ALICE SILVA DUARTE

APLICAÇÕES DE MÉTODOS DE SELEÇÃO DE VARIÁVEIS EM MODELOS DE REGRESSÃO

LAVRAS – MG

2022

ALICE SILVA DUARTE

APLICAÇÕES DE MÉTODOS DE SELEÇÃO DE VARIÁVEIS EM MODELOS DE REGRESSÃO

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

Profa. Izabela Regina Cardoso de Oliveira
Orientadora

Prof. Renato Ribeiro de Lima
Coorientador

LAVRAS – MG
2022

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

Duarte, Alice Silva
Aplicações de Métodos de Seleção de variáveis em Modelos de Regressão
/ Alice Silva Duarte. – Lavras : UFLA, 2022.
70 p. : il.

Dissertação–Universidade Federal de Lavras, 2022.
Orientadora: Profa. Izabela Regina Cardoso de Oliveira.
Bibliografia.

1. Lasso. 2. Random Forest. 3. Regressão Logística. 4. Stepwise. I. Oliveira, Izabela Regina Cardoso de. II. Lima, Renato Ribeiro de. III. Aplicação de Métodos de Seleção de Variáveis em Modelos de Regressão.

ALICE SILVA DUARTE

APLICAÇÕES DE MÉTODOS DE SELEÇÃO DE VARIÁVEIS EM MODELOS DE REGRESSÃO
APPLICATIONS OF VARIABLE SELECTION METHODS IN REGRESSION MODELS

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

APROVADA em 22 de Setembro de 2022.

Prof. Daniel Furtado Ferreira	UFLA
Prof. Gustavo Henrique de Araujo Pereira	UFSCAR

Profa. Izabela Regina Cardoso de Oliveira
Orientadora

Prof. Renato Ribeiro de Lima
Co-Orientador

LAVRAS – MG
2022

À minha mãe, Maria Vicentina Silva Duarte, aos meus irmãos Oswaldo e Maísa, e ao meu noivo, Thassio Moreira Laudomiro.

AGRADECIMENTOS

A realização deste trabalho só foi possível porque tenho comigo pessoas que não mediram esforços para me ajudar no que fosse preciso. A todas elas eu expresso a minha profunda gratidão.

Começo por meu noivo Thassio, obrigada pela confiança e força para seguir em frente dia após dia e por ter sido parceiro e paciente todo este tempo. Obrigado pelos cafés e por aguentar meus momentos de ansiedade. Sem você do meu lado esse trabalho não seria possível.

Agradeço à minha família, em especial à minha mãe Maria Vicentina, aos meus irmãos Maísa e Oswaldo Duarte que apoiaram meus estudos, entenderam minhas ausências e não mediram esforços para que esse sonho se tornasse realidade. Sou grata aos meus afilhados Lívia e Daniel pela força dos sorrisos que, nos momentos difíceis oferecem um amor incondicional, puro e cheio de esperança, obrigada pelo carinho.

A universidade pública me transformou em uma pessoa melhor e ainda me deu oportunidade de conhecer pessoas e torná-las amigas, as quais eu me apoio constantemente quando me encontro em desafios e dificuldades acadêmicas. Dentre todas estas amizades, é enorme a necessidade de enfatizar meus professores e amigos Cláudia Adam Ramos e Pablo Javier Grumnam, que me fizeram acreditar em meu potencial e me apoiaram de maneira fundamental para que eu pudesse realizar minhas primeiras pesquisas. Agradeço também aos meus amigos de graduação e pós graduação Bruno Souza, Matheus Saraiva, Walef Machado, Poliana Beneli e Ana Carolina Orrico, que em muitos momentos me fizeram enxergar desafios sob a luz de outras perspectivas.

À todos os amigos da Mirador Atuarial, em especial aos lideres Fabrízio Krapf, Giancarlo Germany pela oportunidade e por acreditarem no meu trabalho. Obrigada Brenda Trajano pelos conselhos e orientações, sua liderança desperta a busca pelos melhores resultados mesmo nos piores dias. Obrigada Lucas Machado, Patrícia Kipper e Fernanda Alves pela companhia e pelo apoio mesmo nos dias mais desafiadores.

À Universidade Federal de Lavras, todos os seus professores e funcionários quero deixar uma palavra de gratidão.

A orientadora Izabela Regina Cardoso de Oliveira e coorientador Renato Ribeiro de Lima minha gratidão eterna a vocês pelos ensinamentos, pela amizade e principalmente pela dedicação.

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

Quem tiver talento, obterá o êxito na medida que lhe corresponda. Porém, apenas se persistir naquilo que faz.
(Issac Asimov)

RESUMO

Modelos de regressão são técnicas utilizadas para estabelecer relação entre uma variável resposta e uma ou mais variáveis explicativas. Com o avanço tecnológico, o volume e a dimensão dos dados analisados pode ser cada vez maior. Enquanto, por um lado, o grande número de variáveis pode aumentar a capacidade preditiva do modelo, por outro muitas dessas variáveis podem contribuir pouco e gerar um alto custo computacional, fazendo-se necessário a seleção de variáveis e busca por aquelas que têm maior impacto no modelo. O objetivo deste trabalho foi avaliar o uso de métodos de seleção de variáveis em dois estudos de caso. O primeiro trata-se de um estudo de avaliação de frequência e segurança alimentar de pré-escolares do município de Lavras, MG. As respostas analisadas nessa primeira etapa são dados de categorias da Escala Brasileira de Insegurança Alimentar (EBIA) e do Questionário de Frequência Alimentar (QFA), analisados por modelos logísticos. A amostra utilizada envolve dados de 581 pré-escolares caracterizados por cerca de 50 variáveis, de diferentes tipos. Foram considerados os métodos *Stepwise*, Lasso, o *Purposeful Selection of Covariates* (PSV) e *Random Forest* para seleção de variáveis. Posteriormente foram obtidos os modelos logísticos com as variáveis selecionadas por estes métodos. Os modelos foram avaliados em termos de AIC. Dentre os métodos avaliados o que produziu o modelo com melhor desempenho foi o *Stepwise*. A segunda aplicação envolveu um cenário de dados de alta dimensão, obtidos com a utilização de NIRS (*Near infrared spectroscopy*) em um problema de predição de consumo alimentar, a partir de fezes de vacas leiteiras. Foram considerados os métodos *Stepwise*, lasso e *Random Forest* para seleção de variáveis. O lasso apresentou bom desempenho no estudo de validação cruzada. No entanto, esse estudo se limita a utilização dos métodos de forma independente, já que outros autores obtiveram bons resultados aplicando mais de um método simultaneamente. As contribuições deste estudo de caso estão na comparação entre lasso e *Random Forest*, usados separadamente para seleção de variáveis em NIRS e a comparação entre diferentes tipos de validações para os modelos obtidos com o uso do lasso.

Palavras-chave: Alta dimensionalidade. Importância de variáveis. Lasso. Random Forest. Regressão Logística. Stepwise

ABSTRACT

Regression models are applied to study a cause/effect relationship between a response variable and one or more explanatory variables. With technological advances, the volume and dimension of the analyzed data can be increasing. While the large number of variables can increase the predictive capacity of the model many of them variables can contribute little and generate a high computational cost. Then it may be necessary to select variables and search for those that have the greatest impact in the model. In this work we evaluate the use of variable selection methods in two case studies. The first one was carried out to evaluate the frequency and food security of preschoolers in the city of Lavras, MG. The responses analyzed in this first stage are data from categories of the Brazilian Scale of Food Insecurity (EBIA) and the Food Frequency Questionnaire (FFQ), analyzed through logistic models. Data were collected from 581 preschoolers and refer to about 50 variables of different types. The methods Stepwise, Lasso, the Purposeful Selection of Covariates (PSV) and Random Forest were considered for the selection of variables. Subsequently, the logistic models were obtained with the variables selected by these methods. The models were evaluated in terms of AIC. Among the evaluated methods, the one that produced the best performing model was Stepwise. The second application involved a high-dimensional data scenario, obtained with the use of NIRS (Near infrared spectroscopy) in a problem of predicting food consumption, from feces of dairy cows. The methods Stepwise, lasso and Random Forest were considered for the selection of variables. Lasso performed well in the cross-validation study. However, this study is limited to the use of the methods independently. Other authors obtained good results applying more than one method simultaneously. The contributions of this case study are the comparison among lasso and Random Forest, used separately for the selection of variables in NIRS and the comparison between different types of validations for the models obtained using lasso.

Keywords: High dimensionality. Variables importance. Lasso. Random Forest. Logistic Regression. Stepwise

LISTA DE FIGURAS

SUMÁRIO

Parte I

PRIMEIRA PARTE

1 INTRODUÇÃO

A busca por explicar a realidade por meio de modelos matemáticos não é recente. Uma das técnicas mais usadas para isso é a análise de regressão, originada dos experimentos de Francis Galton (1822-1911), na busca da relação explicativa da altura de pais e filhos (SOUSA; ALVES, 2016). A regressão é uma das técnicas mais difundidas em pesquisas aplicadas e tem como objetivo estabelecer uma relação entre as variáveis explicativas e a variável resposta (DRAPER; SMITH, 1998).

Apesar do avanço tecnológico das últimas décadas e do surgimento de diversos outros métodos para modelar e predizer o comportamento de variáveis, os modelos de regressão continuam sendo amplamente utilizados, por serem relativamente simples de serem implementados computacionalmente e pela flexibilidade em relação à predição da média da variável resposta, sendo recomendados para respostas não normais ou normais. Assim, mesmo com o surgimento de diversas outros métodos, os modelos de regressão continuam sendo uma escolha natural em diversas situações (NELDER; WEDDERBURN, 1972).

Em estudos aplicados podem existir diferentes tipos de variáveis respostas de interesse. Um tipo é avariável continua, cuja resposta é um número de um conjunto infinito de resultados possíveis. Essas variáveis representam resultados numéricos de medições, que podem ser preditos por um modelo de regressão (WEISBERG, 2005; BRYK; RAUDENBUSH, 1992). Por outro lado, quando o problema a ser resolvido é o de classificação e associado a uma variável resposta binária, um modelo de regressão amplamente utilizado é o logístico. Este modelo é usado em Ciências médicas, Ciências sociais, em instituições financeiras e também é generalizado para explicar variáveis categóricas de natureza nominal ou ordinal (AGRESTI, 2018; FIGUEIRA, 2006; LAVALLEY, 2008).

Com o avanço da tecnologia o volume de dados cresceu de maneira exponencial sendo necessário o uso de técnicas cada vez mais avançadas para explicar a relação entre estas (muitas) variáveis. Em uma modelagem estatística, o grande número de variáveis pode aumentar a capacidade preditiva do modelo, mas, por outro lado, muitas dessas variáveis podem contribuir pouco e gerar um alto custo computacional, fazendo-se necessário a seleção de variáveis, sendo natural procurar por aquelas que têm maior impacto no modelo (JUNIOR, 2021).

Os métodos mais utilizados no estudo de seleção de variáveis são os métodos *forward*, *backward* e *stepwise*. Esses métodos se baseiam em um algoritmo de decisão que avalia os

modelos por algum critério (AIC, BIC, R^2 Ajustado, F parcial ou T para situações de regressão). O método *stepwise* é o mais usual para seleção de variáveis sendo composto pela união do método *forward* e *backward* (CARSON; CHASE, 2009; GREENLAND, 1989; SULLIVAN et al., 1990).

Enquanto o *forward* parte do modelo sem variáveis e vai acrescentando variáveis uma a uma, o *backward* faz o caminho oposto, começando com o modelo completo e retirando variáveis menos importantes. O *stepwise*, também abordado na literatura como *forward and backward stepwise*, une os dois métodos anteriores verificando a cada passo todas as variáveis anteriores (THOMPSON, 1995).

