RENATA MARTINS DE CARVALHO

Modelagem preditiva da presença e número de gatos nos domicílios brasileiros

São Paulo
2020

RENATA MARTINS DE CARVALHO

Modelagem preditiva da presença e número de gatos nos domicílios brasileiros

Dissertação apresentada ao Programa de Pós-Graduação em Epidemiologia Experimental Aplicada às Zoonoses da Faculdade de Medicina Veterinária e Zootecnia da Universidade de São Paulo para a obtenção do título de Mestre em Ciências.

Departamento:

Medicina Veterinária Preventiva e Saúde Animal

Área de concentração:

Epidemiologia Experimental Aplicada a Zoonoses

Orientador:

Prof. Dr. Oswaldo Santos Baquero

Co-orientadora:

Dra. Mariana Ramos Queiroz

São Paulo

2020

Autorizo a reprodução parcial ou total desta obra, para fins acadêmicos, desde que citada a fonte.

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO

(Biblioteca Virginie Buff D'Ápice da Faculdade de Medicina Veterinária e Zootecnia da Universidade de São Paulo)

T. 4018
FMVZ

Carvalho, Renata Martins de
Modelagem preditiva da presença e número de gatos nos domicílios brasileiros / Renata Martins de Carvalho. – 2020.
59 f. : il.

Dissertação (Mestrado) – Universidade de São Paulo. Faculdade de Medicina Veterinária e Zootecnia. Departamento de Medicina Veterinária Preventiva e Saúde Animal, São Paulo, 2021.

Programa de Pós-Graduação: Epidemiologia Experimental Aplicada às Zoonoses.

Área de concentração: Epidemiologia Experimental Aplicada às Zoonoses.

Orientador: Prof. Dr. Oswaldo Santos Baquero.
Coorientadora: Profa. Dra. Mariana Ramos Queiroz.

1. Felinos. 2. Aprendizado de máquina. 3. Epidemiologia Veterinária. I. Título.

Ficha catalográfica elaborada pela bibliotecária Maria Aparecida Laet, CRB 5673-8, da FMVZ/USP.

São Paulo, 8th December 2020

CERTIFIED

We certify that the Research "Predictive modelling of presence and number of cats in Brazilian households", protocol number CEUAx 7600050418 (ID 000902), under the responsibility Oswaldo Santos Baquero, agree with Ethical Principles in Animal Research adopted by Ethic Committee in the Use of Animals of School of Veterinary Medicine and Animal Science (University of São Paulo), and was approved in the meeting of day May 09, 2018.

Certificamos que o protocolo do Projeto de Pesquisa intitulado "Modelagem preditiva da presença e número de gatos nos domicílios brasileiros", protocolado sob o CEUAx nº 7600050418, sob a responsabilidade de Oswaldo Santos Baquero, está de acordo com os princípios éticos de experimentação animal da Comissão de Ética no Uso de Animais da Faculdade de Medicina Veterinária e Zootecnia da Universidade de São Paulo, e foi aprovado na reunião de 09 de maio de 2018.

Prof. Dr. Marcelo Bahia Labruna
Coordenador da Comissão de Ética no Uso de Animais
Faculdade de Medicina Veterinária e Zootecnia da Universidade de São Paulo

Camilla Mota Mendes
Vice-Coordenadora da Comissão de Ética no Uso de Animais
Faculdade de Medicina Veterinária e Zootecnia da Universidade de São Paulo

FOLHA DE AVALIAÇÃO

Autor: CARVALHO, Renata Martins de

Título: **Modelagem preditiva da presença e número de gatos nos domicílios brasileiros**

Dissertação apresentada ao Programa de Pós-Graduação em Epidemiologia Experimental Aplicada às Zoonoses da Faculdade de Medicina Veterinária e Zootecnia da Universidade de São Paulo para obtenção do título de Mestre em Ciências.

Data: _____/_____/_____

Banca Examinadora

Prof. Dr._____

Instituição:_____ Julgamento:_____

Prof. Dr._____

Instituição:_____ Julgamento:_____

Prof. Dr._____

Instituição:_____ Julgamento:_____

DEDICATÓRIA

Para Gabriel e Zaira, com amor e gratidão.

AGRADECIMENTOS

À minha mãe, que sempre me apoiou em minhas escolhas e que, além de me proporcionar uma infância cheia de amor e carinho, assentou os fundamentos do meu caráter. Obrigada por ser minha referência de tantas maneiras e estar sempre presente em minha vida de uma forma indispensável.

Ao meu orientador, Prof. Dr. Oswaldo Santos Baquero, que muito generosamente me aceitou como sua orientanda, tornando possível a realização deste mestrado. Obrigada por sua confiança em mim, suas valiosas contribuições e incentivos ao longo da elaboração deste projeto.

À minha co-orientadora, Dra. Mariana Ramos Queiroz, grande parceira nessa jornada. Obrigada por sua paciência e dedicação inesgotáveis, pela amizade que construímos e por nunca “soltar minha mão”. Espero poder retribuir à altura tudo que fez por mim.

À minha grande amiga Rosangela Ribeiro Gebara, irmã de coração, companheira de risos e lágrimas, pelo apoio e confiança, sem os quais eu nunca teria chegado até aqui. Um “muito obrigada” nunca será suficiente para demonstrar a minha gratidão pelo que recebo de você.

À Universidade de São Paulo e a todos os professores que me proporcionaram um ensino de alta qualidade.

E a todos os meus amigos que contribuíram, direta e indiretamente, para a realização deste trabalho.

“Feliz é a casa que tem, pelo menos, um gato”

Provérbio italiano

RESUMO

CARVALHO, R.M. **Modelagem preditiva da presença e número de gatos nos domicílios brasileiros**. 2020. 59 f. Dissertação (Mestrado em Ciências) – Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo, São Paulo, 2020.

Os gatos domésticos estão cada vez mais presentes nos domicílios, superando os cães em número em diferentes países. A escolha do gato como animal de companhia pode envolver diferentes aspectos, e é possível que características demográficas, socioeconômicas e geográficas das populações humanas estejam associadas à presença e ao número desses animais nos lares. No presente estudo, os algoritmos de aprendizado de máquina supervisionado Partial Least Square, Random Forest e Extreme Gradient Boosting foram aplicados aos dados demográficos e socioeconômicos provenientes da Pesquisa Nacional de Saúde de 2013 para identificar preditores da presença e número de gatos nos domicílios brasileiros e classificá-los de acordo com sua contribuição ao desempenho preditivo dos modelos construídos. Dentre eles, destacaram-se o número de cães, a zona de localização do domicílio (urbana ou rural) e número de moradores maiores de 18 anos. Porém, os algoritmos só explicaram uma pequena fração da complexidade que determina a coabitação entre humanos e gatos, mesmo incorporando 47 preditores socioeconômicos, geográficos e demográficos da população humana. Pesquisas qualitativas podem identificar preditores mais relevantes e informar estudos preditivos de base populacional.

Palavras-chave: Felinos. Aprendizado de máquina. Epidemiologia Veterinária.

ABSTRACT

CARVALHO, R.M. **Predictive modeling of presence and number of cats in Brazilian households**. 2020. 59 f. Dissertação (Mestrado em Ciências) – Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo, São Paulo, 2020.

Domestic cats' presence in households is increasing, outnumbering dogs in different countries. The choice of keeping a cat as a pet might involve many aspects, and the demographic, socioeconomic, and geographical characteristics of human populations might be associated with the presence and number of those animals in residences. In the present study, we predicted the presence and number of cats in Brazilian households and classified predictors according to their predictive performance. To this end, we used data from the 2013 National Health Survey and three supervised machine learning algorithms: Partial Least Square, Random Forest, and Extreme Gradient Boosting. The number of dogs, the household zone (urban or rural), and the number of residents over 18 years old had the highest predictive performance. However, the algorithms explained only a small fraction of the complexity determining human-cat cohabitation, even with 47 socioeconomic, geographic, and demographic predictors of the human population. Qualitative research might identify more relevant predictors and inform population-based predictive studies.

Keywords: Feline. Machine learning. Veterinary Epidemiology.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

O gato doméstico (*Felis catus*) é um dos animais de companhia mais populares no mundo, superando os cães em número em diversos países, entre eles França, Alemanha e Reino Unido (GFK, 2016). Em 2013, o Brasil possuía a segunda maior população desses felinos, com 21,2 milhões de gatos vivendo em lares, de acordo com estimativa do IBGE (IBGE; ABINPET, 2015), atrás apenas da Rússia, com cerca de 21,7 milhões (FEDIAF, 2018).

Diferentes hipóteses podem explicar a preferência pelo gato como animal de companhia, entre elas a facilidade de cuidado, seu tamanho compatível com pequenos espaços como apartamentos e a capacidade de permanecer longe do tutor por períodos maiores que os cães sem que isso gere grandes problemas (BERNSTEIN, 2007).

É possível, ainda, que características demográficas e socioeconômicas das populações humanas distintas estejam associadas à presença desses animais. Em pesquisas realizadas em diferentes países, a presença dos gatos domiciliados se mostrou ligada a fatores como o nível de educação dos moradores e a presença de cães (SLATER et al., 2008; RAMÓN; SLATER; WARD, 2010; WESTGARTH et al., 2010). Estudos de Downes, Canty e More (2009), Murray et al. (2015) e Carvelli, Iacoponi e Scaramozzino (2016) também encontraram relações entre características dos moradores, como idade e gênero, e dos domicílios, como tipo (casa ou apartamento), e a presença ou ausência de gatos.

No Brasil, estudos que identificam determinantes da presença e do número de gatos nos domicílios são escassos. Martins et al. (2013) encontraram associações entre o número de cães e gatos e as características socioeconômicas dos tutores e dos domicílios na cidade de Pinhais, estado do Paraná. Em estudo semelhante, os dados coletados pelo Inquérito de Saúde do município de São Paulo (ISA Capital), em 2003, foram usados para relacionar o perfil da população domiciliada de cães e gatos às condições dos domicílios e de seus entornos e ao nível socioeconômico dos tutores (MAGNABOSCO, 2006).

Os estudos mencionados observaram fatores ligados à presença e ao número de gatos domiciliados, mas não exploraram seu papel preditivo. Conhecer tal papel pode ajudar na identificação das características que melhor diferenciam os domicílios com gatos daqueles sem gatos, e os domicílios com mais ou menos gatos,

melhorando, assim, a compreensão do que leva as pessoas a manterem esses animais em seus lares.

Nesse contexto, o presente estudo buscou colaborar com o entendimento dos fatores relacionados à guarda de gatos a partir da análise de dados demográficos, socioeconômicos e geográficos dos domicílios brasileiros, provenientes da Pesquisa Nacional de Saúde (PNS) em 2013. Para tal, utilizaram-se técnicas de aprendizado de máquina (AM) para a busca de padrões e predição da presença e do número de gatos nesses domicílios.

2 MATERIAIS E MÉTODOS

2.1 QUESTIONÁRIO

A população-alvo da PNS 2013 foi composta por residentes de domicílios particulares permanentes (DPP) — aqueles construídos para uso exclusivamente como habitação, com a finalidade de servir de moradia a uma ou mais pessoas. Os detalhes do desenho amostral podem ser consultados em Souza Júnior et al. (2015).

A PNS 2013 coletou e analisou informações sobre a situação de saúde e estilo de vida e dados de acesso a serviços relacionados à saúde, bem como informações socioeconômicas dos domicílios acessados. O questionário aplicado foi dividido em três partes. A primeira delas é a domiciliar, com questões sobre características do domicílio e visitas realizadas pela equipe de Saúde da Família e agentes de endemias; tais perguntas foram respondidas pelo responsável pelo domicílio ou pessoa de posse das informações do domicílio no momento da entrevista. A segunda parte investigou dados gerais sobre os moradores do domicílio, incluindo rendimento, nível de educação e trabalho. As informações sobre a saúde foram exploradas na terceira parte do questionário, respondido por um morador maior de 18 anos selecionado pelo pesquisador.

A pesquisa continha perguntas sobre a presença e número cães, gatos, aves e peixes nos domicílios, além do status de vacinação antirrábica de cães e gatos. Até o momento, esses dados foram utilizados pelo IBGE apenas para o cálculo do tamanho dessas populações nas regiões do país e por Baquero e Queiroz (2019) para caracterizar, com mais detalhes, o tamanho e a distribuição espacial da população canina e cobertura de vacinação antirrábica, e por Filho (2020) para calcular as estimativas da população de gatos, sua densidade domiciliar e a cobertura de vacinação antirrábica da espécie.

