PAULA RIBEIRO SANTOS

SELEÇÃO DE VARIÁVEIS PARA REGRESSÃO LOGÍSTICA EM UM EXEMPLO DE SEGURANÇA E FREQUÊNCIA ALIMENTAR

LAVRAS – MG

2020

PAULA RIBEIRO SANTOS

SELEÇÃO DE VARIÁVEIS PARA REGRESSÃO LOGÍSTICA EM UM EXEMPLO DE SEGURANÇA E FREQUÊNCIA ALIMENTAR

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para obtenção do título de Mestre.

Dra. Izabela Regina Cardoso de Oliveira
Orientadora

Dr. Renato Ribeiro de Lima
Coorientador

LAVRAS – MG
2020

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio autor(a).

Santos, Paula Ribeiro

Seleção de variáveis para regressão logística em um exemplo de segurança e frequência alimentar / Paula Ribeiro Santos. – Lavras : UFLA, 2020.

62 p. : il.

Dissertação(metrado)–Universidade Federal de Lavras, 2020.

Orientadora: Dra. Izabela Regina Cardoso de Oliveira.

Bibliografia.

1. Lasso. 2. CART. 3. Árvore de classificação. 4. *stepwise*. I. Cardoso de Oliveira, Izabela Regina. II. Título.

PAULA RIBEIRO SANTOS

SELEÇÃO DE VARIÁVEIS PARA REGRESSÃO LOGÍSTICA EM UM EXEMPLO DE SEGURANÇA E FREQUÊNCIA ALIMENTAR

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para obtenção do título de Mestre.

APROVADA em 24 de Janeiro de 2020.

Dr. Júlio Sílvio de Sousa Bueno Filho	UFLA
Dra. Juliana Petrini	UNIFAL
Dra. Camilla Marques Barroso	UFLA

Dra. Izabela Regina Cardoso de Oliveira
Orientadora

Dr. Renato Ribeiro de Lima
Co-Orientador

LAVRAS – MG
2020

Aos meus pais,
Luiz Alberto e Marlise
À minha irmã,
Gabriela
Ao meu sobrinho,
Luiz Otávio
Dedico.

AGRADECIMENTOS

A Deus, por tudo que faz e por tudo que não tenho consciência que faz por mim.

Aos meus queridos e amados pais, Luiz Alberto e Marlise por tudo que dedicaram a mim.

A minha irmã, Gabriela que me deu o melhor presente que eu pudesse ter, o Luiz Otávio, que sem dúvidas a força que tenho e que renasce em mim vem de vocês.

Agradeço a minha orientadora Profa. Dra. Izabela Regina Cardoso de Oliveira, pela oportunidade de fazer parte deste projeto e por compartilhar comigo tantas experiências e conhecimento, a você professora, serei eternamente grata.

Ao meu coorientador Prof. Dr. Renato Ribeiro de Lima por toda ajuda dispensada à mim e ao meu projeto.

Ao Prof. Dr. Paulo César Lima, pela partilha de conhecimento, conversas e conselhos.

Aos demais professores do Departamento de Estatística desta universidade que fizeram toda diferença na minha formação acadêmica e pessoal, em especial Thelma Sáfadi, Júlio Sílvio de Sousa Bueno Filho e Tales Jesus Fernandes.

Ao Departamento de Estatística pela oportunidade de trilhar minha trajetória desde a iniciação científica, monitoria até o mestrado.

À Universidade Federal de Lavras, da qual tenho orgulho de fazer parte.

À Rosi, Rosália, Tati e Érica, por serem meus maiores exemplos como pesquisadoras e também como pessoas que sempre me incentivaram. Com toda certeza, se estou aqui hoje foi porque tive vocês. Saibam que nunca vou conseguir agradecer por tudo que fizeram por mim.

Ao amigos que estiveram presentes ao longo desta trajetória, em especial Ariane, Fernanda, Denise, Nicásio e Cristian, que foram suporte em todos os momentos, tornando a vida mais leve.

Ao Luiz Felipe de Paiva Lourenção do Departamento de Saúde da UFLA por ter cedido o banco de dados utilizado neste trabalho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

A todos que, de alguma forma, contribuíram para a realização deste trabalho.

Se eu vi mais longe foi por estar de pé sobre ombros de gigantes.
(Isaac Newton)

RESUMO

A análise de regressão linear surgiu no século XIX e, ainda hoje, é uma das técnicas estatísticas mais utilizadas em pesquisas aplicadas quando se deseja relacionar uma variável resposta, Y, com uma ou mais variáveis explicativas, X. Entretanto, quando a variável resposta não segue uma distribuição normal a utilização de modelos lineares generalizados pode ser mais apropriada. Um exemplo com grande aplicação é o modelo logístico para respostas binárias. Nessas análises, quando se tem várias variáveis explicativas faz-se necessário selecionar as que resultariam em um modelo útil e parcimonioso. Uma solução para isso pode ser utilizar a técnica de regularização Lasso, método pelo qual as estimativas dos coeficientes tendem a zero, o que implica que apenas as variáveis que afetam significativamente a variação em Y sejam consideradas no modelo. No entanto, com o aumento do número de variáveis explicativas e da complexidade dos dados, alternativas vêm surgindo, como as técnicas de *Machine Learning*. O objetivo deste trabalho foi utilizar Lasso e árvores de classificação para seleção de variáveis em modelos logísticos, utilizando um exemplo de segurança e frequência alimentar infantil. Os dados foram coletados em 581 crianças de Centros Municipais de Educação Infantil de Lavras, MG. Inicialmente, para a variável resposta frequência alimentar foram consideradas como potenciais preditoras 37 variáveis. Quando aplicadas as técnicas Lasso e árvore de classificação estas foram reduzidas para 3 e 7, respectivamente. Para a variável resposta segurança alimentar foram consideradas 19 variáveis como potenciais preditoras e após aplicação do Lasso e árvore de classificação esse número foi reduzido para 5 e 9, respectivamente. Os modelos obtidos com as variáveis selecionadas foram reduzidos por *stepwise*. Os modelos finais para cada variável resposta foram comparados pelo AIC (Critério de Informação de Akaike) e pela deviance residual. Para a variável resposta frequência alimentar, o modelo obtido a partir do Lasso apresentou menores valores de AIC e deviance residual (AIC= 107,95 e deviance = 101,95) do que aquele obtido a partir da árvore de classificação (AIC = 509,68 e deviance = 489, 68). Esse padrão também ocorreu para a variável resposta segurança alimentar. O AIC do modelo considerando Lasso foi de 273,20 e sua deviance foi 255,20, enquanto que para árvore de classificação o AIC foi 307,37 e a deviance residual foi igual a 283,37. Para esse banco de dados, os modelos que consideraram as variáveis selecionadas pela técnica Lasso apresentaram melhores resultados segundo os critérios estatísticos, mas as árvores de classificação também podem ser consideradas, uma vez que as variáveis selecionadas são de interesse do ponto de vista prático, além de gerarem resultados gráficos intuitivos e de fácil interpretação.

Palavras-chave: Lasso, CART, árvores de classificação, *stepwise*.

ABSTRACT

Linear regression emerged in the nineteenth century and it is one of the most commonly used statistical techniques in applied research when the interest lies on explain a response, Y, based on one or more explanatory variables, X. However, when the response does not follow a normal distribution, generalized linear models may be more appropriate. An example which has broad application is the logistic model for binary responses. In regression analysis, when there are several explanatory variables, it is necessary to select those that would result in a useful and parsimonious model. One solution is the Lasso regularization method, where coefficient estimates shrink to zero, implying that only variables that significantly affect the variation in Y are considered in the model. However, as the number of explanatory variables and data complexity increase, alternatives have emerged, such as Machine Learning techniques. The aim of this study is to use Lasso and Classification Trees for variable selection in logistic models, using an example of food safety and frequency in children. Data were collected from 581 children attending Centros Municipais de Educação Infantil (Municipal Centers of Early Childhood Education), in Lavras, MG, Brazil. The 37 potential predictors of food frequency were reduced to 3 and 7 when Lasso and classification tree, respectively, were applied. For the response food security, the 19 predictors were reduced to 5 and 9 after applying Lasso and classification tree, respectively. The models obtained with the selected variables through both methods were reduced using stepwise. The chosen models for each response variable were compared by AIC (Akaike Information Criterion) and residual deviance. For food frequency, the model obtained from Lasso showed lower values of AIC and residual deviance (AIC = 107.95 and deviance = 101.95) than that obtained from the classification tree (AIC = 509, 68 and deviance = 489, 68). This pattern also occurred for food security. In this case, the AIC of the model considering Lasso was 273.20 and its deviance was 255.20, while for the classification tree the AIC was 307.37 and the residual deviance was 283.37. For this dataset, the models obtained using the variables selected by Lasso presented better results according to the statistical criteria. But classification trees can also be considered, since the selected variables have practical importance and they provide intuitive and easy-to-interpret graphical results.

Keywords: Lasso, CART, classification trees, *stepwise*.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A regressão linear é uma das técnicas estatísticas mais difundidas em pesquisas aplicadas. Seu objetivo é estabelecer uma relação funcional entre uma variável resposta (dependente) e uma ou várias variáveis explicativas (independentes) (DRAPER; SMITH, 1998; GRAYBILL; IYER, 1994). Ela é a classe mais simples, porém mais restritiva de modelos de regressão, cuja aplicação adequada depende dos pressupostos de normalidade, linearidade e homocedasticidade. Mesmo com o desenvolvimento de várias técnicas, principalmente graças ao avanço computacional das últimas décadas, em muitas áreas do conhecimento os modelos normais lineares, ainda são muito utilizados.

Para algumas respostas não normais, os modelos lineares generalizados (MLGs), propostos por Nelder e Wedderburn (1972) são recomendados. Eles dão mais flexibilidade para a relação funcional entre a média da variável resposta e as variáveis preditoras e permitem que a distribuição da variável resposta pertença à família exponencial de distribuições. Um caso especial dos MLGs são os modelos de regressão logística para respostas binárias, cujo nome está associado ao uso da função de ligação logit para conectar a média da resposta ao preditor linear. Esses modelos têm vasta utilização em diversas áreas do conhecimento, mas seu uso na área de medicina e saúde é proeminente, já que possibilita a interpretação das estimativas em termos de razão de chances, uma medida muito utilizada em diversos delineamentos clínicos e epidemiológicos (AGRESTI, 2007; GIOLO, 2017).

De acordo com Ath e Fabricius (2000), a regressão é amplamente utilizada no contexto de um estudo equilibrado bem projetado, sendo eficaz para determinar quais fatores afetam uma resposta. No entanto, à medida que o número de variáveis que afeta a resposta (variáveis explicativas) e a complexidade dos dados aumentam, os modelos lineares se tornam menos eficazes. Adicionalmente, técnicas de aprendizado de máquina (do inglês “Machine Learning”) têm sido cada vez mais aplicadas na busca de associações em dados de alta dimensão, os chamados “big data”, dados não retangulares e outros conjuntos de dados complexos.

Nesse contexto, uma ferramenta muito utilizada são as árvores de classificação. De acordo com James (2013), árvore de classificação é um conjunto de regras que modela uma variável resposta categórica. O processo de construção é dado subdividindo repetidamente os dados, partindo os nós em dois caminhos. O procedimento é então realizado continuamente até que nas folhas os grupos sejam o mais homogêneo possível, cultivando a árvore ou também podando para o tamanho desejado. O CART (do inglês “Classification and Regression Trees”)

é um dos algoritmos de particionamento recursivo binário, que não foi o primeiro a ser introduzido, mas para Wu e Kumar (2009) é o primeiro a ser descrito com rigor analítico e apoiado por uma estatística sofisticada na teoria das probabilidades. Além disso, as árvores são fáceis de serem construídas, apresentam um apelo visual, o que facilita a interpretação, e podem ser usadas para classificação ou como análise exploratória. É possível encontrar na literatura o uso dessa ferramenta na seleção de variáveis (CHO; HONG; HA, 2010; TSAI; CHEN, 2009).

Em modelos de regressão múltipla, diversas variáveis explicativas (preditoras) podem estar associadas à resposta de interesse. Quando há muitas potenciais preditoras, um procedimento comum é buscar um modelo parcimonioso usando técnicas de seleção de variáveis. Para essa finalidade, geralmente utiliza-se o coeficiente de determinação, métodos *stepwise*, *forward* e *backward*, e AIC (PAULA, 2013).

No contexto de modelos de regresão os métodos de regularização ou encolhimento (do inglês "Shrinkage") são muito utilizados. Eles partem de um modelo com todos os preditores e usam uma técnica que regularizam as estimativas dos coeficientes, reduzindo-as em direção a zero. Os métodos de regularização mais comuns são Lasso e Ridge, com pequenas diferenças. O Lasso é indiferente quanto aos preditores correlacionados e tenderá a escolher um e ignorar os demais, enquanto o Ridge reduz os coeficientes preditores relacionados uns em relação aos outros, ambos impondo diferentes penalidades em seu tamanho. Lasso também é amplamente utilizado para seleção de variáveis.

Mesmo com a disponibilidade de diversas técnicas apropriadas para seleção de modelos e de variáveis no contexto de análise de regressão, é muito comum encontrar trabalhos em que as variáveis preditoras para um modelo logístico são selecionadas com base nos resultados de testes marginais, relacionando cada uma delas com a variável resposta (BATALHA et al., 2017; OLIVEIRA et al., 2011). Testes qui-quadrado, de Fisher ou Testes de Wald em modelos logísticos são aplicados separadamente para cada variável e aquelas cujas estimativas satisfazem um nível de significância pré-estabelecido são consideradas candidatas para um modelo logístico múltiplo. Esse procedimento é proposto por Hosmer, Lemeshow e Sturdivant (2013) e conhecido como seleção proposital de covariáveis (do inglês "Purposeful selection of covariates"). Uma desvantagem deste procedimento é que associação entre as variáveis preditoras e a variável resposta são analisadas marginalmente, desconsiderando a presença das demais.

Neste trabalho abordamos o problema de seleção de variáveis em modelos logísticos, considerando um exemplo de segurança e frequência alimentar de crianças. Para isso utiliza-

mos dados coletados em crianças de zero a cinco anos que frequentam os Centros Municipais de Educação Infantil (CMEIs) de Lavras, Minas Gerais. Estes têm um papel fundamental no crescimento e desenvolvimento de crianças pertencentes aos estratos sociais em condições de vulnerabilidade socioeconômica, tornando-se uma estratégia dos países subdesenvolvidos.

Diversas variáveis (fatores socioeconômicos e de saúde) foram avaliadas para cada uma das 581 crianças participantes do estudo e o objetivo é identificar aquelas que estão mais associadas à segurança e frequência alimentar dessas crianças. Esses dados foram inicialmente analisados em colaboração com Lourenção (2019), em que árvores de classificação foram utilizadas para selecionar as variáveis utilizadas nos modelo logístico. No entanto não foi realizado o pré-processamento dos dados nem a pré-seleção das variáveis clinicamente importantes antes da obtenção da árvore e ajuste dos modelos.

O principal objetivo deste trabalho é ajustar modelos logísticos considerando variáveis selecionadas pelos métodos de seleção Lasso e árvore de classificação. Como objetivos específicos, destacam-se:

- Detalhar a etapa de preparação e pré-processamento dos dados, que envolve eliminação de inconsistências, transformação e criação de variáveis;
- Aplicar a técnica de regularização Lasso para seleção de variáveis considerando o modelo logístico para cada um dos desfechos de interesse (segurança e frequência alimentar);
- Obter árvores de classificação para segurança e frequência alimentar;
- Ajustar um modelo logístico para cada desfecho (segurança e frequência alimentar) com as variáveis selecionadas por Lasso e por árvores de classificação;
- Comparar os modelos logísticos ajustados utilizando critérios estatísticos apropriados.