Apesar de esses métodos serem muito difundidos na tarefa de seleção de variáveis, eles podem ser pouco aplicáveis em algumas situações, como quando a base de dados é de alta dimensão. Para este tipo de dados esses métodos tradicionais têm elevado custo computacional e, em algumas situações, não são eficazes.

Como solução para o problema de seleção de variáveis em alta dimensionalidade surgiram os métodos de regularização em regressão, que adicionam uma penalização na equação de mínimos quadrados. O método de regularização lasso (*last absolute shrinkage and selection operator*), proposto por Tibshirani (1996), pode ser usado na estimação dos parâmetros ou para seleção de variáveis (JUNIOR, 2021), pois pode fazer com que as estimativas de alguns parâmetros associados às variáveis preditoras tendam a zero.

Em *Machine Learning* um método que tem sido usado na seleção de variáveis em modelos de regressão, especialmente em grandes bancos de dados, é o *random forests* (florestas aleatórias) ou *random decision forests* (florestas de decisão aleatória). Esse método, em síntese, constrói e agrega um grande volume de árvores de decisão, estabelecendo regras em cada uma delas. A estrutura dessas árvores se assemelha a um fluxograma, verificando uma condição a cada nó e levando assim ao próximo nó, até a finalização da árvore. Este método é uma técnica supervisionada de *machine learning* que utiliza parte dos dados para aprendizado e parte para teste (AGRESTI, 2018). No contexto de seleção de variáveis, pode ser usado o conceito de importância de variáveis, que é uma medida obtida com a aplicação desse método (GENUER; POGGI; TULEAU-MALOT, 2010; CHEN; ISHWARAN, 2012; STROBL et al., 2008).

Por fim, dentro do contexto de regressão logística, tem-se o método proposto por Hosmer, Lemeshow e Sturdivant (2013), cuja proposta é considerar a importância prática da variável

na resposta em questão, estudando marginalmente a associação de cada covariável com ela e incluindo aquelas de maior importância, segundo os valores-p dos testes marginais. Esse método, conhecido como *Purposeful selection of variables* (PVS) e que muito se assemelha ao procedimento de *stepwise*, tem sido usado em estudos nas áreas de ciências da saúde (FUCHS et al., 2013; SANCHEZ-PINTO et al., 2018; TENÓRIO et al., 2020) e em outras áreas (KOGA et al., 2015).

Alguns autores estudaram o desempenho dos métodos supracitados por estudos de simulação e em trabalhos aplicados. Bursac et al. (2008) compararam o método *Purposeful selection of variables* com o método *Stepwise* e Junior (2021) comparou lasso com *Stepwise*.

O objetivo deste trabalho é aplicar e avaliar diferentes métodos de seleção de variáveis em dois diferentes cenários. O primeiro deles envolve um problema de classificação, em que são avaliados os métodos de seleção citados anteriormente para obtenção de modelos logísticos. Para este primeiro estudo aplicado foram comparados os métodos lasso; *stepwise*; *purposeful selection of variables* e *random forest*

A segunda aplicação consiste em selecionar variáveis para um problema de predição a partir de dados de alta dimensão. Nesta segunda aplicação, foram considerados os métodos *stepwise*, lasso, *random forest*.

2 MODELOS DE REGRESSÃO

Modelos de regressão podem ser definidos como técnicas utilizadas para modelar o relacionamento entre duas ou mais variáveis sendo uma delas a variável resposta y (dependente) e as demais variáveis explicativas x (independentes). No modelo de regressão simples, quando há apenas uma variável explicativa, o modelo é conhecido como modelo linear simples. Quando há mais de uma variável explicativa é considerado um modelo linear múltiplo.

O modelo de regressão linear geral é definido como:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \tag{2.1}$$

em que:

y_i: valor da variável resposta da observação i $(i = 1, 2, ..., n)$;

x_{ij}: valor da variável independente j $(j = 1, 2, ..., p)$ do indivíduo i;

β_0 e β_j os parâmetros desconhecidos;

ε_i vetor dos erros associados a cada observação. Os erros ε_i apresentam média zero e variância constante σ^2 e não são correlacionados entre si e com as observações.

A Equação 2.1 assume como pressuposto que os erros são independentes e identicamente distribuídos (iid) com média 0 e variância constante, usualmente assume-se também que os erros possuem distribuição normal, ou seja, $\varepsilon \sim N(0, \sigma^2)$. Outro pressuposto dos modelos lineares de regressão é que a relação entre as covariáveis e a média da variável resposta é linear.

Na equação 2.1, os parâmetros desconhecidos β_0 e β_j representam, respectivamente, o intercepto do modelo e o efeito das variáveis explicativas do modelo. A estimativa desses parâmetros é importante para que o modelo possa ser usado para fins inferenciais (TIBSHIRANI, 2011; **??**).

O método estatístico mais frequente para estimação dos coeficientes de um modelo de regressão é o método de mínimos quadrados ordinários (MQO), que é recomendado por sua precisão (**??**GUIMARÃES, 2008).

A estimação por mínimos quadrados consiste na obtenção de estimativas dos parâmetros β_0 e β_j, de forma que a soma dos quadrados dos erros, ε_i, seja mínima. Assim, busca-se minimizar a soma de quadrados dada por:

$$SQ = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 . \tag{2.2}$$

O mínimo obtido por meio de técnicas de cálculo, mais especificamente por derivadas parciais, resulta nos estimadores dos parâmetros. No entanto, por se tratar de modelos lineares generalizados com j parâmetros a abordagem matricial se mostra mais vantajosa na busca pelo vetor de estimadores β . Pela abordagem matricial a equação 2.1 pode ser reescrita como:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{2.3}$$

ou

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Assume-se que os erros ε_i tem média 0 e variância constante σ^2 além de não correlacionados com quaisquer observações (RENCHER; SCHAALJE, 2008).

Considerando a forma matricial definida na equação 2.3, o estimador do vetor de parâmetros desconhecidos é obtido por meio da maximização da soma de quadrados dos erros.

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y} \tag{2.4}$$

O uso de modelos de regressão vai desde inferências por meio de intervalos de confiança para média da variável resposta, seus parâmetros, testes de hipóteses sobre os coeficientes (β) até modelos para estimar, prever ou predizer valores para a variável resposta y dado as variáveis independentes x_i. No entanto, para que o modelo possa ser usado para fins inferenciais é necessário a realização do diagnóstico do ajuste (ou análise de diagnóstico) que tem como objetivo subsidiar e avaliar a qualidade do modelo.

Quando a variável resposta é do tipo binária o modelo mais utilizado é utilizado a regressão logística. Esses modelos são também chamados de modelos *logit* e simbolizados por $logit(\pi)$. A regressão logística descreve situações em que a relação entre $\pi(x)$ e x é não lineares

(AGRESTI, 2018). Este modelo permite a estimação da probabilidade associada a ocorrência de um evento frente a um conjunto de variáveis.

Sejam $x_1, .., x_n$ observações independentes tomadas para analisar $\pi(x)$. De acordo com Figueira (2006) é razoável assumir como suposição inicial que $\pi(x)$ é monótona com valores entre zero e um quando x varia na reta real.

Uma representação linear simples para π pode não se adequar, fazendo-se necessário a transformação logística de $\pi(.)$, chamada de *transformação logit* e definida como:

$$y_i = logit[P(Y=1)] = log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \sum_{i=1}^{p} \beta_j x_{ij} + \varepsilon_i \tag{2.5}$$

O parâmetro β_i refere-se ao efeito das variáveis explicativas x_i no logit das probabilidades de $Y = 1$.

O modelo de regressão logística é usado em diversas áreas, desde classificação de risco para empresas (BRITO; NETO, 2008) até modelos voltados para a área da saúde (AVANCI et al., 2007). Em epidemiologia esse tipo de modelo tem sido amplamente usado por descrever bons ajustes, ser biologicamente plausível e parcimonioso (HOSMER; LEMESHOW; STURDIVANT, 2013).

Tanto o modelo linear geral quanto o modelo de regressão logística são casos especiais dos modelos lineares generalizados propostos por Nelder e Wedderburn (1972). Esses modelos representam uma classe de modelos de regressão que buscam descrever relações lineares entres variáveis e ampliam as possibilidades de distribuição da variável resposta para distribuições pertencentes à família exponencial, que pode ser expressa por:

$$f(y,\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{2.6}$$

em que $a(\phi), b(\theta)$ e $c(y,\phi)$ são funções específicas. O parâmetro θ é chamado de posição (ou locação) e o ϕ de parâmetro de dispersão, também abordado como σ^2.

Essa família de distribuições possui média e variância dadas por:

$$E(y) = \mu = \frac{d(\theta)}{d\theta} \quad \text{e} \quad V(y) = \frac{d^2(\theta)}{d(\theta)} a(\phi)$$

Pertencem a essa família, por exemplo, as distribuições normal, gama, poisson, binomial e a normal inversa.

Todos os Todos os Modelos Lineares Generalizados (GLM) são construídos através de três componentes: uma componente aleatória, uma componente sistemática e uma função de ligação. A componente aleatória diz respeito à variável resposta, cujas premissas assumidas se referem a aleatoriedade e distribuição de probabilidade pertencente à família exponencial. Já a componente sistemática diz respeito à função do preditor da variável regressora, que deve ser linear nos parâmetros. Por fim, a função de ligação é a componente responsável por conectar as demais componentes, por meio de uma função $g(\mu) = X\beta$, em que $\mu = E(Y)$ (CASELLA; BERGER, 2002).

No modelo de regressão clássico a média é conectada ao preditor linear por meio da função de ligação identidade e assume-se a distribuição normal para a variável resposta. Já na regressão logística a função de ligação utilizada é a logit e assume-se a distribuição Bernoulli para a variável resposta (AGRESTI, 2018; FIGUEIRA, 2006).

Nos modelos lineares generalizados, um dos métodos de diagnóstico mais utilizado é aquele que considera o desvio residual (*deviance residual*) (DAVISON; GIGLI, 1989). Esse método consiste em medir a diferença entre os ajustes do modelo corrente e o modelo saturado, sendo o modelo saturado o mais complexo, com n parâmetros e o modelo corrente com p parâmetros ($p < n$) (AGRESTI, 2018).

A *Deviance* em um modelo linear generalizado é definida por Agresti (2018) por

$$Deviance = -2[L_M - L_S] \tag{2.7}$$

em que L_M e L_S representam, respectivamente, o máximo da função log-verossimilhança (*log-likelihood*) para o modelo M de interesse e para o modelo mais complexo possível (S), denomidado modelo saturado. Assim, menores valores para a *deviance residual* implicam que o modelo ajustado é tão bom quanto o modelo saturado.

2.1 Seleção de variáveis em modelos de regressão

Em muitos problemas aplicados, pesquisadores deparam-se com o problema de escolher, dentre diversas variáveis preditoras potenciais, aquelas que são mais importantes ou que estão realmente associadas à resposta de interesse. Nesta seção serão descritos alguns critérios estatísticos que podem ser usados para selecionar essas variáveis.

2.1.1 Métodos *Stepwise*, *Backward* e *Forward*

Os métodos *forward, backward* e *stepwise* são os métodos mais usuais, quando se trata do problema de seleção de variáveis em modelos de regressão.

O método *forward*, conhecido como inclusão passo a frente, consiste em ordenar as variáveis preditoras de acordo com a intensidade de sua relação com a variável resposta e ajustar o modelo. O modelo é ajustado primeiramente somente com o intercepto e em seguida adicionando as variáveis, uma a uma, e testando a significância dessa adição no modelo. Se a adição for significativa a variável entra no modelo; se não for o procedimento é encerrado (HOCKING, 1976).