As perguntas sobre a presença de animais não permitiram compreender a definição utilizada pela pesquisa para o status da guarda dos gatos. Sendo assim, para fins deste estudo, assumiu-se que as declarações de presença e número de gatos fornecidas pelos respondentes se referiram a animais considerados membros do domicílio.

2.2 BANCO DE DADOS

O banco de dados do questionário da PNS utilizado é composto 1.019 colunas (variáveis) com 64.348 observações (informações dos domicílios pesquisados). Neste estudo, foram consideradas somente as respostas dos responsáveis pelos domicílios, para evitar que um único domicílio fosse contabilizado mais de uma vez.

Os dados e informações complementares da pesquisa podem ser baixados diretamente no site do IBGE (2015).

2.3 APRENDIZADO DE MÁQUINA

Técnicas de aprendizado de máquina (AM) foram aplicadas aos dados para predição de duas variáveis resposta — presença e número de gatos — e para avaliar e classificar variáveis preditoras em relação à sua capacidade de prever as respostas.

As seguintes etapas do AM foram realizadas: (1) análise exploratória; (2) pré-processamento e particionamento dos dados; (3) treinamento dos modelos; e (4) teste dos modelos, descritas a seguir.

2.3.1 Análise exploratória

A análise exploratória foi realizada por meio da leitura do questionário, observação da estrutura geral do banco de dados e construção de gráficos para visualização da distribuição e associação entre variáveis. Isso guiou as etapas subsequentes da metodologia.

2.3.2 Pré-processamento dos dados

O pré-processamento foi realizado para eliminar variáveis sem valor ou com pouco valor preditivo e para representar adequadamente os tipos de variáveis. No primeiro caso, foram eliminadas: variáveis relativas ao desenho amostral; as não presentes no dicionário de variáveis fornecido pelo IBGE; as que tinham variância zero (um valor único) ou variância próxima de zero (razão entre o valor mais frequente e o segundo mais frequente foi maior do que 95/5); e as variáveis em que os valores ausentes (NA) representavam mais de 10% das observações. Nas variáveis com

menos de 10% de NA, os valores ausentes foram imputados com o valor mais frequente da variável.

Para representar adequadamente as variáveis e facilitar sua interpretação pelos algoritmos, as variáveis categóricas (qualitativas) foram transformadas em numéricas, usando-se o método de variáveis fictícias (*dummy variables*).

2.3.2.1 Criação de variáveis

Embora a PNS não forneça informações sobre classificação social baseada em renda, o questionário aplicado coletou alguns dados sobre a estrutura física dos domicílios, o acesso a serviços domiciliares de saneamento e a presença de diferentes bens de conforto.

Entretanto, quando observadas separadamente, tais variáveis não ofereceram uma visão geral sobre as condições físicas do domicílio ou de renda dos moradores, tampouco contribuíram na interpretação dos resultados dos modelos. Por essa razão, optou-se pela criação de duas variáveis a partir dos dados mencionados, com o objetivo de melhorar a compreensão sobre a composição dos domicílios do banco de dados pré-processado e o desempenho dos modelos preditivos aplicados.

A primeira variável criada representa a classificação do domicílio de acordo com os critérios dos Indicadores de Desenvolvimento Sustentável (IDS), uma vez que há uma forte correlação entre pobreza monetária e precariedades e vulnerabilidades nas condições de moradia (IBGE, 2019). Essa classificação considera os materiais utilizados na construção do domicílio, os serviços públicos de água, esgoto e coleta de lixo e densidade de moradores por dormitório (Quadro 1).

Quadro 1 - Classificação dos domicílios segundo os Indicadores de Desenvolvimento Sustentável (IDS) do IBGE (2019).

Classificação	**Materiais usados na construção**	**Serviços de água e esgoto**	**Coleta de lixo**	**Nº de moradores por dormitório**
Adequado	Paredes externas construídas predominantemente de alvenaria (com ou sem revestimento), de taipa revestida, ou de madeira apropriada para construção	Rede geral de abastecimento de água Rede geral de esgoto ou fossa séptica	Coleta de lixo por serviço de limpeza	Até 2 moradores
Inadequado	Paredes de taipa não-revestida, de madeira aproveitada (como tapumes ou madeira retirada de pallets) e de outros materiais	Abastecimento de água proveniente de poço ou nascente ou outra forma sem banheiro e sanitário ou com escoadouro ligado à fossa rudimentar, vala, rio, lago, mar ou outra forma	Lixo queimado, enterrado ou jogado em terreno baldio ou logradouro, em rio, lago ou mar ou outro destino	Mais de 2 moradores
Semiadequado	Pelo menos um serviço inadequado			

Fonte: Carvalho (2020, p. 21).

A segunda variável criada representa a classe social baseada em renda. Apesar do questionário conter perguntas sobre os rendimentos dos moradores, esses dados apresentam muitas informações ausentes e valores inconsistentes. Por essa razão, tornou-se inviável utilizar o critério baseado em número de salários mínimos utilizado pelo IBGE.

Como alternativa, optou-se por adaptar o Critério de Classificação Econômica Brasil (CCEB) da Associação Brasileira das Empresas de Pesquisa. O critério é baseado em uma pontuação atribuída de acordo com a presença e número de bens de conforto, banheiros e empregados domésticos, ao grau de instrução do chefe de família e ao acesso a serviços públicos de água encanada e rua pavimentada (KAMAKURA; MAZZON, 2016).

Do total de quinze itens usados pela pontuação CCEB (ABEP, 2019), somente quatro deles não estão disponíveis no banco de dados da PNS: lava-louças, freezer, secadora de roupa e rua pavimentada. Juntos, esses itens representam 10% (10 pontos) da pontuação total (100 pontos). Para corrigir essa ausência de dados, uma

nova pontuação foi desenvolvida, eliminando-se 10% do intervalo de pontos de cada categoria original do CCEB (Quadro 2).

Quadro 2 - Pontuação adaptada da CCEB para estratificação por classe social.

Pontuação CCEB	Nova pontuação	Classe social
45 - 100	41 - 90	A
38 - 44	34 - 40	B1
29 - 37	26 -33	B2
23 - 28	21 -25	C1
17 - 22	15 - 20	C2
0 - 16	0 - 14	D - E

Fonte: Carvalho (2020, p. 22).

Uma variável categórica binária foi criada para representar a presença ou ausência de pessoas com deficiência no domicílio. Para isso, combinaram-se quatro variáveis: deficiência intelectual, deficiência física, deficiência visual e deficiência auditiva. Havendo o valor 1 (presença) em uma ou mais das variáveis citadas, a nova variável ‘presença de pessoa com deficiência’ recebeu valor 1.

2.3.3 Particionamento dos dados

Os dados foram divididos em dois subconjuntos, um de teste e um de treinamento, utilizando-se a abordagem de *holdout*. O conjunto de treinamento foi utilizado para construir e ajustar o modelo e o de teste, para estimar seu desempenho de predição (KUHN; JOHNSON, 2013).

O conjunto de treinamento foi criado com 80% dos dados e o de teste, com 20%. Para o particionamento, ambas as variáveis resposta (presença e número de gatos) foram testadas; a variável resposta ‘presença de gatos’ foi escolhida como base para partição por apresentar melhor resultado no estudo piloto.

2.3.3.1 Medidas de referência de desempenho (B*enchmark*)

As medidas de referência (*benchmark*) são valores com os quais os resultados dos modelos podem ser comparados para avaliar se estes são melhores ou piores do que uma previsão aleatória (OLSON et al., 2017). *Benchmarks* foram criados para os modelos de regressão e de classificação.

Para o cálculo do *benchmark* dos modelos de regressão, foram gerados valores aleatórios a partir de uma distribuição de Poisson, utilizando-se a média do número de gatos no conjunto de treinamento como parâmetro lambda para sortear o número de gatos em cada observação do conjunto de teste.

No cálculo do *benchmark* de modelos de classificação, os valores aleatórios foram gerados a partir de uma distribuição de binomial, utilizando-se a quantidade de observações do conjunto de teste como número de ensaios e o número médio de sucessos no conjunto de treinamento como probabilidade de sucesso.

2.3.4 Treinamento dos modelos

Escolheram-se três algoritmos de aprendizado de máquina supervisionado: Partial Least Square (PLS), Random Forest (RF) e Extreme Gradient Boosting (XGBoost). Eles permitem ranquear as variáveis de acordo com suas contribuições preditivas em modelos de regressão e classificação. As principais características dos algoritmos escolhidos estão listadas no Quadro 3.

Quadro 3 - Principais características dos três algoritmos de aprendizado de máquina selecionados.

Algoritmo	**Princípio básico**	**Vantagens**	**Desvantagens**
PLS	Extração de variáveis latentes (componentes) que contêm o máximo de informação sobre a variação das variáveis preditoras e têm correlação máxima com a variável resposta.	• Pode ser aplicado em problemas de regressão e de classificação • Tem correlação máxima com a variável resposta • Facilidade de implementação	• Não há um limite determinado para convergência, o que aumenta o tempo computacional • Oferece apenas um hiperparâmetro de ajuste, para o qual não existe um valor padrão determinado •
RF	Agregação de múltiplas árvores de decisão para criar uma regra de predição em um modelo de aprendizado	• Pode ser aplicado em problemas de regressão e de classificação • Pode lidar com grande número de variáveis sem necessidade de eliminação de dados • Gera uma estimativa imparcial do erro de generalização internamente	• Em variáveis categóricas com vários níveis, pode favorecer aquelas que apresentam mais níveis • Pode ocorrer sobreajuste em dados com muita discrepância (noisy) • Número grande de árvores pode tornar o algoritmo lento
XGBoost	Otimização do dos algoritmos de Gradient Boosting Machine (GBM), que agrega múltiplas árvores de decisão	• Pode ser aplicado em problemas de regressão e de classificação • Acurácia, desempenho e velocidade de processamento altos • Combina diferentes metodologias de otimização de aprendizagem de máquina (*boosting*, *bagging*) • Utiliza regularização para prevenir o sobreajuste	• Ajuste mais difícil de ser realizado, pois há muitos hiperparâmetros • Pode ocorrer sobreajuste se os parâmetros não forem ajustados corretamente • Alta complexidade de segmentação dos dados e consumo de memória computacional.

Fonte: Carvalho (2020, p. 24).

2.3.4.1 Princípios do algoritmo Partial Least Squares

O Partial Least Square (PLS) é uma técnica multivariável utilizada para reduzir o número de variáveis explicativas, eliminando problemas de multicolinearidade e otimizando o desempenho preditivo das variáveis criadas (componentes) (MATEOS-APARICIO, 2011).

O modelo generaliza e combina recursos da Análise de Componentes Principais (PCA) e da regressão múltipla para realizar uma predição por meio da extração de um conjunto de variáveis latentes (não observáveis) a partir das variáveis explicativas (observáveis) (ABDI, 2007). Essas variáveis latentes contêm o máximo de informação sobre a variação das variáveis preditoras e têm correlação máxima com a variável resposta, garantindo que a variância residual (erro) das relações preditivas seja mínima(MATEOS-APARICIO, 2011; KUHN; JOHNSON, 2013).

O PLS também pode ser aplicado a problemas de classificação, codificando os membros de uma classe em uma matriz de indicadores, ou ser utilizado como método de redução de dimensionalidade similar ao PCA (ROSIPAL; KRÄMER, 2006).

2.3.4.2 Princípios do algoritmo Random Forest

O Random Forest se baseia na agregação de múltiplas árvores de decisão para criar uma regra de predição em um modelo de aprendizado (BOULESTEIX et al., 2012; CUTLER; CUTLER; STEVENS, 2012).

No Random Forest, cada árvore de decisão é criada a partir de um agrupamento aleatório de variáveis explicativas, que estão posicionadas, inicialmente, no “nó raiz” (*root node*). A partir desse ponto, a árvore divide os dados das variáveis (BOULESTEIX et al., 2012).

A divisão específica que cada árvore utiliza para particionar um nó em seus dois descendentes é escolhida considerando-se cada divisão possível em cada variável explicativa e selecionando-se a melhor divisão de acordo com um critério de “bom ajuste” (regressão) ou de “pureza” (classificação) para um nó (CUTLER; CUTLER; STEVENS, 2012).