2 REFERENCIAL TEÓRICO

2.1 Modelos de regressão e técnicas de regularização

2.1.1 Regressão linear e estimação dos coeficientes

Para Hastie, Tibshirani e Friedman (2008), os modelos lineares foram amplamente desenvolvidos na era das estatísticas dos pré-computadores, mas mesmo na era atual ainda existem boas razões para estudá-los e usá-los. São usados em muitas pesquisas aplicadas para relacionar uma variável resposta, Y, com uma ou mais variáveis preditoras, X, por meio de uma função $f(.)$. São simples e fornecem uma descrição adequada e interpretável de como as entradas (preditoras) afetam a saída (resposta).

Tomando um vetor de entrada , $X^T = (X_1, X_2, \cdots, X_p)$ queremos prever uma saída Y. Então, o modelo de regressão linear tem a forma

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

Sendo assim, o objetivo é estimar os valores dos $\beta's$ tal que os desvios dos valores observados em relação aos estimados sejam mínimos. Existem várias formas de medir a proximidade, a mais comum e popular envolve escolher os coeficientes $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^T$ para minimizar a soma de quadrados dos erros, dado por

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j$$

e

$$SQE(\beta) \quad = \quad \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

Este método é chamado método dos mínimos quadrados. Matricialmente,

$$SQE(\beta) = (y - X\beta)^T (y - X\beta)$$

Fazendo derivada em relação a β, tem-se:

$$\frac{\partial SQE}{\partial \beta} = -2X^T(y - X\beta).$$

Igualando a zero, obtemos a solução do sistema, tendo:

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

em que $\hat{\beta}$ é o estimador de mínimos quadrados do vetor de parâmetros $\beta's$. A matriz X possui posto coluna completo, assim, a matriz $X'X$ possui inversa única.

Porém, quando se tem n pequeno e grande p deve-se reduzir os coeficientes preditores e uma forma de restringir o número de variáveis é impondo uma penalidade em seu tamanho. A solução para isso é utilizar algoritmos de regularização como Ridge e Lasso.

2.1.2 Técnicas de regularização

Ridge e Lasso, também utilizados para a seleção de variáveis, são os métodos mais comuns de regularização que, basicamente, se diferenciam na imposição da penalidade.

A regressão Ridge reduz os coeficientes impondo uma penalidade em seu tamanho. Então, a soma de quadrados dos resíduos penalizada é dada por:

$$\hat{\beta}^{ridge} = \min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{2.1}$$

em que $\lambda \geq 0$ é o parâmetro de complexidade que controla a quantidade de parâmetros. Quanto maior o valor de λ, maior é a quantidade de parâmetros.

Uma outra maneira de escrever a equação (2.1) é:

$$\hat{\beta}^{ridge} = \min_{\beta} \sum_{j=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2,$$

sujeito a $\sum_{j=1}^{p} \beta_j^2 \leq t$, em que t é um número arbitrário utilizado para seleção de variávies, o que torna explícita a restrição de tamanho nos parâmetros solucionando o problema dos coeficientes correlacionados.

Temos que a equação (2.1) é dada matricialmente por:

$$SQR(\lambda) = (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta.$$

Sendo a solução da regressão Ridge da forma:

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

em que I é a identidade $p \times p$. A solução de regressão adiciona uma constante positiva à diagonal de $X^T X$ antes da inversão, sendo de qualquer forma uma função linear de y, com a penalidade quadrática $\beta^T \beta$. O problema torna-se não singular, mesmo que $X^T X$ não seja de posto completo.

O Lasso é um método de seleção como o Ridge, com pequenas diferenças. Por exemplo, o Lasso é indiferente quanto aos preditores correlacionados e tenderá a escolher um e ignorar os demais. Seu estimador é definido por:

$$\hat{\beta}^{lasso} = \min_{\beta} \sum_{j=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} X_{ij} \beta_j)^2,$$

sujeito a $\sum_{j=1}^{p} |\beta_j| \leq t$. Também podemos escrever o Lasso na forma Lagrangeana equivalente

$$\hat{\beta}^{ridge} = \min_{\beta} \left\{ \sum_{j=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} X_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{2.2}$$

De fato, há semelhança entre a regressão Ridge e Lasso. Pode-se notar que a penalidade de Ridge $\sum_1^p \beta_j^2$ é substituída pela penalidade do Lasso $\sum_1^p |\beta_j|$, o que torna as soluções não

lineares e não há expressão de forma fechada como na regressão Ridge. Porém, há algoritmos eficientes disponíveis para a obtenção das soluções.

Em ambos os casos, tanto na regressão Ridge quanto na Lasso, o valor para o parâmetro λ pode ser obtido por meio da validação cruzada. É escolhido uma grade de valores para λ e calculado o erro de validação cruzada para cada valor de λ. Em seguida, o valor do parâmetro é escolhido de tal forma que minimize a estimativa do erro de previsão esperado. Por fim, o modelo é reajustado utilizando o valor selecionado do parâmetro de ajuste (JAMES et al., 2013).

Essas técnicas são empregadas nas mais diversas áreas do conhecimento. Tisbshirani (1997) utilizou para seleção de variáveis em modelos de Cox, baseando no contexto de modelos de regressão linear. Também muito aplicado na genética, como pode-se observar pelos trabalhos de Ayers e Cordell (2010) que exploraram o desempenho de penalização na seleção de SNPs em estudos genéticos; Ogutu, Schulz-Streeck e Piepho (2012), que avaliaram o desempenho dessa técnica para seleção genômica e Waldmann et al. (2013) que utilizaram para estudos de associação genômica. No campo da energia elétrica, Uniejewski, Nowotarski e Weron (2016) selecionou variáveis de previsão de preço de eletricidade

2.2 Árvores de classificação

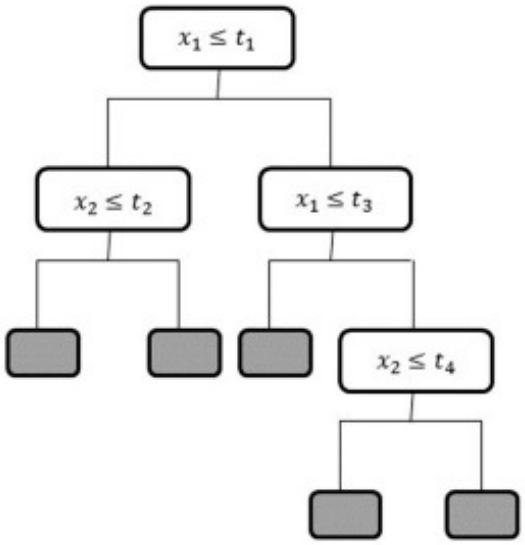
Nas mais diversas pesquisas realizadas, o objetivo, em um conjunto de variáveis que podem ser denotadas como entradas, que são medidas ou predefinidas, verificar se estas têm alguma influência sobre uma ou mais saídas. Na literatura estatística, de acordo com Hastie, Tibshirani e Friedman (2008), as entradas são muitas vezes chamadas de preditoras, classicamente variáveis independentes e as saídas chamadas de respostas ou classicamente as variáveis dependentes. Portanto, prever os valores de uma ou mais variáveis dependentes para um dado conjunto de variáveis independentes é chamado de "aprendizado supervisionado". O contrário, é conhecido por "aprendizado não supervisionado". Neste caso, tem-se um conjunto de N observações $(x_1, x_2, \cdots, x_N)$ tendo por objetivo inferir e fornecer respostas baseando-se apenas nesse vetor aleatório. Diversas técnicas de *Machine Learning* são propostas na literatura. Neste trabalho será descrito uma delas, baseadas em aprendizado supervisionado: árvores de decisão.

Segundo Agresti (2013), o método proposto por Breiman et al. (1984) da árvore de classificação consiste em um processo de decisão que utiliza um conjunto sequencial de perguntas sobre os valores de x a fim de prever uma classificação para y. Graficamente (Figura 2.1), se

resume em divisões da amostra em vários nós, considerando as variáveis, para determinar a previsão.

São aplicadas a grandes bases de dados, em que regularidades implícitas devem ser descobertas automaticamente e expressas, na forma de regras e, de acordo com Wu e Kumar (2009), estão fundamentadas de forma que modelos são obtidos pela identificação de relacionamentos entre variáveis dependentes e independentes. Há duas possibilidades para as variáveis dependentes: variáveis categóricas (árvore de classificação) e variáveis contínuas (árvore de regressão), e o número de objetos (n) deve ser maior que o número de classes (p), ou seja, $n > p$, para permitir a aplicação de testes estatísticos.

Figura 2.1 – Ilustração de uma árvore de decisão com duas variáveis preditoras, X_1 e X_2, e pontos de partição t_1, t_2, t_3 e t_4.



Dessa forma, para construí-las considere o exemplo de um problema de diagnosticar pacientes, retirado de Guimarães (2018). Suponha que o médico precisa diagnosticar um novo paciente que chegou ao consultório. Primeiramente, o médico pode perguntar se o paciente tem sentido dor, que corresponderia ao nó raiz da árvore. Em seguida, dependendo da resposta obtida, ele pode perguntar se ele está tendo febre, enjoos ou se tem notado alguma mancha no corpo. Por meio dessa sequência de perguntas podemos solucionar o problema de classificação até chegar a uma conclusão e diagnosticar o paciente em doente ou saudável.

Essas perguntas e suas possíveis respostas podem ser organizadas na forma de uma árvore de decisão, que tem uma estrutura composta por nós e arestas, parecida com a forma de uma árvore. Dessa forma, a partir do nó raiz e percorrendo cada nó de decisão inicia-se o processo pela raiz da sub-árvore até chegar a um nó folha indicando a classe que corresponde ao

novo paciente. Além disso, toda a trajetória representa uma regra, facilitando a interpretabilidade do modelo.

Na construção, procura-se associar a cada nó de decisão a variável que está causando mais variabilidade nos dados. Assim sendo, segundo James, Witten, Hastie e Tibshirani (2013), dois novos ramos são criados no sentido inferior da árvore, dado que elas fazem divisões dicotômicas. O processo de divisão é continuado resultando em árvores totalmente crescidas até que um critério de parada definido seja alcançado. Uma desvantagem é que elas são instáveis uma vez que um mesmo conjunto de dados pode gerar várias árvores distintas.

A classificação pode ser utilizada tanto para modelagem descritiva, como ferramenta para distinguir diferentes classes visando a compreensão de um determinado exemplo pertencer a uma determinada classe tornando o modelo de classificação fácil de interpretar, quanto para modelagem preditiva, para predizer classes desconhecidas ou novas (TAN; STEINBACH; KUMAR, 2005).

Árvores de decisão são técnicas poderosas de *Machine Learning* e são amplamente utilizadas em problemas de classificação (KRAMER et al., 2001; PAL; MATHER, 2003; TOSCHKE; BEYERLEIN; KRIES, 2005; WAHEED et al., 2006; ZHANG, 1998) por serem expressas em linguagem natural, facilitando o entendimento por parte das pessoas. Além disso, não é necessário nenhuma pressuposição das relações entre as variáveis preditoras e a variável resposta, ou seja, são métodos não-paramétricos.

Para a seleção de variáveis não é tão comum utilizar árvores de decisão, mas Tsai e Chen (2010) utilizaram essa técnica para selecionar variáveis como uma etapa de pré-processamento dos dados. Já Cho, Hong e Ha (2010) utilizaram essa técnica com a finalidade de selecionar as variáveis.

Dentro do contexto de árvores de decisão a técnica mais utilizada para a finalidade de selecionar variáveis é *Random Forest*, usando o conceito de importância de variáveis. Várias árvores são construídas em diferentes amostras, o que sugere uma floresta (“forest”), sob um subconjunto de preditores aleatorizados, por isso o termo “random”. Há trabalhos nas diversas áreas como por exemplo em Genuer, Poggi e Tuleau-Malot (2010); Hapfelmeier e Ulm (2013); Sandri e Zuccolotto (2006); Strobl (2008); e também na saúde, no estudo de Wright, Dankowski e Ziegler (2010).

2.2.1 Algoritmos para crescimento das árvores

Há vários algoritmos na literatura para indução de árvores de decisão, os principais são: ID3 (QUINLAN, 1986), C4.5 (QUINLAN, 1993) e CART (BREIMAN et al., 1984). O ID3, um algoritmo recursivo, divide um nó considerando um único atributo, sendo escolhido pelo quanto informativo é pela propriedade de ganho de informação. Possui algumas limitações: só lida com atributos categóricos não-ordinais, não trabalha com valores ausentes e além disso, para selecionar variáveis utiliza o ganho de informação, método que não considera o número de divisões, o que acarreta em árvores complexas, que não podem ser podadas já que o algoritmo não apresenta nenhum método pós-poda.

Já o algoritmo C4.5 veio para resolver esses problemas e pendências, sendo uma significativa evolução do ID3, uma vez que lida tanto com atributos categóricos como atributos contínuos, trata valores desconhecidos, utilizada a medida de razão de ganho para selecionar o atributo que melhor divide os dados, além de apresentar um método de pós-poda.

O CART (Classification and Regression Trees), algoritmo clássico que foi proposto por Breiman et al. (1984) consiste em uma técnica não-paramétrica que acomoda tanto árvore de classificação (quando o atributo é nominal) quanto de regressão (quando o atributo é contínuo). Segundo Fonseca (1994), a produção de resultados em forma de árvores de decisão simples e legíveis, bem como a capacidade de pesquisa de relações entre os dados, mesmo quando elas não são tão evidentes são as principais virtudes do CART.

As árvores geradas por esse algoritmo são determinadas por um conjunto de condições lógicas do tipo se/então, que podem ser percorridas da raiz até as folhas, expandindo a árvore exaustivamente, realizando pós-poda por meio da redução do fator custo-complexidade produzindo árvores mais simples, precisas e com boa capacidade de generalização (BREIMAN et al., 1984).

2.2.2 Escolha dos atributos

O atributo preditivo em cada nó da árvore é definido pelo critério de seleção e existem diferentes maneiras de ser realizado. A maioria dos algoritmos de indução de árvores de decisão trabalha com partições binárias, apesar de que os nós podem ter múltiplas divisões. Logo cada nó é dividido de acordo com um único atributo, em que um melhor atributo para realizar essa divisão é encontrado.

Os critérios para selecionar a melhor divisão são baseados em medidas como impureza, distância e dependência, com o objetivo de gerar um melhor aumento na acurácia preditiva. A definição das medidas depende de qual algoritmo está sendo utilizado. Por exemplo, para o algoritmo ID3, proposto por Quinlan (1986), as medidas utilizadas são ganho de informação e razão de ganho, já para o algoritmo CART (Breiman et al., 1984), uma medida comumente utilizada é o índice de Gini.

Tomando c classes, o índice de Gini é definido pela equação:

$$\text{gini(nó)} = 1 - \sum_{1}^{c} p(i/\text{nó}),$$

em que $p(i/\text{nó})$ é a fração dos registros pertencentes à classe i no nó. Assim, basta calcular a diferença entre o gini antes e após a divisão. Essa diferença, gini, é apresentada pela equação

$$\text{gini(nó)} = \text{gini(antes)} - \sum_{1}^{n} \frac{N(v_j)}{N} \text{gini}(v_j),$$

em que n é o número de valores do atributo, ou seja, o número de nós-filhos, N é o número total de objetos do nó-pai e $N(v_j)$ é o número de exemplos associados ao nó-filho v_j. Assim, é selecionado o atributo que gerar um maior valor para o índice de gini.