Já o método *backward,* conhecido como eliminação passo atrás, parte pelo caminho oposto. Ele ajusta um modelo com todas as variáveis preditoras e remove variáveis. Caso a remoção não gere perdas de desempenho, o modelo fica sem a variável (EFROYMSON, 1960; KUTNER et al., 2005).

Uma limitação do método *forward*, descrito acima, é que, uma vez que uma variável sai do modelo, ela não entra mais, mesmo que sua contribuição passe a ser significativa com a entrada de outra variável.

O método *stepwise* é um dos métodos mais usados em regressão e é composto pela união dos métodos *forward* e *backward* (NETER et al., 1996). Ele atua nos dois sentidos, adicionando e removendo variáveis. A cada adição é verificado se alguma das variáveis existentes pode ser excluída e a seleção é finalizada quando nenhuma adição ou exclusão melhora o desempenho do modelo (EFROYMSON, 1960; KUTNER et al., 2005).

No entanto esse método possui algumas limitações, principalmente quando há um grande número de variáveis, situação na qual a seleção pode ser lenta ou até impossível de ser executada.

2.1.2 *Purposeful selection of covariates*

Outra forma de selecionar covariáveis e que foi proposta dentro do contexto de regressão logística por Hosmer, Lemeshow e Sturdivant (2013). Conhecido como *Purposeful selection of covariates*, o procedimento é resultante de marginais entre as preditoras e a variável resposta.

Os passos desse algoritmo consistem em:

- **Passo 1:** O método é iniciado com o estudo das associações marginais entre cada variável explicativa candidata e a variável de interesse (resposta). Para variáveis categóricas, as análises são feitas a partir das tabelas de contingência. Para variáveis contínuas, esse estudo envolve o ajuste de um modelo logístico para cada variável. As variáveis candidatas para um primeiro modelo (abordado neste trabalho como "modelo de regressão múltipla") serão aquelas cujos testes marginais para seus efeitos apresentaram valor-p menor que 0,25, dentre todas as variáveis de importância prática conhecida, consideradas.

- **Passo 2:** Ajustar o modelo de regressão múltipla contendo todas as variáveis do passo 1 e avaliar a importância de cada uma pelo valor-p. Variáveis que não contribuem devem ser retiradas e o novo modelo (menor que o anterior) deve ser comparado com o antigo pelo teste de razão de verossimilhança parcial.

- **Passo 3:** Ajustado o modelo reduzido, seus coeficientes estimados devem ser comparados com o modelo de regressão múltipla inicial, atentando-se especialmente às variáveis cujos coeficientes mudaram muito em magnitude. Isso pode indicar que alguma variável excluída era importante, no sentido de fornecer um ajuste necessário das demais variáveis. Este processo de exclusão, ajuste e verificação, continua pelos passos 2 e 3, até que todas as variáveis importantes estejam no modelo e as variáveis excluídas sejam sem importância do ponto de vista prático ou estatístico.

- **Passo 4:** Adicionar cada variável não selecionada no Passo 1 ao modelo obtido, uma de cada vez, verificando sua significância pelo valor-p da estatística de Wald ou pelo teste da razão de verossimilhanças parcial, no caso de uma variável categórica com mais de 2 níveis. Esta etapa identifica variáveis que não estão significativamente relacionadas com a resposta, mas fornecem uma contribuição importante na presença de outras variáveis. O modelo resultante nessa etapa é chamado pelos autores de modelo de efeitos principais preliminares.

- **Passo 5:** Nessa etapa, considera-se que todas as variáveis importantes para a variável resposta de interesse estão consideradas no modelo. Agora, para cada variável contínua do modelo, deve ser verificada a suposição de que o modelo é linear no *logit*. O modelo

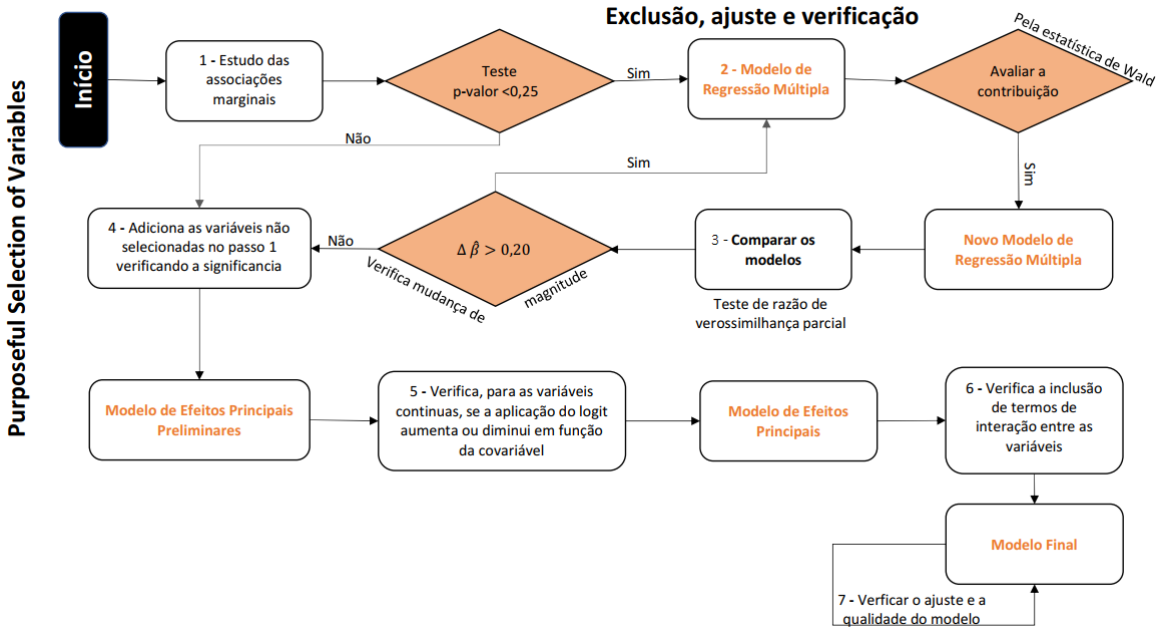
de efeitos principais, resultante do refinamento nesse passo, é chamado de modelo de efeitos principais.

- **Passo 6:** Com o modelo de efeitos principais pronto é verificada a interação entre as variáveis. Em qualquer modelo, a interação entre duas variáveis implica que o efeito de cada variável não é constante ao longo dos níveis da outra variável. Dessa forma, a decisão final sobre a inclusão de um termo de interação deve ser baseada tanto em critérios estatísticos quanto em considerações práticas.

- **Passo 7:** Verificar o ajuste e qualidade do modelo final.

A figura 2.1 exibe o fluxograma dos passos do método PSV citado anteriormente.

Figura 2.1 – Fluxograma do método Purposeful Selection of Variables



Fonte: Da autora (2022).

Bursac et al. (2008) compararam o desempenho deste método com outros três procedimentos de seleção, entre eles o método *Stepwise*. O PVS se mostrou mostrou superior, pois além de selecionar variáveis significativas, o método também incluiu variáveis que eram confundidoras de outras variáveis do modelo.

2.1.3 Regressão Lasso

Um dos maiores problemas nos modelos de regressão é o balanço entre viés e variância. De modo simplificado, o viés mede o quão próximas as estimativas obtidas com o modelo ajustado estão das verdadeiras respostas. A variância é uma medida de dispersão que traduz a distância entre os valores preditos e a média, representando o quanto o modelo consegue se adaptar às mudanças de amostra (JUNIOR, 2021).

Sendo assim, um bom modelo deve apresentar um bom balanço entre viés e variância. Para isso são utilizadas técnicas de regularização, que adicionam uma penalização ao método de mínimos quadrados e têm como objetivo central diminuir a variância, evitando o ajuste excessivo dos dados. Dentre estes métodos, o que será utilizado ao longo deste trabalho é o lasso (*last absolute shrinkage and selection operator*), proposto por Tibshirani (1996).

O lasso minimiza a soma de quadrados adicionando um peso, ou penalização não negativa, de forma a criar esparsidade dentro do modelo, isto é, fazendo muitos coeficientes convergirem para zero. Dessa forma, o lasso, simultaneamente, estima os parâmetros e seleciona covariáveis para o modelo, sendo um método tanto de estimação quanto de seleção de variáveis (JUNIOR, 2021).

O método de penalização lasso contorna problemas de alta dimensionalidade dos dados, ou seja, problemas que surgem quando há um número muito grande de dimensões que chega a ser comparado ao tamanho da amostra (BEUREN, 2010). Este método tem como característica ser indiferente quanto aos preditores correlacionados tendendo a escolher um e ignorar os demais (TIBSHIRANI, 1996). O estimador lasso, para modelos lineares, é dado pela minimização da equação

$$\hat{\beta}^{lasso} = argmin_{\beta} \left\{ \sum_{j=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2.8}$$

sendo:

$\sum_{j=1}^{p} |\beta_j| = \|\beta\|$

sendo $\sum_{j=1}^{p} |\beta_j|$ a penalização lasso. Devido à natureza desta restrição, tornar λ suficientemente grande faz com que alguns coeficientes sejam exatamente zero fazendo uma seleção contínua de subconjuntos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Para determinar a melhor alternativa para o parâmetro de ajuste λ, uma possibilidade é o uso da a técnica de validação cruzada, considerando uma grade de possíveis valores de λ e calculando o erro de validação para cada um, aquele que levar ao menor erro de validação, será o λ utilizado. Assim o modelo é reajustado com o valor de λ encontrado (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A função objetivo descrita pela expressão 2.8, conforme Hastie, Tibshirani e Friedman (2009), pode ser dada pela sua forma lagrangiana como sendo

$$min_{\beta,\beta_0}(y-\mathbf{Z}\beta)^T(y-\mathbf{Z}\beta)+\lambda\cdot\|\beta\|_1 \tag{2.9}$$

em que y representa a variável resposta e $\mathbf{Z}$ a matriz na qual cada linha representa os valores das p covariáveis de cada uma das x observações. A utilização dessa forma lagrangiana apresenta como vantagens o uso da técnica de coordenada descendente, facilitando a solução da função objetivo e a estimação dos parâmetros (JUNIOR, 2021).

Outro ponto que merece destaque na abordagem do método de seleção lasso é o balanço entre viés e variância. O viés do estimador mede a proximidade média entre as estimativas do modelo ajustado e a variável resposta, ou seja, quanto menor o viés mais próximo os valores estimado $\hat{y}$ estão da variável resposta y. No entanto, modelos com viés extremamente baixos podem ter problemas de superajuste, podendo ter resultado restrito apenas à amostra em questão (REID; TIBSHIRANI; FRIEDMAN, 2016; JUNIOR, 2021).

Assim, modelos com viés extremamente baixo (próximo de zero) podem elevar consideravelmente a variância em seus valores estimados reduzindo a capacidade de generalização para outras amostras. Em síntese, quanto menor o valor do parâmetro λ, maior a quantidade de covariáveis no modelo reduzindo o viés e aumentando a variância. O oposto acontece quando aumenta-se o valor de λ, neste caso, menos covariáveis são selecionadas diminuindo a variância e aumentando o viés de estimação.

Dessa forma, a escolha do parâmetro λ deve levar em consideração o balanço entre viés e variância, de forma a garantir que nenhum dos dois seja prejudicial ao modelo. Hastie, Tibshirani e Wainwright (2015) abordam a técnica de validação cruzada como sendo uma boa opção para estimar o parâmetro λ, garantindo o balanço entre viés e variância do modelo.

O método de penalização lasso, além de diminuir a variância do modelo, ainda tem a vantagem, em casos em que há variáveis altamente correlacionadas, de selecionar apenas uma

delas, zerando os coeficientes das outras, de forma a minimizar a penalização, facilitando a interpretação do modelo (SILVA, 2018).