Após a escolha da divisão, várias árvores descendentes passam a ser criadas. O número de árvores de uma floresta e a profundidade de cada árvore de decisão (número de nós descendentes) são definidos por hiperparâmetros (KULLARNI; SINHA, 2013). Quando esses critérios são atendidos, a divisão é finalizada em um "nó terminal". Então, a predição é obtida a partir dos valores dos nós terminais de todas as árvores, calculando-se a média da resposta para problemas de regressão ou a classe mais frequente para problemas de classificação (CUTLER; CUTLER; STEVENS, 2012).

Cerca de 1/3 dos dados originais é deixado de fora durante a seleção da amostra, isto é, não é usado na construção da árvore. Esses dados, chamados de "*out-of-bag*" (OOB – em Português, "fora da sacola") são utilizados como um conjunto de validação interna pela árvore individual para a estimativa de erro — considerada menos otimista e, geralmente, um bom estimador do erro esperado para dados independentes (BOULESTEIX et al., 2012; KULLARNI; SINHA, 2013).

2.3.4.3 Princípios do algoritmo XGBoost

O XGBoost (acrônimo de Extreme Gradient Boosting) é uma otimização dos algoritmos de Gradient Boosting Machine (GBM) utilizados para problemas de regressão e classificação. Enquanto o modelo Random Forest constrói múltiplas árvores de decisão profundas em paralelo e gera uma predição a partir da média dos resultados de todas elas, o GBM cria um conjunto de árvores superficiais em sequência, com cada árvore aprendendo com a anterior (BOEHMKE; GREENWELL, 2019).

No GBM, cada árvore prevê o erro (gradient) da árvore anterior (que deu origem a ela) e melhora (boosting) o resultado com base nesse erro. Portanto, gradient boosting é uma maneira de reduzir o erro de predição de forma gradual (AYYADEVARA, 2018). A ideia principal do boosting é que cada nova árvore na sequência se concentre nas linhas de treinamento nas quais sua antecessora gerou os maiores erros de previsão. Embora as árvores superficiais criadas pelo GBM sejam modelos preditivos bastante fracos individualmente, em conjunto, são capazes de gerar predições mais acuradas (BOEHMKE; GREENWELL, 2019).

O XGBoost é uma variação mais eficiente e escalável do GBM, superando-o na agilidade de construção de árvores, além de fornecer vários hiperparâmetros de regularização para ajudar a reduzir a complexidade do modelo e protegê-lo contra um sobreajuste aos dados (*overfitting*) (BOEHMKE; GREENWELL, 2019). O fato de trabalhar com computação paralela e distribuída torna o aprendizado mais rápido e permite uma exploração mais ágil do modelo (CHEN; GUESTRIN, 2016).

2.3.4.3 Hiperparâmetros utilizados

Para determinar os valores apropriados dos hiperparâmetros de treinamento e ajuste, o método de reamostragem de validação cruzada *k-fold* foi aplicado a todos os modelos. Nesse método, o conjunto de dados é dividido aleatoriamente em uma série de subconjuntos de tamanhos iguais (*k-folds*); o valor de k determina o número de subconjuntos a serem criados (KUHN; JOHNSON, 2013). Neste estudo, utilizou-se o valor de k igual a 10.

No treinamento e ajuste dos modelos PLS de classificação e regressão, utilizou-se o hiperparâmetro número de componentes principais (ncomp). Na classificação e na regressão, foi usado 1 ncomp com incremento de 2 componentes até chegar a 49 componentes principais. O valor de ncomp foi determinado com base no número mínimo e máximo de variáveis preditoras usadas no modelo.

No modelo Random Forest, os hiperparâmetros de treinamento e ajuste usados foram o número de variáveis preditoras selecionadas aleatoriamente para cada divisão da árvore (mtry) e o número de folhas (observações) ao final de cada nó (min.node.size). Os valores para ajuste nos modelos de classificação foram:

- min.node.size = 1 e 5, considerando-se que o valor padrão para modelos de classificação é igual a 1;
- mtry = 5, 15 e 25, considerando-se que o valor padrão é a raiz quadrada do número total de variáveis preditoras usadas pelo modelos; neste caso, raiz quadrada de 49 (7).

Para os modelos de regressão, os valores de ajuste dos hiperparâmetros foram:

- min.node.size = 5 e 12, considerando-se que o valor padrão para modelos de regressão é igual a 5;
- mtry = 5, 15 e 25, considerando-se que o valor padrão é o número total de variáveis preditoras dividido por três variáveis; neste caso, 49 variáveis dividido por 3 (16,3).

Também foram utilizadas três diferentes regras de poda, de acordo com o tipo de modelo. Para os modelos de classificação, usaram-se duas regras de poda: a mensuração da pureza do nó (Índice Gini) e seleção em cada nó da melhor divisão dentre as divisões geradas aleatoriamente (extratrees). Nos modelos de regressão, utilizou-se a redução total da variância da variável resposta (variance) e extratrees.

Os modelos de XGBoost utilizaram sete hiperparâmetros para treinamento e ajuste, de acordo com a documentação do modelo (CHEN; GUESTRIN, 2016), sendo: 1. número máximo de iterações de impulsionamento (nrounds); 2. profundidade máxima de uma árvore (max_depth); 3. taxa de aprendizagem (eta); 4. redução mínima necessária para realizar uma partição adicional em um nó folha da árvore (gamma); 5. proporção da subamostra de colunas (variáveis) ao construir cada árvore (colsample_bytree); 6. mínimo de instâncias (observações) necessárias para cada nó (min_child_weight); e 7. proporção de subamostra das instâncias de treinamento (subsample).

Os valores de hiperparâmetros usados no modelo XGBoost de classificação foram:

- nrounds = 100, 150 e 200, considerando-se que o valor padrão é 100;
- max_depth = 3 e 6, considerando-se que o valor padrão é 6; quanto maior o valor de max_depth, mais complexo será o modelo e a propensão a um sobreajuste será maior;
- eta = 0.1, 0.3 e 0.5, considerando-se que o valor padrão é 0.3; quanto menor o valor de eta, maior o tempo computacional usado pelo algoritmo;
- gamma= 1, considerando-se que o valor padrão é 0; quanto maior for o valor de gama, maior a penalização e a prevenção de sobreajuste.;
- colsample_bytree = 1, considerando-se que o valor padrão é 1;
- min_child_weight = 0, considerando-se que o valor padrão é 1; quanto maior o valor de min_child_weight, mais conservador será o algoritmo.;
- subsample= 0.25, 0.50, 0.75 e 1, que acrescentam 4 pontos de aleatoriedade ao algoritmo.

Outros dois parâmetros foram acrescentados para aumentar a eficiência do modelo em bancos de dados desbalanceados:

- max_delta_step = 1 (valor padrão 0); quando definido com um valor positivo, pode ajudar a tornar a etapa de atualização mais conservadora;
- tree_method = hist (o valor padrão é auto, que pode selecionar entre approx, hist e gpu_hist).

O treinamento e ajuste do modelo XGBoost de regressão utilizou os mesmos valores de hiperparâmetros usados no modelo de classificação, exceto para min_child_weight, para qual utilizou-se o valor padrão (1).

2.3.5 Avaliação do desempenho dos modelos

Neste estudo, foram comparados três diferentes algoritmos para selecionar o de melhor desempenho preditivo, bem como o modelo de melhor desempenho do espaço de hipóteses do algoritmo.

2.3.5.1 Modelos de Classificação

Os modelos de classificação (variável resposta “presença de gatos”) foram avaliados com base na curva ROC-AUC (*Receiver Operating Characteristic*) e valores de MCC (*Matthews Correlation Coefficient*).

O gráfico ROC é composto por dois eixos, X e Y, que representam, respectivamente, a taxa de falsos positivos (1 menos o valor da especificidade) e a taxa de verdadeiros positivos (sensibilidade), e uma linha diagonal formada a partir de seus valores, a curva ROC. O desempenho preditivo de um algoritmo pode ser avaliado pela medida da área abaixo da curva ROC, a AUC (*Area Under de Curve*), que gera valores entre 0 e 1 (FACELI et al., 2017). Quanto maior a AUC, melhor desempenho preditivo.

O coeficiente de correlação de Matthews (MCC) é uma forma de descrever a matriz de confusão em único valor. O coeficiente leva em consideração verdadeiros positivos, falsos positivos, verdadeiros negativos e verdadeiros positivos e é, geralmente, considerado uma medida equilibrada, que pode ser usada mesmo se as classes forem de tamanhos muito diferentes. Os valores de MCC variam entre 1 e -1, sendo 1 uma previsão perfeita, 0 uma previsão não melhor do que uma previsão aleatória e -1 uma discordância total entre a predição e a observação (CHICCO; JURMAN, 2020).

2.3.5.2 Modelos de Regressão

Para a avaliação de desempenho dos modelos de regressão (variável resposta “número de gatos”), utilizou-se a medida de Raiz Quadrada do Erro Médio (RMSE). Ela corresponde à raiz quadrada da média das diferenças ao quadrado entre as predições feitas pelo modelo e as observações reais elevada ao quadrado (resíduos). Seu valor pode ser interpretado como a distância média entre os valores observados

e as previsões feitas pelo modelo, podendo variar de zero (nenhum erro) a infinito (KUHN; JOHNSON, 2013).

2.4 SOFTWARES UTILIZADOS

O software utilizado para a análise e construção dos modelos foi o R versão 3.5.0 (R CORE TEAM, 2020), com os pacotes tidyverse versão 1.3.0 (WICKHAM, 2019), caret versão 6.0-86 (KUHN et al., 2020); corrplot versão 0.84 (WEI et al., 2017), mice versão 3.11.0 (BUUREN et al., 2019), ggplot2 versão 3.3.2 (WICKHAM et al., 2020), mltools versão 0.3.5 (GORMAN, 2018), data.table versão 1.13.2 (DOWLE et al., 2020) e ggpubr versão 0.4.0 (KASSAMBARA, 2020).

Para a análise descritiva que deu origem às tabelas, utilizou-se o software Tableau Desktop versão 10.3.21.

3 RESULTADOS

3.1 ANÁLISE EXPLORATÓRIA

Nesta subseção, apresentam-se estatísticas descritivas do banco de dados pré-processado utilizado na modelagem preditiva. Assim, os dados diferem dos dados brutos da PNS 2013 devido ao pré-processamento. É importante ressaltar que não se trata de estimativas populacionais

3.1.1 Características dos domicílios

A maioria dos domicílios do banco de dados pré-processado era do tipo casa (86,9%) e estava localizada em zona urbana (82,19%). Eles se encontravam distribuídos em todas as regiões do país, de acordo com o desenho amostral da PNS 2013, e os estados com maior concentração de domicílios eram São Paulo (8,74%), Minas Gerais (6,11%) e Rio de Janeiro (5,9%).

A região sudeste apresentou a maior quantidade de moradias adequadas (84,43%) e a norte, a maior porcentagem de domicílios semiadequados (61,73%). A maior parte dos domicílios tinha pelo menos cinco cômodos (75,29%) e era habitada por dois a quatro moradores.

As casas também representaram o tipo majoritário entre os domicílios que tinham pelo menos um gato (95,84%). Quando comparados em sua própria categoria, os domicílios do tipo cabeça-de porco tiveram maior percentual de presença de gatos (10,53% do total) do que os apartamentos (5,72%).

As regiões Nordeste e Norte concentraram, juntas, mais da metade dos domicílios com pelo menos um gato neste banco de dados pré-processado. A maioria deles era semiadequada para moradia e estava localizada em zona urbana. São Paulo, Ceará e Rio Grande do Sul foram os estados com maior presença de gatos nos domicílios e Espírito Santo teve o menor número de domicílios com gatos.

A Tabela 1 apresenta a comparação entre as principais características de todos os domicílios do banco de dados pré-processado com aqueles que têm pelo menos um gato e os que não têm gato.

Tabela 1 – Principais características demográficas e estruturais dos domicílios (% e número) dos domicílios com pelo menos um gato e dos domicílios sem gatos. Banco de dados pré-processado da PNS 2013.

