1.38		AL	71.62	68.59	3.03
São Roque do Canaã	ES	73.16	74.53	-1.37	Feira Nova	PE	73.43	70.02	3.41
Guariba	SP	73.63	75.00	-1.37	Olindina	BA	73.60	70.02	3.58
Janaúba	MG	72.74	74.11	-1.37	Alto Alegre	RR	73.78	70.02	3.76

Top 100 under and overachievers for the **third** tertile of life expectancy at birth.

Underachievers					Overachievers				
Municipality	State	Observed value	SuperLearner predicted value	Difference	Municipality	State	Observed Value	SuperLearner predicted value	Difference
Itabirito	MG	74.68	76.69	-2.01	Uberlândia	MG	78.09	76.61	1.48
São Ludgero	SC	75.01	76.71	-1.70	Comodoro	MT	75.37	73.89	1.48
Santa Rita do Sapucaí	MG	74.80	76.49	-1.69	Rodeio	SC	78.34	76.85	1.49
São Sebastião do Paraíso	MG	75.15	76.80	-1.65	Guapiaçu	SP	76.93	75.43	1.50
Criciúma	SC	75.76	77.36	-1.60	Colombo	PR	77.17	75.67	1.50
Lagoa da Prata	MG	74.89	76.47	-1.58	Caconde	SP	76.45	74.94	1.51
Biguaçu	SC	75.17	76.73	-1.56	Aguaí	SP	76.49	74.98	1.51
Erechim	RS	74.95	76.50	-1.55	Lima Duarte	MG	76.68	75.16	1.52
Nova Ponte	MG	75.08	76.56	-1.48	Nova Londrina	PR	76.36	74.84	1.52
Timbó	SC	76.36	77.80	-1.44	Charqueadas	RS	77.22	75.70	1.52
Ouro Preto	MG	75.03	76.45	-1.42	Perdizes	MG	77.80	76.27	1.53
Juiz de Fora	MG	75.65	77.06	-1.41	Alto Paraná	PR	75.76	74.23	1.53
Itapiranga	SC	75.50	76.89	-1.39	Brasília	DF	77.35	75.82	1.53
Luz	MG	75.00	76.35	-1.35	São Sepé Santo Antônio da Patrulha	RS	77.23	75.69	1.54
Garuva	SC	74.82	76.16	-1.34	Ilhota	SC	76.97	75.41	1.56
Indaiatuba	SP	75.22	76.55	-1.33	Barbacena	MG	77.95	76.39	1.56
Uberaba	MG	75.71	77.03	-1.32	Valença	RJ	77.88	76.31	1.57
Contagem	MG	74.94	76.24	-1.30	Mariana	MG	75.85	74.28	1.57
Alfenas	MG	75.45	76.74	-1.29	Guaratinguetá	SP	77.43	75.85	1.58
Timóteo	MG	75.14	76.43	-1.29	Mirandópolis	SP	78.17	76.59	1.58
Três de Maio	RS	75.02	76.26	-1.24	Porto União	SC	77.12	75.54	1.58
Ituiutaba	MG	75.37	76.60	-1.23	Taubaté	SP	78.43	76.85	1.58
Três Coroas	RS	74.75	75.96	-1.21	Juscimeira	MT	77.98	76.40	1.58
Laranjal Paulista	SP	74.73	75.91	-1.18	Brasnorte	MT	75.54	73.96	1.58
São José do Cedro	SC	74.79	75.97	-1.18	Serro	MG	76.00	74.41	1.59
Jaguaruna	SC	74.80	75.97	-1.17	Ronda Alta	RS	75.01	73.42	1.59
Aparecida	SP	74.67	75.83	-1.16	Seberi	RS	77.03	75.43	1.60
Divinópolis	MG	75.63	76.78	-1.15	Fronteira	MG	76.91	75.31	1.60
Guaíba	RS	74.99	76.13	-1.14	Fartura	SP	77.28	75.66	1.62
Pinhais	PR	75.15	76.29	-1.14	Jataizinho	PR	77.02	75.40	1.62
Ubá	MG	75.45	76.56	-1.11	Canelinha	SC	75.70	74.08	1.62
Veranópolis	RS	75.29	76.39	-1.10	Mairiporã	SP	76.89	75.27	1.62
Carlos Barbosa	RS	75.07	76.16	-1.09	Abelardo Luz	SC	77.86	76.24	1.62
Jacareí Santo Antônio do Monte	SP	75.20	76.29	-1.09	Jaguari Teodoro	RS	76.10	74.47	1.63
Guarulhos	SP	75.37	76.46	-1.09	Sampaio	SP	76.93	75.29	1.64
Lajeado	RS	74.83	75.91	-1.08	Igrejinha	RS	76.38	74.74	1.64
		75.41	76.48	-1.07			77.53	75.89	1.64

Salto	SP	75.19	76.26	-1.07	São Caetano do Sul	SP	78.20	76.55	1.65
Tapera	RS	75.07	76.13	-1.06	Brasilândia de Minas	MG	75.87	74.21	1.66
Sete Lagoas	MG	75.37	76.42	-1.05	Pedro Afonso	TO	75.77	74.11	1.66
Jaguarão	RS	74.89	75.93	-1.04	Mondaí	SC	77.52	75.84	1.68
Palmitos	SC	75.13	76.15	-1.02	Tietê	SP	78.04	76.36	1.68
Videira	SC	76.42	77.43	-1.01	Guaraciaba	SC	77.28	75.59	1.69
Itatiba	SP	75.64	76.65	-1.01	Suzano Nova	SP	77.36	75.66	1.70
Valinhos	SP	76.01	77.02	-1.01	Bandeirantes	MT	75.51	73.81	1.70
Socorro	SP	74.66	75.67	-1.01	Alvorada	RS	77.41	75.71	1.70
Carazinho	RS	75.42	76.43	-1.01	Holambra	SP	77.66	75.95	1.71
Antônio Prado	RS	75.07	76.07	-1.00	Alto Rio Doce	MG	74.70	72.97	1.73
Batatais	SP	75.41	76.41	-1.00	Passos	MG	78.15	76.42	1.73
Praia Grande	SP	75.04	76.02	-0.98	Águas Lindas de Goiás	GO	75.85	74.12	1.73
São José do Rio Preto	SP	75.74	76.71	-0.97	Canápolis	MG	77.32	75.58	1.74
Casa Branca	SP	75.08	76.05	-0.97	Bom Jesus de Goiás	GO	76.59	74.84	1.75
Camanducaia	MG	75.22	76.19	-0.97	Rio Branco do Sul	PR	75.84	74.08	1.76
Otacílio Costa	SC	75.12	76.09	-0.97	PariqueraAçu	SP	76.70	74.94	1.76
Sapiranga	RS	74.93	75.88	-0.95	Nanuque	MG	76.02	74.24	1.78
Araçatuba	SP	75.46	76.41	-0.95	Cosmópolis	SP	77.57	75.78	1.79
Tupanciretã	RS	74.67	75.61	-0.94	Novo Gama	GO	76.06	74.27	1.79
Guaporé	RS	75.22	76.16	-0.94	Altônia	PR	75.88	74.09	1.79
Ibirama	SC	75.87	76.81	-0.94	Taquara	RS	77.82	76.01	1.81
Araranguá	SC	76.17	77.09	-0.92	Itabirinha	MG	75.30	73.48	1.82
Vacaria	RS	75.25	76.16	-0.91	Santiago	RS	77.93	76.06	1.87
Raposos	MG	74.90	75.80	-0.90	Orlândia	SP	77.89	76.02	1.87
Sabará	MG	74.99	75.89	-0.90	Tamarana	PR	74.76	72.86	1.90
Nova Prata	RS	75.36	76.25	-0.89	Andradina	SP	78.07	76.15	1.92
Palmeira das Missões	RS	74.83	75.71	-0.88	Vargem Grande Paulista	SP	78.02	76.10	1.92
Tupi Paulista	SP	75.08	75.96	-0.88	Triunfo	RS	77.35	75.43	1.92
Esteio	RS	75.57	76.44	-0.87	Jaú	SP	78.13	76.20	1.93
Governador Valadares	MG	75.06	75.93	-0.87	Rio Brilhante	MS	76.67	74.74	1.93
Osasco	SP	75.42	76.29	-0.87	Rio Verde de Mato Grosso	MS	76.11	74.17	1.94
Lençóis Paulista	SP	75.19	76.06	-0.87	Miguelópolis	SP	76.78	74.84	1.94
Cássia	MG	75.22	76.09	-0.87	Estância Velha	RS	78.23	76.28	1.95
Lagoa Formosa	MG	74.91	75.77	-0.86	Jaguariaíva	PR	76.54	74.59	1.95
Santa Fé do Sul	SP	75.41	76.27	-0.86	Cafelândia	SP	77.01	75.04	1.97
Sorocaba	SP	75.59	76.44	-0.85	Monte Santo de Minas	MG	77.46	75.46	2.00
Ribeirão Preto	SP	75.65	76.50	-0.85	São Sebastião	SP	77.51	75.51	2.00
Santos	SP	76.13	76.98	-0.85	Sobral	CE	74.93	72.92	2.01