Mas, de forma geral, esses algoritmos buscam dividir os dados de um nó-pai de forma a minimizar o grau de impureza dos nós-filhos. Quanto menor o grau de impureza, mais desbalanceada é a distribuição de classes. A impureza é nula se todos os exemplos nele pertencerem à mesma classe. Analogamente, o grau de impureza é máximo no nó se houver o mesmo número de exemplos para cada classe possível.

2.2.3 Métodos de poda

Depois que as árvores estão construídas, inicia-se o processo de poda. Dado que muitas das arestas ou sub-árvores podem refletir ruídos ou erros, acarretando em um problema de sobreajuste, são utilizados métodos de poda (HAN, 2001) da árvore para melhorar a taxa de acerto do modelo. Assim, busca-se, ao final, um modelo parcimonioso, com baixa heterogeneidade em seus nós finais e permitindo generalizar, tornando a árvore podada mais simples e facilitando a sua interpretação.

Assim como existem diferentes algoritmos para o método de seleção, há diversas formas de realizar poda em uma árvore de decisão. O processo de poda é análogo ao método *forward* e *backward*, em seleção de modelos de regressão. Por meio desses decide-se quando parar de construir ou quando uma árvore maximal é reduzida por meio da poda, respectivamente.

De forma geral, o processo acontece da seguinte forma: para cada nó interno da árvore, o algoritmo calcula a taxa de erro, caso a sub-árvore abaixo desse nó seja podada. Em seguida, é calculada a taxa de erro caso não haja a poda. Se a diferença entre essas duas for menor que um valor pré-estabelecido, a árvore é podada. Caso contrário, não ocorre a poda. Esse processo se repete progressivamente, gerando um conjunto de árvores podadas, até que a árvore que obtiver a melhor acurácia é a escolhida.

2.3 Modelos lineares generalizados

Apesar da disponibilidade de ferramentas estatísticas modernas e poderosas, segundo Paula (2013) os modelos normais lineares foram utilizados por muito tempo para descrever fenômenos aleatórios. Isso se deve ao fato deles envolverem propriedades simples, porém restritivas, como independência, normalidade, homogeneidade de variâncias e linearidade. Portanto, quando a resposta do fenômeno sob estudo não atende a essas pressuposições, algum tipo de transformação é sugerida a fim de alcançá-las.

Entretanto, há situações nas quais esses modelos não são apropriados, como por exemplo quando se tem uma informação a respeito da forma da relação, $f(.)$, entre a resposta e as variáveis preditoras. Com o desenvolvimento computacional nos últimos anos, com alta capacidade de processamento e desenvolvimento dos pacotes estatísticos, novos modelos foram propostos. Para Paula (2013), a proposta mais interessante e inovadora no assunto foi apresentada por Nelder e Wedderburn (1972), que propuseram os modelos lineares generalizados (MLGs). Esses são extensões dos modelos tradicionais de regressão linear e permitem que a variável resposta Y tenha várias opções de distribuições, permitindo que a mesma pertença à família exponencial de distribuições.

Os modelos lineares generalizados são formados por três componentes: o componente aleatório, que identifica a variável resposta Y e assume uma distribuição para ela; o componente sistemático, que especifica as variáveis explicativas, e a que especifica uma função do valor esperado de Y com as variáveis explicativas por meio de uma equação, a função de ligação.

A variável resposta Y é identificada como um componente aleatório de um MLG e é adotada uma distribuição de probabilidade. Tome as observações independentes em Y por $(Y_1, \cdots, Y_n)$. Se as observações de Y são binárias, como por exemplo "sucesso" ou "falha" ou se cada Y_i é o número de sucessos de um determinado número fixo de tentativas assumimos uma distribuição binomial para Y. Se tivermos que cada observação é uma contagem assumimos uma distribuição que se aplica a todos os números inteiros não negativos, como Poisson ou binomial negativa. Ou se ainda cada observação for contínua, podemos assumir uma distribuição normal para Y.

As variáveis explicativas são especificadas em um MLG como componente sistemático. Elas são consideradas linearmente como preditores na equação do modelo, tal que

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

e, assim, essa combinação linear é chamada de preditor linear.

A função de ligação $g(.)$, terceiro componente de um MLG, conecta os componentes aleatórios e sistemáticos. Logo, se $\mu = E(Y)$ é a média da distribuição de probabilidade do valor esperado de Y, então $g(.)$ relaciona μ ao preditor linearmente como

$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

A função de ligação $g(\mu) = \mu$ é a mais simples, modela a média diretamente e é chamada de ligação de identidade. Há outras funções de ligação que não sejam linearmente relacionadas aos preditores. Quando temos dados de contagem, por exemplo, a função de ligação mais aplicada é a log, $g(\mu) = \log(\mu)$, que modela o log da média. Um MLG que utiliza essa função de ligação é chamado de modelo log-linear.

Já quando temos "sucesso" ou "falha", em que a probabilidade de μ está entre 0 e 1, uma função de ligação possível é a logit, $g(\mu) = \log[\frac{\mu}{1-\mu}]$ que modela o log de uma probabilidade. Um MLG que usa a ligação logit é chamado de modelo de regressão logística.

2.3.1 O modelo de regressão logística

Em Agresti (2007), quando se tem dois resultados possíveis, 1 (“sucesso”) ou 0 (“falha”) para a variável resposta Y, sua distribuição é especificada pelas probabilidades $P(Y=1)=\pi$, de sucesso e $P(Y=0)=1-\pi$, de falha. Para n observações independentes, o número de sucessos possui distribuição binomial, de forma que

$$P(y)=\frac{n!}{y!(n-y)!}\pi^{y}(1-\pi)^{n-y},$$

sendo que cada observação binária é uma variável binomial com $n=1$.

O valor de π varia conforme o valor da variável explicativa x muda, e substituímos π por $\pi(x)$ quando queremos descrever a dependência desse valor. Essa relação entre $\pi(x)$ e x é geralmente não-linear de modo que $\pi(x)$ aumenta geralmente continuamente ou diminui continuamente em forma de S à medida que x aumenta. A função matemática com esta forma tem fórmula

$$\pi(x)=\frac{e^{\beta_0+\beta x}}{1+e^{\beta_0+\beta x}}.$$

que é chamada de função de regressão logística. Para a fórmula correspondente ao modelo de regressão logística suponha que exista uma única variável explicativa x. Para uma variável resposta binária Y, a probabilidade de “sucesso” $\pi(x)$ dado o valor x é definida tal que

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right)=\beta_0+\beta x, \tag{2.3}$$

Esse modelo de regressão logística é um caso especial de um MLG. O componente aleatório para os resultados tem uma distribuição binomial. A função de ligação é o *logit* da função $\log(\frac{\pi}{1-\pi})$ de π, em que logit pode ser qualquer número real, enquanto π é restrito ao intervalo de 0 a 1. Os preditores lineares $(\beta_0+\beta x)$ formam o componente sistemático de um MLG.

O parâmetro β determina a taxa de aumento ou diminuição da curva em forma de S para $\pi(x)$. O sinal de β indica se a curva é crescente ($\beta>0$) ou decrescente ($\beta<0$) e a taxa de

alteração aumenta à medida que $|\beta|$ aumenta. Quando $\beta = 0$, simplifica para uma constante, então, $\pi(x)$ é idêntico em todo x. Assim, a curva se reduz a uma linha reta horizontal, o que implica que a resposta binária Y é então independente de x.

Nesse modelo de regressão logística, uma das grandes vantagens é a interpretação direta dos coeficientes como medidas de associação, que utiliza as probabilidades e a razão de chances. Para as chances de resposta 1 (ou seja, "sucesso"), no modelo 2.3, são

$$\frac{\pi(x)}{1-\pi(x)} = exp(\beta_0 + \beta x) = e^{\beta_0}(e^{\beta})^x.$$

Logo, para cada aumento de 1 unidade em x, as probabilidades são multiplicadas por e^{β}.

Diversas áreas do conhecimento utilizam regressão logística, sendo proeminentes n a área da saúde. Pedraza, Rocha e Sousa (2013) ajustaram modelos logísticos para obter as razões de chances a fim de verificar a deficiência de micronutrientes em crianças assistidas pelo Núcleo de Creches, no estado da Paraíba. Batalha et al. (2017) utilizaram para predizer as chances dos fatores que estão diretamente ligados ao consumo de alimentos processados e ultraprocessados por crianças em São Luís, no Maranhão. Toloni et al. (2011) a empregaram para calcular as razões de chances de variáveis, como escolaridade materna e renda *(*per capita), em função da introdução de alimentos industrializados. Por meio de regressão logística, Durão et al. (2015) verificaram a associação entre responsabilidade materna e práticas de alimentação infantil em relação a dietas. Já Oliveira et al. (2011), a aplicaram para associação entre diversas variáveis e desnutrição em crianças cadastradas no Bolsa Família, em um município de Minas Gerais

2.4 Critérios de seleção de modelos de regressão

Determinadas as variáveis ou fatores que devem ser incluídas no modelo de regressão, resta saber qual é o melhor e qual apresenta as variáveis e interações mais importantes (PAULA, 2013). Há vários procedimentos para seleção de modelos, em que os mais conhecidos são maior R_p^2, menor s_p^2, *forward*, *backward*, *stepwise* e AIC.

O método *forward* começa com um modelo que não tem preditores, em seguida são adicionados e testados preditores, um por vez, da seguinte forma:

1. Tome o modelo nulo $\mu = \alpha$.

2. Ajustamos então um modelo para a primeira variável $\mu = \beta_0 + \beta_j x_j$ em que $(j = 1, \cdots, q)$

Testamos $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$. Sendo P o menor nível descritivo dentre os q testes. Se $P \leq P_E$, a variável correspondente entra no modelo.

3. Se, por exemplo, x_1 tenha sido escolhida, então o próximo modelo é ajustado considerando essa variável, $\mu = \beta_0 + \beta_1 x_1 + \beta_j x_j$ em que $(j = 2, \cdots, q)$

   Testamos $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$. Sendo P o menor nível descritivo dentre os $(q-1)$ testes. Se $P \leq P_E$, a variável correspondente entra no modelo.

4. O procedimento é repetido até que ocorra $P > P_E$.

O método *backward* é o contrário de *forward*. Nesse caso começa com o modelo completo, contendo todos os preditores e iterativamente remove-se os preditores um a um tal que,

1. Tome o modelo completo $\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$.

   Testamos $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$, $j = 1, \cdots, q$. Sendo P o maior nível descritivo dentre os q testes. Se $P > P_S$, a variável correspondente sai no modelo.

2. Se, por exemplo, x_1 tenha saído do modelo, então o próximo modelo é ajustado desconsiderando essa variável, $\mu = \beta_0 + \beta_2 x_2 + \cdots + \beta_q x_q$.

   Testamos $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$, $j = 2, \cdots, q$. Sendo P o maior nível descritivo dentre os $(q-1)$ testes. Se $P > P_S$, a variável correspondente sai no modelo.

3. O procedimento é repetido até que ocorra $P \leq P_S$.

Os procedimentos descritos acima consideram o nível descritivo P como critério de inclusão ou exclusão da variável no modelo. Outros critérios, como o AIC, podem ser considerados.

O método *stepwise* baseia-se nos dois procedimentos, *forward* e *backward*, numa mistura de inclusão e eliminação de variáveis explicativas até que nenhuma variável seja incluída ou retirada do modelo. Exemplos do uso dessa metodologia de seleção por *stepwise* são encontrados em Zhang (2016), que utilizou um banco de dados (bwt) presente no pacote MASS do R para exemplificar a seleção de variáveis por *stepwise*. Alves, Lotufo e Lopes (2013) aplicaram em um modelo de regressão linear múltipla para prever a demanda de carga para um melhor planejamento e distribuição de energia elétrica.

No trabalho de Bursac et al. (2008), ele afirma que em muitas situações o principal problema é escolher dentre um grande conjunto de covariáveis, aquelas que devem ser incluídas

a fim de obter o "melhor" modelo. Nesse estudo ele realiza uma simulação para comparar o desempenho do procedimento proposto por Hosmer, Lemeshow e Sturdivant (2013) , conhecido como seleção propositalde covariáveis, com três bem fundamentados métodos de seleção de variáveis, *forward*, *backward* e *stepwise*.

O método proposto por Akaike (1974), baseado no máximo da função de verossimilhança, busca o modelo que seja parcimonioso, ou seja, que tenha um número reduzido de parâmetros e que explique bem o comportamento da variável resposta. Quanto menor o seu valor, melhor o ajuste, logo, o ideal seria encontrar um número p de parâmetros que minimizasse a função $AIC = -L(\hat{\beta}) + p$, uma vez que o logaritmo da função de verossimilhança $L(\beta)$ cresce com o aumento do número de parâmetros do modelo.

De acordo com Agresti (2007) uma maneira de avaliar a qualidade do ajuste de um MLG é utilizar a deviance. Essa função é uma medida de diferença dos ajustes dos modelos corrente e saturado. O modelo corrente é o proposto, com p $<$ n parâmetros, que está entre os modelos nulo e saturado. O saturado é o modelo mais complexo, completo, com n parâmetros. Logo, um pequeno valor para a deviance significa que o modelo proposto têm um ajuste quase tão bom quanto o modelo saturado, indicando evidência de bom ajuste.

3 METODOLOGIA

3.1 Material

Os dados utilizados neste trabalho foram obtidos em parceria com o Departamento de Ciências da Saúde (DSA) da Universidade Federal de Lavras e são resultantes de um estudo transversal realizado entre abril e novembro de 2018 com crianças pré-escolares.

A seleção da amostra foi realizada proporcionalmente em cada Centro Municipal de Educação Infantil (CEMEI), sendo estes classificados e agrupados em tercis de nível socioeconômico, considerando: a) condições socioeconômicas do bairro, para obter um número homogêneo de crianças; b) cálculo do estrato por faixas etárias similares em cada grupo; c) utilizada a mesma proporção da faixa etária por cada CMEI inserido na pesquisa. Os CMEIs foram alocados em três grupos, sendo o mesmo número de escolas com piores condições socioeconômicas, de condições intermediárias e de melhores condições.

As 650 crianças foram avaliadas para elegibilidade. Dessas foram excluídas 36, por não atenderem aos critérios de inclusão, não autorizados pelos pais e desistências. Então 614 crianças foram alocadas para a coleta de dados. Porém, devido a desistência ao longo do estudo e ausência no momento da coleta, outras 33 crianças deixaram de participar do estudo.

A amostra final foi composta por 581 pré-escolares de zero a 5 anos de idade, matriculados em CMEIs de Lavras-MG. Todas as crianças receberam um questionário semi-estruturado com 49 questões (Apêndice A) para ser respondido por seus pais ou responsáveis. Este questionário visou recolher informações sobre nível socioeconômico, características relacionadas à gestação, condições de nascimento da criança e hábitos alimentares pregressos e atuais. Os tipos de variáveis encontrados em cada um desses blocos com suas respectivas quantidades são apresentados na Tabela 3.1.

Tabela 3.1 – Quantidade de variáveis no questionário aplicado segundo bloco de perguntas (socioeconômicas, gestação, condições de nascimento e hábitos alimentares) para o estudo com crianças pré-escolares matriculadas em CMEIs de Lavras, MG, realizado no período de abril a novembro de 2018.