Característica	**Todos os domicílios** n = 64.348	**Domicílios com gatos** n = 12.168	**Domicílios sem gatos** n = 52.180
Região			
Nordeste	29,3% (19.431)	35,2% (4.282)	29,0% (15.149)
Sudeste	23,8% (15.250)	16,6% (2.021)	25,4%(13.229)
Norte	21,5% (13.846)	26,7% (3.246)	20,3% (10.600)
Sul	13,0% (7.839)	11,4% (1.383)	12,4% (6.456)
Centro-Oeste	12,2% (7.982)	10,1 (1.236)	12,9% (6.746)
Zona			
Urbana	82,2% (52.888)	65,0% (7.912)	86,2% (44.974)
Rural	17,8% (11.460)	34,0% (4.256)	13,8% (7.204)
Tipo			
Casa	86,9% (55.917)	95,8% (11.662)	84,8% (44.255)
Apartamento	12,3% (7.937)	3,8% (454)	14,4% (7.483)
Cabeça-de-porco	0,8% (494)	0,4% (52)	0,8% (442)
Adequação para moradia			
Adequado	62,0% (39.874)	44,6% (5.431)	66,0% (34.443)
Semiadequado	38,0% (24.474)	55,4% (6.737)	34,0% (17.737)
Número de cômodos			
≥ 5 (máx. de 29)	75,3% (48.449)	76,5% (9.305)	75,0% (39.144)
4	14,5% (9.370)	15,0% (1.828)	14,4% (7.542)
3	7,2% (4.626)	5,9% (720)	7,4% (3.906)
2	2,4% (1.542)	2,1% (251)	2,7% (1.291)
1	0,6% (361)	0,5% (64)	0,5% (297)
Número de moradores			
1	13,5% (8.739)	8,4% (1.022)	14,8%(7.717)
2	23,5% (15.139)	22,2% (2.699)	23,8% (12.440)
3	25,2% (16.187)	23,6% (2.869)	25,5% (13.318)
4	20,1% (12.922)	20,9% (2.546)	19,9% (10.376)
≥ 5 (máx. de 22)	17,7% (11.361)	24,9% (3.032)	16,0% (8.329)
Nº de moradores com 18+ anos			
Nenhum	0,1% (40)	0,1% (7)	0,1% (33)
1	19,7% (12.617)	13,8% (1.684)	21,0% (10.987)
2	50,5% (32.540)	48,8% (5.933)	51,0% (26.607)
3	18,3% (11.754)	21,4% (2.602)	17,5% (9.152)
≥ 4	11,4% (7.343)	15,9% (1.942)	10,4% (5.401)

Fonte: Carvalho (2020, p. 32).

3.1.2 Características dos moradores

No total, mais da metade dos responsáveis pelo domicílio era do sexo masculino (55,68%). A maioria dos respondentes (ambos os sexos) declarou-se da cor parda,

tinha entre 40-59 anos, era casada e não tinha instrução ou não completou o Ensino Fundamental. Na distribuição por classe social pelo critério ABEP adaptado, os moradores estavam mais concentrados nas classes B2, C1 e C2.

Também entre os domicílios que tinham pelo menos um gato, mais da metade dos responsáveis era do sexo masculino (55,18%). As características dos moradores quanto a raça, idade, estado civil e nível de educação eram semelhantes às do banco de dados pré-processado geral. Os moradores dos domicílios com gatos estavam mais concentrados nas classes C2 e C1 da classificação ABEP adaptada.

A Tabela 2 apresenta a comparação entre as principais características dos moradores de todos os domicílios do banco de dados pré-processado com aqueles que têm pelo menos um gato e os que não têm gato.

Tabela 2 – Características socioeconômicas dos moradores (% e número) dos domicílios com presença de gato e sem gatos. Banco de dados pré-processado da PNS 2013 (continua).

Característica	**Todos os domicílios** n = 64.348	**Domicílios com gatos** n = 12.168	**Domicílios sem gatos** n= 52.180
Sexo			
Feminino	44,3% (28.519)	44,2% (5.381)	44,3% (23.138)
Masculino	55,7% (35.829)	55,8% (6.787)	55,7% (29.042)
Raça			
Parda	49,0% (31.555)	53,4% (6.499)	48,0% (25.056)
Branca	39,3% (25.311)	35,5% (4.315)	40,2% (20.996)
Preta	10,0% (6.456)	9,5% (1.154)	10,2% (5.302)
Amarela	0,9% (550)	0,7% (91)	0,9% (459)
Indígena	0,8% (476)	0,9% (109)	0,7% (367)
Faixa etária			
Até 17 anos	0,2% (157)	0,2% (24)	0,2% (133)
18 a 29 anos	13,2% (48.519)	8,8% (1.064)	14,3% (7.455)
30 a 39 anos	22,1% (14.194)	19,0% (2.318)	22,8% (11.876)
40-59 anos	41,6% (26.758)	45,0% (5.478)	40,8% (21.280)
Acima de 60 anos	22,9% (14.720)	27,0% (3.284)	21,9% (11.436)
Nível de educação			
Sem instrução/Fundamental incompleto	44,6% (28.683)	58,9% (7.164)	41,2% (21.519)
Médio completo/Superior incompleto	28,2% (18.175)	20,1% (2.451)	30,1% (15.724)
Fundamental Completo/Médio incompleto	14,2% (19.141)	13,5% (1.643)	14,4% (7.498)
Superior completo	13,0% (78.349)	7,5% (910)	14,3% (7.439)
Estado civil			
Casado	44,7% (28.731)	46,0% (5.601)	44,3% (23.130)
Solteiro	36,6%(23.557)	36,4% (4.429)	36,7% (19.128)
Viúvo	9,6% (6.198)	10,3% (1.254)	9,5% (4.944)
Divorciado/Separado	9,1% (5.862)	7,3% (884)	9,5% (7.978)

Tabela 2 – (conclusão) Características socioeconômicas dos moradores (% e número) dos domicílios com presença de gato e sem gatos. Banco de dados pré-processado da PNS 2013.

Característica	**Todos os domicílios** n = 64.348	**Domicílios com gatos** n = 12.168	**Domicílios sem gatos** n= 52.180
*Classe social**			
A	6,6% (4.236)	4,1% (497)	7,1% (3.739)
B1	10,9% (7.033)	6,4% (785)	12,0% (6.248)
B2	24,0% (15.470)	17,8% (2.171)	25,5% (13.299)
C1	21,7% (13.952)	20,5% (2.494)	22,0% (11.458)
C2	26,9% (17.266)	32,7% (3.984)	25,5% (13.282)
D-E	9,9% (6.391)	18,4% (2.237)	7,9% (4.154)
Acesso à internet			
Sim	41,7% (26.798)	30,0% (3.659)	44,3% (23.139)
Não	58,3% (37.550)	70,0% (8.509)	55,7% (29.041)
Presença de empregado doméstico mensalista			
Sim	5,9% (3.818)	3,0% (446)	4,0% (3.372)
Não	94,1% (60.530)	97,0% (11.722)	96,0 (48.808)
Responsável é aposentado			
Sim	23,9% (15.392)	27,5% (3.352)	23,1% (12.040)
Não	76,1% (48.956)	72,5% (8.816)	76,9% (40.140)
Presença de morador com deficiência (intelectual, física, auditiva ou visual)			
Sim	9,7% (6.220)	12,1% (10.699)	9,1% (47.429)
Não	90,3% (58.128)	87,9% (1.466)	90,9% (4.751)

Fonte: Carvalho (2020, p. 33 e 34).
* Adaptação do Critério de Classificação Econômica Brasil (CCEB) da Associação Brasileira das Empresas de Pesquisa (ABEP).

3.1.3 Bens de conforto nos domicílios

A maioria dos domicílios do banco de dados pré-processado tinham um refrigerador, uma máquina de lavar roupas, uma televisão em cores e dois telefones celulares. Esse cenário foi similar nos domicílios que tinham pelo menos um gato, com exceção do item máquina de lavar roupa, que estava ausente em mais da metade desses lares (57,86%). Computadores, carros e motocicletas também não estavam presentes na maior parte dos domicílios do banco de dados pré-processado, tivessem eles gatos ou não.

A Tabela 3 apresenta a comparação entre o número de bens de todos os domicílios do banco de dados pré-processado com aqueles que têm pelo menos um gato e os que não têm gato.

Tabela 3 – Comparativo do número de bens de conforto (% e número) dos domicílios com presença de gato e domicílios sem gatos. Banco de dados pré-processado da PNS 2013.

Número de bens	**Todos os domicílios** n = 64.348	**Domicílios com gatos** n = 12.168	**Domicílios sem gatos** n= 52.180
Refrigeradores			
Nenhum	3,8% (2.441)	5,2% (636)	3,5% (1.805)
1	89,0% (57.295)	86,3% (10.498)	89,6% (46.797)
2	6,6% (4.257)	7,7% (939)	6,4% (3.318)
≥ 3 (máx. de 9)	0,6% (275)	0,8% (74)	0,5% (201)
Micro-ondas			
Nenhum	55,1% (35.462)	67,2% (8.181)	52,3% (27.281)
1	44,3% (28.483)	32,2% (3.922)	47,1% (24.561)
2 ou 3	0,6% (349)	0,5% (58)	0,5% (291)
Máquina de lavar roupa			
Nenhum	46,4% (29.855)	57,9% (7.041)	43,72% (22.814)
1	52,2% (33.616)	40,6% (4.941)	54,95% (28.675)
≥ 2 (máx. de 5)	1,4% (877)	1,5% (177)	1,3% (663)
Televisores em cores			
Nenhum	3,5% (2.288)	4,3% (522)	3,4% (1.766)
1	53,1% (34.155)	57,5% (7.001)	52,0% (27.154)
2	29,5% (19.015)	27,1% (3.300)	30,1% (15.715)
3	10,0% (6.419)	8,2% (1.000)	10,4% (5.419)
≥ 4 (máx. de 9)	3,8% (2.471)	2,9% (345)	41% (297)
Computador			
Nenhum	54,4% (34.961)	65,4% (7.952)	51,8% (27.009)
1	33,3% (21.444)	26,4% (3.212)	34,9% (18.232)
2	8,3% (5.344)	5,4% (659)	9,0% (4.685)
≥ 3 (máx. de 9)	4,0% (1.777)	2,8% (227)	4,3% (1.550)
Telefone celular			
Nenhum	9,5% (6.127)	12,0% (1.463)	8,9% (4.664)
1	27,7% (17.802)	28,1% (3.420)	27,6% (14.382)
2	32,7% (21.032)	29,7% (3.614)	33,4% (17.418)
3	17,0% (10.948)	16,5% (2.003)	17,1% (8.945)
≥ 4 (máx. de 9)	13,1% (8.439)	13,7% (1.668)	13,0% (6.771)
Carro			
Nenhum	60,0% (38.621)	67,8% (8.250)	58,2% (30.371)
1	31,4% (20.184)	25,9% (3.149)	32,6% (17.035)
2	7,02 (4.416)	5,0% (608)	7,5% (3.908)
≥ 3 (máx. de 9)	1,6% (1.027)	1,3% (161)	1,7% (866)
Motocicleta			
Nenhum	77,7% (49.972)	71,6% (8.716)	79,0% (41.256)
1	19,8% (12.766)	24,6% (2.993)	18,7% (9.773)
2	2,2% (1.419)	3,3% (399)	2,0% (1.020)
≥ 3 (máx. de 9)	0,3% (191)	0,5% (60)	0,3% (131)

Fonte: Carvalho (2020, p. 35).

3.1.4 Presença e número de animais

Quanto à existência de animais, a maioria dos entrevistados (54,6%) afirmou ter pelo menos um animal no domicílio. No total de domicílios com e sem animais, os cães foram a espécie mais presente (44,29%), seguida pelo gato (18,9%), ave (9,67%) e peixe (2,06%). A presença de cães foi ainda mais expressiva quando observada somente entre os domicílios com animais; destes, 81,1% tinham cão, 34,6%, gato, 17,7%, ave e 3,8%, peixe (Tabela 4).

Tabela 4. Presença de animais por espécie (% e número) entre os domicílios em que há pelo menos um animal. Banco de dados pré-processado da PNS 2013.

Espécie	**% de domicílios**	**Nº de domicílios** n = 35.133
Cão	81,1	28.502
Gato	34,6	12.168
Ave	17,7	6.223
Peixe	3,8	1.328

Fonte: Carvalho (2020, p. 36).

Os cães também formam o maior grupo em relação ao número de indivíduos, e as aves ultrapassaram o gatos em quantidade (Tabela 5).

Tabela 5. Quantidade total de animais por espécie. Banco de dados pré-processado da PNS 2013.

Espécie	**Nº de animais**
Cão	52.266
Gato	23.322
Ave	32.777
Peixe	14.488

Fonte: Carvalho (2020, p. 36).