2.2 *Random Forest*

Com o avanço na tecnologia e o aumento no volume de dados, algumas situações exigem técnicas automatizadas de seleção de variáveis e os métodos de *machine learning* (aprendizado de máquina) impulsionam metodologias que podem ser aplicadas nestas situações (XIN et al., 2018; AZEVEDO, 2019).

Os métodos de *machine learning* podem ser categorizados basicamente de três formas: aprendizagem supervisionada, não supervisionada ou semi-supervisionada (JAMES et al., 2013). As técnicas de aprendizado supervisionado são aquelas que necessitam de acompanhamento humano na fase de saída de dados, ou seja, precisa de ajuda externa para ser executado, sendo os dados são treinados e testados (MAHESH, 2020). Esses algoritmos contam apenas com dados rotulados, ou seja, para cada observação $x_1, x_2, ..., x_p$ há uma resposta associada (y_i).

As técnicas de aprendizado não supervisionado são aquelas que não possuem supervisão e nem resposta correta, ou seja, os dados não são rotulados (IZBICKI; SANTOS, 2020).

Já o aprendizado semi-supervisionado trata-se de uma mistura de técnicas supervisionadas e não supervisionadas, ou seja, uma combinação dos métodos anteriores. Nesta combinação é usual que se tenha uma pequena parcela de dados rotulados, ou seja, que possuem variáveis de entrada e suas respectivas saídas e variáveis não rotuladas que possuem apenas a entrada (ENGELEN; HOOS, 2020).

Random Forest é um dos métodos de aprendizagem supervisionada baseado em diversas árvores de decisão e amplamente utilizados em problemas de classificação ou regressão. Ele combina a simplicidade das árvores de decisão com a flexibilidade e a aleatoriedade, para melhorar a precisão do modelo. Esse método adiciona aleatoriedade aos dados para resolver um problema comum das árvores de decisão, o *overfitting*.

O método cria diversas árvores de decisão para predizer melhor o resultado. A aleatoriedade está presente na seleção de atributos que serão utilizados na criação das árvores que serão também treinadas com conjuntos de dados distintos, garantindo que cada árvore levará a um modelo diferente (BREIMAN, 2001).

As árvores de decisão são a menor unidade de um modelo *Random Forest* (RF) cuja estrutura depende da particularidade de cada aplicação. Zhang e Singer (2010) descrevem a estrutura de uma árvore de decisão a partir de camadas. A primeira camada é composta pela raiz da árvore, local de início do processo de particionamento dos dados. Na raiz de uma árvore aleatória encontram-se os dados amostrais de treinamento ou o conjunto total dos dados, portanto, qualquer estrutura subsequente possui uma menor quantidade de dados. Cada decisão (ou nó) é criado a medida que uma amostra atravessa a árvore (WEST; BHATTACHARYA, 2016).

A quantidade de camadas que uma árvore possui depende da quantidade de conclusões binárias realizadas. Dessa forma, da raiz de uma árvore podem surgir um nó interno, que leva a uma nova decisão, ou um terminal. Assim, o particionamento da árvore é repetido até que todas as decisões cheguem ao estado terminal, que representa a conclusão (ZHANG; SINGER, 2010).

O algoritmo de *random forest* seleciona parte dos dados originais de maneira aleatória e, em seguida, são criadas as árvores da florestas, a partir desse conjuntos.Sendo assim, a primeira árvore é criada a partir de p características (variáveis) selecionadas aleatoriamente para a criação da árvore. A partir de um processo de avaliação da impureza, é selecionada a característica que ficará no topo da árvore. As árvores são construídas considerando apenas os subconjuntos de atributos selecionados e não foi realizado processo de poda (BREIMAN, 2001; IZBICKI; SANTOS, 2020).

O objetivo é ter árvores de tamanhos distintos e com variáveis distintas para gerar modelos diferentes e mais robustos. A partir daí o algoritmo de *Random Forest* compara os diversos modelos, sempre partindo do topo de cada árvore e tomando a classificação ou predição da variável. Esse processo é chamado de *bagging*. Em síntese, o algoritmo de *Random Forest* pode ser descrito nos 4 passos a seguir:

- **Passo 1:** Criação do *bootstrap dataset* (método de criar diversos bancos de dados por reamostragem);
- **Passo 2:** A cada passo são selecionadas p variáveis para montar a árvore;
- **Passo 3:** Diversas árvores são criadas a partir de subconjuntos diferentes;

- **Passo 4:** O teste deve percorrer todas as árvores e a classificação será aquela mais frequente para modelos de classificação. Em modelos de regressão, a saída desse algoritmo representa a média das saídas de todas as árvores da floresta aleatória.

Esse método tem como vantagem ser robusto e menos propenso a sofrer *overffiting*, quando comparado a uma única árvore de decisão, além de permitir descobrir as variáveis que mais impactaram no modelo.

Usar o método de *random forest* para selecionar variáveis requer a definição de uma medida de importância para identificar as variáveis que mais colaboram para a decisão e também um critério de avaliação dos subconjuntos que serão gerados (FONTES, 2020).

Uma limitação desse método para seleção de variáveis é o custo computacional, que pode ser elevado na etapa de treinamento, dependendo do conjunto de dados que está sendo trabalhado.

Neste trabalho, com o objetivo de selecionar variáveis, o uso de *random forest* foi associado à eliminação recursiva de características (RFE), por meio do pacote *VSURF* do software R (GENUER; POGGI; TULEAU-MALOT, 2015). Esse algoritmo seleciona um pequeno conjunto possivelmente não colinear de preditores para ajuste do modelo. O procedimento implementado no pacote *VSURF* realiza seleção *backward* das variáveis, classificando a importância de cada uma e construindo vários modelos de calibração com base nos preditores mais importantes $x_1, x_2, x_3, ..., x_p$ dos possíveis preditores, produzindo o melhor modelo dentre os candidatos.

2.3 Métodos de Validação Cruzada

São diversos os métodos de avaliar e identificar um bom modelo. No que diz respeito ao aprendizado supervisionado, destaca-se a validação cruzada como sendo uma forma de comparar o aprendizado do algoritmos, por meio da divisão da amostra em duas partes. A primeira parte é a parcela de treino, que será utilizada para o aprendizado e calibração do modelo e a segunda parte é a amostra de teste que, posteriormente, será comparada com os resultados preditos pelo modelo treinado (REFAEILZADEH; TANG; LIU, 2009).

Ao longo deste trabalho serão utilizados três tipos de validação cruzada, a validação cruzada *k-fold*, a validação cruzada *leave-one-animal-out* (LOAO) e a validação *leave-one-experiment out* (LOEO).

De acordo com Refaeilzadeh, Tang e Liu (2009), o método de validação cruzada mais usual é a validação cruzada *k-fold*. Essa validação particiona os dados em k subconjuntos de tamanhos similares e repete k iterações de treinamento e validação de forma que, para cada iteração uma parte é mantida para validação (teste) e as $k-1$ restantes mantidas para o treinamento do algoritmo.

A validação cruzada requer o reajuste do modelo com conjuntos diferentes de treinamento. Porém algumas situações podem acabar prejudicando a validação,. Uma delas ocorre quando há diversas observações de um mesmo indivíduo/animal observado em períodos diferentes, levando a parcela de treino, aprender com as informações daquele próprio indivíduo/animal, que pode estar também na amostra de teste. Uma forma de contornar essa situação é utilizando a técnicas de validação cruzada *leave-one-out* (LOO) (VEHTARI; GELMAN; GABRY, 2017).

Assim, considerando como exemplo a validação LOAO, em uma amostra de N animais, $N-1$ seriam utilizados para treinamento e 1 para teste, ou seja, o modelo será ajustado com um animal a menos e posteriormente testado para este animal que ficou de fora da etapa de treinamento. Esse processo é repetido até que todos os animais tenham sido utilizados para validação, totalizando N repetições. Esse método possui como vantagem a diminuição do viés para experimentos aplicados por não tem influência da aleatorização das parcelas (CHENG; GARRICK; FERNANDO, 2017).

A validação LOO pode ser entendida como um caso particular da validação *k-fold* onde a amostra é particionada em k grupos, sendo k o total de animais (para LOAO) ou experimentos (para LOEO) envolvidos. A estimativa Bayesiana LOO fora da amostra é dada por:

$$elpd_{loo} = \sum_{i=1}^{n} log(p(y_i|y_{i-1})) \tag{2.10}$$

onde

$$p(y_i|y_{i-1}) = \int p(y_i|\theta)p(\theta|y_{-i})d\theta \tag{2.11}$$

é a densidade preditiva da exclusão, ou seja, os dados sem a i-ésima parte. Assim o conjunto de dados tem interpretação mais simples da escala definindo uma precisão preditiva para os n pontos tomados um a um (VEHTARI; GELMAN; GABRY, 2017).

3 CONSIDERAÇÕES GERAIS

O primeiro estudo de caso teve como objetivo avaliar diferentes métodos de seleção de variáveis para regressão logística em um estudo de caso de segurança e frequência alimentar de pré escolares. Os métodos utilizados foram *stepwise*, lasso, *random forest* e PSV. Dentre os modelos avaliados o que produziu melhor desempenho foi o modelo com as variáveis selecionadas por *stepwise*.

O primeiro estudo teve como resultado a associação do grau de insegurança alimentar com a situação socieconômica. Os resultados aqui apresentados foram comparados com outros autores que analisaram a mesma base de dados anteriormente, no entanto, o método *stepwise* não havia sido aplicado.

Assim, as contribuições desta primeira etapa estão em acrescentar os métodos PSV, *random forest* e *stepwise* às análises de segurança e frequência alimentar identificando as variáveis que causam maiores impactos na alimentação de crianças, uma etapa importante para identificação de situações de risco e planejamento de ações voltadas para alimentação e nutrição.

O segundo estudo de caso apresentado teve como objetivo avaliar o uso de métodos de seleção de variáveis em uma base de dados de alta dimensão obtidos com a utilização de *Near-Infrared Spectroscopy* (NIRS) em um problema de predição de consumo alimentar de vacas leiteiras. Neste contexto não foi possível a aplicação do método PSV por incluir etapa manual de seleção e o método stepwise se mostrou ineficiente devido ao alto custo computacional na sua execução.

Nessa segunda análise, utilizando como critério de seleção de modelos o RMSE, o uso da regressão lasso se mostrou a mais eficiente utilizando 10 preditores, reduzindo assim a dimensão dos dados em 99,54%. Como o conjunto de dados é proveniente de um mesmo rebanho apresenta como desvantagem a possibilidade do modelo ser menos preciso do que quando usado em rebanhos e ambientes diferentes. Métodos de *machine learning* também podem ser usados para este objetivo no entanto o custo computacional na etapa de seleção de variáveis é alto

As contribuições deste segundo estudo estão na comparação entre os métodos lasso e *random forest* usados separadamente para seleção de variáveis em NIRS, tendo como diferencial a comparação entre diferentes validações para o lasso. Estes resultados são restritos a esta base de dados, mas estudos futuros podeme ser feitos considerando a união entre os métodos, outras aplicações, validação fora da amostra e a comparação entre lasso e *random forest* com

os métodos tradicionalmente utilizados, como *Principal Component Analysis* (PCA) e *Partial Least Squares* (PSL).

O lasso apresentou bom desempenho em ambos os casos apesar de não ter sido o melhor método na primeira situação foi o método mais eficiente no cenário de alta dimensão além do baixo custo computacional.

De forma geral o presente trabalho apresentou diferentes métodos de seleção de variáveis em dois estudos de caso, o primeiro deles envolvendo um problema de classificação comparando modelos logísticos elaborados por diferentes métodos de seleção e o segundo envolvendo uma base de dados de alta dimensão. O método stepwise que apresentou bom desempenho na primeira situação não foi suficiente quando se trata de um problema de alta dimensão.