	Tipos de variáveis			
Bloco do questionário	Binária	Nominal	Ordinal	Discreta
Socioeconômico	1	6	4	1
Gestação	4	-	-	-
Condições de nascimento	7	6	1	-
Hábitos alimentares	15	2	1	1

Fonte: Da autora (2019)

Além dessas questões investigadas no questionário, por intermédio da Escala Brasileira de Insegurança Alimentar (EBIA), proposta e validada para o Brasil por Segall-Corrêa et al. (2003), foi obtida a situação de segurança alimentar das famílias, que é uma das variáveis respostas de interesse (Questão 49 - Apêndice A). Essa escala é composta por 14 perguntas centrais dicotômicas e, em função do número de respostas sim, a criança era classificada da seguinte forma: nenhuma resposta afirmativa (segurança alimentar) de 1 a 5 (insegurança leve), de 6 a 9 (insegurança moderada) e de 10 a 14 (insegurança grave), que abordam a percepção de insegurança alimentar, relativa aos três meses precedentes à entrevista. Sendo assim, a escala tem por objetivo avaliar a preocupação de a comida acabar antes de se poder comprar mais, bem como a situação de ausência total de alimentos, na qual um morador pode permanecer um dia inteiro sem comer.

A condição de insegurança alimentar é classificada quando existe preocupação ou incerteza quanto à disponibilidade de alimentos no futuro, em quantidade e qualidade adequada. A segurança alimentar ocorre quando há acesso regular e permanente a alimentos de qualidade e em quantidade suficiente, sem incerteza quanto a sofrer restrição no futuro próximo. Dessa forma, cada criança foi classificada em uma dessas quatro categorias e para 57 participantes há ausência dessa informação. As frequências observadas segundo categoria foram: segurança alimentar (298), insegurança alimentar leve (185), insegurança alimentar moderada (26) e insegurança grave (15).

Outra variável resposta de interesse é sobre a frequência alimentar dos participantes (Questão 48 - Apêndice A). Informações sobre alimentação foram obtidas a partir de um questionário de frequência alimentar (QFA) com 19 itens alimentares divididos em seis grupos: leite e derivados, carnes e ovos, óleos/gorduras, cereais/leguminosas, doces e frutas/verduras/legumes. A frequência (1-2 vezes ao dia, 3 vezes ou mais ao dia, 1-2 vezes por semana, 3 vezes ou mais por semana, nunca/raramente) era anotada e em seguida a criança recebia uma pontuação baseado em estudos realizados no Brasil (MONDINI et al., 2007). De acordo com essa pontuação, cada criança foi classificada em uma dessas três categorias: baixa qualidade, qualidade intermediária, boa qualidade e para 55 participantes há ausência dessa informação. As frequências observadas segundo categoria foram: baixa qualidade (168), qualidade intermediária (83), boa qualidade (275).

O objetivo do estudo de Lourenção (2019) foi avaliar quais variáveis potenciais do questionário estão associadas à segurança e frequência alimentar das crianças.

3.1.1 Pré-processamento dos dados

Algumas das variáveis preditoras foram categorizadas inicialmente em vários níveis, havendo uma frequência relativamente baixa em determinadas categorias. Dessa forma, foram realizadas algumas transformações com o intuito de redução de categorias. Isso ocorreu para as seguintes variáveis: idade, situação econômica, habitação, profissão do chefe da família, local do parto, peso ao nascer, tipo de leite.

Alguns temas foram abordados em duas questões do questionário, em que a segunda era condicionada à resposta da primeira. Um exemplo são as questões sobre o teste do pezinho, em que a primeira investigava se o teste foi realizado e a segunda se o resultado foi positivo ou negativo. O mesmo aconteceu para questões do teste do olhinho, audição e em duas questões sobre a saúde de quem fica a maior parte do tempo com a criança. Em todos esses casos uma nova variável, que é a combinação das respostas das duas questões foi criada.

Além disso, a variável resposta EBIA, que a princípio tinha quatro categorias (segurança alimentar, insegurança leve, moderada e grave) foi tratada como binária (segurança ou insegurança). Isso também ocorreu para a outra variável resposta, QFA, que anteriormente tinha três categorias (baixa qualidade, qualidade intermediária, boa qualidade) e nesse estudo foi tratada como binária (boa qualidade, qualidade intermediária ou ruim).

Uma pré-seleção das variáveis foi feita junto com o pesquisador responsável pelo banco de dados, sendo consideradas como potenciais preditoras apenas aquelas que tinham associação prática com os desfechos. Para a variável resposta frequência alimentar (QFA) foram consideradas 37 variáveis, já para a variável resposta segurança alimentar (EBIA) foram consideradas 19. A codificação e descrição de todas as variáveis é apresentada no Apêndice B.

3.2 Métodos

De forma geral, para cada um dos desfechos, foi feita a seleção de variáveis por Lasso e árvore de classificação. Em seguida, os modelos logísticos foram ajustados e o modelo final foi selecionado por *stepwise*. A qualidade de ajuste dos modelos foi avaliada por envelope simulado e esses foram comparados utilizando critérios estatísticos apropriados. A descrição de cada etapa é apresentada nas seções seguintes.

3.2.1 Seleção de variáveis por Lasso

O pacote *glmnet* (Friedman, Hastie, Tibshirani, 2010) do R (Development Core Team, 2019) foi usado na aplicação do Lasso. A escolha do valor do parâmetro de complexidade, λ, foi feita por validação cruzada, conforme Zeviani e Ferreira (2017). As variáveis que apresentaram menor erro de validação cruzada por meio do λ mínimo foram selecionadas.

3.2.2 Seleção de variáveis por árvore de classificação

Árvores de classificação para os desfechos foram obtidas com base no algorimo CART, que está implementado no pacote *rpart* (THERNEAU; ATKINSON, 2018) do R (R CORE TEAM, 2019). Os gráficos das árvores foram construídos pelo pacote rpart.plot (MILBORROW, 2018). O critério utilizado na divisão dos nós foi o índice de Gini. As variáveis presentes nas árvores obtidas foram, então, selecionadas para os modelos.

3.2.3 Ajuste dos modelos logísticos

Depois de selecionadas as variáveis preditoras para os desfechos por Lasso e árvores de classificação, modelos logísticos foram ajustados. A forma geral dos modelos é dada por:

$$\text{logit}[P(Y=1)] = \log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \sum_{j=1}^{n} \beta_j x_j,$$

em que a probabilidade de "sucesso", $P(Y=1)$ é segurança e a probabilidade de "fracasso", $P(Y=0)$ é insegurança para o desfecho de segurança alimentar. Para o desfecho de frequência alimentar, a probabilidade de "sucesso", $P(Y=1)$ é boa qualidade e a probabilidade de "fracasso", $P(Y=0)$ é qualidade intermediária ou ruim. Temos ainda, que β_0 é o intercepto e β_j são os efeitos relacionados às variáveis selecionadas pelo Lasso e pela árvore de classificação que serão consideradas no modelo.

Com os modelos ajustados, foi aplicado o método *stepwise*, com o objetivo de selecionar o melhor modelo. O critério adotado no método foi menor AIC. As estimativas dos parâmetros dos modelos foram interpretadas em termos de razão de chances (RC). Intervalos a 95% de confiança para essa medida foram obtidos aplicando-se a função exponencial aos limites dos intervalos baseados na razão de verossimilhança, como apresentado em Agresti (2007).

3.2.4 Comparação das modelos

Para verificar a qualidade do ajuste foram gerados gráficos de envelope simulado, utilizando o pacote *hnp* (MORAL; HINDE; DEMÉTRIO, 2017) do R (R CORE TEAM, 2019). Para comparar os modelos ajustados com as variáveis selecionadas por ambos os métodos (Lasso e árvores de classificação) foram utilizados o Critério de Informação de Akaike (AIC) e também a deviance residual (AGRESTI, 2007).

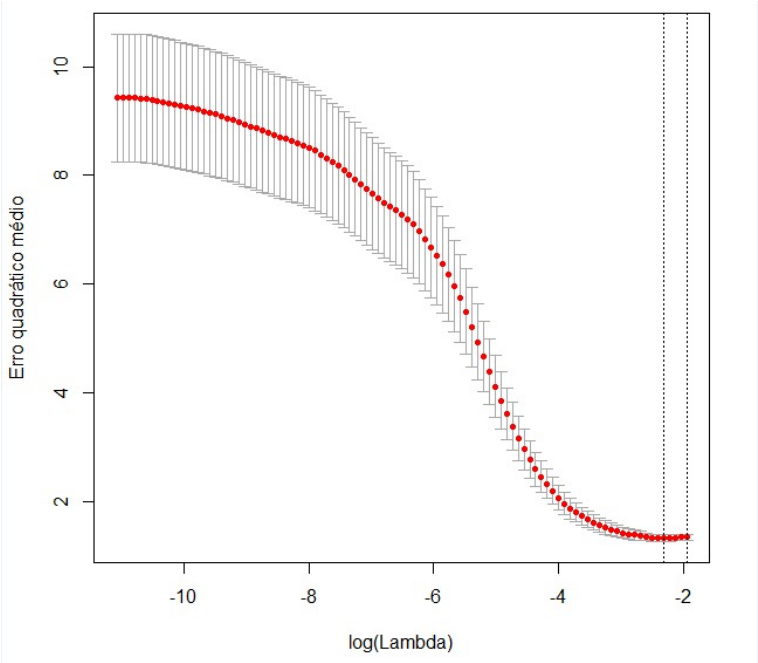
4 RESULTADOS E DISCUSSÃO

4.1 Preditores associados à frequência alimentar

4.1.1 Lasso aplicado à frequência alimentar

Para a variável resposta frequência alimentar (QFA), nota-se, pelo gráfico apresentado na Figura 4.1 que o valor de λ que minimiza o erro de validação cruzada é 0,09794 (log(0,09794) $\approx$ - 2,3234) para o modelo com penalização Lasso, ainda o maior valor de λ que está no intervalo de um erro padrão do erro quadrático médio de validação é 0,14209 (log(0,14209) $\approx$ - 1,9513).

Figura 4.1 – Valores de λ para o erro quadrático médio de validação cruzada para a variável resposta frequência alimentar.



Fonte: Da autora (2019)

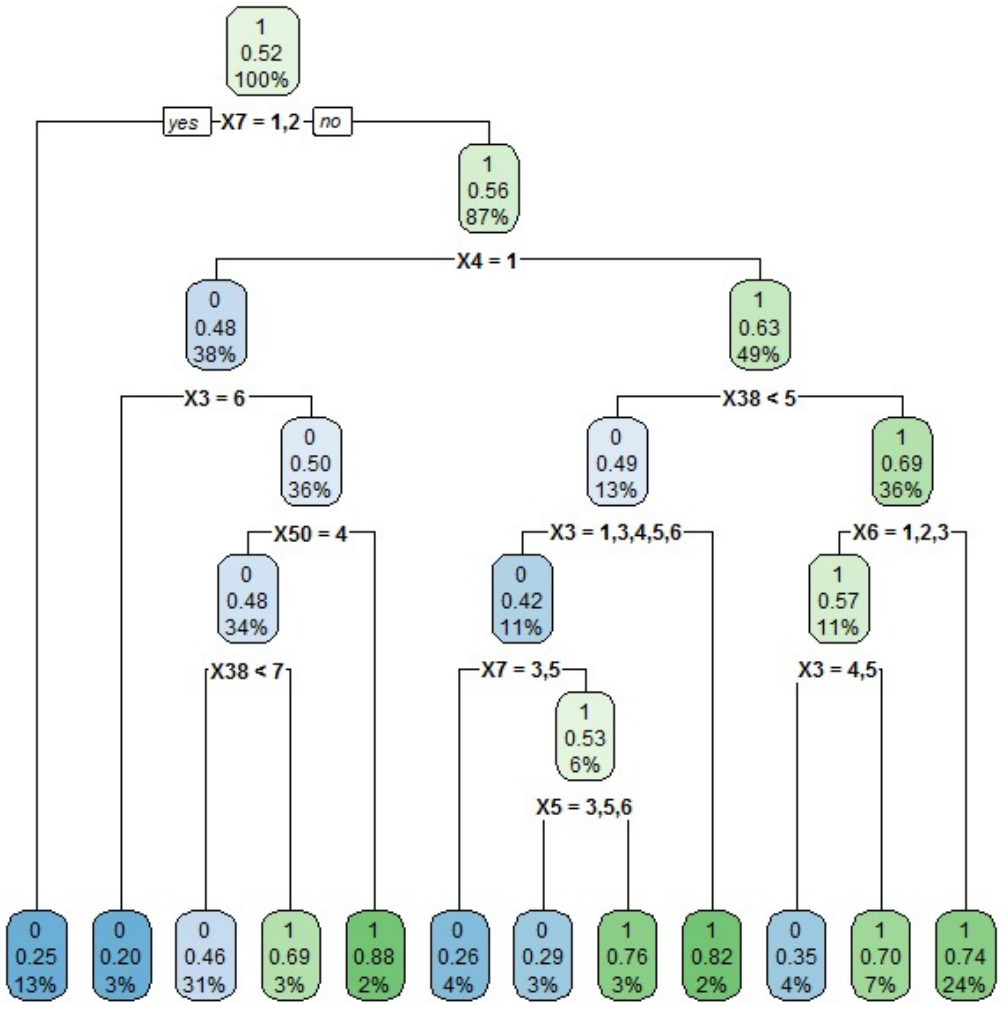
Tomando o valor de λ mínimo as variáveis selecionadas foram: idade (X3), número de refeições da criança durante o dia (X38) e uso de vitaminas nos dois primeiros anos de vida (X39).

4.1.2 Árvore de classificação aplicada à frequência alimentar

A árvore gerada é apresentada na Figura 4.2. As variáveis consideradas para ajuste do modelo logístico são aquelas presentes nos nós da árvore, sendo: idade (X3), situação econômica (X4), número de pessoas na família (X5), escolaridade do pai (X6), escolaridade da mãe

(X7), número de refeições (X38) e se quem convive com a criança quando ela não está na creche tem problemas de saúde mental (X50).

Figura 4.2 – Árvore de classificação para a variável resposta frequência alimentar.



Fonte: Da autora (2019)

A variável responsável pela primeira partição é escolaridade da mãe (X7), que é a variável que mais causa variabilidade na amostra. As interpretações dos ramos seguintes são feitas com base nas combinações de condicionantes presentes nesta amostra.

A interpretação acontece da seguinte forma: o primeiro nó é escolaridade da mãe (X7), então se ela tiver nos níveis 1 (primário incompleto) ou 2 (ginásio incompleto) então a criança é classificada em qualidade alimentar intermediária ou ruim. Caso contrário, se ela tiver nos níveis 3 (2^o grau incompleto) ou acima então o próximo nó a ser interpretado é relacionado à situação econômica (X4). Se a situação econômica da família é até um salário mínimo então a

criança é classificada em qualidade alimentar intermediária ou ruim e a próxima variável a ser estudada é idade (X3). Por outro lado, se a família recebe mais de um salário mínimo então a criança é classificada em boa qualidade alimentar e o próximo nó a ser interpretado é número de refeições (X38). Logo, percorre-se a árvore pelos nós e folhas sucessivamente até os nós de término.

4.1.3 Modelos logísticos para frequência alimentar

Os resultados dos modelos logísticos ajustados para a variável resposta frequência alimentar após aplicação do *stepwise* são apresentados na Tabela 4.1.

Tabela 4.1 – Estimativas e erros padrões para os parâmetros dos modelos logísticos, estimativas de razões de chances (RC) e respectivos intervalos de confiança a 95% para a variável resposta frequência alimentar com variáveis selecionadas por Lasso e árvore de classificação.