3.1.5 Presença e número de gatos

Os cães estavam presentes na maioria dos domicílios com gatos do banco de dados pré-processado (52%), e aproximadamente metade desses domicílios (50,56%) tinha apenas um gato e um cão (Tabela 6). O número de gatos por domicílio variou entre 1 (7.429 domicílios do banco de dados pré-processado) a 51 (um domicílio do banco de dados pré-processado). Os lares com apenas um gato foram maioria (Tabela 7).

Tabela 6. Presença de animais de outras espécies em domicílios com gatos. Banco de dados pré-processado da PNS 2013.

Gatos x outras espécies	**% de domicílios***	**Nº de animais**
Somente gato	32,1	3.913
Gato e cão	52,0	6.333
Gato e ave	3,0	367
Gato e peixe	0,4	51
Gato, cão e ave	10,5	1.281
Gato, cão e peixe	1,0	124
Gato, ave e peixe	0,1	12
Gato, cão, ave e peixe	0,7	87

Fonte: Carvalho (2020, p. 37).
* n = 12.168

Tabela 7. Número de gatos por domicílio em domicílios com gatos. Banco de dados pré-processado da PNS 2013.

Gatos x outras espécies	**% de domicílios***	**Nº de animais**
1 gato	61,0	7.429
2 gatos	20,0	2.430
3 gatos	7,6	932
4 gatos	4,3	518
≥ 5 gatos	7,0	859

Fonte: Carvalho (2020, p. 37).
* n = 12.168

Nos domicílios com até quatro gatos, o número de cães diminui à medida que o número de gatos aumentou. Aproximadamente metade deles (50,56%) tinha apenas um gato e um cão (Tabela 8).

Tabela 8. Porcentagem do números de gatos versus número de cães em domicílios com gatos do banco de dados*.

Nº de gatos	**1 cão**	**2 cães**	**3 cães**	**4 cães**	**5 cães**	**≥ 6 cães**
1 gato	50,56	27,35	12,01	5,16	2,45	2,47
2 gatos	35,54	29,78	17,47	8,91	2,85	5,46
3 gatos	32,69	28,51	20,0	8,96	4,03	5,82
4 gatos	32,43	28,11	15,95	10,81	6,49	6,22
5 gatos	29,26	23,58	17,90	8,73	10,92	9,61
≥ 6 gatos	27,30	24,86	15,68	11,89	8,65	11,62

Fonte: Carvalho (2020, p. 38).
* n = 12.168

3.3 BENCHMARK

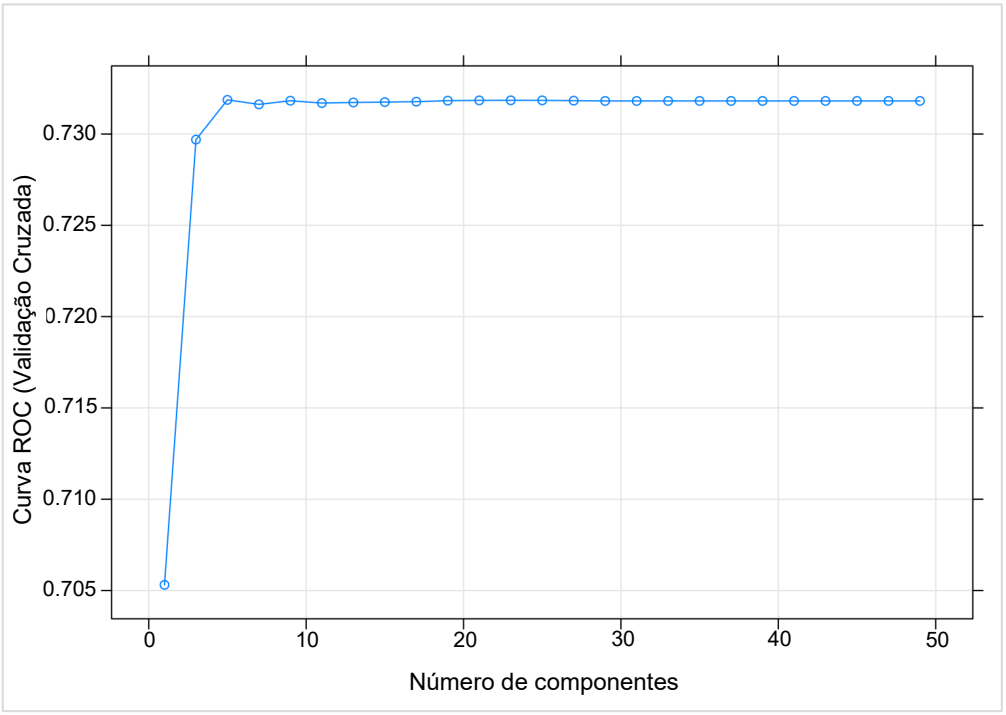
O valor obtido de AUC-ROC para o modelo de classificação foi 0,4989051 e o de RMSE para o modelo de regressão foi de 1,310875.

3.4 TREINAMENTO DOS ALGORITMOS

3.4.1 Modelos PLS

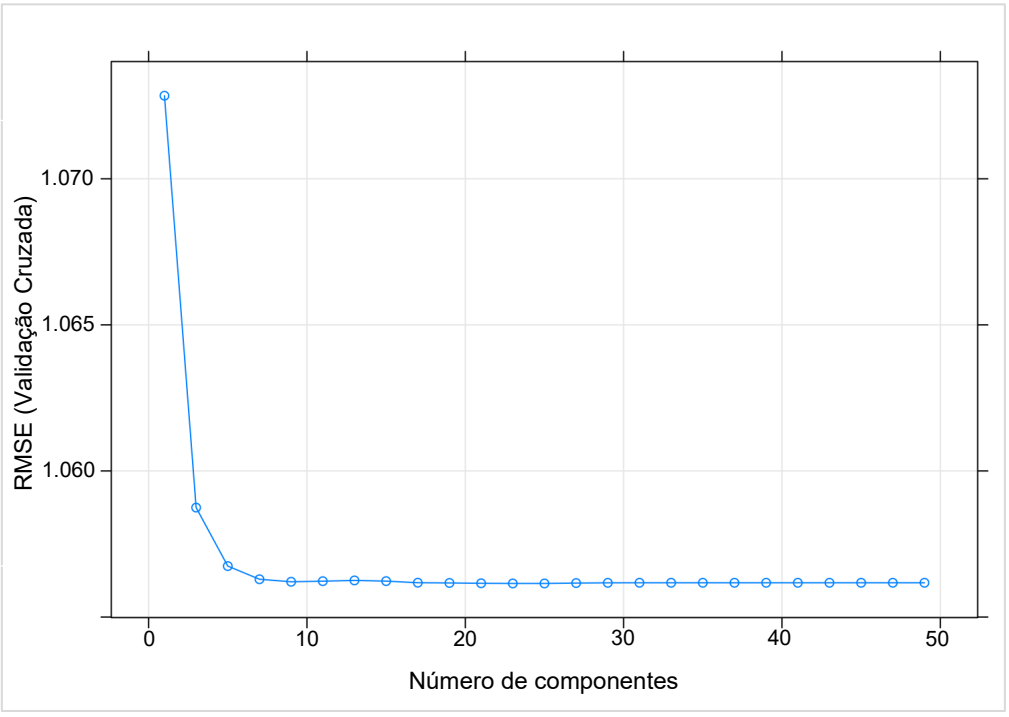
O modelo PLS de classificação (presença de gatos) teve melhor desempenho com o uso de 5 componentes principais. O valor de AUC-ROC obtido para 5 componentes foi de 0,7318704 (Figura 1). O uso de 25 componentes principais mostrou menor erro para o modelo de regressão (número de gatos) na fase de treinamento, com RMSE igual a 1,056149 (Figura 2).

Figura 1. Valores de AUC-ROC e número de componentes principais durante treinamento e ajuste do hiperparâmetro do modelo PLS de classificação.



Fonte: Carvalho (2020, p. 39).

Figura 2. Valores de RMSE e número de componentes principais durante treinamento e ajuste do hiperparâmetro do modelo PLS de regressão.

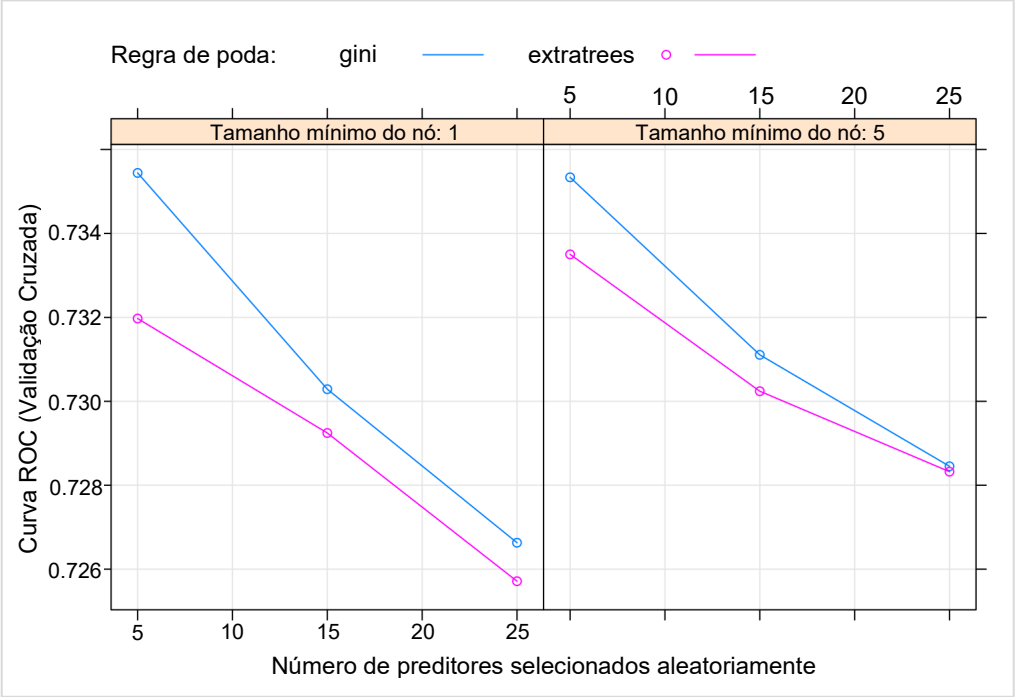


Fonte: Carvalho (2020, p. 39).

3.4.2 Modelos Random Forest

No treinamento do modelos Random Forest de classificação, o valor de AUC-ROC foi superior com o uso de 5 variáveis preditoras selecionadas randomicamente, tamanho mínimo de nó igual a 1 e técnica de poda gini. O valor de AUC-ROC com esses hiperparâmetros foi de 0,7354413 (Figura 3).

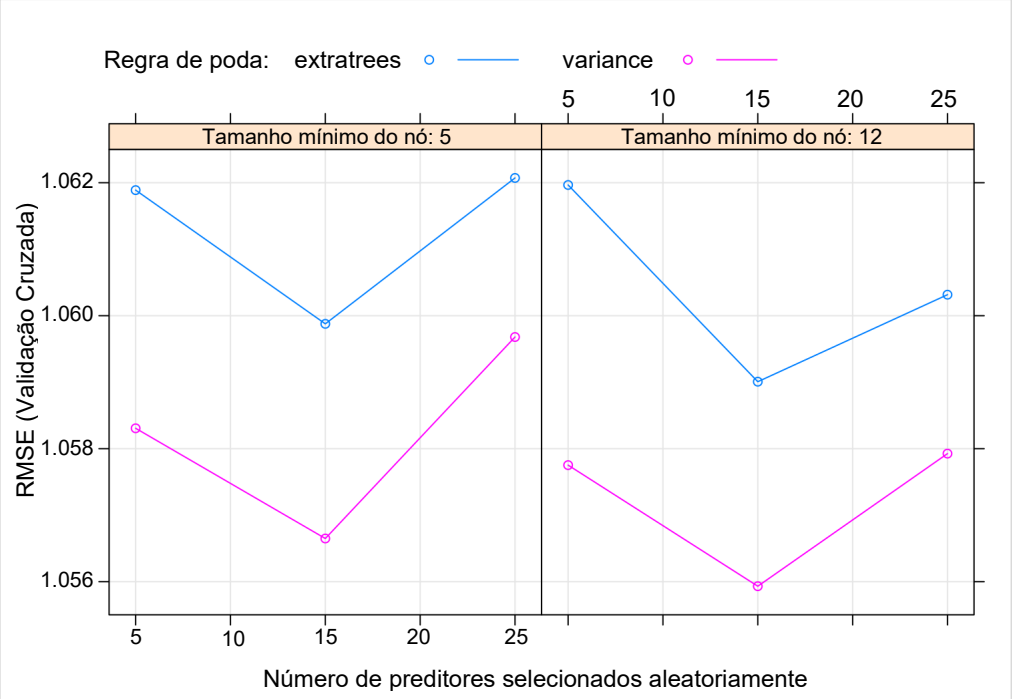
Figura 3. Valores de AUC-ROC e número de variáveis preditoras e valor mínimo de nós durante treinamento e ajuste de hiperparâmetros do modelo Random Forest de classificação com uso das técnicas de divisão gini e extratrees.