Tanabi	SP	74.73	75.58	-0.85	Santa Rita do Passa Quatro	SP	78.22	76.17	2.05
Piracicaba Bom	SP	75.88	76.71	-0.83	Arraias	TO	74.73	72.67	2.06
Princípio Cachoeira Paulista	RS	75.18	76.01	-0.83	Tapiratiba	SP	77.18	75.12	2.06
Santana de Parnaíba	SP	75.19	76.01	-0.82	Piracanjuba Cruz	GO	77.21	75.09	2.12
Quatro Barras	PR	75.92	76.74	-0.82	Machado São Félix do	PR	75.67	73.53	2.14
Jaraguá do Sul	SC	74.87	75.68	-0.81	Araguaia Nova	MT	75.86	73.70	2.16
Bento Gonçalves	RS	76.92	77.73	-0.81	Petrópolis Lagoa Vermelha	RS	78.38	76.21	2.17
Caldas Novas	RS	75.52	76.31	-0.79	Vermeilha	RS	78.10	75.90	2.20
Londrina	GO	74.89	75.68	-0.79	Maracaju Cerro Grande do Sul	MS	77.35	75.13	2.22
Tapejara	PR	75.19	75.97	-0.78	Açucena São Gonçalo do Amarante	RS	76.28	74.04	2.24
Restinga Seca	RS	75.23	76.01	-0.78	Eldorado	MG	75.66	73.38	2.28
Macaé	RJ	74.70	75.47	-0.77	Piraquara	RN	74.72	72.42	2.30
Itaguara	RJ	74.66	75.43	-0.77	Bananal	SP	75.82	73.51	2.31
Pirapora	MG	75.14	75.90	-0.76	Espumoso Nova Laranjeiras	PR	77.15	74.80	2.35
Ponta Grossa	MG	74.65	75.41	-0.76	Pires do Rio Vila Bela da Santíssima Trindade	SP	77.29	74.94	2.35
Nhandeara Campo Limpo Paulista	PR	75.22	75.97	-0.75	Encruzilhada do Sul	RS	78.04	75.65	2.39
Goiânia	SP	75.08	75.81	-0.73	Parobé Santana do Paraíso	PR	74.96	72.56	2.40
Dores do Indaiá	SP	75.40	76.13	-0.73	Porto Xavier São Domingos	GO	77.79	75.31	2.48
Dracena	GO	75.28	76.00	-0.72		MT	75.57	73.07	2.50
Arapongas Presidente Venceslau Volta Redonda	MG	75.40	76.11	-0.71		RS	77.50	74.75	2.75
	SP	75.51	76.22	-0.71		RS	78.18	75.41	2.77
	PR	75.02	75.72	-0.70		MG	77.66	74.88	2.78
	SP	75.19	75.89	-0.70		RS	77.53	74.69	2.84
	RJ	74.98	75.67	-0.69		GO	74.77	71.85	2.92

eAppendix 6

Spatial distribution of the top 100 under and overachievers by tertile of life expectancy at birth.

eAppendix 7

We performed the same analyses for identifying the under and overachievers, but instead of using the SuperLearner to predict the expected value of life expectancy at birth, we used the individual algorithm with the highest predictive performance (random forests). The difference between the final results is small, and the results for random forests are presented below.

Healthcare variables	1st Tertile				2nd Tertile				3rd Tertile			
	Over	Under	Difference*	CI (95%)	Over	Under	Difference*	CI (95%)	Over	Under	Difference*	CI (95%)
Caesarean deliveries (%)	30,24	31,87	-2.97	-6.52;0.65	35,86	59,27	-22.01	-26.20;-17.94	56,76	62,44	-6.62	-10.71;-2.60
Family Health Strategy teams per 10,000 residents	3,50	3,10	0,10	-0.10;0.40	3,00	2,05	1,10	0.70;1.40	2,00	1,40	0,60	0.20;0.90
Hospital beds per 10,000 residents	12,50	14,25	-0.80	-4.10;1.20	15,35	16,40	-1.80	-5.40;0.90	15,75	14,35	0,80	-1.90;3.60
Life support equipment per 10,000 residents	0,80	1,20	-0.40	-0.80;-0.00	1,40	3,60	-2.00	-2.60;-1.50	2,90	3,00	0,10	-0.50;0.60
Low birth weight (%)	6,88	6,74	0,43	-0.14;0.99	6,96	8,05	-0.85	-1.54;-0.21	8,35	8,70	-0.58	-1.14;0.00
Mammographies per 100 women	2,00	2,50	-0.00	-1.00;0.00	3,00	12,50	-8.00	-10.00;-7.00	10,50	14,50	-4.00	-6.00;-2.00
Oral health coverage	88,18	84,38	-0.00	-2.06;0.16	81,23	32,37	34,72	23.28;44.69	42,88	16,88	14,29	3.87;25.00
Primary health coverage for poor residents (last year)	100,00	93,32	0,00	-0.00;0.15	87,28	70,39	15,00	6.85;21.99	64,85	51,17	13,89	5.09;21.96
Ultrasound machines per 10,000 live births	23,35	27,10	-0.00	-9.80;1.90	25,80	48,20	-21.10	-35.30;-6.20	37,05	37,55	0,00	-7.80;10.60
Vaccination coverage (%)	81,30	78,73	1,80	-1.44;4.82	78,08	75,12	2,41	0.16;4.77	75,61	72,99	2,04	0.25;3.75
X-ray machines per 10,000 residents	0,45	0,45	0,00	-0.00;0.01	0,60	1,00	-0.50	-0.70;-0.30	1,00	1,10	-0.00	-0.20;0.20

* For Mann-Whitney U tests, the difference in this case is the median of the difference between a sample from the underachievers and a sample from the overachievers, not the difference of the median between the two groups.

eAppendix 8

We also performed another sensitivity test, for which we ran the same analyses but excluding the first, second and third variables that most contributed to improve predictive performance, i.e. residing in Minas Gerais State, the illiteracy rate and proportion of households with automobile ownership. The MSE on each of these cases were 0.186, 0.183 and 0.178, respectively, which was similar to the one with all the variables included 0.175.

eAppendix 9

Available on <https://github.com/Hellengeremias/Tese-HellenGeremias>

6 CONCLUSÕES E CONSIDERAÇÕES FINAIS

Neste trabalho, técnicas supervisionadas de *machine learning* (ML) foram aplicadas a problemas de classificação e de regressão para prever respostas categóricas (binárias) e quantitativas de interesse para a saúde pública e a medicina. Portanto, a associação entre os preditores (variáveis independentes ou explicativas) e a resposta de interesse (variável dependente ou desfecho) com o propósito inferencial não foi explorado, embora esta seja etapa importante para o desenvolvimento de modelos preditivos mais precisos, pois contribui para a seleção de preditores informativos. No caso das pesquisas desenvolvidas na presente tese, essa seleção foi realizada de acordo com a literatura referente a cada objeto de pesquisa (óbito em idosos, qualidade de vida e sobrevivência de pacientes com câncer e expectativa de vida ao nascer).

De modo geral, todas as aplicações seguiram as etapas do roteiro para utilização de algoritmos de ML em análise preditiva, apresentado no capítulo 2 da tese. Na etapa de pré-processamento, variáveis quantitativas foram padronizadas e, as qualitativas, organizadas como variáveis indicadoras. Na etapa de aprendizado, para todos os algoritmos, exceto a regressão linear e a logística, hiperparâmetros foram otimizados por técnica de validação cruzada.

Entre os artigos que compõem a tese, os dois primeiros correspondem a problemas de classificação e o último a um problema de regressão. A principal diferença entre os mesmos está no tipo da variável resposta e na avaliação da performance preditiva dos modelos ajustados. Para problemas de regressão, medidas correspondentes à diferença entre o valor predito e o observado, como o erro quadrático médio (EQM), são as mais utilizadas. Para os de classificação, as referentes à habilidade discriminatória dos modelos, como a área abaixo da curva (AUC) ROC (*Receiver Operating Characteristic*), a sensibilidade e a especificidade, são as aplicadas com mais frequência em pesquisas de saúde. Adicionalmente, em problemas de classificação, a avaliação da calibração da probabilidade predita pelo modelo ajustado também é interessante, em especial, se houver a intenção de implementá-lo na prática. São exemplos de medidas de calibração a curva de calibração, a distribuição da probabilidade predita e o teste de Hosmer e Lemeshow que avalia diferenças entre a frequência de resposta predita e observada.

No artigo 1, dados do estudo “Saúde, Bem-estar e Envelhecimento” – estudo SABE – foram utilizados para prever óbito em até 5 anos para idosos, com o objetivo de exemplificar as etapas envolvidas na aplicação de ML em análises preditivas em saúde. A resposta de

interesse apresentou frequência de 15% e 37 variáveis foram utilizados como preditores (características socioeconômicas, demográficas, comportamentais e relacionadas ao estado de saúde, morbidades autorreferidas, condições de mobilidade e estado funcional). Todos os algoritmos avaliados (regressão logística, regressão logística penalizada, redes neurais, *gradient boosted trees* e *random forest*) apresentaram AUC ROC superior a 0,70 nos dados de teste, e os melhores resultados foram observados para os modelos resultantes da aplicação de redes neurais, regressão logística com penalização *lasso* e regressão logística simples. Em relação à calibração da probabilidade predita, foram avaliadas sua distribuição e curvas de calibração para os modelos com melhor discriminação. Os resultados evidenciaram calibração insatisfatória, em especial para o modelo de redes neurais.

A calibração inadequada da probabilidade predita pode ser decorrente da representação limitada, na fase de construção do modelo, da subpopulação de indivíduos com resposta de interesse presente (MATHENY; OHNO-MACHADO, 2014). No entanto, na prática, as probabilidades preditas são agregadas em níveis de risco para ajudar a priorizar indivíduos em decisões relacionadas a ações de tratamento ou prevenção e, mesmo que um modelo não apresente boa calibração, o mesmo pode ainda 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 (MEURER; TOLLES, 2017).