REFERÊNCIAS

AGRESTI, A. **An introduction to categorical data analysis**. [S.l.]: John Wiley & Sons, 2018.

AVANCI, J. Q. et al. Fatores associados aos problemas de saúde mental em adolescentes. **Psicologia: Teoria e Pesquisa**, SciELO Brasil, v. 23, n. 3, p. 287–294, 2007.

AZEVEDO, R. B. d. Métodos de machine learning para seleção de variáveis com aplicações ao rugby sevens feminino. 2019.

BEUREN, G. M. Análise de dados de altas dimensões. 2010.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BRITO, G. A. S.; NETO, A. A. Modelo de classificação de risco de crédito de empresas. **Revista Contabilidade & Finanças**, SciELO Brasil, v. 19, n. 46, p. 18–29, 2008.

BRYK, A. S.; RAUDENBUSH, S. W. **Hierarchical linear models: Applications and data analysis methods.** [S.l.]: Sage Publications, Inc, 1992.

BURSAC, Z. et al. Purposeful selection of variables in logistic regression. **Source code for biology and medicine**, Biomed central, v. 3, n. 1, p. 1–8, 2008.

CARSON, R. L.; CHASE, M. A. An examination of physical education teacher motivation from a self-determination theoretical framework. **Physical Education and Sport Pedagogy**, Taylor & Francis, v. 14, n. 4, p. 335–353, 2009.

CASELLA, G.; BERGER, R. L. Statistical inference. Cengage Learning, Pacific Grove, CA, 2002.

CHEN, X.; ISHWARAN, H. Random forests for genomic data analysis. **Genomics**, Elsevier, v. 99, n. 6, p. 323–329, 2012.

CHENG, H.; GARRICK, D. J.; FERNANDO, R. L. Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. **Journal of animal science and biotechnology**, Springer, v. 8, n. 1, p. 1–5, 2017.

DAVISON, A. C.; GIGLI, A. Deviance residuals and normal scores plots. **Biometrika**, Oxford University Press, v. 76, n. 2, p. 211–221, 1989.

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. [S.l.]: John Wiley & Sons, 1998. v. 326.

EFROYMSON, M. **Mathematical Methods for Digital Computers, chapter Multiple regression analysis**. [S.l.]: Wiley, New York, NY, 1960.

ENGELEN, J. E. V.; HOOS, H. H. A survey on semi-supervised learning. **Machine Learning**, Springer, v. 109, n. 2, p. 373–440, 2020.

FIGUEIRA, C. V. Modelos de regressão logística. 2006.

FONTES, J. d. A. Abordagens de seleção de variáveis para classificação e regressão em dados espectrais para controle da qualidade. 2020.

FUCHS, P. A. et al. Purposeful variable selection and stratification to impute missing fast data in trauma research. **The journal of trauma and acute care surgery**, NIH Public Access, v. 75, n. 1 0 1, p. S75, 2013.

GENUER, R.; POGGI, J.-M.; TULEAU-MALOT, C. Variable selection using random forests. **Pattern recognition letters**, Elsevier, v. 31, n. 14, p. 2225–2236, 2010.

GENUER, R.; POGGI, J.-M.; TULEAU-MALOT, C. Vsurf: an r package for variable selection using random forests. **The R Journal**, v. 7, n. 2, p. 19–33, 2015.

GREENLAND, S. Modeling and variable selection in epidemiologic analysis. **American journal of public health**, American Public Health Association, v. 79, n. 3, p. 340–349, 1989.

GUIMARÃES, P. R. B. Métodos quantitativos estatísticos. **Curitiba: Iesde Brasil SA**, v. 1, p. 252, 2008.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. An introduction to statistical learning. 2009.

HASTIE, T.; TIBSHIRANI, R.; WAINWRIGHT, M. Statistical learning with sparsity. **Monographs on statistics and applied probability**, v. 143, p. 143, 2015.

HOCKING, R. R. A biometrics invited paper. the analysis and selection of variables in linear regression. **Biometrics**, JSTOR, p. 1–49, 1976.

HOSMER, D. W. J.; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. [S.l.]: John Wiley & Sons, 2013. v. 398.

IZBICKI, R.; SANTOS, T. M. dos. **Aprendizado de máquina: uma abordagem estatística**. [S.l.: s.n.], 2020. ISBN 978-65-00-02410-4.

JAMES, G. et al. **An introduction to statistical learning**. [S.l.]: Springer, 2013. v. 112.

JUNIOR, G. P. d. A. Avaliação do lasso e métodos alternativos em modelos de regressão logística. Universidade Federal de São Carlos, 2021.

KOGA, G. K. C. et al. Fatores associados a piores níveis na escala de burnout em professores da educação básica. **Cadernos Saúde Coletiva**, SciELO Brasil, v. 23, n. 3, p. 268–275, 2015.

KUTNER, M. H. et al. **Applied linear statistical models**. [S.l.]: McGraw-Hill Irwin Boston, 2005. v. 5.

LAVALLEY, M. P. Logistic regression. **Circulation**, Am Heart Assoc, v. 117, n. 18, p. 2395–2399, 2008.

MAHESH, B. Machine learning algorithms-a review. **International Journal of Science and Research (IJSR).[Internet]**, v. 9, p. 381–386, 2020.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. **Journal of the Royal Statistical Society: Series A (General)**, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972.

NETER, J. et al. Applied linear statistical models. Irwin Chicago, 1996.

REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. **Encyclopedia of database systems**, Springer, v. 5, p. 532–538, 2009.

REID, S.; TIBSHIRANI, R.; FRIEDMAN, J. A study of error variance estimation in lasso regression. **Statistica Sinica**, JSTOR, p. 35–67, 2016.

RENCHER, A. C.; SCHAALJE, G. B. **Linear models in statistics**. [S.l.]: John Wiley & Sons, 2008.

SANCHEZ-PINTO, L. N. et al. Comparison of variable selection methods for clinical predictive modeling. **International journal of medical informatics**, Elsevier, v. 116, p. 10–17, 2018.

SILVA, C. B. P. d. A técnica lasso e suas potencialidades na seleção de variáveis para modelos lineares. 2018.

SOUSA, G. C. de; ALVES, J. M. S. A regressão linear de galton: Atividades históricas para função afim e estatística básica usando planilhas eletrônicas. **Conexões-Ciência e Tecnologia**, v. 9, n. 4, p. 26–36, 2016.

STROBL, C. et al. Conditional variable importance for random forests. **BMC bioinformatics**, Springer, v. 9, n. 1, p. 1–11, 2008.

SULLIVAN, D. H. et al. Impact of nutrition status on morbidity and mortality in a select population of geriatric rehabilitation patients. **The American journal of clinical nutrition**, Oxford University Press, v. 51, n. 5, p. 749–758, 1990.

TENÓRIO, L. R. et al. Preditores de dificuldade em traqueostomia percutânea à beira do leito: estudo piloto. **Revista do Colégio Brasileiro de Cirurgiões**, SciELO Brasil, v. 47, 2020.

THOMPSON, B. **Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial**. [S.l.]: Sage Publications Sage CA: Thousand Oaks, CA, 1995. 525–534 p.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. JOURNAL OF THE ROYAL STATISTICAL SOCIETY: SERIES B (STATISTICAL METHODOLOGY), 2011.

VEHTARI, A.; GELMAN, A.; GABRY, J. Practical bayesian model evaluation using leave-one-out cross-validation and waic. **Statistics and computing**, Springer, v. 27, n. 5, p. 1413–1432, 2017.

WEISBERG, S. **Applied linear regression**. [S.l.]: John Wiley & Sons, 2005. v. 528.

WEST, J.; BHATTACHARYA, M. Intelligent financial fraud detection: a comprehensive review. **Computers & security**, Elsevier, v. 57, p. 47–66, 2016.

XIN, Y. et al. Machine learning and deep learning methods for cybersecurity. **Ieee access**, IEEE, v. 6, p. 35365–35381, 2018.

ZHANG, H.; SINGER, B. H. **Recursive partitioning and applications**. 2. ed. [S.l.]: Springer-Verlag New York, 2010. ISBN 978-1-4419-6823-4.

Parte II

SEGUNDA PARTE

4 SELEÇÃO DE VARIÁVEIS PARA REGRESSÃO LOGÍSTICA: AVALIAÇÃO DE DIFERENTES MÉTODOS EM UM ESTUDO DE SEGURANÇA E FREQUÊNCIA ALIMENTAR DE PRE-ESCOLARES

Alice Silva Duarte [1], Paula Ribeiro Santos[2], Luiz Felipe de Paiva Lourenção[3], Renato Ribeiro de Lima [4], Izabela Regina Cardoso de Oliveira[5]

RESUMO: A tarefa de selecionar variáveis vem sendo cada vez mais tratada em estudos de diversas áreas frente ao aumento no volume de dados disponíveis. Avaliar e identificar quais as variáveis contribuem mais e quais contém algum tipo de ruído exige cuidados. Esse problema pode surgir em diversos tipos de estudos, como naqueles que envolvem obtenção de dados por *survey*, com aplicação de questionários. Neste trabalho, nós avaliamos diferentes métodos de seleção de variáveis para regressão logística em um estudo de frequência e segurança alimentar de pré-escolares. Os dados utilizados ao longo da pesquisa são resultados da aplicação da Escala Brasileira de Insegurança Alimentar (EBIA) e do Questionário de Frequência Alimentar (QFA), com o objetivo de buscar variáveis (socioeconômicas, de gestação, condições de nascimento e hábitos alimentares) mais associadas com seus resultados. A amostra utilizada envolve 581 pré-escolares de 0 a 5 anos de idade matriculados em CMEIs (Centro Municipal de Educação Infantil) do município de Lavras-MG. Este banco de dados possui 27 variáveis binárias, 14 variáveis nominais, 6 ordinais e 2 discretas. Foram considerados os métodos Stepwise, lasso, o *Purposeful Selection of Covariates* (PSV) e *Random Forest* para de seleção de variáveis e, posteriormente, foram elaborados modelos logísticos com as variáveis selecionadas por estes métodos. Os modelos foram avaliados pelo critério de Akaike (AIC). Dentre os métodos avaliados, o que produziu o modelo com melhor desempenho foi o Stepwise. Os resultados apontam uma relação entre segurança alimentar e variáveis associadas a classe socioeconômica da família destacando a importância de intervenções educativas e preventivas.

Palavras-chave: Stepwise, Purposeful Selection of Covariates, Lasso, Alimentação pré-escolar.

[1] Mestranda em Estatística e Experimentação Agropecuária (UFLA). Email: alice.sduarte15@gmail.com

[2] Doutoranda em Estatística e Experimentação Agronômica, Departamento de Ciências Exatas, Universidade de São Paulo, Campus Luiz de Queiroz. E-mail: paullasant_s@hotmail.com

[3] Professor Substituto pela Faculdade Federal de Alfenas - UNIFAL-MG. E-mail: luizfelipepaiva03@gmail.com

[4] Professor do Departamento de Estatística/ICET da Universidade Federal de Lavras. Email: rrlima@ufla.br

[5] Professora do Departamento de Estatística/ICET da Universidade Federal de Lavras. Email: izabela.oliveira@ufla.br

4.1 INTRODUÇÃO

Regressão pode ser entendida como a busca por explicar a relação entre uma variável resposta por meio de uma ou mais variáveis explicativas. Quando o problema a ser resolvido está relacionado a uma variável resposta categórica o modelo mais usual é a Regressão Logística (AGRESTI, 2018). O uso deste modelo atrai pesquisadores de diversas áreas, sobretudo da saúde, por possibilitar a interpretação das estimativas em termos de razão de chance.