Fonte: Carvalho (2020, p. 40).

No modelo de regressão, as árvores de decisão que utilizaram 15 variáveis preditoras selecionadas aleatoriamente, tamanho mínimo de nós igual a 12 nós e técnica de poda *variance* apresentaram melhor desempenho, com RMSE igual a 1,055931 (Figura 4).

Figura 4. Valores de RMSE e número de variáveis preditoras e valor mínimo de nós durante treinamento e ajuste de hiperparâmetros do modelo Random Forest de regressão com uso das técnicas de poda gini e extratrees.



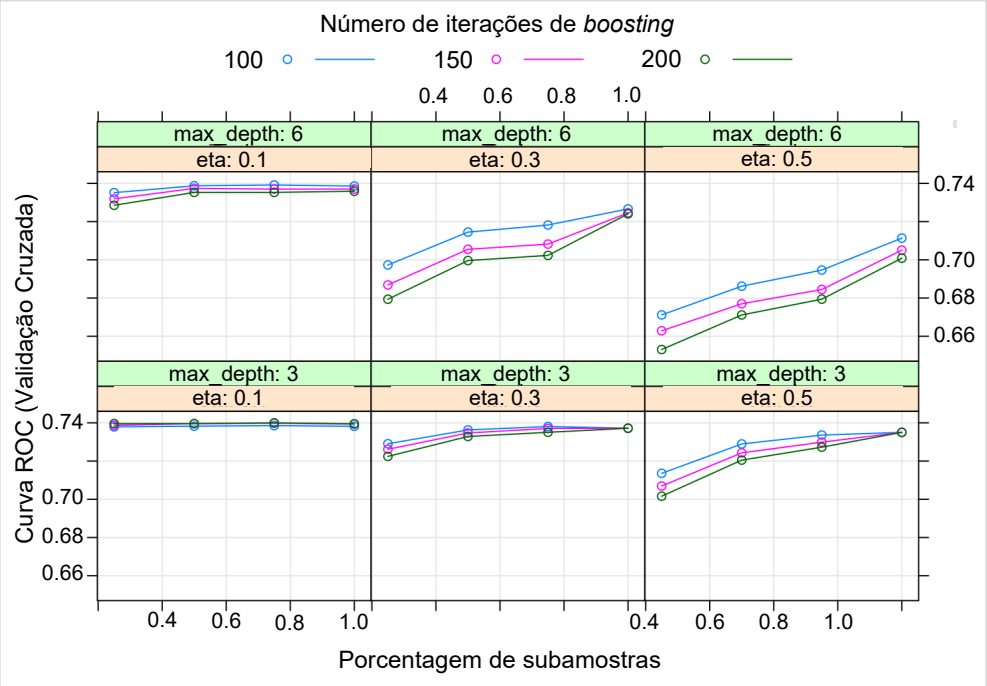
Fonte: Carvalho (2020, p. 41).

3.4.3 Modelos XGBoost

O modelo XGBoost de classificação apresentou melhor desempenho preditivo com os hiperparâmetros nrounds = 200 , max_depth = 3, eta = 0,1 e subsample = 0,75. Para o ajuste de escolha, o valor de AUC-ROC foi de 0,7400088 (Figura 5).

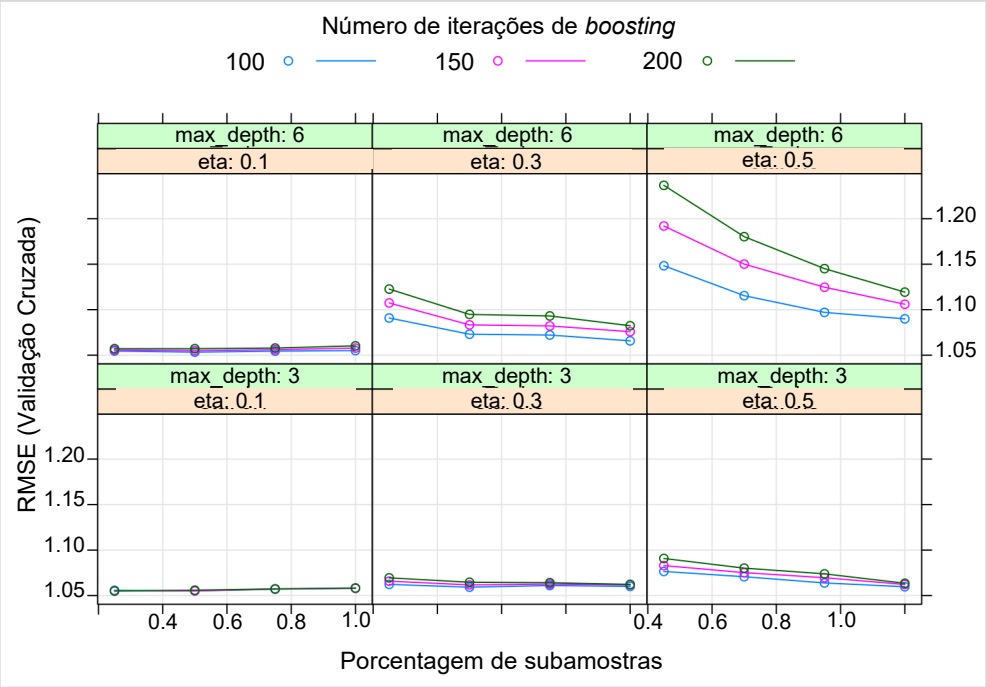
No treinamento do modelo XGBoost de regressão, os valores para hiperparâmetros que geraram melhor desempenho do modelo foram nrounds = 100, max_depth = 6, eta = 0.1e subsample = 0,5. O melhor ajuste apresentou valor de RMSE de 1,053319 (Figura 6).

Figura 5. Valores de AUC-ROC e hiperparâmetros durante treinamento e ajuste de hiperparâmetros do modelo XGBoost de classificação.



Fonte: Carvalho (2020, p. 42).

Figura 6. Valores de RMSE e hiperparâmetros durante treinamento e ajuste de hiperparâmetros do modelo XGBoost de regressão.



Fonte: Carvalho (2020, p. 42).

3.3 TESTE DOS MODELOS

Tabela 9 - Comparativo de valores de AUC-ROC entre os algoritmos PLS, Random Forest e XGBoost de classificação.

Modelo de classificação	**Teste**
PLS	0,53
Random Forest	0,54
XGBoost	0,57
Benchmark	0,50

Fonte: Carvalho (2020, p. 43).

Tabela 10 - Comparativo de valores de MCC entre os algoritmos PLS, Random Forest e XGBoost de classificação.

Modelo de classificação	**Teste**
PLS	0,20
Random Forest	0,15
XGBoost	0,23
Benchmark	0,00

Fonte: Carvalho (2020, p. 43).

Tabela 11 - Comparativo de valores de RMSE entre os algoritmos PLS, Random Forest e XGBoost de regressão.

Modelo de regressão	**Teste**
PLS	1,12
Random Forest	1,12
XGBoost	1,12
Benchmark	1,30

Fonte: Carvalho (2020, p. 43).

3.4 IMPORTÂNCIA DAS VARIÁVEIS

3.4.1 Variáveis utilizadas na construção dos modelos

Após o pré-processamento do banco de dados, 49 variáveis foram utilizadas para a construção dos modelos. Foram elas:

- Adequação do domicílio para moradia
- Aderência do responsável ao serviço de saúde
- Classe social
- Domicílio é cadastrado na unidade de saúde da família
- Domicílio tem acesso à internet
- Domicílio tem água canalizada em pelo menos um cômodo
- Estado civil do responsável
- Estado de saúde do responsável
- Faixa etária do responsável
- Frequência com que responsável vai a consultas médicas
- Frequência com que responsável vai a consultas odontológicas
- Frequência quem o domicílio recebe visita de agente de endemias
- Gênero do responsável
- Nível de educação
- Número de aparelhos de vídeo/DVD
- Número de cães
- Número de carros
- Número de celulares
- Número de cômodos
- Número de computadores
- Número de dormitórios
- *Número de gatos* (variável resposta)
- Número de linhas de telefone fixo
- Número de máquinas de lavar
- Número de micro-ondas
- Número de moradores
- Número de motocicletas
- Número de refrigeradores

- Número de residentes com mais de 18 anos
- Número de televisores a cores
- Origem da água usada para beber
- Presença de cães no domicílio
- *Presença de gatos no domicílio* (variável resposta)
- Presença de morador com algum tipo de deficiência
- Raça do responsável
- Região do país em que o domicílio está localizado
- Responsável tem doença crônica
- Responsável é beneficiário do INSS
- Responsável estava empregado
- Responsável foi internado nos últimos 12 meses
- Responsável mora com cônjuge/parceiro
- Responsável recebeu atendimento médico nas duas últimas semanas
- Responsável se afastou de atividades por motivos de saúde nas últimas duas semanas
- Responsável tem plano de saúde
- Responsável tem rendimentos de investimento
- Responsável já teve dengue
- Situação do domicílio (urbana/rural)
- Tipo de domicílio
- Unidade federativa

3.4.1 Modelos de classificação

Tabela 12 – Dez primeiros preditores selecionados pelo modelo Partial Least Squares de classificação, organizadas por ordem decrescente de importância.

Posição	Variável preditora	Importância (%)
1ª	Número de cães	100
2ª	Situação do domicílio (rural ou urbana)	89,76
3ª	Presença de cães	87,47
4ª	Adequação do domicílio para moradia	63,05
5ª	Nível de instrução do responsável	50,96
6ª	Número de moradores	47,88
7ª	Domicílio do tipo apartamento	47,44
8ª	Classe social E	46,98
9ª	Número de moradores com mais de 18 anos	46,85
10ª	Acesso à internet no domicílio	45,33

Fonte: Carvalho (2020, p. 46).

Tabela 13 – Dez primeiros preditores selecionados pelo modelo Random Forest de classificação, organizadas por ordem decrescente de importância.

Posição	Variável preditora	Importância (%)
1ª	Número de cães	100
2ª	Número de cômodos	85,18
3ª	Número de moradores	76,06
4ª	Número de aparelhos celulares	67,36
5ª	Número de moradores com mais de 18 anos	57,51
6ª	Número de dormitórios	50,19
7ª	Presença de cães	49,59
8ª	Situação do domicílio	47,63
9ª	Número de televisores em cores	46,40
10ª	Número de aparelhos de DVD	39,48

Fonte: Carvalho (2020, p. 46).

Tabela 14 – Dez primeiros preditores selecionados pelo modelo XGBoost de classificação, organizadas por ordem decrescente de importância.

Posição	Variável preditora	Importância (%)
1ª	Número de cães	100
2ª	Situação do domicílio (rural ou urbana)	44,02
3ª	Tipo de domicílio	15,74
4ª	Adequação do domicílio para moradia	15,62
5ª	Número de moradores	10,97
6ª	Responsável possui plano de saúde	9,16
7ª	Nível de instrução do responsável	8,94
8ª	Número de moradores com mais de 18 anos	8,58
9ª	Número de micro-ondas	7,37
10ª	Número de dormitórios	6,45

Fonte: Carvalho (2020, p. 47).

Tabela 15 - Preditores selecionados por dois ou mais modelos, classificados por valor da média ponderada calculada com o valor de AUC-ROC como peso.