Para o Artigo 2, dados do estudo “QALY – Modelo para prever sobrevida ajustada para a qualidade em pacientes com câncer admitidos em unidade de terapia intensiva” foram utilizados para prever o risco de tempo de vida com qualidade de até 30 dias em pacientes com câncer admitidos em UTI. Essa resposta apresentou frequência de 44,7% na amostra estudada e 27 características coletadas na admissão do paciente à UTI foram utilizadas como preditores (perfil sociodemográfico, hábitos de vida, presença de morbidades, história de saúde e complicações relacionadas ao câncer).

Na etapa de treinamento dos algoritmos selecionados (regressão logística, regressão logística penalizada, redes neurais, *gradient boosted trees* e *random forest*) utilizou-se validação cruzada adicional (externa), caracterizando a técnica de validação cruzada aninhada, para comparar a performance preditiva dos diferentes algoritmos. Observou-se AUC ROC superior a 0,80 para todos os algoritmos, exceto a árvore de classificação simples, evidenciando habilidade de discriminação satisfatória. Em relação às medidas de calibração, avaliada para os modelos com bom poder discriminatório, os gráficos demonstraram boa calibração, embora o

teste de Hosmer-Lemeshow tenha indicado calibração adequada apenas para os modelos de redes neurais, *gradient boosted trees* e *random forest*.

De modo geral, os resultados dos Artigos 1 e 2 reforçam a importância da medida AUC ROC na comparação de dois ou mais modelos com diferentes preditores, hiperparâmetros ou mesmo classificadores decorrentes de algoritmos diferentes e, em particular, para os bancos de dados analisados, evidenciam que modelos menos complexos, como a regressão logística (com e sem penalização) e aqueles mais flexíveis (redes neurais, *gradient boosted trees* e *random forest*) apresentam desempenho semelhante em discriminar indivíduos com e sem a resposta de interesse.

É importante destacar que mesmo um modelo preditivo que apresente boa discriminação e que seja bem calibrado pode não se traduzir em melhores cuidados, pois uma predição precisa não diz o que fazer para modificar o desfecho sob análise, ou nem mesmo é possível assumir que o desfecho predito possa ser modificado (CHEN; ASCH, 2017). Adicionalmente, outras medidas podem ser utilizadas para avaliar modelos preditivos em circunstâncias específicas. Por exemplo, medidas de utilidade clínica, como as curvas de decisão, contribuem para avaliar a aplicabilidade de modelos preditivos, trazendo informações que vão além das medidas tradicionais de calibração e discriminação (VICKERS; ELKIN, 2008).

O Artigo 3 objetivou prever a expectativa de vida ao nascer de municípios brasileiros, utilizando como preditores apenas características não controladas diretamente por gestores de saúde (socioeconômicas e demográficas), e, posteriormente, comparar as características de saúde de municípios com expectativa de vida ao nascer superior à esperada (*overachievers*) com as daqueles com expectativa de vida inferior à esperada (*underachievers*), visando identificar práticas de saúde pública bem-sucedidas.

Na etapa de predição, o algoritmo combinatório (*ensemble*) *super learner* foi utilizado para ajustar o modelo preditivo. Sua principal vantagem está na possibilidade de incluir diversos algoritmos como candidatos à combinação das predições, e a garantia teórica de que seu desempenho é, ao menos, tão bom quanto a escolha ótima (desconhecida) entre os algoritmos candidatos (VAN DER LAAN; POLLEY; HUBBARD, 2007), oferecendo, dessa forma, uma alternativa flexível, não só quando comparado a métodos convencionais, como o modelo de regressão linear, mas também a alternativas não paramétricas (PIRRACCHIO et al., 2015).

De modo geral, não foram observadas diferenças substanciais no desempenho preditivo dos algoritmos analisados (*super learner* e algoritmos candidatos) e, com exceção das árvores de regressão, todos apresentaram performance preditiva satisfatória (EQM de 0,17 para o *super*

learner e entre 0,18 e 0,20 para os demais algoritmos da coleção), assinalando que é possível prever com boa precisão a expectativa de vida de municípios brasileiros utilizando apenas características socioeconômicas e demográficas. Nesse cenário, o presente estudo empregou a seguinte estratégia para identificar iniciativas bem-sucedidas de saúde: “prever para identificar *overachievers*”. Espera-se que a mesma seja aplicada em outras pesquisas da área da saúde à medida que a disponibilidade de dados e a utilização de ML continue a se expandir nessa área de conhecimento.

Na etapa de comparação das características de saúde, o presente estudo mostra que avaliar o valor esperado de um desfecho de saúde pode trazer importantes *insights* sobre boas práticas de saúde pública. Os *overachievers* apresentaram melhor desempenho em características relacionadas à atenção primária à saúde e os *underachievers*, por sua vez, em características relacionadas à atenção secundária, resultados que sugerem a importância da alocação de recursos em ações de saúde desenvolvidas na atenção primária, especialmente em países em desenvolvimento. Nesse cenário, sabe-se que os serviços de saúde enfrentam crescentes pressões para melhorar a qualidade do atendimento e, ao mesmo tempo, reduzir custos. Portanto, a análise dos dados gerados por esses serviços pode melhorar a utilização de recursos, a gestão de ações de saúde e, conseqüentemente, eficiência dos cuidados de saúde (OZAYDIN; HARDIN; CHHIENG, 2016).

O emprego de métodos estatísticos e de ML para produzir conhecimento a partir de dados tornou-se componente integral da pesquisa em saúde. As técnicas para a construção e avaliação de modelos estão em constante evolução, e há poucas justificativas teóricas para se preferir um algoritmo de aprendizagem em lugar de outro (MATHENY; OHNO-MACHADO, 2014). Em especial, modelos decorrentes da aplicação de regressão logística são frequentemente utilizados na área da saúde e oferecem um grau relativamente alto de transparência, pois são expressos em fórmulas que permitem compreender a contribuição de cada preditor no cálculo do risco predito. No entanto, ao contrário de métodos mais flexíveis, modelos de regressão convencionais não consideram automaticamente interações entre preditores ou relações não-lineares entre um preditor e a resposta de interesse, além de serem sensíveis à presença de preditores irrelevantes ou a maior quantidade de preditores. Por outro lado, métodos flexíveis requerem suposições mínimas com relação aos dados e a maioria realiza seleção de variáveis, sobretudo em cenários que apresentam um número relativamente grande de preditores, e compara a capacidade preditiva entre variáveis, ajustando para possíveis interações (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

Embora nem todos os algoritmos de ML aplicados a pesquisas em saúde tenham sido utilizados para modelar os problemas de classificação e de regressão apresentados no presente trabalho, os resultados comparativos dos algoritmos aplicados evidenciam que não há um único algoritmo que tenha o melhor desempenho em todas as situações e que, não necessariamente, quanto mais complexo for o modelo preditivo, melhor será a sua performance preditiva. Haverá casos em que modelos de regressão convencionais apresentarão desempenho tão bom quanto algoritmos mais flexíveis. No entanto, a menos que um conjunto de algoritmos seja avaliado, não é possível afirmar *a priori* que um método convencional irá atingir desempenho preditivo comparável ao de um algoritmo mais flexível.

Diferenças na performance de modelos preditivos podem ser atribuídas não só às técnicas utilizadas para sua construção e avaliação, mas também às características dos dados, como o número de observações disponíveis, sobretudo de observações com desfecho presente no caso de problemas de classificação, ou de variáveis candidatas a preditores. Nesse sentido, a expansão do conjunto de preditores pode resultar em modelos com desempenho preditivo ainda melhor. E, embora o aumento no número de preditores estimule a aplicação de abordagens não paramétricas e de algoritmos do tipo *ensemble*, como o *super learner*, o *gradient boosted trees* e o *random forest*, tal situação irá exigir um equilíbrio entre complexidade e desempenho preditivo para que os modelos sejam aplicáveis na prática clínica e de saúde pública (PIRRACCHIO et al., 2015).

Em muitos casos, dentro de uma margem de variabilidade, alguns modelos serão equivalentes em termos de performance, como no caso das comparações apresentadas no presente trabalho, em que os dados são estruturados, o número de observações e de variáveis predictoras não é grande e cujas variáveis selecionadas para a predição da resposta de interesse são conhecidas como potenciais preditores (fatores de risco) para essa resposta. Portanto, em cenários como esses, é possível ponderar complexidade computacional, facilidade de implementar a função de predição estimada, interpretabilidade entre outras características dos algoritmos, para a escolha do modelo final (KUHN; JOHNSON, 2013).

A análise preditiva está presente em muitas atividades diárias, em especial nas relacionadas a sistemas de recomendação, como por exemplo, de mensagens a serem apresentadas em aplicativos de rede social, de resultados de pesquisas realizadas na *web*, de filmes e músicas em aplicativos que utilizam tecnologia *streaming*, de respostas em atendimentos virtuais, entre outras. Em saúde, dados são rotineiramente coletados e disseminados por organizações e pesquisadores de serviços públicos e privados. De modo geral,

registros de estatísticas vitais e de admissões hospitalares são os mais frequentemente analisados, pois contribuem com a descrição dos perfis de morbidade e mortalidade da população. Porém, mais recentemente, registros médicos eletrônicos, dados genéticos, sistemas de posicionamento geográfico, dados de imagens, entre outros, passaram a ser produzidos por esses serviços. No entanto, como esses dados estão armazenados em diferentes formatos e bancos de dados, permanece o desafio da sua integração, visando criar uma base única com potencial para aprimorar sistemas de aprendizagem (WYBER et al., 2015).