Lasso					
Variável	Estimativa	Erro padrão	Valor-p	RC	$IC_{95\%}$ (RC)
Refeições diárias (X38)	0,3925	0,1827	0,0317	1,481	[1,050;2,166]
Uso de vitaminas (X39)					
Se fez uso de vitaminas	1,4301	0,6656	0,0317	4,179	[1,188;17,124]
Árvore de classificação					
Variável	Estimativa	Erro padrão	Valor-p	RC	$IC_{95\%}$ (RC)
Idade (X3)					
12 a 23 meses	0,1906	0,5582	0,73276	1,210	[0,390;3,566]
24 a 35 meses	-0,6255	0,5318	0,23949	0,535	[0,180;1,487]
36 a 47 meses	0,5559	0,5248	0,28953	0,574	[0,196;1,573]
48 a 59 meses	-1,0312	0,6329	0,10324	0,357	[0,099;1,207]
mais que 60 meses	-0,7775	0,6701	0,24597	0,460	[0,119;1,680]
Situação econômica (X4)					
de 2 a 3 salários mínimos	0,6381	0,2311	0,00575	1,893	[1,205;2,985]
mais que 3 salários mínimos	1,2088	0,4580	0,00830	3,349	[1,411;8,644]
Refeições diárias (X38)	0,2458	0,0830	0,00307	1,279	[1,089;1,509]
Saúde mental do cuidador (X50)					
Não tem problemas	1,4565	0,8571	0,08925	4,291	[0,904;30,965]

Fonte: Da autora (2019)

Considerando o modelo obtido por meio do método de seleção Lasso e o nível de significância de 5%, as chances de ter uma alimentação de boa qualidade estão estatisticamente associadas ao uso de vitaminas nos 2 primeiros anos de vida e ao número de refeições diárias. Nesse último, a cada aumento de uma unidade, ou seja, uma refeição diária a mais, as chances estimadas de estar em boa qualidade aumentam 1,481 vezes. De acordo com o intervalo reportado, esse aumento é no mínimo 5% e no máximo o dobro das chances. Para as crianças

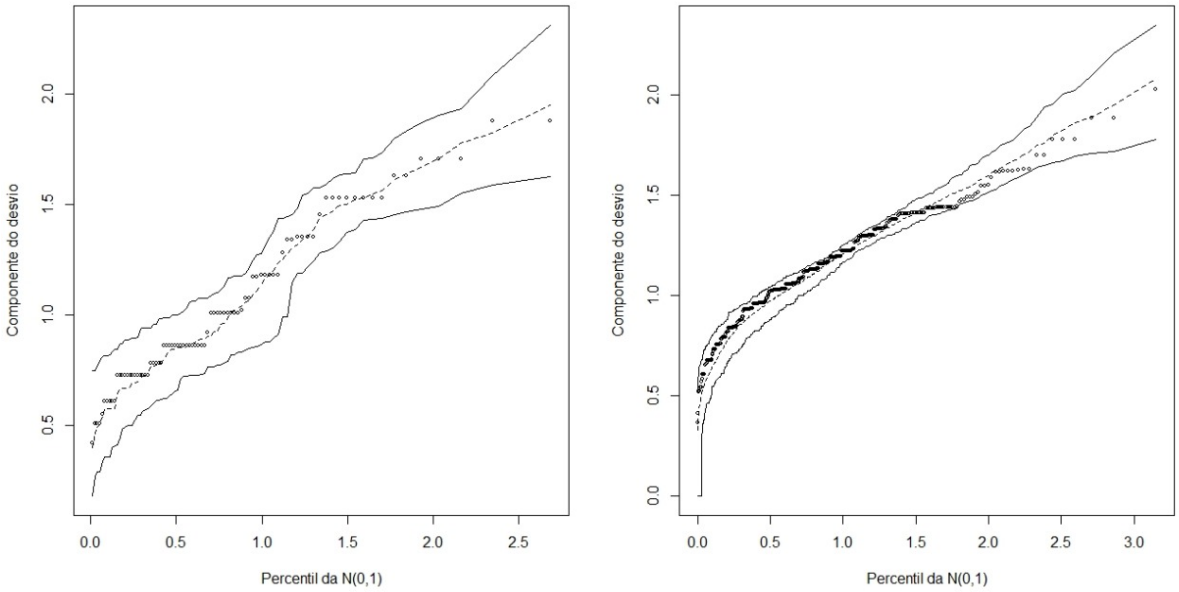
que fazem uso de vitaminas, as chances estimadas de estarem em boa qualidade aumentam em 4,179 vezes quando comparado àquelas que não utilizam.

Em relação ao modelo obtido a partir da árvore de classificação, as chances estimadas de ter uma alimentação de boa qualidade estão estatisticamente associadas à situação econômica da família e ao número de refeições diárias. Nesse último, a cada aumento de uma unidade, ou seja, uma refeição diária a mais, as chances de estar em segurança alimentar aumentam 1,279 vezes. Já em relação à situação econômica da família, as chances estimadas da criança estar em boa qualidade são 1,893 e 3,349 vezes quando a família recebe 2 ou 3, ou mais que 3 salários mínimos, respectivamente, quando comparadas a crianças cujas famílias recebem até 1 salário mínimo. Com base no intervalo, para as famílias com a situação econômica mais favorável as chances podem ser cerca de oito vezes se comparadas as crianças de famílias com situação econômica mais desfavorável.

Para o método de regularização Lasso, temos que AIC = 107,95 e a deviance residual foi de 101,95. Para árvore de classificação, AIC = 509,68 e deviance residual foi de 489,68, o que indica que o modelo ajustado considerando as variáveis selecionadas pelo Lasso, sendo elas refeições diárias e uso de vitaminas, apresentou menores valores de AIC e deviance residual.

Pelos gráfico 4.3 pode-se notar que ambos os modelos foram bem ajustados, pois todos os pontos situam-se dentro do intervalo definido pelo envelope simulado.

Figura 4.3 – Gráficos de diagnóstico referentes aos modelos logísticos ajustados aos dados de frequência alimentar com variáveis selecionadas pelo método Lasso (a) e pelo método da árvore de classificação (b).



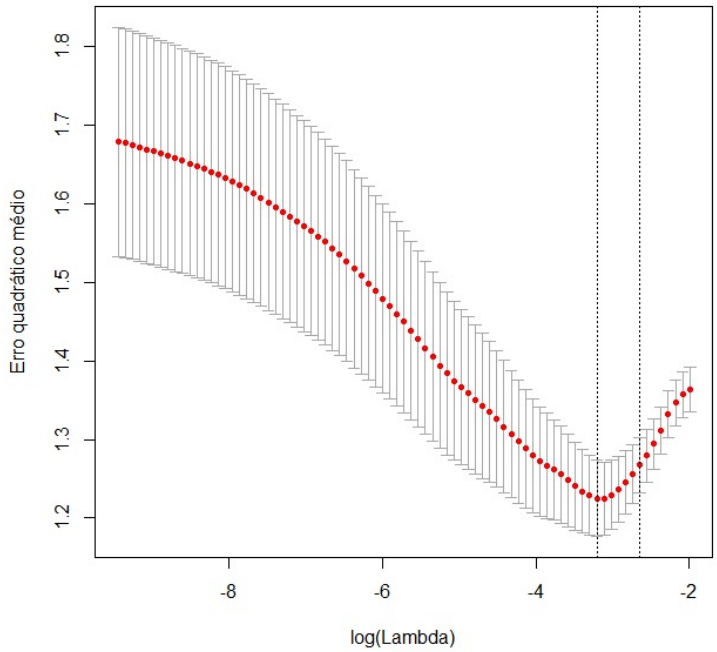
Fonte: Da autora (2019)

4.2 Preditores associados à segurança alimentar

4.2.1 Lasso aplicado à segurança alimentar

Em relação à variável resposta segurança alimentar (EBIA), nota-se, pelo gráfico apresentado na Figura 4.4 que o valor de λ que minimiza o erro de validação cruzada é 0,04051 (log(0,04051) $\approx$ - 3,2062) para o modelo com penalização Lasso, ainda o maior valor de λ que está no intervalo de um erro padrão do erro quadrático médio de validação é 0,07079 (log(0,07079) $\approx$ - 2,6480).

Figura 4.4 – Valores de λ para o erro quadrático médio de validação cruzada para a variável resposta segurança alimentar.



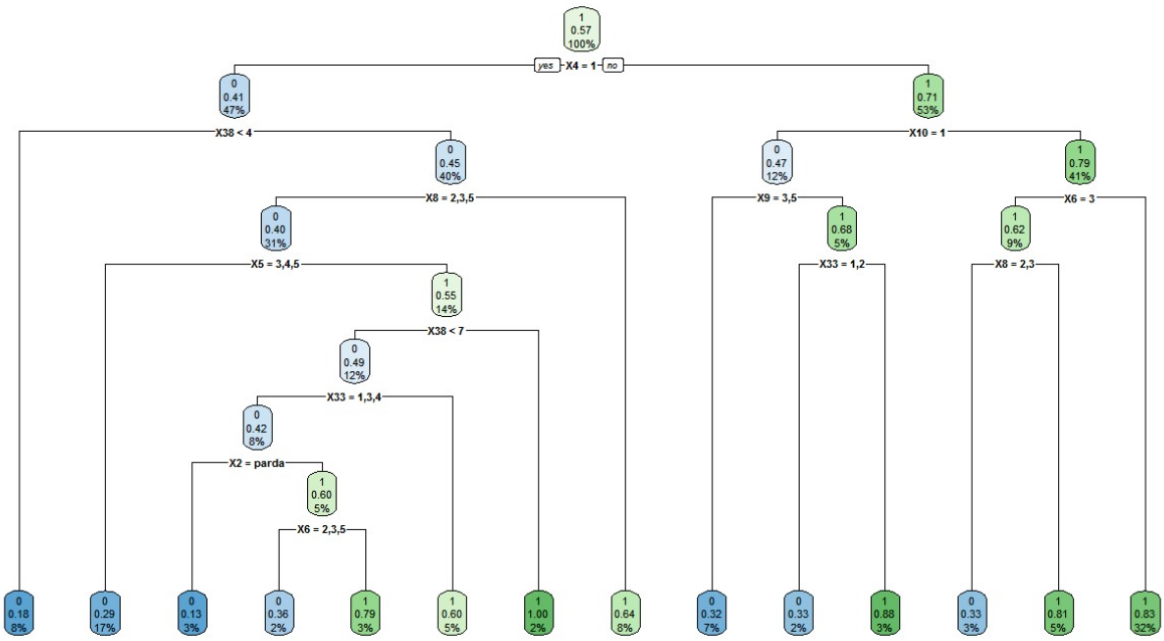
Fonte: Da autora (2019)

Tomando o valor de λ mínimo as variáveis selecionadas foram: cor/raça (X2), situação econômica (X4), escolaridade paterna (X6), profissão do chefe da família (X9) e posse de automóvel (X10).

4.2.2 Árvore de classificação aplicada à segurança alimentar

A árvore gerada é apresentada na Figura 4.5. As variáveis consideradas para ajuste do modelo logístico são aquelas presentes nos nós da árvore, sendo: cor/raça (X2), situação econômica (X4), número de pessoas na família (X5), escolaridade do pai (X6), habitação (X8), profissão do chefe da família (X9), posse de automóvel (X10), idade do ínicio do desmame (X33) e número de refeições (X38).

Figura 4.5 – Árvore de classificação para a variável resposta segurança alimentar.



Fonte: Da autora (2019)

A variável responsável pela primeira partição é situação econômica (X4), ou seja, é a variável que mais causa variabilidade na amostra. Como na seção 4.1.2, as interpretações dos ramos seguintes são feitas com base nas combinações de condicionantes presentes nesta amostra.

A interpretação acontece da seguinte forma: o primeiro nó é situação econômica da família (X4), se a família recebe até um salário mínimo então a criança é classificada em insegurança alimentar e o próximo nó a ser interpretado é número de refeições (X38). Caso contrário, se a família recebe mais de um salário mínimo então a criança é classificada em segurança alimentar e o próximo nó a ser considerado é posse de automóvel (X10).

Analisando, primeiramente, número de refeições (X38), se a criança faz menos que 4 refeições em um dia ela é classificada em insegurança alimentar. Para a posse de automóvel a criança é classificada em insegurança alimentar se a família possui 1 automóvel, se possui mais de 1 é classificada em segurança alimentar. Logo, percorre-se a árvore pelos nós e folhas sucessivamente até os nós de término.

4.2.3 Modelos logísticos para segurança alimentar

Os resultados dos modelos logísticos ajustados para a variável resposta segurança alimentar após aplicação do *stepwise* são apresentados na Tabela 4.2.

Tabela 4.2 – Estimativas e erros padrões para os parâmetros dos modelos logísticos, estimativas de razões de chances (RC) e respectivos intervalos de confiança a 95% para a variável resposta segurança alimentar.

Lasso					
Variável	Estimativa	Erro padrão	Valor-p	RC	$IC_{95\%}$ (RC)
Cor/Raça (X2)					
Negra	-1,0308	0,4407	0,01933	0,357	[0,148;0,840]
Parda	-0,5746	0,3466	0,09733	0,563	[0,282;1,104]
Situação econômica (X4)					
de 2 a 3 salários mínimos	1,7459	0,3411	<0,0001	5,731	[2,977;11,386]
mais que 3 salários mínimos	3,1117	0,8088	<0,0001	22,460	[5,565;153,644]
Escolaridade paterna (X6)					
Ginasial incompleto	-1,4867	0,7648	0,05190	0,226	[0,047;0,981]
2º grau incompleto	-1,6195	0,7287	0,02626	0,198	[0,044;0,800]
2º grau completo e superior incompleto	-0,8069	0,7254	0,26600	0,446	[0,099;1,794]
Superior completo	-1,3571	0,8743	0,12060	0,257	[0,044;1,402]
Árvore de classificação					
Variável	Estimativa	Erro padrão	Valor-p	RC	$IC_{95\%}$ (RC)
Cor/Raça (X2)					
Negra	-0,8634	0,4241	0,0417	0,422	[0,182;0,964]
Parda	-0,6062	0,3271	0,0638	0,545	[0,285;1,031]
Situação econômica (X4)					
de 2 a 3 salários mínimos	1,6570	0,3255	<0,0001	5,244	[2,802;10,075]
mais que 3 salários mínimos	3,2085	0,8181	<0,0001	24,741	[6,022;171,754]
Escolaridade paterna (X6)					
Ginasial incompleto	-1,6572	0,7463	0,0264	0,191	[0,040;0,789]
2º grau incompleto	-1,7486	0,7257	0,0160	0,174	[0,038;0,692]
2º grau completo e superior incompleto	-1,0767	0,7134	0,1312	0,341	[0,077;1,324]
Superior completo	-1,4887	0,8554	0,0818	0,226	[0,039;1,173]
Habitação (X8)					
Residência própria quitada	0,6533	0,4422	0,1396	1,922	[0,810;4,620]
Própria com financiamento a pagar	-0,4375	0,4361	0,3157	0,646	[0,272;1,509]
Residência alugada	0,0385	0,4163	0,9263	1,039	[0,457;2,354]

Fonte: Da autora (2019)

Considerando o modelo obtido pelo método de seleção Lasso e o nível de significância de 5%, as três variáveis do modelo tem associação significativa com a condição de segurança alimentar. Para esse banco de dados e mantendo todas as outras variáveis constantes, observou-se que a chance estimada de uma criança negra estar em segurança alimentar é 0,357 da chance