Variável preditora	Média ponderada	PLS	RF	XGBoost
Número de cães	100	100	100	100
Situação do domicílio (rural ou urbana	59,95	89,76	47,6	44,02
Presença de cães	44,62	87,47	49,59	-
Número de moradores com 18+anos	37,12	46,85	57,5	8,58
Adequação do domicílio para moradia	25,73	63,05	-	15,62
Tipo de domicílio	20,74	47,44	-	15,74
Nível de instrução do responsável	19,52	50,96	-	8,94
Número de dormitórios	18,86	-	50,19	6,45

Fonte: Carvalho (2020, p. 47).
Legenda: PLS: Partial Least Squares; RF: Random Forest; XGB: XGBoost.

3.4.3 Modelos de regressão

Tabela 16 - Dez primeiros preditores selecionados pelo modelo Partial Least Squares de regressão, organizadas por ordem decrescente de importância.

Posição	Variável preditora	Importância (%)
1ª	Número de cães	100
2ª	Situação do domicílio (rural ou urbano)	59,18
3ª	Presença de cães	55,13
4ª	Adequação do domicílio para moradia	37,29
5ª	Nível de instrução do responsável	30,93
6ª	Nº de moradores com mais de 18 anos	27,73
7ª	Domicílio do tipo apartamento	27,48
8ª	Classe social ABEP	26,26
9ª	Acesso à internet no domicílio	26,16
10ª	Número de micro-ondas	24,75

Fonte: Carvalho (2020, p. 48).

Tabela 17 - Dez primeiros preditores selecionados pelo modelo Random Forest de regressão, organizadas por ordem decrescente de importância.

Posição	Variável preditora	Importância (%)
1ª	Número de cães	100
2ª	Número de cômodos	48,41
3ª	Número de moradores	36,63
4ª	Número de aparelhos celulares	32,43
5ª	Número de moradores com mais de 18 anos	31,19
6ª	Situação do domicílio - urbana	28,50
7ª	Número de dormitórios	21,03
8ª	Presença de cães	20,79
9ª	Número de televisores a cores	20,75
10ª	Número de refrigeradores	17,89

Fonte: Carvalho (2020, p. 48).

Tabela 18 - Dez primeiros preditores selecionados pelo modelo XGBoost de regressão, organizadas por ordem decrescente de importância.

Posição	Variável preditora	Importância (%)
1ª	Número de cães	100
2ª	Situação do domicílio (rural ou urbana)	19,59
3ª	Número de moradores com mais de 18 anos	13,71
4ª	Número de cômodos	13,32
5ª	Número de moradores	9,51
6ª	Número de refrigeradores	6,24
7ª	Número de dormitórios	5,84
8ª	Número de carros	5,55
9ª	Responsável com mais de 60 anos	5,29
10ª	Número de aparelhos celulares	5,00

Fonte: Carvalho (2020, p. 49).

Tabela 19 - Preditores selecionados por dois ou mais modelos, classificados por valor da média ponderada calculada com o valor de RSME como peso.

Variável preditora	Média ponderada	PLS	RF	XGBoost
Número de cães	100	100	100	100
Situação do domicílio (rural ou urbana)	35,75	59,18	28,5	19,59
Número de moradores com 18+anos	24,21	27,73	31,19	13,71
Presença de cães	25,30	55,13	20,79	-
Número de cômodos	20,58	-	48,41	13,32
Número de moradores	15,38	-	36,63	9,51
Número de celulares	12,48	-	32,43	5
Número de dormitórios	8,96	-	21,03	5,84
Número de refrigeradores	8,04	-	17,89	6,24

Fonte: Carvalho (2020, p. 49).
Legenda: XGB: XGBoost; RF: Random Forest; PLS: Partial Least Squares.

4 DISCUSSÃO

Por meio do uso de três algoritmos de aprendizado de máquina supervisionado, este estudo identificou preditores da presença e número de gatos nos domicílios brasileiros e os classificou de acordo com sua contribuição ao desempenho preditivo dos modelos construídos.

Os resultados mostram que todos os algoritmos utilizados tiveram desempenho preditivo superior ao de *benchmarks* aleatórios. Em comparação com seus respectivos *benchmarks*, as máximas reduções de erro dos modelos preditivos foram de 2,33 pontos percentuais para AUC-ROC e 0,12 para RMSE, e valor de MCC de 0,19.

Ao pressupor que tais reduções sejam insatisfatórias, duas hipóteses decorrem: a primeira, de que, mesmo que os alguns preditores sejam os mais influentes nos modelos gerados, eles não são os principais determinantes para a presença e o número de gatos nos domicílios; e a segunda, de que esses preditores mais influentes estejam entre os principais determinantes da presença e do número de gatos nos domicílios, porém, por serem tantos os determinantes relevantes, um subconjunto tão pequeno como o utilizado só permite aos modelos uma melhora modesta na capacidade de discriminar domicílios com ou sem gatos ou de predizer a quantidade de gatos. Por outro lado, pode-se presumir que, frente à complexidade que determina o convívio entre humanos e gatos, os preditores avaliados representam apenas um pequeno subconjunto, mas, mesmo assim, conseguiram reduzir o erro, o que representa um ganho importante para o entendimento dos fatores que diferenciam domicílios com ou sem gatos, bem como aqueles com menor ou maior número desses animais.

Independentemente de qual interpretação seja adotada, este estudo evidenciou que a coabitação entre humanos e gatos é definida por mais fatores do que apenas as variáveis consideradas neste trabalho e, por mais que estas mudem, o número de domicílios com gatos e a quantidade desses animais por domicílio não mudarão substancialmente. Ainda assim, alguns preditores selecionados pelos modelos podem contribuir na discussão sobre os determinantes da presença e número de gatos nos lares brasileiros.

Todos os modelos de classificação e de regressão compartilharam três dos dez preditores de maior importância para cada um deles: ‘número de cães’, ‘situação

do domicílio' e 'número de moradores com mais de 18 anos'. Destes, a variável 'número de cães' foi a mais importante nos três modelos.

A coabitação entre gatos e cães nos domicílios já foi observada em estudos anteriores. Downes, Canty e More (2009) relataram a presença de cães como um possível preditor da presença de gatos em um estudo realizado na Irlanda, no qual as famílias com cães de estimação se mostraram mais propensas a ter um gato e vice-versa. Uma observação semelhante foi feita por Carvelli, Iacoponi e Scaramozzino (2016) em uma região central da Itália. Lá, cerca de 58% dos lares que possuíam pelo menos um gato também abrigavam pelo menos um cão. E, embora Westgath et al. (2010) tenham constado uma divisão na preferência entre as duas espécies em casas com apenas um animal no Reino Unido, também notaram que nos domicílios com dois ou mais cachorros ou dois ou mais gatos, era comum encontrar ambos.

Os motivos para esse cenário, entretanto, não estão totalmente esclarecidos. Desconhece-se, por exemplo, se uma espécie pode determinar a presença da outra — ou se causas comuns determinam a presença de ambas, ou se o cão tem alguma influência sobre a crescente popularidade do gato como animal de estimação e de qual maneira. Além disso, os determinantes da presença de cães e gatos nos domicílios são complexos e envolvem aspectos culturais, históricos e comportamentais dos tutores (BRANSON et al., 2019; CROWLEY; CECCHETTI; MCDONALD, 2020). A maneira como as pessoas valorizam e consideram os animais de estimação, em particular, vem mudando desde a metade do século passado e, em muitos casos, cães e gatos são considerados membros da família e desfrutam de múltiplas relações sociais e naturais com os humanos (IRVINE; CILIA, 2017). Ademais, em um país cultural, socioeconômica e geograficamente diverso como o Brasil, é possível que os papéis desempenhados por cães e gatos domiciliados variem entre as regiões, e que esses animais nem sempre sejam mantidos em um mesmo domicílio por motivos semelhantes.

Soma-se a isso o fato de que as pessoas podem ter diferentes entendimentos sobre o status da guarda do gato. Uma parte deles é mantida nos lares por uma ampla gama de motivos, enquanto outra simplesmente vive ao lado de pessoas, com vários graus de interação. Estudos em diferentes países mostraram que, em muitas situações, pessoas que alimentam e oferecem cuidados aos gatos não os reconhecem como seu animal de estimação, mesmo quando esses animais frequentam a casa para se alimentar e dormir (TOUKHSATI; BENNETT; COLEMAN,

2007; SLATER et al., 2008; DOWNES; CANTY; MORE, 2009; TOUKHSATI et al., 2012). Como citado anteriormente, assumiu-se que os dados sobre presença e número de gatos utilizados neste estudo se referem a animais considerados membros do domicílios pelos respondentes. Deste modo, não é possível garantir que todos os animais declarados sejam realmente domiciliados ou, ainda, que todos aqueles animais que frequentam o domicílio foram considerados na pesquisa, fato que pode interferir na identificação de fatores domiciliares tanto para a presença como para o número de gatos.

O segundo preditor de maior importância para os modelos foi ‘situação do domicílio’, que teve importância composta acima de 50% nos modelos de classificação e acima de 35% nos de regressão. Apesar de sua relevância para os modelos preditivos, não foi possível compreender de qual maneira essa variável determina a presença e número de gatos nos domicílios apenas com base nos dados utilizados neste estudo. Notou-se, porém, que as situações rural e urbana são diferentes de forma relevante em termos da proporção de domicílios com gatos. Enquanto 65% dos lares com gatos do banco de dados pré-processado estavam localizados em zona urbana, a proporção domicílios com gatos foi maior na zona rural (0,15 gato/domicílio na zona urbana contra 0,34 gato/domicílio na zona rural). Carvelli, Iacoponi e Scaramozzino (2016) também encontraram uma maior proporção domicílios com gatos em uma região semirrural na Itália, em que os gatos domiciliados viviam, principalmente, ao ar livre em propriedades que possuíam outras espécies de animais. No Brasil, essa questão necessita de mais investigações, pois os determinantes da presença e número de gatos nos domicílios rurais podem variar desde características estruturais (por exemplo, casas maiores e presença de outros animais) ao papel do gato no domicílio (como, por exemplo, controle de pragas).

Os resultados dos modelos também permitem pressupor que os preditores ‘número de moradores’ e ‘número de moradores com mais de 18 anos’ possam ter alguma influência sobre a presença e o número de gatos. No banco de dados pré-processado, a presença de gatos foi menor nos domicílios com apenas um morador e maior nos domicílios com cinco ou mais moradores; já o número de gatos não aparentou ser influenciado pelo número de moradores. Esse padrão também foi observado na variável ‘número de moradores com mais de 18 anos’, um subconjunto de 'número de moradores'. Em estudos realizados anteriormente, a associação entre número de moradores e presença de gatos variou e a questão permanece não

esclarecida. Baquero et al. (2018), no Brasil, Murray et al. (2015), no Reino Unido, e Carvelli, Iacoponi e Scaramozzino (2016), na Itália, não encontraram correlações entre o número de moradores e a presença de gatos, ao contrário de Westgath et al. (2010), no Reino Unido.

Outros preditores relacionadas a características do domicílio e a bens de conforto selecionadas pelos modelos sugerem, ao menos, duas possibilidades que podem coocorrer. Primeiro, essas variáveis, assim como a presença e número de gatos, podem estar parcialmente determinadas pelo tamanho do domicílio (por exemplo, em domicílios maiores, tende a haver mais cômodos, mais moradores e mais celulares). Segundo, podem estar parcialmente determinadas por um padrão de consumo nos domicílios, que se reflete na presença de diferentes itens, entre eles, os gatos.

Os achados aqui apresentados provêm de um estudo quantitativo e populacional, limitado por poucas opções fechadas de resposta, e pelo fato de a PNS 2013 não ter como objetivo principal a compreensão dos determinantes do convívio domiciliar entre humanos e gatos. Os estudos qualitativos possibilitam compreensões diferentes, mais detalhadas e diversas, embora não populacionais — a rigor, estudos qualitativos podem ser feitos em amostras representativas, mas isso, muitas vezes, não é viável. Sendo assim, a realização de estudos qualitativos e alguns dos seus achados podem ser levados a estudos amostrais para revelar sua dimensão populacional. Por exemplo, estudos como de Zito et al. (2015), que encontrou associações entre a percepção de guarda e a provisão de cuidados e interações, e de Staats, Wallace e Anderson (2008), que identificou alguns motivos para manutenção de animais de companhia, poderiam oferecer novas pistas sobre os determinantes da convivência entre pessoas e gatos.