Por fim, ainda que a disponibilidade de um ambiente computacional no processo de atendimento ao usuário não seja uma realidade em muitos dos serviços de saúde brasileiros, avanços tecnológicos e movimentos recentes de aproximação desses serviços com a tecnologia da informação e comunicação, suscitam a possibilidade de implementação de modelos preditivos como ferramenta *web*, como descrito em pesquisas recentes (OLIVERA et al., 2017; PIRRACCHIO et al., 2015), o que pode representar o caminho para que profissionais de saúde os utilizem na prática e para que pesquisadores os validem externamente.

Espera-se que este estudo tenha contribuído para a compreensão sobre possibilidades de aplicações de ML em análises preditivas de interesse para a saúde pública e a medicina e que tenha possibilitado uma reflexão sobre seus respectivos desafios e as perspectivas no contexto de pesquisas em saúde.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRESTI, A. **Categorical data analysis**. 3. ed. Hoboken, New Jersey: Wiley-Interscience, 2013.

ALMEIDA FILHO, N. DE; ROUQUAYROL, M. Z. **Introdução à epidemiologia**. 4. ed. Rio de Janeiro: Guanabara- Koogan, 2006.

ASSUNÇÃO, F. **Estratégias para tratamento de variáveis com dados faltantes durante o desenvolvimento de modelos preditivos**. 2012. 74f. Dissertação (Mestrado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2012.

AUSTIN, P. C. et al. Regression trees for predicting mortality in patients with cardiovascular disease: What improvement is achieved by using ensemble-based methods? **Biometrical Journal**, v. 54, n. 5, p. 657–673, 2012.

BARATA, R. B. Epidemiologia e políticas públicas. **Rev Bras Epidemiol**, v. 16, n. 1, p. 3–17, 2013.

BARNATO, A. E.; ANGUS, D. C. Value and role of intensive care unit outcome prediction models in end-of-life decision making. **Critical Care Clinics**, v. 20, n. 3, p. 345–362, 2004.

BHARDWAJ, R.; NAMBIAR, A. R.; DUTTA, D. A Study of Machine Learning in Healthcare. **Proceedings of 41st Annual Computer Software and Applications Conference (COMPSAC)**, v. 2, p. 236–241, jul. 2017.

BREIMAN, L. Stacked regressions. **Machine Learning**, v. 24, n. 1, p. 49–64, 1996.

CARUANA, R. et al. Intelligible Models for HealthCare : Predicting Pneumonia Risk and Hospital 30-day Readmission. **Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15**, p. 1721–1730, 2015.

CAVALCANTI, A. B. **QALY - Modelo para prever sobrevida ajustada para a qualidade em pacientes com câncer admitidos em unidade de terapia intensiva**. Instituto do Câncer do Estado de São Paulo, 2009.

CESAR, G. L. C. O “enfoque de risco” em saúde pública. In: BARRETO, M. L. et al. (Eds.). **Epidemiologia, serviços e tecnologias em saúde**. Rio de Janeiro: Editora FIOCRUZ, 1998. p. 79–92.

CHEN, J. H.; ASCH, S. M. Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. **New England Journal of Medicine**, v. 376, n. 26, p. 2507–2509, jun. 2017.

CORONA, L. P.; DUARTE, Y. A. DE O.; LEBRÃO, M. L. Prevalence of anemia and associated factors in older adults: Evidence from the SABE Study. **Revista de Saude Publica**, v. 48, n. 5, p. 723–731, 2014.

D’AGOSTINO, R. B. et al. Validation of the Framingham Coronary Heart Disease prediction

score - Results of a Multiple Ethnic Groups Investigation. **JAMA: The Journal of the American Medical Association**, v. 286, n. 2, p. 180–187, 2001.

DAI, W. et al. Prediction of hospitalization due to heart diseases by supervised learning methods. **International Journal of Medical Informatics**, v. 84, n. 3, p. 189–197, mar. 2015.

DOLAN, P. Modeling valuations for EuroQol health states. **Medical care**, v. 35, n. 11, p. 1095–108, nov. 1997.

ELFIKY, A. A. et al. A machine learning approach to predicting short-term mortality risk in patients starting chemotherapy. **bioRxiv**, p. 1–29, 2017.

ESTEVA, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. **Nature**, v. 542, n. 7639, p. 115–118, 2017.

EXPERT PANEL ON DETECTION, EVALUATION, AND T. OF H. B. C. IN A. 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: The Journal of the American Medical Association**, v. 285, n. 19, p. 2486–2497, 2001.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge: MIT Press, 2016.

GORELICK, M. H. Bias arising from missing data in predictive models. **Journal of Clinical Epidemiology**, v. 59, n. 10, p. 1115–1123, 2006.

GULSHAN, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. **JAMA - Journal of the American Medical Association**, v. 316, n. 22, p. 2402–2410, 2016.

GUSMAO-FLORES, D. et al. The validity and reliability of the Portuguese versions of three tools used to diagnose delirium in critically ill patients. **Clinics (São Paulo, Brazil)**, v. 66, n. 11, p. 1917–22, 2011.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference and Prediction**. New York: Springer New York, 2008.

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

ICAZA, M. G.; ALBALA, C. **Minimental State Examination (MMSE) del estudio de demencia en Chile: análisis estadístico**. Washington: Organizacioin Panamericana de la Salud, 1999.

IZBICKI, R.; SANTOS, T. M. DOS. **Machine Learning sob a ótica estatística - Uma abordagem preditivista para a estatística com exemplos em R**, 2018. Disponível em: <<https://rizbicki.files.wordpress.com/2016/09/main2.pdf>>

JAMEI, M. et al. Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. **PLoS ONE**, v. 12, n. 7, p. 1–14, 2017.

JAMES, G. et al. **An Introduction to Statistical Learning**. New York: Springer New York, 2014.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 6. ed. New Jersey: Pearson Education, Inc., 2007.

KANNEL, W. B.; MCGEE, D.; GORDON, T. A general cardiovascular risk profile: The Framingham study. **The American Journal of Cardiology**, v. 38, n. 1, p. 46–51, 1976.

KEEGAN, M. T.; SOARES, M. What every intensivist should know about prognostic scoring systems and risk-adjusted mortality. **Revista Brasileira de Terapia Intensiva**, v. 28, n. 3, p. 264–269, 2016.

KESSLER, R. C. 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**, v. 13, n. 3, p. 265–274, out. 2014.

KNAUS, W. A. et al. The APACHE III prognostic system: Risk prediction of hospital mortality for critically III hospitalized adults. **Chest**, v. 100, n. 6, p. 1619–1636, 1991.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. New York: Springer, 2013.

KUTNER, M. H.; NACHTSHEIM, C.; NETER, J. **Applied linear regression models**. 4. ed. Chicago: McGraw Hill/Irwin Series, 2004.

LE GALL, J.-R.; LEMESHOW, S.; SAULNIER, F. A New Simplified Acute Physiology Score (SAPS II) Based on a European / North American Multicenter Study. **JAMA : the journal of the American Medical Association**, v. 270, n. 24, p. 2957–2963, 1993.

LEBRÃO, M.; DUARTE, Y. Desafios de um estudo longitudinal: o Projeto SABE. **Saúde Coletiva**, v. 5, n. 24, p. 166–67, 2008.

LEBRÃO, M. L.; LAURENTI, R. Saúde, bem-estar e envelhecimento : o estudo SABE no Município de São Paulo. **Rev Bras Epidemiol**, v. 8, n. 2, p. 127–141, 2005.

LEMESHOW, S. et al. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. **JAMA : the journal of the American Medical Association**, v. 270, n. 20, p. 2478–2486, 1993.

LITTLE, R. J. A.; RUBIN, D. B. **Statistical Analysis with Missing Data**. 2. ed. New Jersey: John Wiley & Sons, Inc., 2002.

LUIZ, O. D. C.; COHN, A. Sociedade de risco e risco epidemiológico Risk society and epidemiological risk. **Medicina**, v. 22, n. 11, p. 2339–2348, 2006.

MATHENY, M. E.; OHNO-MACHADO, L. Generation of Knowledge for Clinical Decision Support. In: GREENES, R. A. (Ed.). . **Clinical Decision Support: Clinical Decision Support The Road to Broad Adoption**. 2. ed. United States: Elsevier Inc., 2014. p. 309–337.

MCCULLAGH, P. (PETER); NELDER, J. A. **Generalized linear models**. Boca Raton: Chapman and Hall, 1989.

MENA, L. J. et al. Machine learning approach to extract diagnostic and prognostic thresholds: Application in prognosis of cardiovascular mortality. **Computational and Mathematical Methods in Medicine**, v. 2012, 2012.

METNITZ, P. G. H. et al. SAPS 3-From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description. **Intensive Care Medicine**, v. 31, n. 10, p. 1336–1344, 2005.

MEURER, W. J.; TOLLES, J. Logistic Regression Diagnostics. **JAMA**, v. 317, n. 10, p. 1068–9, 2017.

OBERMEYER, Z.; EMANUEL, E. J. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. **New England Journal of Medicine**, v. 375, n. 13, p. 1216–1219, 2016.