Com o avanço da tecnologia o volume de dados cresceu de maneira exponencial, sendo necessário o uso de técnicas cada vez mais avançadas para explicar a relação entre estas (muitas) variáveis. Em uma modelagem estatística, por um lado, o grande número de variáveis pode aumentar a capacidade preditiva do modelo, mas, muitas dessas variáveis podem contribuir pouco e gerar um alto custo computacional fazendo-se necessário a seleção de variáveis, sendo natural procurar por aquelas que têm maior impacto no modelo (JUNIOR, 2021).

Os métodos mais clássicos e frequentemente utilizados no processo de seleção de variáveis são os métodos *forward, backward* e *stepwise*. Esses métodos baseiam-se em um algoritmo de decisão que checa a importância das variáveis (THOMPSON, 1995). O método *stepwise* é o mais usual para seleção de variáveis sendo amplamente utilizado na saúde, na nutrição, na educação física e em diversas outras áreas (CARSON; CHASE, 2009; GREENLAND, 1989; SULLIVAN et al., 1990).

Outro método utilizado é o método de regularização lasso (last absolute shrinkage and selection operator), proposto por Tibshirani (1996). Este pode ser usado na estimação dos parâmetros ou na seleção de variáveis, pois faz com que as estimativas de alguns parâmetros associados às variáveis preditoras tendam a zero. O lasso tem sido uma escolha natural, principalmente quando os dados tratados são de alta dimensão por reduzir consideravelmente o tempo de processamento por meio de seleção de variáveis (LOZANO et al., 2018; VILOR-TEJEDOR et al., 2018; MEINSHAUSEN; BÜHLMANN, 2006).

Outro método que pode ser usado na seleção de variáveis em modelos de regressão, especialmente em grandes bancos de dados, é o *random Forests* (florestas aleatórias) ou *random decision forests* (florestas de decisão aleatória). Esse método é um método de *Machine Learning* que, em síntese, constrói e agrega um grande volume de árvores de decisão estabelecendo regras em cada uma delas. A estrutura dessas árvores se assemelha a um fluxograma, verificando uma condição a cada nó e levando assim ao próximo nó, até a finalização da árvore. Este é uma técnica supervisionada de machine learning que utiliza parte dos dados para aprendizado e parte para teste (AGRESTI, 2018). No contexto de seleção de variáveis pode ser usado o conceito de importância de variáveis, que é uma medida obtida com a aplicação deste método (GENUER; POGGI; TULEAU-MALOT, 2010; CHEN; ISHWARAN, 2012; STROBL et al., 2008).

Por fim, dentro do contexto de regressão logística, tem-se o método proposto por Jr, Lemeshow e Sturdivant (2013), cuja proposta é considerar a importância prática da variável resposta em questão, estudando marginalmente a associação de cada covariável com a variável

resposta de interesse e incluindo aquelas de maior importância, segundo os valores-p dos testes marginais. Esse método, conhecido como P*urposeful selection of variables* e que muito se assemelha ao procedimento de *stepwise*, tem sido aplicado em diversas áreas (FUCHS et al., 2013; SANCHEZ-PINTO et al., 2018).

Neste trabalho utilizamos uma base de dados de um estudo cujo objetivo era encontrar preditores da segurança e frequência alimentar de crianças. Esses dois fatores são importantes para qualificar a alimentação dos indivíduos no que diz respeito à presença de alimentos básicos de qualidade e em quantidades satisfatórias. Os fatores que influenciam a segurança e a qualidade alimentar são diversos. Assim saber quais as variáveis têm mais impacto na qualidade e na segurança alimentar é um desafio tanto para os profissionais da saúde quanto para a população de forma geral.

A busca por medir a segurança e a frequência alimentar resultou na criação da Escala Brasileira de Insegurança Alimentar (EBIA), proposta e validada por Segall-Corrêa et al. (2003) e no Questionário de Frequência Alimentar (QFA) desenvolvido por Mondini et al. (2007). Assim a base de dados utilizada neste artigo consiste no resultado da aplicação desses questionários e da coleta de diversas outras variáveis em uma amostra de 581 pré-escolares matriculados em Centros Municipais de Educação Infantil de Lavras, Minas Gerais.

Dessa forma, o presente estudo tem como principal objetivo avaliar diferentes métodos de seleção de variáveis para regressão logística nessa aplicação. Foram considerados os métodos *stepwise,* lasso, o PSV e *random forest* para de seleção de variáveis e, posteriormente foram elaborados e avaliados os modelos logísticos obtidos com as variáveis selecionadas por cada um desses métodos.

4.2 SELEÇÃO DE VARIÁVEIS EM MODELO DE REGRESSÃO

Está seção tem a finalidade de abordar, conceitualmente, os métodos de seleção de variáveis utilizados ao longo deste artigo.

4.2.1 Método *Stepwise*

Os métodos *forward, backward* e *stepwise* são os métodos mais usuais, quando se trata do problema de seleção de variáveis em modelos de regressão.

O método *forward*, conhecido como inclusão passo a frente, consiste em ordenar as variáveis preditoras de acordo com a intensidade de sua relação com a variável resposta e ajustar o modelo. O modelo é ajustado primeiramente somente com o intercepto e em seguida adicionando as variáveis, uma a uma, e testando a significância dessa adição no modelo. Se a adição for significativa a variável entra no modelo; se não for o procedimento é encerrado (HOCKING, 1976).

Já o método Backward, conhecido como eliminação passo atrás, parte pelo caminho oposto. Ele ajusta um modelo com todas as variáveis preditoras e remove variáveis, caso a

remoção não gere perdas de desempenho o modelo fica sem a variável (EFROYMSON, 1960; NETER et al., 1996). Uma limitação dos métodos *Forward* e *Backward* é que, uma vez que uma variável é descartada do modelo ela não é mais considerada, mesmo que sua contribuição passe a ser significativa com a entrada de outra variável.

O método *stepwise* é um dos métodos mais usados em regressão e é composto pela união do método *forward* e *backward* (NETER et al., 1996). Ele atua nos dois sentidos, ele adiciona e remove variáveis. A cada adição é verificado se alguma das variáveis existentes pode ser excluída. A seleção é finalizada quando nenhuma adição ou exclusão melhora o desempenho do modelo (EFROYMSON, 1960). No entanto, esse método possui limitações. Quando há muitas variáveis a seleção pode ser lenta ou até impossível de ser executada (SENRA et al., 2007).

A seleção de variáveis pelos métodos *stepwise* exploram conjuntos de modelos avaliando por algum critério de seleção (AIC, BIC, R^2, F parcial ou T para situações de regressão). O método *stepwise* é a versão hibrida entre os métodos *Forward* e *Backward*, esse método adiciona variáveis sequencialmente a cada adição é checado se a remoção de alguma variável fornece melhorias no modelo até que nenhuma variável possa ser adicionada nem retirada do modelo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Este trabalho adotou como métrica de seleção de modelos para o método *Forward and Backward Stepwise* o critério de informação de Akaike (AIC). Esse critério mensura a qualidade a simplicidade de um modelo por meio de uma medida de forma que, menores valores de AIC representam uma mais qualidade e simplicidade.

4.2.2 Lasso

Um dos maiores problemas nos modelos de regressão é o balanço entre viés e variância. De modo simplificado o viés mede quão próximas as estimativas do modelo ajustado estão das verdadeiras respostas, enquanto a variância é uma medida de dispersão que traduz a distância entre os valores preditos e a média traduzindo o quanto o modelo consegue se adaptar às mudanças de amostra. O lasso (Least Absolute Shrinkage and Selection Operator) é um método de encolhimento e seleção de variáveis que considera ambos os problemas (RANSTAM; COOK, 2018; TIBSHIRANI, 2011). Dessa forma, o lasso, simultaneamente, estima os parâmetros e seleciona covariáveis para o modelo sendo um método tanto de estimação quanto de seleção de variáveis (RANSTAM; COOK, 2018).

Sendo assim, um bom modelo deve apresentar um bom balanço entre viés e variância. Para isso são utilizadas técnicas de regularização, que adicionam uma penalização ao método de mínimos quadrados e têm como objetivo central diminuir a variância, evitando o ajuste excessivo dos dados.

Em regressão logística, o uso do lasso tem sido utilizado na literatura por ser capaz de selecionar covariáveis podendo ser usado de forma independente ou combinado com outros métodos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; ZHANG; HUANG, 2008).

Em contexto de alta dimensão, ou seja $p > n$, a solução encontrada pelo método tradicional de mínimos quadrados pode não ser única criando modelos com muitos parâmetros e de difícil interpretação. O método lasso resolve esse problema por penalizar coeficientes levando vários deles a zero. Sua função objetivo é dada pela minimização da equação abaixo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009):

$$\hat{\beta}^{lasso} = argmin_{\beta} \left\{ \sum_{j=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{4.1}$$

sendo:
$\sum_{j=1}^{p} |\beta_j| = \|\beta\|$
sendo $\sum_{j=1}^{p} |\beta_j|$ a penalização lasso. Devido à natureza desta restrição, tornar λ suficientemente grande faz com que alguns coeficientes sejam exatamente zero fazendo uma seleção contínua de subconjuntos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Para determinar a melhor alternativa para o parâmetro de ajuste λ foi utilizada a técnica de validação cruzada encontrado em uma grade de possíveis valores de λ calculando o erro de validação para cada um, o menor erro de validação será o λ utilizado. Assim o modelo e reajustado parametrizado com o valor de λ encontrado (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

4.2.3 *Random Forest*

Com o avanço na tecnologia e o aumento no volume de dados, algumas situações exigem técnicas automatizadas de seleção de variáveis. Para essas situações, os métodos de aprendizado de máquina impulsionam metodologias que podem ser aplicadas em diversas situações (XIN et al., 2018). Para o aprendizado supervisionado[1] um dos algoritmos utilizados para tarefas de classificação são as florestas aleatórias (*random forest*).

Random forest é um método de *machine learning* utilizado para tarefas de regressão e classificação. Esse método, para classificação, se baseia em várias árvores de decisão de forma que cada árvore depende apenas de um vetor aleatório amostrado com a mesma distribuição. A cada passo são selecionadas variáveis para montar as árvores e cada árvore é treinada de forma independente, classificando a variável resposta de interesse. A classificação final será a mais frequente de todas as árvores.

Assim a primeira árvore é criada a partir de p características (variáveis) selecionadas aleatoriamente para a criação da árvore, sendo assim a partir de uma medida de impureza é selecionada a característica que ficará no topo da árvore. As árvores são construídas considerando apenas os subconjuntos de atributos selecionados, sendo este subconjunto diferente de uma árvore para outras (BREIMAN, 2001).

[1] Em *Machine Learning* o aprendizado supervisionado ocorre quando para cada observação $x_1, x_2, ..., x_p$ há uma resposta associada y_i (JAMES et al., 2013).

Para apurar a eficácia da predição dos modelos é utilizada a validação cruzada, que, em síntese, divide os dados em duas amostras, uma de treino e uma de validação. A partir da amostra de treino um modelo é ajustado e então sua capacidade preditiva é testada no conjunto de validação. O modelo final é aquele que se ajusta melhor aos dados de validação (CUTLER; CUTLER; STEVENS, 2012; BREIMAN, 2001).

Random Forests é uma melhoria do método de *bagged trees* por meio de um ajuste que elimina a correlação entre as árvores. Esse método constrói árvores de decisão em amostras selecionadas por bootstrap utilizando m dos p preditores da base de dados ($m \leq p$), usualmente escolhe-se $m \approx \sqrt{p}$. Cada divisão dos dados considera um subconjunto dos preditores reduzindo a correlação entre as árvores e diminuindo a variabilidade entre as árvores (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Então, adotando como critério $m \approx \sqrt{p}$, foram utilizadas 5 variáveis reamostradas aleatóriamente entre as variáveis explicativas em casa árvore(mtry = 5), sendo $p_EBIA = 19$ e $p_QFA = 18$. Uma limitação deste método é o custo computacional que pode ser elevado na etapa de treinamento dependendo da massa de dados que está sendo analisada.