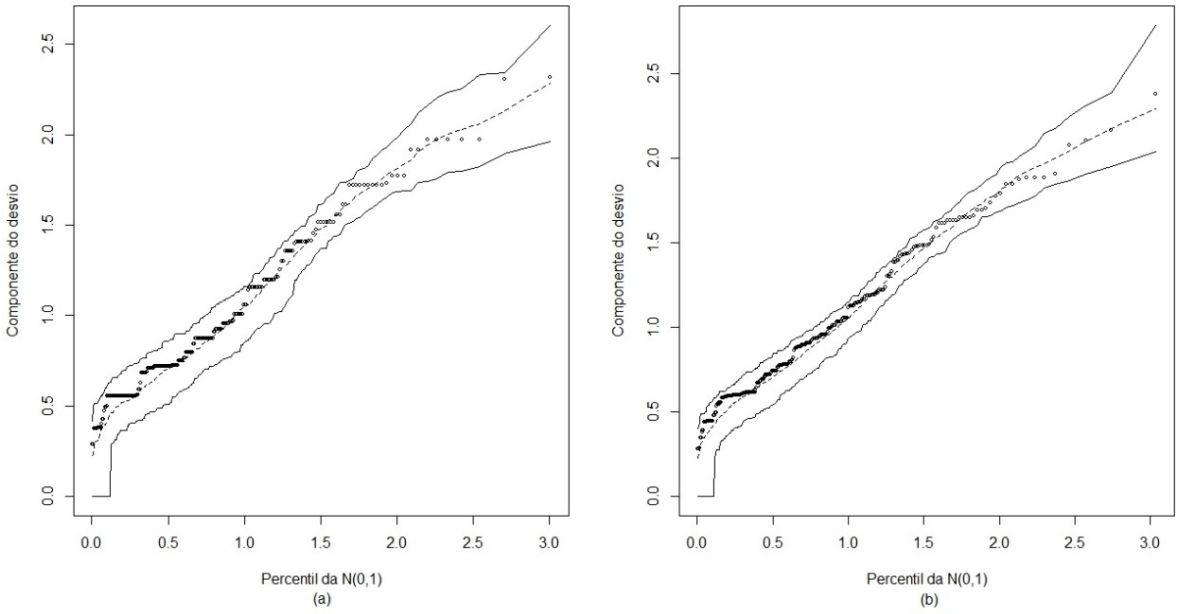
estimada de uma criança branca, ou seja, é aproximadamente um terço da chance da criança branca. Em relação a situação econômica da família, as chances estimadas da criança estar em segurança alimentar são 5,731 e 22,460 vezes para quem recebe 2 ou 3, ou mais que 3 salários mínimos, respectivamente, se comparadas a famílias que recebem até 1 salário mínimo. Com base no intervalo, para as famílias com a situação econômica mais favorável as chances das crianças estarem em segurança alimentar podem ser cerca de cinquenta e três vezes se comparadas as crianças de famílias com situação econômica mais desfavorável. Por fim, para esses dados, a chance estimada da criança cujo pai possui 2º grau incompleto é 0,198 vezes da chance de uma criança cujo pai possui primário incompleto.

Em relação ao modelo obtido por meio da árvore de classificação,observou-se que, para essa amostra e considerando as demais variáveis constantes, a chance estimada de uma criança negra estar em segurança alimentar é 0,422 da chance estimada de uma criança branca. Para a situação econômica da família, as chances estimadas da criança estar em segurança alimentar são 5,244 e 24,741 vezes para quem recebe 2 ou 3, ou mais que 3 salários mínimos, respectivamente, se comparadas a famílias que recebem até 1 salário mínimo. Por fim, as chances estimadas da criança cujo pai possui ginasial incompleto e 2º grau incompleto é 0,191 e 0,174, respectivamente, das chances estimadas de uma criança cujo pai possui primário incompleto.

Para o método de regularização Lasso, temos que AIC = 273,20 e deviance residual foi de 255,20. Para árvore de classificação AIC = 307,37 e deviance residual foi de 283,37, o que indica que o modelo ajustado considerando as variáveis selecionadas pelo Lasso, sendo elas cor/raça, situação econômica e escolaridade paterna, apresentou menor valor de AIC e deviance residual.

Pelo gráfico 4.6 e pode-se notar que ambos os modelos foram bem ajustados, pois todos os pontos situam-se dentro do intervalo definido pelo envelope simulado.

Figura 4.6 – Gráficos de diagnóstico referentes aos modelos logísticos ajustados aos dados de segurança alimentar com variáveis selecionadas pelo método Lasso (a) e pelo método da árvore de classificação (b).



Fonte: Da autora (2019)

4.3 Discussão

Do ponto de vista prático, pode-se dizer que foram encontrados bons resultados, isto é, as variáveis selecionadas estão diretamente ligadas aos desfechos de interesse. Para a variável resposta frequência alimentar, as chances para as crianças que fazem uso de vitaminas aumentam ocasionando uma boa qualidade alimentar. Em Pedraza, Rocha e Sousa (2013), a deficiência das vitaminas A e zinco estão diretamente ligadas a esse desfecho corroborando para o crescimento das crianças.

O nível de escolaridade dos pais é uma variável a ser considerada já que são eles os maiores responsáveis pela educação da criança. Na variável resposta segurança alimentar, a escolaridade do pai foi um fator significativo. De acordo com Batalha et al. (2017) e Toloni et al. (2011), que também estudaram a escolaridade materna, afirmam que quanto menor o grau de escolaridade ou anos que frequentaram a escola, menor é o grau de instrução e consequentemente o acesso a informação sobre o consumo alimentar adequado na vida infantil.

Para ambas as variáveis respostas, um fator que está relacionado a uma boa qualidade alimentar é a situação econômica. Tanto neste trabalho como nos de Batalha et al. (2017) e

Toloni et al. (2011), esse fator que está associado ao poder aquisitivo das famílias implicam a adquirir alimentos processados e ultra-processados de baixo custo agravando a saúde da criança.

Neste trabalho, o desempenho dos modelos obtidos a partir da seleção de variáveis via Lasso foram superiores àqueles obtidos a partir da árvore, considerando os critérios estatísticos AIC e deviance residual.

Deve-se observar que algumas variáveis coincidem em ambas técnicas de seleção de variáveis. Por exemplo, para a variável resposta frequência alimentar, as variáveis idade e número de refeições diárias aparecem em ambos os modelos. Já para a variável resposta segurança alimentar há mais variáveis coincidentes, como cor/raça, situação econômica, escolaridade paterna, profissão do chefe da família e posse de automóvel.

Essas variáveis também são recorrentes nos trabalhos apresentados. Batalha et al. (2017) ajusta o modelo para idade da mãe, anos que frequentou escola, remuneração da mãe, estado civil e número de pessoas na família que também é uma variável apresentada na árvore de classificação para segurança alimentar. No trabalho de Toloni et al. (2011) as variáveis consideradas são escolaridade materna, idade materna e situação econômica familiar, sendo essas apresentadas tanto na técnica Lasso quanto na árvore de classificação para os dois desfechos. No trabalho de Oliveira et al. (2011) que utilizou o procedimento proposto por Hosmer, Lemeshow e Sturdivant (2013), em que todas as variáveis que apresentaram $p < 0,25$ foram consideradas no modelo, sendo elas: água de consumo (fervida, filtrada ou clorada, sem tratamento) e idade da criança (> 48 meses, < 48 meses), que também foi uma variável apresentada nos métodos de Lasso e da árvore de classificação para a variável resposta frequência alimentar.

A técnica de regularização Lasso é muito utilizada em diversas áreas do conhecimento, como por exemplo na genética. Os autores que utilizaram essa técnica para seleção de variáveis encontraram bons resultados. Em Tisbshirani (1997), a seleção por Lasso foi mais precisa quando comparada a seleção por *stepwise*; Ayres e Cordell (2010) disseram que os métodos de regressão penalizada oferecem uma alternativa atraente, já que reduzem para zero os coeficientes que têm pouco efeito aparente sobre a característica de interesse; Uniejewski, Nowotarski e Weron (2016) afirmaram que Lasso proporcionou, em média, melhor desempenho quando comparado aos modelos cujas variáveis foram selecionadas considerando o conhecimento do especialista, ou seja, de forma empírica. Já em Ogutu, Schulz-Streeck e Piepho (2012) os resultados são discrepantes, uma vez que Lasso apresentou comportamento inferior aos comparados no trabalho.

Já árvore de classificação, há poucos trabalhos utilizando essa técnica com o objetivo de selecionar variáveis. Porém, os autores que a utilizaram obtiveram bons resultados. Em Cho, Hang e Ha (2010) as variáveis selecionadas pelas árvores de decisão tenderam a ter uma interação quando comparadas aos produzidos pelas abordagens de regressão, logo os resultados obtidos indicam que a abordagem que foi proposta supera algumas técnicas atualmente em uso. Tsai e Chen (2009), que utilizaram árvores para selecionar variáveis como uma etapa de pré-processamento dos dados, afirmaram que poucos estudos consideram-o durante a mineração de dados, cujo objetivo é filtrar dados ou informações não representativas. Essa técnica é mais utilizada em termos de problemas de classificação, como podemos ver nos trabalhos de Kramer et al. (2001); Pal e Mather (2003); Toschke, Beyerlein e Kries (2005); Waheed et al. (2006); Zhang (1998).

5 CONCLUSÃO

Os dois métodos de seleção de variáveis foram aplicados com sucesso aos dados de segurança e frequência alimentar. Dos modelos de regressão logística ajustados, os que apresentaram melhores resultados, ou seja, menor valor de AIC e menor diferença entre deviance residual e graus de liberdade, foram aqueles que utilizaram como método de seleção de variáveis a regularização Lasso.

Em ambos os métodos, as variáveis selecionadas assemelham-se com aquelas apresentadas em trabalhos aplicados, o que mostra, de fato, a importância clínica para os desfechos estudados. Além disso, nos trabalhos apresentados as variáveis coincidem mais com as que foram mostrados nas árvores de classificação. Logo, estas podem ser utilizadas já que apresentaram resultados semelhantes e fornecem um resultado gráfico atraente e de fácil interpretação.

Os resultados encontrados nesse trabalho são restritos a esse conjunto de dados. Em trabalhos futuros, estudos de simulação podem ser feitos para considerar outros cenários, com diferentes tamanhos amostrais, números e tipos de variáveis explicativas. Além disso, a técnica de *Random Forest*, que já vem sendo usada no contexto de seleção de variáveis, pode ser considerada.

REFERÊNCIAS

AGRESTI, A. **An introduction to categorical data analysis**. 2. ed. New York: J. Wiley & Sons, 2007. 394 p.

_____. **Categorical data analysis**. 3. ed. New York: J. Wiley & Sons, 2013. 714 p.

ALVES, M. F.; LOTUFO, A. D. P.; LOPES, M. L. M. Seleção de variáveis stepwise aplicadas em redes neurais artificiais para previsão de demanda de cargas elétricas. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, São Carlos, v. 1, n. 1, p. 1-6, 2013.

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, New York, v. 19, n. 6, p. 716-723, Dec. 1974.

ATH, G.; FABRICIUS, K. Classification and Regression trees: a powerful yet simple tech-nique for ecological data analysis. **Ecology**, London, v. 81. n. 11, p. 3178-3192, Nov. 2000.

AYERS, K. L.; CORDELL, H. J. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. **Genetic Epidemiology**, Philadelphia, v. 34, n. 8, p. 879-891, Dec. 2010.

BATALHA, M. A. et al. Processed and ultra-processed food con- sumption among children aged 13 to 35 months and associated factors. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 33, n. 11, p. 1-16, nov. 2017.

BREIMAN, L. et al. **Classification and regression trees**. New York: Taylor & Francis, 1984. 368 p.

BURSAC, Z. et al. Purposeful selection of variables in logistic regression. **Source Code for Biology and Medicine, Expert Systems with Applications**, London, n. 3, p. 17, Dec. 2008.

CHO, S.; HONG, H.; HA, B.-C. A hybrid approach based on the com- bination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance: For bankruptcy prediction. **Expert Systems with Applications**, London, v. 37, n. 4, p. 3482-3488, Apr. 2010.

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. New YorK: Wiley-Intercience, 1998. 736 p.

DURÃO, C. et al. Maternal child-feeding practices and dietary inadequacy of 4-year-old children. **Appetite**, London, n. 92, p. 15-23, Sept. 2015.

FONSECA, J. **Indução de árvores de decisão**. Tese (Doutorado) - Universidade Nova de Lisboa, Lisboa, 1994.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for generalized linear. **Journal of Statistical Software**, California, v. 33, n. 1, p. 1-22, 2010.

GENUER, R.; POGGI, J.-M.; TULEAU-MALOT, C. Variable selection using random forests. **Pattern Recognition Letters**, Amsterdam, v. 31, p. 2225-2236, 2010.

GIOLO, S. R. **Introdução à análise de dados categóricos com aplicações**. São Paulo: Blucher, 2017. 256 p.

GRAYBILL, F. A.; IYER, H. K. **Regression analysis:** concepts e applications. Oxford: Duxbury Press, 1994. 701 p.

GUIMARÃES, P. H. S. Uma introdução ao Machine Learning com o R. In: SEMANA DA ESTATÍSTICA DO DES-UFLA, 1., 2018, Lavras. **Minicursos...** Lavras: UFLA, 2018.

HAN, J. **Data mining:** concepts and techiniques. Canadá: Morgan Kaufmann Publishers, 2001. 550 p.

HAPFELMEIER, A.; ULM, K. A new variable selection approach using Random Forests. **Computational Statistics and Data Analysis**, Amsterdam, v. 60, p. 50-69, Apr. 2013.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning:** data mining, inference and prediction. 2. ed. Amsterdam: Springer, 2008. 764 p.

HOSMER, D.; LEMESHOW, S.; STURDIVANT, R. **Applied logistic regression**. New York: J. Wiley & Sons, 2013. 528 p.

JAMES, G. et al. **An introduction to statistical learning:** with applications in R. London: Springer Natura, 2013. 426 p.

KRAMER, S. et al. Prediction of ordinal classes using regression trees. **Fundamenta Informaticae**, Amsterdam, v. 47, n. 1/2, p. 1-13, 2001.

LOURENÇÃO, L. F. P. **Avaliação nutricional de pré-escolares e implementação de um programa educativo voltado nutricional para servidores de educação infantil.** 2019. 120 p. Dissertação (Mestrado em Ciências da Saúde) - Universidade Federal de Lavras, Lavras, 2019.

MILBORROW, S. **rpart.plot**: plot “rpart” models: an enhanced version of “plot.rpart’. [S.l.: s.n.], 2018. Disponível em: <https://CRAN.R-project.org/package=rpart.plot>.

MONDINI, L. et al. Prevalência de sobrepeso e fatores associados em crianças ingressantes no ensino fundamental em um município da região metropolitana de São Paulo, Brasil. **Caderno de Saúde Pública**, Rio de Janeiro, v. 23, n. 8, p. 1825-1834, ago. 2007.

MORAL, R. A.; HINDE, J.; DEMÉTRIO, C. G. B. Half-normal plots and overdis- persed models in R: the hnp package. **Journal of Statistical Software**, California, v. 81, n. 10, p. 1-23, 2017.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society**, London, v. 135, n. 3, p. 370-384, 1972.

OGUTU, J. O.; SCHULZ-STREECK, T.; PIEPHO, H.-P. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their exten sions. **BMC Proceedings**, London, v. 6, Supp. 2, p. 10, 2012.

OLIVEIRA, F. C. C. et al. Estado nutricional e fatores determinantes do déficit estatural em crianças cadastradas no Programa Bolsa Família. **Epidemiologia e Serviços de Saúde**, Brasília, v. 20, n. 1, p. 7-18, mar. 2011.