OKEN, M. M. et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. **American journal of clinical oncology**, v. 5, n. 6, p. 649–55, 1982.

OLIVERA, A. R. et al. Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes - ELSA-Brasil: accuracy study. **Sao Paulo Medical Journal**, v. 135, n. 3, p. 234–246, 2017.

OZAYDIN, B.; HARDIN, J. M.; CHHIENG, D. C. Data Mining and Clinical Decision Support Systems. In: BERNER, E. S. (Ed.). . **Clinical Decision Support Systems**. Health Informatics. Cham: Springer International Publishing, 2016. p. 45–68.

PAN, I. et al. Machine learning for social services: A study of prenatal case management in Illinois. **American Journal of Public Health**, v. 107, n. 6, p. 938–944, 2017.

PENCINA, M. J.; D'AGOSTINO, R. B. Evaluating Discrimination of Risk Prediction Models. **Jama**, v. 314, n. 10, p. 1063–4, 2015.

PETERSEN, M. L. et al. Super Learner Analysis of Electronic Adherence Data Improves Viral Prediction and May Provide Strategies for Selective HIV RNA Monitoring. **JAIDS Journal of Acquired Immune Deficiency Syndromes**, v. 69, n. 1, p. 109–118, 2015.

PIRRACCHIO, R. et al. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): A population-based study. **The Lancet Respiratory Medicine**, v. 3, n. 1, p. 42–52, 2015.

POLLEY, E. C.; LAAN, M. J. VAN DER. **Super Learner In Prediction**. Berkeley: [s.n.]. Disponível em: <<http://biostats.bepress.com/ucbbiostat/paper266>>.

POLLEY, E. C.; ROSE, S.; VAN DER LAAN, M. J. Super Learning. In: VAN DER LAAN, M. J.; ROSE, S. (Eds.). . **Target Learning: Causal Inference for Observational and Experimental Data**. New York: Springer, 2011. p. 43–66.

QUINLAN, J. R. LEARNING WITH CONTINUOUS CLASSES. **Machine Learning**, v. 92, p. 343–348, 1992.

RASCHKA, S. **Python Machine Learning**. 2. ed. Birmingham: Packt Publishing Ltd, 2017.

ROSE, S. Mortality risk score prediction in an elderly population using machine learning.

American Journal of Epidemiology, v. 177, n. 5, p. 443–452, 2013.

ROUQUAYROL, M. Z.; GOLDBAUM, M. Epidemiologia, história natural e prevenção de doenças. In: ROUQUAYROL, M. Z.; ALMEIDA FILHO, N. DE (Eds.). . **Epidemiologia & Saúde**. 6. ed. Rio de Janeiro: Medsi, 2003. p. 15–30.

SCHMIDT, M. I.; DUNCAN, B. B. Epidemiologia clínica e medicina baseada em evidências. In: ROUQUAYROL, M. Z.; ALMEIDA FILHO, N. DE. (Eds.). . **Epidemiologia & Saúde**. 6. ed. [s.l.] Medsi, 2003. p. 193–227.

SCHONBERG, M. A. et al. Index to predict 5-year mortality of community-dwelling adults aged 65 and older using data from the national health interview survey. **Journal of General Internal Medicine**, v. 24, n. 10, p. 1115–1122, 2009.

SEMENZA, J. C.; SUK, J.; MANISSERO, D. Intervening on high-risk or vulnerable populations? **American Journal of Public Health**, v. 98, n. 8, p. 1351–1352, 2008.

SILVA, N. N. DA. Aspectos metodológicos: processo de amostragem. In: LEBRÃO, M. L.; DUARTE, Y. A. DE O. (Eds.). . **O Projeto SABE no Município de São Paulo: uma abordagem inicial**. OPAS/MS ed. Brasília: [s.n.]. p. 47–57.

STEYERBERG, E. W. **Clinical Prediction Models**. New York, NY: Springer New York, 2009. v. 53

SUEMOTO, C. K. 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. **The Journals of Gerontology Series A: Biological Sciences and Medical Sciences**, v. 00, n. 00, p. glw166, 2016.

SZENDE, A.; OPPE, M.; DEVLIN, N. **EQ-5D Value Sets: Inventory, Comparative Review and User Guide**. [s.l.] Springer, 2007.

SZKLO, M.; NIETO, J. F. **Epidemiology: beyond the basics**. Burlington: Jones & Bartlett Learning, 2014.

THE CRITERIA COMMITTEE OF THE NEW YORK HEART. **Nomenclature and Criteria for Diagnosis of Diseases of the Heart and Great Vessels**. 9. ed. Boston: Little, Brown & Co, 1994.

VAN DER HEIJDEN, G. J. M. G. et al. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. **Journal of Clinical Epidemiology**, v. 59, n. 10, p. 1102–1109, 2006.

VAN DER LAAN, M. J.; POLLEY, E. C.; HUBBARD, A. E. Super Learner. **Statistical Applications in Genetics and Molecular Biology**, v. 6, n. 1, p. Article 25, 2007.

VARMA, S.; SIMON, R. Bias in error estimation when using cross-validation for model selection. **BMC Bioinformatics**, v. 7, n. 1, p. 91, 2006.

VICKERS, A. J.; ELKIN, E. B. Decision curve analysis: a novel method for evaluating prediction models. **Medical Decision Making**, v. 26, n. 6, p. 565–574, 2008.

VINCENT, J. L. et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. **Intensive Care Med**, v. 22, n. 7, p. 707–710, 1996.

WENG, S. F. et al. Can Machine-learning improve cardiovascular risk prediction using routine clinical data? **PLoS ONE**, v. 12, n. 4, p. 1–14, 2017.

WOLPERT, D. H. Stacked Generalization (Stacking). **Neural Networks**, v. 5, n. 505, p. 241–259, 1992.

WU, X.; KUMAR, V. **The Top Ten Algorithms in Data Mining**. Boca Raton: Chapman and Hall, 2009.

WYBER, R. et al. Big data in global health: Improving health in low- and middle-income countries . **Bulletin of the World Health Organization**, v. 93, n. 3, p. 203–208, 2015.

YOURMAN, L. C. et al. Prognostic Indices for Older Adults: A Systematic Review Lindsey. **JAMA**, v. 307, n. 2, p. 182, 11 jan. 2012.

ZHANG, P. Multiple imputation: Theory and method. **International Statistical Review**, v. 71, n. 3, p. 581–592, 2003.

APÊNDICES

Apêndice A – Dicionário de variáveis socioeconômicas e demográficas utilizadas para a predição da expectativa de vida ao nascer de municípios brasileiros.

Nome da variável	Descrição	Notas técnicas	Fonte	Número da tabela Sidra / Seleção de variáveis DATASUS
Expectativa de vida	Número médio de anos que as pessoas deverão viver a partir do nascimento, se permanecerem constantes o nível e o padrão de mortalidade por idade, prevalentes no ano do Censo.	Dados calculados a partir dos microdados do Censo demográfico de 2010 – IBGE.	Atlas do Desenvolvimento Humano do Brasil	(Link: http://www.atlasbrasil.org.br/2013/ ; Seção: Download)
População Residente	Total de residentes no município.	1 - Dados do Universo.	IBGE - Censo Demográfico 2010	Tabela 1378
Densidade domiciliar	Número de moradores por domicílio.	Número de habitantes/ Total de domicílios.	Sidra, IBGE - Censo Demográfico 2010	Tabela 2009 e Tabela 3504
Densidade populacional	Densidade populacional segundo município (Habitantes/Km ²).	Número de habitantes / Km ² do local de residência.	IBGE - Censo Demográfico 2010	Link: http://www.censo2010.ibge.gov.br/sinopse/index.php?uf=12&dados=21
Idade mediana	Mediana= limite inferior + (n/2 - frequência acumulada anterior) / frequência absoluta *amplitude da classe.		Sidra, IBGE - Censo Demográfico 2010	Tabela 200 - População residente por sexo, situação e grupos de idade - Amostra - Características Gerais da População
Nascimentos per capita	Número de nascidos vivos / População residente.		DATASUS e Sidra	DATASUS - nascidos vivos segundo município e Tabela 200 (ano de referência: 2008-2012)
Razão de dependência	Razão de dependência segundo município.	População de menor de 15 e maior de 60 anos / População entre 15 e 59 anos.		Tabela 200 - População residente por sexo, situação e grupos de idade - Amostra - Características Gerais da População