4.2.4 *Purposeful Selection of Covariates (PSV)*

Outra forma de selecionar covariáveis para um modelo logístico é pelo procedimento que parte dos resultados de testes marginais, proposto por Jr, Lemeshow e Sturdivant (2013) e conhecido como "Purposeful selection of covariates". Este método busca o modelo mais parcimonioso e que reflita a especificidade de cada situação. Bursac et al. (2008) comparou o desempenho do PSV com outros três procedimentos de seleção, entre eles o método *stepwise* citado anteriormente. O método apresenta vantagens, pois além de selecionar variáveis significativas possibilita a inclusão de variáveis que eram confundidoras de outras variáveis do modelo.

Esse método vem sendo explorado principalmente em diversas áreas. Ele possibilita a seleção intencional de variáveis, permitindo ao analista decidir se determinada variável fica ou não no modelo, com base na importância da variável pelo ponto de vista prático, não apenas estatístico. Este método de seleção de variáveis agrega informações ao modelo principalmente quando no que diz respeito a modelagem de fatores de risco e retenção de variáveis confundidoras podendo resultar em um modelo mais rico (BURSAC et al., 2008; FUCHS et al., 2013; LEDERER et al., 2019).

A seleção de variáveis pelo algoritmo PSV, proposto por Jr, Lemeshow e Sturdivant (2013) é dividida em sete principais etapas, são elas:

Etapa 1: Estudo das associações marginais entre as variáveis com o objetivo de identificar o primeiro modelo de regressão múltipla, esse modelo é composto pelas variáveis cujo teste uni variado apresentou valor de p inferior a 0,25.

Etapa 2: Ajuste do modelo de regressão múltipla avaliando a importância de cada uma das variáveis não selecionadas na etapa anterior pela estatística de Wald eliminando as variáveis

não significativas e realizando um novo modelo de regressão múltipla. Esse novo modelo será comparado com o anterior pelo teste de razão de verossimilhança parcial.

Etapa 3: Os coeficientes do novo modelo devem ser comparados com os coeficientes do modelo anterior observando mudanças de magnitude, caso essa mudança seja maior do que 20% ($\Delta(\hat{\beta}) \geq 0,2$) é possível que alguma variável importante tenha sido removida, então deve-se voltar a segunda etapa ajustando novamente o modelo de regressão múltipla. Esse processo deve ser repetido até a inclusão de todas as variáveis importantes no modelo e excluídas a que não apresentam significado estatístico nem clínico.

Etapa 4: As variáveis não selecionadas no primeiro passo devem ser incluídas uma a uma e seu valor verificado pela estatística de Wald ou pelo teste de razão (para as variáveis categóricas) identificando as variáveis que, sozinhas não estão relacionadas ao resultado, mas contribuem na presença de outras variáveis. O modelo resultante desta etapa será chama de "Modelo de Efeitos Principais Preliminares".

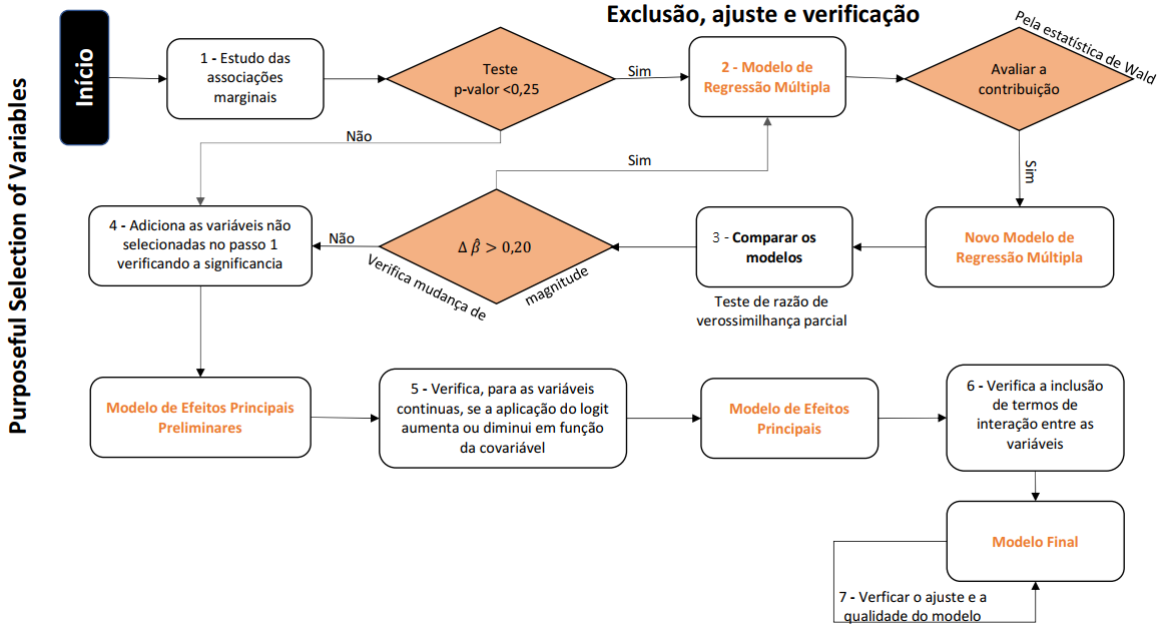
Etapa 5: As variáveis do modelo de efeitos principais preliminares devem ser examinadas verificando se a aplicação do logit aumenta ou diminui em função da covariável. Resultado assim no "Modelo de Efeitos Principais".

Etapa 6: Verifica-se nesta etapa a relação entre as variáveis e o sentido prático desta relação para o contexto em questão.

Etapa 7: Verificar o ajuste do modelo antes de usa-lo para fins inferenciais.

Os passos deste algoritmo, propostos por (JR; LEMESHOW; STURDIVANT, 2013) estão resumidos na Figura 3.

Figura 4.1 – Fluxograma do método *Purposeful selection of Variables*



Fonte: a autora (2022)

4.3 MATERIAL: A BASE DE DADOS

A amostra analisada neste estudo envolve dados de 581 pré-escolares de 0 a 5 anos de idade matriculados em CMEIs (Centro Municipal de Educação Infantil) do município de Lavras-MG. Este banco de dados possui 27 variáveis binárias, 14 variáveis nominais, 6 ordinais e 2 discretas cujos detalhes estão no Anexo A. Lourenção et al. (2021) analisaram esses dados com o objetivo de encontrar as variáveis que estão mais associadas às variáveis respostas qualidade (1: boa ou intermediária 0: ruim) e segurança alimentar (1: segurança ou 0: insegurança) das crianças. Os autores utilizaram lasso para selecionar as variáveis mais importantes. Os efeitos nos modelos logísticos resultantes foram estimados por máxima verossimilhança e interpretados em termos de razões de chances. No presente estudo, além do lasso, outros métodos de seleção de variáveis em modelos logísticos serão avaliados.

A insegurança alimentar foi avaliada por meio da Escala Brasileira de Insegurança Alimentar (EBIA), proposta por Segall-Corrêa et al. (2003). A escala considera para classificação 14 perguntas centrais dicotômicas e classifica a segurança alimentar por meio da quantidade de respostas "sim" podendo a criança ser 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 (SANTOS, 2020). Para 57 participantes há ausência dessa informação. As frequências observadas segundo categoria foram: segurança alimentar (289), insegurança alimentar leve (185), insegurança alimentar moderada (26) e insegurança grave (15).

A outra variável resposta de interesse é a frequência alimentar dos participantes (Questão 48 - Anexo 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) foi avaliada 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 de três categorias: baixa qualidade alimentar, qualidade intermediária, boa qualidade. Para 55 participantes há ausência dessa informação. As frequências observadas segundo categoria foram: baixa qualidade (168), qualidade intermediária (83) e boa qualidade (275).

4.4 MÉTODOS: Forma geral do modelo de regressão logístico

Os modelos de regressão logística são também chamados de modelos logit e simbolizados por $logit(\pi)$ (AGRESTI, 2018).

A forma geral dos modelos ajustados nesse trabalho é dada por:

$$y_i = logit[P(Y=1)] = log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \sum_{i=1}^{p} \beta_j x_{ij} \tag{4.2}$$

em que:
y_i: valor da variável resposta da observação i $(i = 1,2,...,n)$;
x_{ij}: valor da variável independente j $(j = 1,2,...,p)$ do indivíduo i;
β_0 e $\beta = (\beta_1,...,\beta_p)^T$ os parâmetros desconhecidos;

Neste estudo de caso a probabilidade de "sucesso", $P(Y=1)$ é segurança e a probabilidade de "fracasso", $P(Y=0)$ é insegurança para a variável resposta segurança alimentar. Para a variável resposta 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 explicativas, que serão selecionadas pelos métodos descritos na sequência.

4.4.1 Implementação Computacional

Todas as análises foram feitas utilizados o software R (R Core Team, 2021). Para o lasso foi usado o pacote "glmnet" (FRIEDMAN; HASTIE; TIBSHIRANI, 2010) do software R (R Core Team, 2021). O parâmetro λ foi de 0,0405 para a variável resposta segurança alimentar (EBIA) e de 0,0604 para frequência alimentar (QFA). A escolha do parâmetro λ foi feita por validação cruzada.

A implementação de *random forest* foi feito utilizando pacote RandomForest (LIAW; WIENER et al., 2002) e o pacote VSURF (GENUER; POGGI; TULEAU-MALOT, 2019). Foram utilizadas 5 variáveis amostradas aleatoriamente entre as variáveis explicativas como candidatas em cada árvore (mtry=5).

Para a seleção foi Stepwise foi utilizado o pacote MASS (VENABLES; RIPLEY, 2002) utilizando como métrica de comparação de modelos o critério de AIC.

O algoritmo de seleção de variáveis pelo método PSV foi realizado utilizando o pacote purposeful (LAROUR, 2022). Foi utilizado como critério de seleção das variáveis para o modelo de regressão múltipla p-valor $< 0,25$.

4.5 RESULTADOS E DISCUSSÃO

A base de dados foi submetida aos quatro métodos de seleção de variáveis, stepwise, lasso, random forest e PSV. A escolha do parâmetro de ajuste λ na implementação do Lasso foi feita utilizando a técnica de validação cruzada e os resultados são apresentados no Anexo B. As variáveis selecionadas por cada método se encontram no Quadro 3.1.

4.5.1 Comparação dos métodos de seleção de variáveis

As únicas variáveis selecionadas por todos os métodos considerando a variável resposta EBIA foram X4 – situação econômica e X6 – escolaridade do pai. Além disso, as variáveis X2 - raça e X10 – posse de automóvel não foram selecionados apenas da seleção por *random forest* e a variável X8 – habitação foi selecionada por todos os métodos com exceção do lasso.

Quadro 4.1 – Variáveis selecionadas pelos diferentes métodos de seleção de variáveis para as respostas da Escala Brasileira de Insegurança Alimentar (EBIA) e do Questionário de Frequência Alimentar (QFA).