PAULA, G. A. **Modelos de regressão com apoio computacional**. São Paulo: IME-USP, 2013. 441 p.

PAL, M.; MATHER, P. M. An assessment of the effectiveness of decision tree methods for land cover classification. **Remote Sensing of Environment**, New York, v. 86, n. 4, p. 554-565, Aug. 2003.

PEDRAZA, D. F.; ROCHA, A. C. D.; SOUSA, C. P. C. Crescimento e deficiências de micronutrientes: perfil das crianças assistidas no núcleo de creches do governo da Paraíba, Brasil. **Ciência & Saúde Coletiva**, Rio de Janeiro, v. 18, n. 11, p. 3379-3390, Nov. 2013.

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, New York, v. 1, n. 1, p. 81-106, 1986.

QUINLAN, J. R. **C4.5:** programs for machine learning. San Francisco: Morgan Kaufmann, 1993. 302 p.

R CORE TEAM. **R:** a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2019. Disponível em: <http://www.R-project.org/>. Acesso em: 16 ago. 2019.

SANDRI, M.; ZUCCOLOTTO, P. Variable selection using random forests. In: ZANI, S. et al. (Ed.). **Data analysis, classification and the forward search.** Berlin: Springer, 2006. p. 263-270.

SEGALL-CORRÊA, A. M. et al. **Acompanhamento e avaliação da segurança alimentar de famílias brasileiras:** validação de metodologia e de instrumento de coleta de informação. Cam- pinas: UNICAMP, 2003. 33 p.

STROBL, C. et al. Conditional variable importance for random forests. **BMC Bioinformatics**, London, n. 9, p. 307, July 2008.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. San Francisco: Pearson, 2005. 202 p.

TIBSHIRANI, R. The Lasso method for variable selection in the cox model. **Statistics in Medicine**, Chichester, v. 16, p. 385-395, 1997.

THERNEAU, T.; ATKINSON, B. **rpart:** Recursive Partitioning and Regression Trees. [S.l.: s.n.], 2018. Disponível em: <http://https://CRAN.R-project.org/package=rpart>. Acesso em: 23 out. 2019.

TOLONI, M. H. A. et al. Introdução de alimentos industrializados e de alimentos de uso tradicional na dieta de crianças de creches públicas no município de São Paulo. **Revista de Nutrição**, Campinas, v. 24. n. 1, p. 61-70, jan./fev. 2011.

TOSCHKE, A. M.; BEYERLEIN, A.; KRIES, R. von. Children at high risk for overweight: a classification and regression trees analysis approach. **Obesity Research**, Baton Rouge, v. 13, n. 7, p. 1270-1274, July 2005.

TSAI, C.-F.; CHEN, M.-Y. Variable selection by association rules for customer churn prediction of multimedia on demand. **Expert Systems with Applications**, v. 37, n. 3, p. 2006-2015, Mar.

2009.

UNIEJEWSKI, B.; NOWOTARSKI, J.; WERON, R. Automated variable selec- tion and shrinkage for day-ahead electricity price forecasting. **Energies**, London, v. 9, n. 621, p. 1-22, 2016.

WAHEED, T. et al. Measuring performance in precision agriculture: CART a decision tree approach. **Agricultural Water Management**, Geneva, v. 84, n. 1/2, p. 173-185, July 2006.

WALDMANN, P. et al. Evaluation of the lasso and the elastic net in genome-wide association studies. **Frontiers in Genetics**, Lausanne, n. 4, n. 270, Dec. 2013.

WRIGHT, M. N.; DANKOWSKI, T.; ZIEGLER, A. Unbiased split variable selection for random survival forests using maximally selected rank statistics. **Statistics in Medicine**, Hoboken, v. 36, n. 8, p. 1272-1284, Apr. 2010.

WU, X.; KUMAR, V. **The top ten algorithms in data mining**. Boca Raton: Chapman & Hall/CRC, 2009. 214 p.

ZEVIANI, W. M.; FERREIRA, E. V. **Regularização**. [S.l.: s.n.], 2017. Notas de aula. Disponível em: <http://leg.ufpr.br/ walmes/ensino/ML/tutorials/03-regularization.html>. Acesso em: 23 out. 2019.

ZHANG, H. Classification trees for multiple binary responses. **Journal of the American Statistical Association**, Washington, v. 93, n. 441, p. 180-193, Mar. 1998.

ZHANG, Z. Variable selection with stepwise and best subset approaches. **Annals of Translational Medicine**, Hong Kong, v. 4, n. 7, p. 2-6, Apr. 2016.

APÊNDICE A – Questionário aplicado na coleta de dados socioeconômicos, de saúde e segurança alimentar de pré-escolares que frequentam CMEIs em Lavras, MG.

QUESTIONÁRIO SOCIOECONÔMICO E CLÍNICO Nº _____

Nome da criança: _____

1. Sexo: 1. ( ) Masculino 2. ( ) Feminino **2. Cor/Raça:** _____

Data de Nascimento:___/___/___. **3.Idade:** _____

Endereço: _____ Telefone:( ) _____

4. SITUAÇÃO ECONÔMICA DA FAMÍLIA (Renda familiar mensal)

1. ( ) até 1 salário mínimo
2. ( ) de 2 a 3 salários mínimos
3. ( ) de 4 a 5 salários mínimos
4. ( ) de 5 a 6 salários mínimos
5. ( ) de 6 a 7 salários mínimos
6. ( ) de 7 a 8 salários mínimos
7. ( ) Acima de 9 salários mínimos

5. NÚMERO DE PESSOAS NA FAMÍLIA (Residentes na mesma casa)

1.( ) até 2 pessoas 2.( ) 3 pessoas 3.( ) 4 pessoas
4.( ) 5 pessoas 5.( ) 6 pessoas 6. ( ) acima de 6 pessoas

GRAU DE INSTRUÇÃO DOS PAIS OU RESPONSÁVEIS

	6. PAI	**7. MÃE**	
1.	( )	( )	Não alfabetizado
2.	( )	( )	Alfabetizado
3.	( )	( )	1ª a 4ª série incompleta (antigo Primário)
4.	( )	( )	1ª a 4ª série completa (antigo Primário)
5.	( )	( )	5ª a 8ª série incompleta (antigo Ginasial)
6.	( )	( )	5ª a 8ª série completa (antigo Ginasial)
7.	( )	( )	2º Grau incompleto (antigo Colegial)
8.	( )	( )	2º Grau completo (antigo Colegial)
9.	( )	( )	Superior incompleto
10.	( )	( )	Superior completo

8. HABITAÇÃO (Moradia)

1.( ) Residência própria quitada
2.( ) Residência própria com financiamento a pagar
3.( ) Residência cedida pelos pais ou parentes por não ter onde morar
4.( ) Residência cedida em troca de trabalho
5.( ) Residência alugada
6.( ) Residência cedida

9. PROFISSÃO DO CHEFE DA FAMÍLIA (Mencionar mesmo que desempregado) : _____

10. POSSE DE AUTOMÓVEL:

1.( ) Não possui 2. ( ) Possui um automóvel 3. ( ) Possui 2 ou mais automóveis

11. A ÁGUA UTILIZADA NA CASA É DE ABASTECIMENTO DA REDE PÚBLICA (ENCANADA ATÉ O DOMICÍLIO): 1. ( ) sim 2. ( ) não.

12. A ÁGUA DE BEBER É:

1.( ) Filtrada 2. ( ) Fervida 3. ( ) Clorada
4. ( ) comprada de galão (mineral) 5. ( ) Não tratada.

GESTAÇÃO

13. Consultas de Pré-natal? 1. ( ) sim 2. ( ) não

14. A gravidez foi planejada? 1. ( ) sim 2. ( ) não

15. Teve algum tipo de problema de saúde durante o pré-natal: 1. ( ) sim 2. ( ) não.

16. A gravidez foi considerada como de alto risco: 1. ( ) Sim 2. ( ) Não

CONDIÇÕES DO NASCIMENTO DA CRIANÇA

17. Idade Gestacional: _____ semanas

18. Local do parto: 1. ( ) Hospitalar 2. ( ) Domicílio 3. ( ) Outro: Qual? _____

19. Tipo de parto: 1. ( ) normal 2. ( ) cesárea 3. ( ) fórceps

20. A criança precisou ficar na UTI/CTI neonatal: 1. ( ) sim 2. ( ) não

21. Apgar de 5 minutos: _____

22. Peso as nascer: _____

23. Estatura ao nascer: _____

24. Perímetro Cefálico ao nascer: _____

25. Teste pezinho: 1. ( ) sim 2. ( )não 3. ( ) Não sabe informar

26. Resultado do teste de pesinho: 1. ( ) Positivo para: _____2. ( ) Negativo

27. Teste do olhinho realizado. 1.( ) sim 2. ( )não **28. Resultado:** 1.( ) Positivo 2. ( ) Negativo

29. Triagem auditiva neonatal: 1.( ) sim 2. ( )não **30. Resultado:** 1.( ) Positivo 2. ( ) Negativo

DADOS GERAIS DE ALIMENTAÇÃO E SAÚDE DA CRIANÇA

31. A criança está em aleitamento materno exclusivo? 1. ( ) sim 2. ( ) Não

32. Aleitamento materno exclusivo até 6 meses de idade: 1. ( ) sim 2. ( ) Não

33. Idade de início do desmame: _____ (dias) ou _____(meses)

34. Fez ou faz uso de leite tipo Nan? 1. ( ) sim 2. ( ) não.

35. SE SIM NA PERGUNTA ANTERIOR: Com que idade começou a usar? _____

QUAL A MARCA/TIPO? _____

37. Idade de introdução de outros alimentos que não leite: _____

38. Número de refeições da criança durante o dia: _____

39. Fez uso de vitaminas nos dois primeiros anos de vida? 1. ( ) sim 2. ( ) não

40. Fez uso de sulfato ferroso nos dois primeiros anos de vida? 1. ( ) sim 2. ( ) não

41. Faz acompanhamento de rotina no posto ou com algum pediatra? 1. ( ) sim 2. ( ) não

42. Alguém da equipe de saúde da família faz visitas domiciliares para avaliar a criança?

1. ( ) sim 2. ( ) não

43. A criança tem alguma doença grave ou infecção de repetição? 1. ( ) sim 2. ( ) não

44 Quem fica com a criança na maior parte do tempo quando ela não está na creche?

1. ( ) Pai/Mãe 2. ( ) Avô/Avó 3. ( ) Irmão/Irmã menor de idade 4. ( ) Outro: _____

45. Alguém que convive com a criança é usuária de álcool ou droga? 1. ( ) sim 2. ( ) não

SE SIM, é quem fica a maior parte do tempo com a criança fora da creche? 3. ( ) sim 4. ( ) não

46. Alguém que convive com a criança tem problemas de saúde mental? 1. ( ) sim 2. ( ) não

SE SIM, é quem fica a maior parte do tempo com a criança fora da creche? 3. ( ) sim 4. ( ) não

47. A criança está com a vacinação em dia 1. ( ) sim 2. ( ) não

AVALIAÇÃO DE DADOS NUTRICIONAIS **Nº** _____

Nome da criança: _____

48. QUESTIONÁRIO DE FREQUÊNCIA ALIMENTAR (QFA): _____

PRODUTOS	FREQUÊNCIA				
	1 - 2 vezes ao dia	3 vezes ou mais ao dia	1 - 2 vezes por semana	3 vezes ou mais por semana	Nunca/raramente
LEITE E DERIVADOS					
Leite					
Iogurte					
Queijo					
CARNES E OVOS					
Ovo					
Carne de boi					
Carne de frango					
Frango					
Peixe					
Embutidos (salame, salsicha, presunto, mortadela)					
Vísceras (fígado, coração)					
ÓLEOS / GORDURAS					
Óleo					
Manteiga					
Margarina					
Maionese					
CEREAIS / LEGUMINOSAS					
Arroz					
Biscoito					
Bolos					
Macarrão					
Feijão					
Pães					
Cereais (sucrilhos, linhaça, etc.)					
DOCES					
Balas					
Chocolates					
Sobremesas					
Refrigerante					
FRUTAS, VERDURAS E LEGUMES					
Frutas					
Verduras					
Legumes					

49. ESCALA BRASILEIRA DE INSEGURANÇA ALIMENTAR – EBIA _____

1 - Nos últimos TRÊS MESES, os moradores deste domicílio tiveram preocupação de que os alimentos acabassem antes de poderem comprar ou receber mais comida?	( ) SIM ( )NÃO
2 - Nos últimos TRÊS MESES, os alimentos acabaram antes que os moradores deste domicílio tivessem dinheiro para comprar mais comida?	( ) SIM ( )NÃO
3 - Nos últimos TRÊS MESES, os moradores deste domicílio ficaram sem dinheiro para ter uma alimentação saudável e variada?	( ) SIM ( )NÃO
4 - Nos últimos TRÊS MESES, os moradores deste domicílio comeram apenas alguns alimentos que ainda tinham porque o dinheiro acabou?	( ) SIM ( )NÃO
5 - Nos últimos TRÊS MESES, algum morador de 18 anos ou mais de idade deixou de fazer uma refeição porque não havia dinheiro para comprar comida?	( ) SIM ( )NÃO
6 - Nos últimos TRÊS MESES, algum morador de 18 anos ou mais de idade, alguma vez comeu menos do que devia porque não havia dinheiro para comprar comida?	( ) SIM ( )NÃO
7 - Nos últimos TRÊS MESES, algum morador de 18 anos ou mais de idade, alguma vez sentiu fome, mas não comeu, porque não havia dinheiro para comprar comida?	( ) SIM ( )NÃO
8 - Nos últimos TRÊS MESES, Algum morador de 18 anos ou mais de idade, alguma vez, fez apenas uma refeição ao dia ou ficou um dia inteiro sem comer porque não havia dinheiro para comprar comida?	( ) SIM ( )NÃO
9 - Nos últimos TRÊS MESES, algum morador com menos de 18 anos de idade, alguma vez, deixou de ter uma alimentação saudável e variada porque não havia dinheiro para comprar comida?	( ) SIM ( )NÃO
10 - Nos últimos TRÊS MESES, algum morador com menos de 18 anos de idade, alguma vez, não comeu quantidade suficiente de comida porque não havia dinheiro para comprar comida?	( ) SIM ( )NÃO
11 - Nos últimos TRÊS MESES, alguma vez, foi diminuída a quantidade de alimentos das refeições de algum morador com menos de 18 anos de idade, porque não havia dinheiro para comprar comida?	( ) SIM ( )NÃO
12 - Nos últimos TRÊS MESES, alguma vez, algum morador com menos de 18 anos de idade deixou de fazer alguma refeição, porque não havia dinheiro para comprar comida?	( ) SIM ( )NÃO
13 - Nos últimos TRÊS MESES, alguma vez, algum morador com menos de 18 anos de idade, sentiu fome, mas não comeu porque não havia dinheiro para comprar comida?	( ) SIM ( )NÃO
14 - Nos últimos TRÊS MESES, alguma vez, algum morador com menos de 18 anos de idade, fez apenas uma refeição ao dia ou ficou sem comer por um dia inteiro porque não havia dinheiro para comprar comida?	( ) SIM ( )NÃO

APÊNDICE B – Codificação das variáveis do banco de dados

Tabela 1 – Código, nomes e níveis das variáveis presentes no banco de dados.