PIB per capita		1. Valores do PIB per capita estão em reais correntes. Os demais valores estão em milhares de reais correntes. 2. Os valores do PIB per capita podem divergir do publicado em outras fontes, caso haja diferença nos valores estimados da população. 3. Série histórica construída com a referência 2010. Para mais detalhes sobre a metodologia de cálculo, consulte o sítio do IBGE.	IBGE, em parceria com os Órgãos Estaduais de Estatística, Secretarias Estaduais de Governo e Superintendência da Zona Franca de Manaus	Seleção das variáveis= Linha: Município; Coluna: Não ativa; Conteúdo: PIB per capita; Período: 2010. (Link: http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/cnv/pibmunbbr.def)
Índice de Gini	Índice de Gini da renda domiciliar per capita segundo município.	Os dados foram processados pelo Instituto de Pesquisas Econômicas e Aplicadas (IPEA), de acordo com os critérios do indicador de Índice de Gini da renda domiciliar per capita - B.9, dos Indicadores e Dados Básicos da Rede Interagencial de Informações para a Saúde (RIPSA).	IBGE - Censo Demográfico 2010	(Link: http://www2.datasus.gov.br/DATASUS/index.php?area=0206&id=8065372&VObj=http://tabnet.datasus.gov.br/cgi/ibge/censo/cnv/gini); Abrangência geográfica: Brasil por município
Renda média	Valor do rendimento nominal médio mensal das pessoas de 10 anos ou mais de idade (em Reais).	1 - Dados da Amostra.	IBGE - Censo Demográfico 2010	Tabela 3548
Percentual de crianças em situação domiciliar de baixa renda	Percentual de crianças (até 14 anos) em situação domiciliar de baixa renda (até meio salário mínimo).	O salário mínimo do último ano para o qual a série está sendo calculada torna-se a referência para toda a série. Esse valor é corrigido para todos com base no INPC de julho de 2010, alterando o valor da linha de pobreza e consequentemente a proporção de pobres. Nesta tabela, o valor de referência, salário mínimo de 2010, é de R\$ 510,00.	IBGE - Censos Demográficos 2010	Seleção das variáveis= Linha: Município; Coluna: Não ativa; Conteúdo: % crianças renda dom < 1/2 SM; Período: 2010 (Link: http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/censo/cnv/crianpobr.br.def)
Taxa de alfabetização	Percentual da população de 10 anos ou mais de idade alfabetizadas.	1 - Dados do Universo.	IBGE - Censo Demográfico 2010	Tabela 1383
Taxa de desemprego	Pessoas com 16 anos ou mais economicamente ativas e desocupadas/ Total de pessoas com 16 anos ou mais *100.	1 - Dados da Amostra.	IBGE - Censo Demográfico 2010	Tabela 3573

Taxa de trabalho infantil	Percentual da população de 10 a 15 anos ocupada.		DATASUS	DATASUS>Demográficas e Socioeconômicas>Trabalho e Renda>Taxa de trabalho infantil>Brasil por município (Link: http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/censo/cnv/trabinfbr.def). Seleção das variáveis: Linha-Município, Coluna-não ativa, Conteúdo-Taxa de trabalho infantil, Período-2010.
Percentual de aposentados	Percentual de pessoas com 10 anos ou mais de idade na condição de aposentado ou pensionista de instituto de previdência social.	1 - Dados da Amostra.	IBGE - Censo Demográfico 2010	Tabela 3580
Percentual de brancos	(Número de residentes que se declararam brancos/ Total da população) *100.	1 - Dados do Universo.	IBGE - Censo Demográfico 2010	Tabela 3175/ Tabela 1378
Percentual de casados	Percentual de casados na população residente segundo município.	(População de casados / População total) *100.	Sidra, IBGE - Censo Demográfico 2010	Tabela 1206 - Pessoas de 10 anos ou mais de idade, por estado conjugal e estado civil, segundo o sexo e a religião
Percentual de idosos	Percentual de idosos na população residente segundo município.	(População com 60 anos ou mais / População total) *100.	Sidra, IBGE - Censo Demográfico 2010	Tabela 200 - População residente por sexo, situação e grupos de idade - Amostra - Características Gerais da População
Percentual de mulheres	Percentual de mulheres na população residente segundo município.	(População de mulheres / População total) *100.	Sidra, IBGE - Censo Demográfico 2010	Tabela 200 - População residente por sexo, situação e grupos de idade - Amostra - Características Gerais da População
Percentual de evangélicos	Percentual de evangélicos na população residente segundo município.	(População de evangélicos / População total) *100.	Sidra, IBGE - Censo Demográfico 2010	Tabela 1489 - População residente, por cor ou raça, segundo o sexo a

				religião - Resultados Gerais da Amostra
Percentual de pessoas com alguma deficiência	Percentual de pessoas com pelo menos uma deficiência, investigadas na população residente segundo município.	(População de deficientes / População total) *100.	Sidra, IBGE - Censo Demográfico 2010	Tabela 3428 - População residente, por existência ou não de pelo menos uma das deficiências investigadas e frequência à escola ou creche, segundo o sexo e os grupos de idade - Amostra - Características Gerais da População
Percentual de domicílios com eletricidade	Percentual de domicílios com eletricidade segundo município.	(Número de domicílio com eletricidade / Total de domicílios) *100.	Sidra, IBGE - Censo Demográfico 2010	Tabela 3503 - Domicílios particulares permanentes, por tipo de material das paredes externas, segundo o número de cômodos, a existência de água canalizada e forma de abastecimento de água, o destino do lixo, a existência de energia elétrica e a existência de alguns bens duráveis
Percentual de domicílios com automóvel	Percentual de domicílios com automóvel para uso particular segundo município.	(Número de domicílio de domicílios com carro / Total de domicílios) *100.	Sidra, IBGE - Censo Demográfico 2010	Tabela 2249 - Domicílios particulares permanentes, por alguns bens duráveis existentes no domicílio - Resultados Gerais da Amostra
Percentual de domicílios com computador	Percentual de domicílios com microcomputador segundo município.	(Número de domicílios com microcomputador / Total de domicílios) *100.	Sidra, IBGE - Censo Demográfico 2010	Tabela 2249 - Domicílios particulares permanentes, por alguns bens duráveis existentes no domicílio - Resultados Gerais da Amostra
Percentual de domicílios com geladeira	Percentual de domicílios com geladeira segundo município.	(Número de domicílios com geladeira / Total de domicílios) *100.	Sidra, IBGE - Censo Demográfico 2010	Tabela 2249 - Domicílios particulares permanentes, por alguns bens duráveis existentes no domicílio - Resultados Gerais da Amostra
Proporção de residentes em áreas arborizadas	Número de residentes em áreas arborizadas (em áreas urbanas com ordenamento regular) / Total da população residente no município.	Foi pesquisado se na face ou na sua face confrontante ou no canteiro central, existia arborização, ou seja, existia árvore ao longo da	IBGE - Censo Demográfico 2010	Tabela 3357/ Tabela 1378

		calçada/passeio e/ou em canteiro que divide pistas de um mesmo logradouro, mesmo que apenas em parte. Considerou-se também a arborização quando existente em logradouros sem pavimentação e/ou sem calçada/passeio. 1 - Dados do Universo. 2 - Inclui somente os domicílios em áreas com ordenamento urbano regular (informação sobre arborização).		
Proporção de residentes em áreas pavimentadas	Pessoas residentes em áreas urbanas pavimentadas com ordenamento regular/ total da população residente no município.	A face percorrida foi considerada como tendo pavimentação, quando existisse cobertura da via pública com asfalto, cimento, paralelepípedos, pedras, etc. 1 - Dados do Universo. 2 - Inclui somente os domicílios em áreas com ordenamento urbano regular.	IBGE - Censo Demográfico 2010	Tabela 3357/ Tabela 1378
Percentual de cobertura do Programa Bolsa Família	Percentual de cobertura do programa Bolsa Família com base nas famílias totalmente acompanhadas.		DATASUS	Link: http://bolsafamilia.datasus.gov.br/w3c/consol_regiao_cobertura_bfa.asp?gru=5&vigencia=19&vigatual=N&uf=TD&regional=00&regiaosaud e=00&cob=1&brsm=1
Percentual de residentes com ensino médio completo	Percentual de pessoas de 25 anos ou mais de idade, residentes em domicílios particulares, com o ensino médio completo.	1 - Dados da Amostra.	IBGE - Censo Demográfico 2010	Tabela 3541
Percentual de residentes com ensino superior completo	Percentual de pessoas de 25 anos ou mais de idade, residentes em domicílios particulares, que concluiu o ensino superior.	1 - Dados da Amostra.	IBGE - Censo Demográfico 2010	Tabela 3541
Percentual de residentes que trabalham em	Percentual de pessoas com 10 anos ou mais de idade com exercício do trabalho principal em outro município.	1 - Dados da Amostra.	IBGE - Censo Demográfico 2010	Tabela 3602

outro município				
Percentual de residentes com carga horária de trabalho $\geq$ 45 horas	Percentual de pessoas de 10 anos ou mais de idade, ocupadas na semana de referência, com carga horária de trabalho de 45 horas ou mais.	1 - Dados da Amostra.	IBGE - Censos Demográficos	Tabela 3582
Percentual de residentes em favela	Percentual de residentes em casa de cômodos, cortiço ou cabeça de porco segundo município.	(População de moradores em casa de cômodos, cortiço ou cabeça de porco / População total) *100.	Sidra, IBGE - Censo Demográfico 2010	Tabela 3513 - Domicílios particulares permanentes e moradores em domicílios particulares permanentes por adequação da moradia, segundo o tipo de domicílio, a condição de ocupação do domicílio, o tipo de material das paredes externas e a existência de energia elétrica
Percentual de residentes em zona urbana	Percentual da população residentes em domicílios em zona urbana segundo município.	(População residente em zona urbana / População total) *100.	Sidra, IBGE - Censo Demográfico 2010	Tabela 200 - População residente por sexo, situação e grupos de idade - Amostra - Características Gerais da População
Percentual de residentes no município há menos de 10 anos	Percentual de pessoas de 15 anos ou mais que residiam há menos de 10 anos ininterruptos no município.	1 - Dados da Amostra.	IBGE - Censo Demográfico 2010	Tabela 3181
Percentual de estrangeiros	(População de estrangeiros / População total) *100.	1 - Dados da Amostra.	IBGE - Censo Demográfico 2010	Tabela 3180
Estado	Sigla do estado a que o município pertence.			

Apêndice B – Dicionário de características de saúde utilizadas na comparação de municípios overachievers e underachievers.