Este estudo se limitou ao uso de três algoritmos de aprendizado de máquina supervisionado e não foram testadas todas as possíveis combinações de hiperparâmetros disponíveis para cada um deles. Ainda assim, os desempenhos preditivos dos modelos criados foram semelhantes, sugerindo que eventuais ganhos de desempenho seriam modestos no caso de novas combinações de hiperparâmetros e não teriam efeito sobre as conclusões. Deve-se pontuar que não existem regras definidas para seleção de um algoritmo e não é possível supor que uma técnica terá um desempenho melhor na resolução de um determinado tipo de problema. Como demonstraram Fernandez-Delgado et al. (2014), o desempenho de vários algoritmos

de aprendizado supervisionado é, muitas vezes, semelhante, e as variáveis preditoras relevantes tendem a tornar o desempenho preditivo insensível à escolha do algoritmo.

Outro importante ponto de atenção é que a criação de novos preditores e avaliação de diferentes subconjuntos de variáveis não foram exploradas à exaustão neste estudo e, sendo assim, é possível que novas estratégias de pré-processamento dos dados levem à identificação de preditores mais relevantes, ou mesmo à melhora do desempenho preditivo dos modelos construídos.

5 CONCLUSÕES

O presente estudo identificou e classificou diferentes preditores da presença e número de gatos nos domicílios brasileiros a partir do uso de três algoritmos de aprendizado de máquina.

O desempenho pouco satisfatório dos modelos preditivos pode estar associado ao fato de que os preditores mais relevantes podem não ser os principais determinantes da variável resposta ou, de forma contrária, que estes estejam entre os principais determinantes, mas por serem muitos os determinantes relevantes, o subconjunto utilizado não foi suficiente para que a predição dos modelos fosse muito superior à uma predição aleatória. Isso oferece uma maior compreensão sobre os achados de estudos anteriormente, pois alguns dos preditores aqui identificados foram apontados previamente na literatura científica, mas sem informações sobre seus desempenhos preditivos. Esta pesquisa mostrou que tais preditores têm capacidade bastante limitada para discriminar domicílios com e sem gatos ou com diferente número de gatos. Contudo, seus desempenhos representam um avanço no entendimento dos determinantes do convívio domiciliar entre humanos e gatos.

O livre acesso aos dados coletados pela Pesquisa Nacional de Saúde do IBGE em 2013, que contém informações sobre a presença e número de gatos, além de informações socioeconômicas e demográficas da população brasileira, permitiu a construção dos modelos preditivos apresentados. Esses dados podem ser utilizados, futuramente, para explorar novos subconjuntos de preditores ou algoritmos, bem na condução de estudos regionais.

Pesquisas qualitativas podem expandir os conhecimentos sobre os determinantes relacionados à presença e número dos gatos nos domicílios brasileiros, como os identificados neste estudo, ou, ainda, seus achados podem ser utilizados em novos estudos preditivos.

REFERÊNCIAS

ABDI, H. Partial Least Square Regression. In: SALKIND, N. (Ed.). **Encyclopedia of Measurement and Statistics - Volume 1**. 1ª ed. Los Angeles, Califórnia: Sage Publishing, 2007. p. 1416.

ABEP. **Alterações na aplicação do Critério Brasil**. Disponível em: <http://www.abep.org/criterioBr/01_cceb_2019.pdf>. Acesso em: 16 jul. 2020.

AYYADEVARA, V. K. **Pro Machine Learning Algorithms: A Hands-On Approach to Implementing Algorithms in Python**. 1ª ed. Berkeley, CA: Apress, 2018.

BAQUERO, O. S. et al. Companion animal demography and population management in Pinhais, Brazil. **Preventive Veterinary Medicine**, v. 158, n. January, p. 169–177, 2018. Disponível em: <https://doi.org/10.1016/j.prevetmed.2018.07.006>.

BAQUERO, O. S.; QUEIROZ, M. R. Size, spatial and household distribution, and rabies vaccination coverage of the Brazilian owned-dog population. **Transboundary and Emerging Diseases**, p. 1–8, 2019.

BERNSTEIN, P. L. The human-cat relationship. In: ROCHLITZ, I. (Ed.). **The welfare of cats**. 1ª ed. Dordrecht, Holanda: Springer, 2007. p. 47–89.

BOEHMKE, B.; GREENWELL, B. **Hands-On Machine Learning with R**. 1ª ed. Flórida: Chapman and Hall/CRC, 2019.

BOULESTEIX, A. L. et al. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 2, n. 6, p. 493–507, 2012.

BRANSON, S. M. et al. Biopsychosocial factors and cognitive function in cat ownership and attachment in community-dwelling older adults. **Anthrozoos**, v. 32, n. 2, p. 267–282, 2019.

BUUREN, T. Van et al. **Package "mice": Multivariate Imputation by Chained Equations**, 2019. Disponível em: <https://cran.r-project.org/web/packages/mice>.

CARVELLI, A.; IACOPONI, F.; SCARAMOZZINO, P. A Cross-Sectional Survey to Estimate the Cat Population and Ownership Profiles in a Semirural Area of Central Italy. **BioMed Research International**, v. 2016, 2016.

CHEN, T.; GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System**. Disponível em: <https://arxiv.org/abs/1603.02754>. Acesso em: 18 jul. 2020.

CHICCO, D.; JURMAN, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. **BMC Genomics**, v. 21, n. 1, p. 1–13, 2020.

CROWLEY, S. L.; CECCHETTI, M.; MCDONALD, R. A. Our wild companions: domestic cats in the anthropocene. **Trends in Ecology and Evolution**, v. 35, n. 6, p. 477–483, 2020. Disponível em: <https://doi.org/10.1016/j.tree.2020.01.008>.

CUTLER, A.; CUTLER, D. R.; STEVENS, J. R. Random Forest. In: ZHANG, C.; MA, Y. (Ed.). **Ensemble Machine Learning: Methods and Applications**. 1ª ed. Nova Iorque: Springer Science & Business Media, 2012. p. 157–175.

DOWLE, M. et al. **Package "data.table": Extension of "data.frame"**, 2020. . Disponível em: <https://cran.r-project.org/web/packages/data.table/data.table.pdf>.

DOWNES, M.; CANTY, M. J.; MORE, S. J. Demography of the pet dog and cat population on the island of Ireland and human factors influencing pet ownership. **Preventive Veterinary Medicine**, v. 92, n. 1–2, p. 140–149, 2009.

FACELI, K. et al. **Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina**. 1ª Edição ed. Rio de Janeiro: Gen LTC, 2017.

FEDIAF. **European Facts & Figures**, 2018. Disponível em: <http://www.fediaf.org/images/FEDIAF_Facts__and_Figures_2018_ONLINE_final.pdf>. Acesso em: 1 jul. 2019.

FERNÁNDEZ-DELGADO, M. et al. Do we need hundreds of classifiers to solve real world classification problems? **Journal of Machine Learning Research**, v. 15, p. 3133–3181, 2014.

FILHO, A. P. S. **Tamanho , distribuição espacial e cobertura vacinal de gatos domiciliados no Brasil**. 2020. Universidade de São Paulo, 2020.

GFK. Pet ownership global survey. n. May, p. 1–82, 2016. Disponível em: <http://www.gfk.com/fileadmin/user_upload/website_content/Global_Study/Documents/Global-GfK-survey_Pet-Ownership_2016.pdf>.

GORMAN, B. **Package "mltools": Machine Learning Tools**, 2018. Disponível em: <https://cran.r-project.org/web/packages/mltools/mltools.pdf>.

IBGE. **Pesquisa Nacional de Saúde - PNS**. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?edicao=9161&t=sobre>. Acesso em: 20 jul. 2018.

IBGE. **Síntese de indicadores sociais: uma análise das condições de vida da população brasileira: 2019**. [s.l: s.n.]. Disponível em: <https://servicodados.ibge.gov.br/Download/Download.ashx?http=1&u=biblioteca.ibge.gov.br/visualizacao/livros/liv101678.pdf>.

IBGE; ABINPET. **População de animais de estimação no Brasil**. Disponível em: <http://www.agricultura.gov.br/assuntos/camaras-setoriais-tematicas/documentos/camaras-tematicas/insumos-agropecuarios/anos-anteriores/ibge-populacao-de-animais-de-estimacao-no-brasil-2013-abinpet-79.pdf>.

IRVINE, L.; CILIA, L. More-than-human families: Pets, people, and practices in multispecies households. **Sociology Compass**, v. 11, n. 2, p. 1–13, 2017.

KAMAKURA, W.; MAZZON, J. A. Critérios de estratificação e comparação de classificadores socioeconômicos no Brasil. **Revista de Administração de Empresas**, v. 56, n. 1, p. 55–70, 2016.

KASSAMBARA, A. **Package "ggpubr": "ggplot2" Based Publication Ready Plots**, 2020. . Disponível em: <https://cran.r-project.org/web/packages/ggpubr/ggpubr.pdf>.

KUHN, M. et al. **Package "caret": Classification and Regression Training**, 2020. . Disponível em: <https://cran.r-project.org/web/packages/caret>.

KUHN, M.; JOHNSON, K. **Applied predictive modeling**. 1ª ed. New York: Springer, 2013.

KULLARNI, V. Y.; SINHA, P. K. Random Forest Classifier: A Survey and Future Research Directions. **International Journal of Advanced Computing**, v. 36, n. 1, p. 1144–1156, 2013.

MAGNABOSCO, C. **População domiciliada de cães e gatos em São Paulo: perfil obtido através de um inquérito domiciliar multicêntrico**. 2006. Universidade de São Paulo, 2006.

MARTINS, C. M. et al. Impact of demographic characteristics in pet ownership: Modeling animal count according to owners income and age. **Preventive Veterinary Medicine**, v. 109, n. 3–4, p. 213–218, 2013.

MATEOS-APARICIO, G. Partial least squares (PLS) methods: Origins, evolution, and application to social sciences. **Communications in Statistics - Theory and Methods**, v. 40, n. 13, p. 2305–2317, 2011.

MURRAY, J. K. et al. Assessing changes in the UK pet cat and dog populations: Numbers and household ownership. **Veterinary Record**, v. 177, n. 10, p. 259, 12 set. 2015.

OLSON, R. S. et al. PMLB: A large benchmark suite for machine learning evaluation and comparison. **BioData Mining**, v. 10, n. 1, p. 1–13, 2017.

R CORE TEAM. **R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna**Viena, 2020.

RAMÓN, M. E.; SLATER, M. R.; WARD, M. P. Companion animal knowledge, attachment and pet cat care and their associations with household demographics for residents of a rural Texas town. **Preventive Veterinary Medicine**, v. 94, n. 3–4, p. 251–263, maio 2010.

ROSIPAL, R.; KRÄMER, N. Overview and recent advances in partial least squares. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 3940 LNCS, p. 34–51, 2006.

SLATER, M. R. et al. Cat and dog ownership and management patterns in central Italy. **Preventive Veterinary Medicine**, v. 85, n. 3–4, p. 267–294, 2008.

SOUZA-JÚNIOR, P. R. B. de et al. Desenho da amostra da Pesquisa Nacional de Saúde 2013. **Epidemiologia e Serviços de Saúde**, v. 24, n. 2, p. 207–216, 2015.

STAATS, S.; WALLACE, H.; ANDERSON, T. Reasons for companion animal guardianship (pet ownership) from two populations. **Society and Animals**, v. 16, n. 3, p. 279–291, 2008.

TOUKHSATI, S. R. et al. Semi-ownership and sterilisation of cats and dogs in Thailand. **Animals**, v. 2, n. 4, p. 611–627, 2012.

TOUKHSATI, S. R.; BENNETT, P. C.; COLEMAN, G. J. Behaviors and attitudes towards semi-owned cats. **Anthrozoos**, v. 20, n. 2, p. 131–142, 2007.

WEI, T. et al. **Package "corrplot": Visualization of a Correlation Matrix**, 2017. Disponível em: <https://cran.r-project.org/web/packages/corrplot>.

WESTGARTH, C. et al. Factors associated with cat ownership in a community in the UK. **Veterinary Record**, v. 166, n. 12, p. 354–357, 20 mar. 2010.

WICKHAM, H. **Easily Install and Load the "Tidyverse"**, 2019. Disponível em: <https://cran.r-project.org/web/packages/tidyverse/tidyverse.pdf>.

WICKHAM, H. et al. **Package "ggplot2": Create Elegant Data Visualisations Using The Grammar of Graphics**CRAN R Project, , 2020. Disponível em: <https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf>.

ZITO, S. et al. Cat ownership perception and caretaking explored in an internet survey of people associated with cats. **PLoS ONE**, v. 10, n. 7, p. 1–21, 2015.