Código	Variável	Níveis
X1	Sexo	1 = Masculino, 2 = Feminino
X2	Raça	Amarela, branca, negra, parda
X3	Idade	0 = <6 meses; 1 = 6 a 11 meses; 2 = 12 a 23 meses; 3 = 24 a 35 meses; 4 = 36 a 47 meses; 5 = 48 a 59 meses; 6 = 60 a 71 meses; 7 = >72 meses
X4	Situação econômica	1 = até 1 salário mínimo; 2 = 2 a 3 salários mínimos; 3 = 4 a 5 salários mínimos; 4 = 5 a 6 salários mínimos; 5 = 6 a 7 salários mínimos; 6 = 7 a 8 salários mínimos; 7 = acima de 9 salários mínimos
X5	Número de pessoas na família	1 = até 2 pessoas; 2 = 3; 3 = 4; 4 = 5; 5 = 6; 6 = acima de 6 pessoas
X6	Escolaridade do pai	1 = Não alfabetizado, 2 = Alfabetizado, 3 = primário incompleto, 4 = primário completo, 5 = ginasial incompleto, 6 = ginasial completo, 7 = colegial incompleto, 8 = colegial completo, 9 = superior incompleto, 10 = superior completo
X7	Escolaridade da mãe	1 = Não alfabetizado, 2 = Alfabetizado, 3 = primário incompleto, 4 = primário completo, 5 = ginasial incompleto, 6 = ginasial completo, 7 = colegial incompleto, 8 = colegial completo, 9 = superior incompleto, 10 = superior completo
X8	Habitação	1 = Residência própria quitada, 2 = residência própria com financiamento a pagar, 3 = residência cedida pelos pais ou parentes, 4 = residência cedida em troca de trabalho, 5 = alugada, 6 = cedida
X9	Profissão do chefe da família	Grau I e II, III, IV, V
X10	Posse de automóvel	1 = Não possui, 2 = possui um, 3 = possui 2 ou mais
X11	Abastecimento de água da rede pública	1 = Sim, 2 = Não
X12	Água de beber	1 = Filtrada, 2 = fervida, 3 = clorada, 4 = mineral, 6 = não tratada
X13	Consultas pré-natal	1 = Sim, 2 = Não
X14	Gravidez planejada	1 = Sim, 2 = Não
X15	Problema de saúde durante o pré-natal	1 = Sim, 2 = Não
X16	Gravidez de alto risco	1 = Sim, 2 = Não
X17	Idade gestacional	1 = Pré-termo, 2 = termo, 3 = pós-termo
X18	Local do parto	1 = Hospitalar, 2 = domicílio, 3 = outro
X19	Tipo de parto	1 = Normal, 2 = cesárea, 3 = fórceps
X20	UTI/CTI neonatal	1 = Sim, 2 = Não
X21	Apgar	1 = Reanimação, 2 = asfixia moderada, 3 = vitalidade normalidade
X22	Peso	1 = <1000g, 2 = <1500g, 3 = <2500g, 4 = 2501-4000g, 5 = >4001g
X23	Estatura	1 = Pequeno para idade gestacional, 2 =adequado para idade gestacional, 3 = grande para idade gestacional
X24	Perímetro cefálico	1 = Abaixo do esperado, 2 = adequado, 3 = acima do esperado
X25	Teste pezinho	1 = Sim, 2 = não, 3 = não sabe informar
X26	Resultado do teste pezinho	1 = Positivo, 2 = negativo, 3 = traço anemia falciforme
X27	Teste olhinho	1 = Sim, 2 = Não
X28	Resultado do teste olhinho	1 = Positivo, 2 = negativo
X29	Teste audição	1 = Sim, 2 = Não
X30	Resultado do teste audição	1 = Positivo, 2 = negativo
X31	Aleitamento exclusivo	1 = Sim, 2 = Não
X32	Aleitamento até 6 meses de idade	1 = Sim, 2 = Não
X33	Idade do ínicio do desmame	0 = Ainda mama, 1 = <2 meses, 2 = 2 a 6 meses, 3 = >6 meses, 4 = 1 a 2 anos, 5 = >2 anos
X34	Uso de leite artificial	1 = Sim, 2 = Não
X35	Idade de uso de leite artificial	1 = <6 meses, 2 = >6 meses
X36	Tipo de leite	1 = Leite de vaca, 2 = fórmulas infantis, 3 = fórmulas de necessidade dietoterápica, 4 = à base de soja, 5 = outros
X37	Idade de introdução de outros alimentos	1 = <6 meses, 2 = >6 meses
X38	Número de refeições	1,2, ⋯, 10
X39	Uso de vitaminas	1 = Sim, 2 = Não
X40	Uso de sulfato ferroso	1 = Sim, 2 = Não
X41	Acompanhamento de rotina	1 = Sim, 2 = Não
X42	Equipe de saúde em visitas domiciliares	1 = Sim, 2 = Não
X43	Doença grave ou infecção de repetição	1 = Sim, 2 = Não
X44	Quem fica com a criança na maior parte do tempo	1 = mãe/pai, 2 = avó/avô, 3 = irmã/irmão menor de idade, 4 = outro
X45	Alguém que convive com a criança é usuária de álcool ou droga	1 = Sim, 2 = Não
X45.1	Se sim, é quem fica com a criança quando ela não está na creche?	3 = Sim, 4 = Não
X46	Alguém que convive com a criança tem problemas de saúde mental	1 = Sim, 2 = Não
X46.1	Se sim, é quem fica com a criança quando ela não está na creche?	3 = Sim, 4 = Não
X47	Vacinação em dia	1 = Sim, 2 = Não
X48	Frequência alimentar	1 = Baixa qualidade, 2 = qualidade intermediária, 3 = boa qualidade
X49	Segurança alimentar	1 = Segurança, 2 = Segurança leve, 3 = Insegurança moderada, 4 = Insegurança grave

Tabela 2 – Código, nomes e níveis das variáveis criadas e das categorias agrupadas.

Código	Variável	Níveis
X1	Sexo	1 = Masculino, 2 = Feminino
X2	Raça	Amarela, branca, negra, parda
X3	Idade	1 = <6 meses a 11 meses; 2 = 12 a 23 meses; 3 = 24 a 35 meses; 4 = 36 a 47 meses; 5 = 48 a 59 meses; 6 = >60 meses
X4	Situação econômica	1 = até 1 salário mínimo; 2 = 2 a 3 salários mínimos; 3 = mais que 3 salários mínimos
X5	Número de pessoas na família	$1,2,\cdots,6$
X6	Escolaridade do pai	1 = Primário incompleto, 2 = Ginasial incompleto, 3 = 2^o grau incompleto, 4 = 2^o grau completo e superior incompleto, 5 = Superior completo
X7	Escolaridade da mãe	1 = Primário incompleto, 2 = Ginasial incompleto, 3 = 2^o grau incompleto, 4 = 2^o grau completo e superior incompleto, 5 = Superior completo
X8	Habitação	1 = Residência própria quitada, 2 = própria com financiamento a pagar, 3 = cedida, 5 = alugada
X9	Profissão do chefe da família	Grau I e II, III, IV, V
X10	Posse de automóvel	1 = Não possui, 2 = possui um, 3 = possui 2 ou mais
X11	Abastecimento de água da rede pública	1 = Sim, 2 = Não
X12	Água de beber	1 = Filtrada, 2 = fervida, 3 = clorada, 4 = mineral, 5 = não tratada
X13	Consultas pré-natal	1 = Sim, 2 = Não
X14	Gravidez planejada	1 = Sim, 2 = Não
X15	Problema de saúde durante o pré-natal	1 = Sim, 2 = Não
X16	Gravidez de alto risco	1 = Sim, 2 = Não
X17	Idade gestacional	1 = Pré-termo, 2 = termo, 3 = pós-termo
X18	Local do parto	1 = Hospitalar, 2 = outro
X19	Tipo de parto	1 = Normal, 2 = cesárea, 3 = fórceps
X20	UTI/CTI neonatal	1 = Sim, 2 = Não
X21	Apgar	1 = Reanimação, 2 = asfixia moderada, 3 = vitalidade normalidade
X22	Peso	1 = <2500g, 4 = 2501-4000g, 5 = >4000g
X23	Estatura	1 = Pequeno para idade gestacional, 2 = adequado para idade gestacional, 3 = grande para idade gestacional
X24	Perímetro cefálico	1 = Abaixo do esperado, 2 = adequado, 3 = acima do esperado
X25	Teste pezinho	1 = Positivo, 2 = Negativo
X27	Teste olhinho	1 = Positivo, 2 = Negativo
X29	Teste audição	1 = Positivo, 2 = Negativo
X31	Aleitamento exclusivo	1 = Sim, 2 = Não
X32	Aleitamento até 6 meses de idade	1 = Sim, 2 = Não
X33	Idade do ínicio do desmame	0 = Ainda mama, 1 = <2 meses, 2 = 2 a 6 meses, 3 = >6 meses, 4 = 1-2 anos, 5 = >2anos
X34	Uso de leite artificial	1 = Sim, 2 = Não
X35	Idade de uso de leite artificial	1 = <6 meses, 2 = >6 meses
X36	Tipo de leite	1 = Outros, 2 = fórmulas infantis
X37	Idade de introdução de outros alimentos	1 = <6 meses, 2 = >6 meses
X38	Número de refeições	$1,2,\cdots,10$
X39	Uso de vitaminas	1 = Sim, 2 = Não
X40	Uso de sulfato ferroso	1 = Sim, 2 = Não
X41	Acompanhamento de rotina	1 = Sim, 2 = Não
X42	Equipe de saúde em visitas domiciliares	1 = Sim, 2 = Não
X43	Doença grave ou infecção de repetição	1 = Sim, 2 = Não
X44	Quem fica com a criança na maior parte do tempo	1 = mãe/pai, 2 = avó/avô, 3 = irmã/irmão menor de idade, 4 = outro
X51	Quem convive com a criança é usuário de álcool ou droga	3 = Sim, 4 = Não
X50	Quem convive com a criança tem problemas de saúde mental	3 = Sim, 4 = Não
X47	Vacinação em dia	1 = Sim, 2 = Não
X48	Frequência alimentar	1 = Boa qualidade, 2 = qualidade intermediária ou ruim
X49	Segurança alimentar	1 = Segurança, 2 = Insegurança

APÊNDICE C – Códigos em R Markdown para seleção de variáveis (LASSO e Árvore de classificação) e ajuste dos modelos logísticos para a variável segurança alimentar (EBIA)

Carregue os pacotes necessários:

```
library(glmnet)
library(plotmo)
library(rpart)
library(rpart.plot)
library(hnp)
```

Carregue o banco de dados:

```
setwd("~/Desktop/Paula Santos/Mestrado/Izabela/Dissertacao/Analises finais/lasso")
dados <- read.csv("EBIA.csv", sep = ";", header = TRUE)
```

A variável resposta de interesse é X49.bin (Classificação EBIA em duas categorias: 1=segurança, 2=insegurança).

As variáveis estão codificadas como numéricas, então é necessário que sejam declaradas como fatores.

```
dados$X1 <- as.factor(dados$X1)
dados$X2 <- as.factor(dados$X2)
dados$X3 <- as.factor(dados$X3)
dados$X4 <- as.factor(dados$X4)
dados$X5 <- as.factor(dados$X5)
dados$X6 <- as.factor(dados$X6)
dados$X7 <- as.factor(dados$X7)
dados$X8 <- as.factor(dados$X8)
dados$X9 <- as.factor(dados$X9)
dados$X10 <- as.factor(dados$X10)
dados$X11 <- as.factor(dados$X11)
dados$X12 <- as.factor(dados$X12)
dados$X22 <- as.factor(dados$X22)
dados$X31 <- as.factor(dados$X31)
dados$X32 <- as.factor(dados$X32)
dados$X33 <- as.factor(dados$X33)
dados$X37 <- as.factor(dados$X37)
dados$X38 <- as.numeric(dados$X38)
dados$X49.bin <- as.factor(dados$X49.bin)
```

Consideraremos para análise somente as variáveis que fazem sentido prático do ponto de vista do pesquisador da área aplicada.

```
dd <- dados[,c(1,2,3,4,5,6,7,8,9,10,11,12,22,28,29,30,34,35,45)]
```

Técnica de regularização Lasso

Todos os valores ausentes no banco de dados serão excluídos.

```
dt <- na.exclude(dd)
```

```
x <- model.matrix(~ . - X49.bin, data = dt)[,-1]
y <- cbind(dt$X49.bin)

# Ajuste com penalização lasso
mlasso <- glmnet(x = x, y = y, family = "binomial", alpha = 1, grouped = FALSE)

# Traços
plot(mlasso, xvar = "lambda")
```

```r
# Validação cruzada para escolha do lambda
cv.lasso <- cv.glmnet(x,
                      y,
                      family = "binomial",
                      alpha = 1)

plot(cv.lasso)
```

```r
cv.lasso$lambda.min
cv.lasso$lambda.1se
coef(cv.lasso, s = cv.lasso$lambda.min)
```

```r
plot(mlasso, xvar = "lambda")
abline(v = log(cv.lasso$lambda.min), lty = 2)
```

As variáveis selecionadas foram X2, X4, X6, X9, X10. Utilizando como critério o menor valor de lambda que minimiza o erro de validação cruzada.

Dessa forma, o modelo logístico foi ajustado.

```r
m1 <- glm(X49.bin ~ X2 + X4 + X6 + X9 + X10, data = dt, family = binomial)
```

Em seguida, aplicou-se stepwise com o objetivo de selecionar o melhor modelo.

```r
backwards <- step(m1)
```

Sendo as variáveis X2, X4, X6 as que apresentaram o modelo com menor AIC, esse foi definido como modelo final.

Sendo assim, os níveis de referência foram definidos com a finalidade de calcular as razões de chance.

```r
dt$X2 <- relevel(dt$X2, "branca")
dt$X4 <- relevel(dt$X4, "1")
dt$X6 <- relevel(dt$X6, "1")
```

O modelo final foi:

```r
m2 <- glm(X49.bin ~ X2 + X4 + X6, data = dt, family = binomial)
summary(m2)
```

As razões de chances e seus respectivos intervalos de confiança.

```r
exp(coef(m2))
exp(confint(m2))
```

Por fim, gerando o envelope simulado para verificar a qualidade do ajuste.

```r
hnp(m2)
```

Árvore de classificação

Inicialmente uma árvore foi gerada.

```r
marvore <- rpart(X49.bin ~ ., data = dd, na.action = na.rpart)
rpart.plot(marvore)
```

Considerando apenas as variáveis que apareceram no nós, um modelo foi ajustado.

```r
ebia.bin <- dd[,c(3,4,5,6,8,9,10,16,18,19)]
```

```
ebia.bin <- na.exclude(ebia.bin)

m3 <- glm(X49.bin ~ ., data = ebia.bin, family = binomial)
```

Em seguida, aplicou-se stepwise com o objetivo de selecionar o melhor modelo.

```
backwards <- step(m2)
```

Sendo as variáveis X2, X4, X6, X8 as que apresentaram o modelo com menor AIC, esse foi definido como modelo final.

Sendo assim, os níveis de referência foram definidos com a finalidade de calcular as razões de chance.

```
ebia.bin$X2 <- relevel(ebia.bin$X2, "branca")
ebia.bin$X4 <- relevel(ebia.bin$X4, "1")
ebia.bin$X6 <- relevel(ebia.bin$X6, "1")
ebia.bin$X8 <- relevel(ebia.bin$X8, "3")
```

O modelo final foi:

```
m4 <- glm(X49.bin ~ X2 + X4 + X6 + X8, data = dt, family = binomial)
summary(m4)
```

As razões de chances e seus respectivos intervalos de confiança.

```
exp(coef(m4))
exp(confint(m4))
```

Por fim, gerando o envelope simulado para verificar a qualidade do ajuste.

```
hnp(m4)
```