Nome da variável	Descrição	Fonte	Link para acesso ao DATASUS
Percentual de baixo peso ao nascer	Número de nascidos vivos com <2.500g segundo local de residência da mãe/ Total de nascidos vivos.	DATASUS/ Tabnet/ Fonte: CNES e SINASC Nascidos vivos/ Peso ao nascer	http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sinasc/cnv/nvbr.def
Percentual de parto cesárea	Número de nascidos vivos por cesárea segundo local de residência da mãe / Total de nascidos vivos.	DATASUS/ Tabnet/ Fonte: CNES e SINASC Nascidos vivos / tipo de parto	http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sinasc/cnv/nvbr.def
Cobertura das equipes da Estratégia Saúde da Família	Cobertura populacional estimada pelas equipes de atenção básica.	DATASUS/ Tabnet/ Fonte: CNES e ESF	http://tabnet.datasus.gov.br/cgi/tabcgi.exe?pacto/2012/cnv/pactbr.def
Cobertura vacinal	Coberturas vacinais segundo o município. (Número de doses aplicadas da dose indicada/ população alvo) *100.	DATASUS/Tabnet/ Fonte: PNI	http://tabnet.datasus.gov.br/cgi/tabcgi.exe?pni/cnv/cpnibr.def
Cobertura das equipes de Saúde Bucal	Cobertura das equipes de Saúde Bucal segundo o município.	DATASUS/ Tabnet/ Fonte: Histórico de Cobertura da Saúde da Família/DAB/SAS/MS	http://tabnet.datasus.gov.br/cgi/tabcgi.exe?pacto/2010/cnv/pactbr.def
Número de equipes da Estratégia Saúde da Família/ 10.000 habitantes	(Número de equipes da Estratégia de Saúde da Família por 1000 habitantes segundo o município) *10.	DATASUS/ Tabnet/ Fonte: CNES e Censo Demográfico de 2010	http://tabnet.datasus.gov.br/cgi/deftohtm.exe?cnes/cnv/equipebr.def
Número de leitos/ 10.000 habitantes	(Número total de leitos por 1000 habitantes segundo o município) *10.	DATASUS/ Tabnet/ Fonte: CNES e Censo Demográfico de 2010	http://tabnet.datasus.gov.br/cgi/deftohtm.exe?cnes/cnv/leintbr.def
Número de mamografias realizadas/ 100 mulheres	(Razão de mamografias realizadas segundo o município) * 100.	DATASUS/Tabnet/ Fonte: SIA/SUS	http://tabnet.datasus.gov.br/cgi/tabcgi.exe?pacto/2010/cnv/pactbr.def
Número de equipamentos de ultrassom/ 10.000 nascidos vivos	(Número de equipamentos de ultrassom por 1000 nascidos vivos segundo o município) *10.	DATASUS/ Tabnet/ Fonte: CNES e SINASC	http://tabnet.datasus.gov.br/cgi/deftohtm.exe?cnes/cnv/equipobr.def

Número de equipamentos de Raio X/ 10.000 habitantes	(Número de equipamentos de Raio X / por 1000 habitantes segundo o município) *10.	DATASUS - Recursos Físicos / Equipamentos	http://tabnet.datasus.gov.br/cgi/tabcgi.exe?cnes/cnv/equipobr.def
Número de equipamentos para manutenção da vida/ 10.000 habitantes	(Número de equipamentos para manutenção da vida / por 1000 habitantes segundo o município) *10. Equipamentos para manutenção da vida: berço aquecido, desfibrilador, equipamento de fototerapia, incubadora, marca-passo temporário, monitor de eco cardiograma, monitor de pressão invasivo, monitor de pressão não invasivo, oxímetro, reanimador pulmonar, respirador/ventilador adulto, respirador/ventilador infantil.	DATASUS - Recursos Físicos / Equipamentos	http://tabnet.datasus.gov.br/cgi/tabcgi.exe?cnes/cnv/equipobr.def

ANEXOS

Anexo A – Aprovação da Comissão Nacional de Ética em Pesquisa, Estudo SABE 2000.

MINISTÉRIO DA SAÚDE
Conselho Nacional de Saúde
Comissão Nacional de Ética em Pesquisa - CONEP

PARECER Nº 315/99

Processo nº 25000.024350/99-80 Registro CONEP = 675 (Protocolo CEP :118)
Projeto de Pesquisa: "As condições de saúde dos idosos na América do Sul e Caribe".
Pesquisador Responsável: Dr. Ruy Lauro
Instituição: Faculdade de Saúde Pública / USP
Área Temática Especial: Pesquisa com cooperação estrangeira

Ao se proceder à análise do protocolo em questão, cabem as seguintes considerações:

a) as informações enviadas atendem aos aspectos fundamentais das Res. CNS 196/96 e 251/97, sobre Diretrizes e Normas Regulamentadoras de Pesquisas Envolvendo Seres Humanos;

b) o projeto foi aprovado pelo Comitê de Ética em Pesquisa – CEP da instituição.

Diante do exposto, a Comissão Nacional de Ética em Pesquisa – CONEP, de acordo com as atribuições da Res. CNS 196/96, manifesta – se pela aprovação do projeto de pesquisa proposto, com a seguinte recomendação a ser acompanhada pelo CEP:

1) Orçamento financeiro detalhado, especificando a remuneração do pesquisador (Res. CNS 196/96, VI, "j").

Situação : Projeto aprovado com recomendação.

Brasília, 17 de junho de 1999.

WILLIAM SAAD HOSSNE
Coordenador da CONEP-MS

Anexo B – Aprovação do Comitê de Ética em Pesquisa da Faculdade de Saúde Pública, Estudo SABE 2006.

Anexo C – Aprovação do Comitê de Ética em Pesquisa da Faculdade de Saúde Pública, Estudo SABC 2010.

COMITÊ DE ÉTICA EM PESQUISA – COEP/FSP

Universidade de São Paulo
Faculdade de Saúde Pública

OF.COEP/23/10

5 de março de 2010.

Prezado(a) Pesquisador(a) e Orientador(a),

O Comitê de Ética em Pesquisa da Faculdade de Saúde Pública da Universidade de São Paulo – COEP/FSP, **analisou**, de acordo com a Resolução N.º 196/96 do Conselho Nacional de Saúde – CNS e suas complementares, o protocolo de pesquisa n.º **2044**, intitulado **"ESTUDO SABC 2010: SAÚDE, BEM-ESTAR E ENVELHECIMENTO - ESTUDO LONGITUDINAL SOBRE AS CONDIÇÕES DE VIDA E SAÚDE DOS IDOSOS NO MUNICÍPIO DE SÃO PAULO"**, área temática **GRUPO III**, sob responsabilidade do(a) pesquisador(a) **Maria Lucia Lebrao**, e considerou que a pendência anteriormente apresentada por este COEP foi atendida. Protocolo de pesquisa **APROVADO "AD-REFERENDUM"**.

Cabe lembrar que conforme Resolução CN /196/96, são deveres do (a) pesquisador (a): **1. Comunicar**, de imediato, qualquer alteração no projeto e aguardar manifestação deste CEP (Comitê de Ética em Pesquisa), para dar continuidade à pesquisa; **2. Manter sob sua guarda e em local seguro**, pelo prazo de 5 (cinco) anos, os dados da pesquisa, contendo fichas individuais e todos os demais documentos recomendados pelo CEP, no caso eventual auditoria; **3. Comunicar**, formalmente a este Comitê, quando do encerramento deste projeto; **4. Elaborar e apresentar relatórios parciais e final**; **5. Justificar**, perante o CEP, interrupção do projeto ou a não publicação dos resultados.

Atenciosamente,

Cláudio Leone
Professor Titular

Coordenador do Comitê de Ética em Pesquisa - COEP

Anexo D – Aprovação do Comitê de Ética para Análise de Projetos de Pesquisa da Diretoria Clínica do Hospital das Clínicas e da Faculdade de Medicina da Universidade de São Paulo, Estudo QALY.

APROVAÇÃO

A Comissão de Ética para Análise de Projetos de Pesquisa - CAPPesq da Diretoria Clínica do Hospital das Clínicas e da Faculdade de Medicina da Universidade de São Paulo, em sessão de 07/04/2010, **APROVOU** o Protocolo de Pesquisa nº **0161/10**, intitulado: **"QALY - MODELO PARA PREDIZER SOBREVIDA AJUSTADA PARA A QUALIDADE EM PACIENTES COM CÂNCER ADMITIDOS EM UNIDADE DE TERAPIA INTENSIVA"** apresentado pelo Departamento de **CIRURGIA**, inclusive o Termo de Consentimento Livre e Esclarecido.

Cabe ao pesquisador elaborar e apresentar à CAPPesq, os relatórios parciais e final sobre a pesquisa (Resolução do Conselho Nacional de Saúde nº 196, de 10/10/1996, inciso IX.2, letra "c").

Pesquisador (a) Responsável: **Alexandre Biasi Cavalcanti**

Pesquisador (a) Executante: **Juliana Martins Giorgi**

CAPPesq, 08 de Abril de 2010

Prof. Dr. Eduardo Massad
Presidente da Comissão
de Ética para Análise de
Projetos de Pesquisa

Comissão de Ética para Análise de Projetos de Pesquisa do HCFMUSP e da FMUSP Diretoria Clínica do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo Rua Ovídio Pires de Campos, 225, 5º andar - CEP 05403 010 - São Paulo - SP Fone: 011 3069 6442 Fax: 011 3069 6482 e-mail: cappesq@hcnet.usp.br / secretariacappesq2@hcnet.usp.br

Anexo E – Aprovação do Comitê de Ética em Pesquisa do Hospital de câncer de Barreto, Estudo QALY.

Comitê de Ética em Pesquisa
CEP

Para: Ulysses Vasconcellos Silva

De: Dr. Eduardo José Paton
Vice-Coordenador do Comitê de Ética em Pesquisa

Data: 05/10/2010

Projeto de Pesquisa: **337/2010**

Prezado (a) Senhor (a),

Vimos, por meio desta, informar que o Comitê de Ética em Pesquisa da Fundação Pio XII – Hospital de Câncer de Barretos aprovou a resposta as pendências projeto de pesquisa **337/2010** intitulada **“QALY- Modelo para predizer sobrevida ajustada para a qualidade em pacientes com câncer admitidos em unidade de terapia intensiva”**.

Atenciosamente,

Dr. Eduardo José Paton
Vice-Coordenador do Comitê de Ética em Pesquisa
Hospital de Câncer de Barretos

Rua Antenor Duarte Vilela, 1331 – B. Dr. Paulo Prata – Barretos – SP CEP 14794-400 – CNPJ 49.150.352/0001-12
Tel. 17 33216600 R. 6894 – cep@hcancoerbarretos.com.br

Anexo F – Comprovante de submissão, Artigo 1

13/07/2018

E-mail de Universidade de São Paulo - Novo artigo (CSP_0508/18)

Hellen Geremias dos Santos <hellengeremias@usp.br>

Novo artigo (CSP_0508/18)

1 mensagem

Cadernos de Saude Publica <cadernos@fiocruz.br>

13 de março de 2018 17:20

Para: hellengeremias@usp.br

Prezado(a) Dr(a). Hellen Geremias dos Santos:

Confirmamos a submissão do seu artigo "Machine learning para análises preditivas em saúde: exemplo de aplicação para prever óbito em idosos de São Paulo" (CSP_0508/18) para Cadernos de Saúde Pública. Agora será possível acompanhar o progresso de seu manuscrito dentro do processo editorial, bastando clicar no *link* "Sistema de Avaliação e Gerenciamento de Artigos", localizado em nossa página <http://www.ensp.fiocruz.br/csp>.

Em caso de dúvidas, envie suas questões através do nosso sistema, utilizando sempre o ID do manuscrito informado acima. Agradecemos por considerar nossa revista para a submissão de seu trabalho.

Atenciosamente,

Profª. Marília Sá Carvalho
Profª. Claudia Medina Coeli
Profª. Luciana Dias de Lima
Editoras

Cadernos de Saúde Pública / Reports in Public Health

Escola Nacional de Saúde Pública Sergio Arouca
Fundação Oswaldo Cruz
Rua Leopoldo Bulhões 1480
Rio de Janeiro, RJ 21041-210, Brasil
Tel.: +55 (21) 2598-2511, 2508 / Fax: +55 (21) 2598-2737
cadernos@ensp.fiocruz.br
<http://www.ensp.fiocruz.br/csp>

Anexo G – Comprovante de submissão, Artigo 2

Hellen Geremias dos Santos <hellengeremias@usp.br>

ICME-D-18-01232 - Submission Confirmation

1 mensagem

Intensive Care Medicine - Editorial Office (ICME) <em@editorialmanager.com>

25 de agosto de 2018 19:07

Responder a: "Intensive Care Medicine - Editorial Office (ICME)" <atchaya.durai@springer.com>

Para: Hellen Geremias dos Santos <hellengeremias@usp.br>

Dear Mrs. Santos,

Thank you for submitting your manuscript, "Machine learning to predict 30-day quality-adjusted survival in critically ill patients with cancer", to Intensive Care Medicine.

The submission id is: ICME-D-18-01232

Please refer to this in any future correspondence.

Your paper is now being processed by our editors.

During that time, you can track the status of your manuscript by accessing the journal's website.

Anexo H – Comprovante de submissão, Artigo 3

13/07/2018

E-mail de Universidade de São Paulo - Verify your authorship and copyright for a submission to Epidemiology

Hellen Geremias dos Santos <hellengeremias@usp.br>

Verify your authorship and copyright for a submission to Epidemiology

1 mensagem

Epidemiology <em@editorialmanager.com>

5 de julho de 2018 15:32

Responder a: Epidemiology <timothy.lash@epidemiology-journal.com>

Para: Hellen Geremias dos Santos <hellengeremias@usp.br>

EPIDEMIOLOGY

05 Jul 2018

RE: Submission of "Overachievers in public health: a machine learning approach" (the "Work") by Prof Alexandre Dias Porto Chiavegatto Filho to Epidemiology

Dear Dr. dos Santos;

Please click the link below to confirm that you are a contributing author of this paper. Upon clicking the link, you will be taken to a web page to answer the terms of the copyright agreement. Note that the corresponding author will receive a decision letter if the paper is rejected without review, usually within seven days. Therefore, you may wish to wait seven days before completing this process.

[Yes, I am affiliated.](#)

If you are not a contributing author for this article, please click the link below to notify the EPID Editorial Office.

[No, I am not affiliated.](#)

Best Regards,

A handwritten signature in blue ink, appearing to read 'Timothy Lash'.

Timothy L. Lash, DSc
 Editor-in-Chief
 Epidemiology
<http://www.epidem.com>

Anexo I – Declaração de ciência e aceitação de utilização de artigo em caso de coautoria

Os autores do manuscrito intitulado “*Overachievers in public health: a machine learning approach*”, submetido ao periódico “*Epidemiology*”: Alexandre Dias Porto Chiavegatto Filho, Hellen Geremias dos Santos, Carla Ferreira do Nascimento, Kaio Massa e Ichiro Kawachi, por meio deste suficiente instrumento, declaram, para os devidos fins, que concordam com a utilização do artigo como resultado da Tese de doutorado intitulada “*Machine Learning para análises preditivas em saúde: desafios e perspectivas*”, apresentada pela aluna Hellen Geremias dos Santos ao programa de pós-graduação em epidemiologia da Faculdade de Saúde Pública da Universidade de São Paulo. Os autores declaram ainda que aceitam a condição de que o referido manuscrito não seja utilizado em nenhuma outra Tese ou Dissertação.

São Paulo, 13 de julho de 2018.

Assinaturas:

Alexandre Dias Porto Chiavegatto Filho

Hellen Geremias dos Santos

Carla Ferreira do Nascimento

Kaio Henrique Correa Massa

Ichiro Kawachi

CURRÍCULO LATTES – ALUNA

Hellen Geremias dos Santos

Endereço para acessar este CV: <http://lattes.cnpq.br/4319629810006649>

Última atualização do currículo em 03/07/2018

Hellen Geremias dos Santos é doutoranda em Ciências (Habilitação: Epidemiologia) na Faculdade de Saúde Pública da Universidade de São Paulo. Fez mestrado em Saúde Coletiva na Universidade Estadual de Londrina (2012), graduação em Estatística na Universidade Estadual de Maringá (2015) e em Enfermagem na Universidade Estadual de Londrina (2009). Atualmente, estuda métodos e técnicas para o ajuste de modelos preditivos em pesquisas de saúde. **(Texto informado pelo autor)**

Identificação

Nome Hellen Geremias dos Santos

Nome em citações bibliográficas SANTOS, H. G.; Santos, Hellen Geremias dos; DOS SANTOS, HELLEN GEREMIAS

CURRÍCULO LATTES – ORIENTADOR

Alexandre Dias Porto Chiavegatto Filho

Endereço para acessar este CV: <http://lattes.cnpq.br/5517850224634709>

Última atualização do currículo em 10/07/2018

Possui graduação em Economia pela FEA/USP, doutorado direto em Saúde Pública pela FSP/USP e pós-doutorado na Universidade de Harvard. É professor doutor do Departamento de Epidemiologia da FSP/USP na área de estatísticas de saúde, orientador dos programas de pós-graduação em Saúde Pública e Saúde Global da USP e coordenador de cursos sobre o uso do R para a análise de dados e sobre machine learning em saúde. Atuou como professor convidado (2016) e pesquisador visitante (2017) na Universidade de Harvard. Atualmente é o Pesquisador Principal de pesquisas financiadas pela FAPESP, CNPq e Fundação Lemann. Em 2015-2016 foi responsável pelo curso online Big Data em Saúde no Brasil da parceria USP-Coursera, que teve mais de 8.500 alunos matriculados e representantes de todos os Estados brasileiros. É o diretor do Laboratório de Big Data e Análise Preditiva em Saúde (Labdaps) da FSP/USP. Tem experiência em pesquisas na área de saúde pública, com ênfase em estatísticas de saúde. **(Texto informado pelo autor)**

Identificação

Nome Alexandre Dias Porto Chiavegatto Filho

Nome em citações bibliográficas Chiavegatto Filho, A.D.P.;CHIAVEGATTO FILHO, A. D. P.;Chiavegatto Filho, Alexandre Dias Porto;Filho, Alexandre Dias Porto Chiavegatto;PORTO CHIAVEGATTO FILHO, ALEXANDRE DIAS;Chiavegatto Filho, Alexandre DP;Chiavegatto, Alexandre Dias Porto;Chiavegatto Filho, Alexandre;Filho, Alexandre Chiavegatto;CHIAVEGATTO FILHO, ALEXANDRE D P;CHIAVEGATTO FILHO, A D P