EBIA				
	Random Forest	Lasso	Stepwise	PSV
X1 - Sexo				
X2 - Raça		x	x	x
X3 - Idade				x
X4 - Situação econômica	x	x	x	x
X5 - Número de pessoas na família				x
X6 - Escolaridade do pai	x	x	x	x
X7 - Escolaridade da mãe				x
X8 - Habitação	x		x	x
X9 - Profissão do chefe da família		x		x
X10 - Posse de automóvel		x	x	x
X12 - Água de beber				x
X22 - Peso				
X31 - Aleitamento exclusivo				x
X38 - Número de refeições				
QFA				
X1 - Sexo				
X2 - Raça				
X3 - Idade		x	x	x
X4 - Situação econômica				x
X5 - Número de pessoas na família				
X6 - Escolaridade do pai				
X7 - Escolaridade da mãe				x
X8 - Habitação				x
X9 - Profissão do chefe da família				
X10 - Posse de automóvel			x	x
X12 - Água de beber		x		x
X22 - Peso				
X31 - Aleitamento exclusivo				
X32 - Aleitamento até 6 meses de idade			x	x
X33 - Idade do início do desmame				
X37 - Idade de introdução de outros alimentos				
X38 - Número de refeições	x	x	x	x

Fonte: a autora (2022)

Na avaliação da frequência e qualidade alimentar a única variável selecionada por todos os métodos foi a variável X38 – número de refeições. Além disso, destaca-se as variáveis X10 – Posse de automóvel e X12 – água de beber selecionada por dois dos métodos.

Após a seleção foi feito o modelo de regressão logístico considerando as variáveis selecionadas em cada um dos métodos. Os resultados dos ajustes, em termos de AIC encontram-se na Tabela 4.1.

Tabela 4.1 – Comparativo entre os diferentes métodos de seleção de variáveis para as respostas da Escala Brasileira de Insegurança Alimentar (EBIA) e do Questionário de Frequência Alimentar (QFA).

EBIA	
Método	AIC
Random Forest	315,14
Lasso	313,59
Stepwise	309,85
PSV	334,77
QFA	
Método	AIC
Random Forest	304,92
Lasso	304,42
Stepwise	300,02
PSV	311,93

Fonte: Da autora (2022)

Para ambas as respostas e pelos critérios considerados, observamos, para essa aplicação, uma superioridade dos modelos obtidos com variáveis selecionadas pelo clássico método stepwise.

Cabe destacar que todos os métodos foram usados para seleção de variáveis e, porteriormente, foi ajustado o modelo de regressão com estas variáveis. Sendo assim, o uso de *random forest* e lasso poderiam apresentar resultados diferentes se utilizados na parte de ajuste do modelo.

4.5.2 Modelos logísticos estimados para EBIA e QFA

As estimativas para os efeitos dos modelos para segurança alimentar (EBIA) e para frequência alimentar (QFA) com variáveis selecionadas pelo método *stepwise* são apresentadas na Tabela 4.2.

Tabela 4.2 – Estimativas, erro padrão e teste de Wald dos efeitos dos modelos logísticos para as respostas EBIA e QFA, com variáveis selecionadas pelo método stepwise.

EBIA			
Variável	Estimativa	Erro padrão	Valor-p
Intercepto	0,786	0,6861	0,25197
X2 - Raça			
X2 - Negra	-1,1332	0,4349	0,00918
X2 - Parda	-0,4898	0,338	0,14726
X4 - Situação Econômica			
X4.2 - 2 a 3 Salários mínimo	1,7361	0,3381	0,00000
X4.3 - 4 a 5 Salários mínimos	3,305	0,8315	0,00007
X6 - Escolaridade do pai			
X6.2 - Alfabetizado	-1,3377	0,7199	0,06314
X6.3 - Primário incompleto	-1,4808	0,6773	0,02880
X6.4 - Primário completo	-0,6518	0,6659	0,32768
X6.5 - Ginasial incompleto	-1,2228	0,8286	0,14001
X8 - Habitação			
X8.2 - Residência própria com financiamento a pagar	-1,1086	0,4213	0,00850
X8.3 - Residência cedida pelos pais ou parentes	-0,2088	0,4482	0,64134
X8.5 - Alugada	-0,3589	0,4081	0,37915
X10 - Posse de automóvel			
X10.2 - Possui um	0,3618	0,3305	0,27363
X10.3 - Possui 2 ou mais	1,3994	0,6987	0,04519
QFA			
Variável	Estimativa	Erro padrão	Valor-p
Intercepto	-0,4305	0,9924	0,6644
X3 - Idade			
X3.2 - de 6 a 11 meses	0,321	0,9351	0,7314
X3.3 - 12 a 23 meses	-1,0055	0,8748	0,2504
X3.4 - 24 a 35 meses	-1,1171	0,869	0,1986
X3.5 - 36 a 47 meses	-1,5265	0,9844	0,121
X3.6 - 48 a 59 meses	-1,1753	1,035	0,2561
X10 - Posse de automóvel			
X10.2 - Possui um	0,3547	0,3078	0,2491
X10.3 - Possui 2 ou mais	-0,7696	0,5738	0,1798
X32 - Aleitamento até 6 meses de idade			
X32.2 - Não	0,4694	0,3067	0,1259
X38 - Número de refeições	0,3086	0,1199	0,0101

Fonte: Da autora (2022)

Considerando a segurança alimentar as variáveis selecionadas para o modelo final foram raça, situação econômica, escolaridade do pai, habitação e posse de automóvel. Destes indicadores aqueles que influenciam negativamente para a segurança alimentar foram: raça - negra, escolaridade do pai - alfabetizado e primário incompleto e habitação - própria com financiamento a pagar. Estas variáveis estão associadas a escassez financeira que pode reduzir a quantidade e a qualidade de alimentos consumidos pelos familiares. As variáveis que influenci-

aram positivamente o modelo foram: situação economia – acima de 2 salários-mínimos e posse de automóvel – 2 ou mais.

Dados da Pesquisa de Orçamentos Familiares 2017-2018 (POF), divulgados pelo Instituto Brasileiro de Geografia e Estatística (IBGE) apontam insegurança alimentar em aproximadamente 49,9% dos lares com crianças menores de 5 anos, destes 15,7 se referem a insegurança grave ou moderada (IBGE, 2019)(IBGE, 2019). A base explorada neste trabalho apresenta resultados preocupantes em relação as estatísticas nacionais com aproximadamente 43,13% referentes a insegurança grave ou moderada.

Esta mesma base de dados foi analisada anteriormente por Santos (2020) e Lourenção et al. (2021).O grau de insegurança alimentar tem sido associado a situação socioeconômica, uma vez que a escassez financeira reduz a quantidade e a frequência alimentar e leva famílias a escolhas não saudáveis de alimentação optando por alimentos mais baratos.

A Escala Brasileira de Insegurança Alimentar (EBIA) é considerada uma escala importante por nutricionistas e estudiosos da saúde por possibilitar a identificação da situação de insegurança alimentar. Destaca-se a importância social, econômica e demográfica da identificação dos fatores relacionados a insegurança alimentar com a finalidade de embasar possíveis intervenções para a inclusão de alimentos mais saudáveis, principalmente em classes socioeconômicas mais vulneráveis (AIRES et al., 2012; SANTOS et al., 2014; ALEXANDRE et al., 2018).

Para a frequência alimentar, resultante dos Questionário de Frequência Alimentar (QFA) foram selecionadas as variáveis: idade, posse de automóveis, aleitamento até seis meses de idade e a quantidade de refeições. Destas, a única que apresentou efeito significativo a 5% foi a quantidade de refeições diárias da criança. Resultados similares foram encontrados previamente por Santos (2020) e Lourenção et al. (2021). O QFA é comumente utilizado por profissionais da saúde com o objetivo de identificar a qualidade e a frequência alimentar por apresentar praticidade no preenchimento reduzindo o viés de informação (COLUCCI; PHILIPPI; SLATER, 2004; MOLINA et al., 2010).

Santos (2020) utilizou lasso e arvores de classificação na seleção de variáveis para os modelos logísticos, que posteriormente foram reduzidos por *Stepwise*. Os métodos foram comparados entre si em termos de AIC e de deviance residual. Lourenção et al. (2021) teve como objetivo avaliar o estado nutricional, o consumo alimentar e a situação de insegurança alimentar destas crianças identificando os fatores associados a variável resposta de segurança alimentar e nutricional. Para isso foi utilizado a técnica de penalização lasso para selecionar as variáveis.

O modelo apresentado pela Tabela 4.2 é resultado da aplicação do método *stepwise*, esse método não foi considerado nos estudos anteriores para a mesma base de dados, assim, apenas para fins comparativos com os estudos anteriores, será considerado a seleção de modelos por lasso de acordo com as variáveis selecionadas e indicadas na Tabela 4.1. Para a variável resposta segurança alimentar (EBIA) o modelo gerado por lasso inclui duas variáveis além das selecionadas pelo mesmo método por Santos (2020) e Lourenção et al. (2021), a posse

de automóveis e a profissão do chefe da família, sendo a posse de automóvel uma variável significativa para o modelo.

Na avaliação da qualidade e frequência alimentar (QFA) o uso de lasso selecionou, além da variável X38 (número de refeições), identificada também pelos estudos anteriores, as variáveis X3 idade e X12 água de beber. Tais variáveis, apesar de identificadas pelo método, não foram significativas ao nível de 5%.

As variáveis selecionadas pelo método lasso nos estudos de Santos (2020) e Lourenção et al. (2021) diferem-se daquelas encontradas neste trabalho, aplicando a mesma técnica. A razão disso pode estar na forma com que o pré-processamento foi realizado nos dados. Como recomendado (PAULAUSKAS; AUSKALNIS, 2017; MALLEY; RAMAZZOTTI; WU, 2016) um cuidado especial foi tomado nessa etapa. Antes da aplicação dos métodos, foi feita uma pré-seleção das variáveis, levando-se em conta a importância prática de cada, e foram checadas inconsistências na base de dados.

Desta forma, as contribuições deste trabalho estão em acrescentar os métodos PSV, *Random Forest* e *Stepwise*. Além disso, a associação entre os fatores socioeconômicos associados a baixa qualidade da alimentação das crianças tem sido destaque em estudos aplicados em saúde e nutrição (MOLINA et al., 2010; LOURENÇÃO et al., 2021). A identificação das variáveis que causam impacto na alimentação, sobretudo de crianças, é uma etapa importante para o estabelecer situações de risco e para o planejamento de ações voltadas para alimentação e nutrição.

4.5.3 CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo avaliar diferentes métodos de seleção de variáveis para regressão logística em um estudo de frequência e segurança alimentar de pré-escolares. Foram considerados os métodos *stepwise*, lasso, o PSV e *random forest* para de seleção de variáveis e posteriormente foram elaborados modelos logísticos com as variáveis selecionadas por estes métodos. Os modelos foram avaliados por meio do critério de Akaike (AIC).

Para ambas as respostas avaliadas (segurança e frequência alimentar) o método que se mostrou mais eficiente para selecionar variáveis para o modelo foi o clássico *stepwise*. Destaca-se que esta mesma base de dados foi analisada anteriormente e o método que aqui se mostrou mais eficiente não havia sido aplicado.

Os resultados apontam uma relação entre segurança alimentar com as variáveis associadas a classe socioeconômica da família e a qualidade e segurança alimentar associada, principalmente, ao número de refeições realizadas pelas crianças, destacando a importância de intervenções educativas e preventivas.

REFERÊNCIAS

AGRESTI, A. **An introduction to categorical data analysis**. [S.l.]: John Wiley & Sons, 2018.

AIRES, J. d. S. et al. (in) segurança alimentar em famílias de pré-escolares de uma zona rural do ceará. **Acta Paulista de Enfermagem**, SciELO Brasil, v. 25, p. 102–108, 2012.

ALEXANDRE, D. da R. et al. Correlação da segurança alimentar com o estado nutricional de crianças escolares. **Motricidade**, Edições Desafio Singular, v. 14, n. 1, p. 164–169, 2018.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BURSAC, Z. et al. Purposeful selection of variables in logistic regression. **Source code for biology and medicine**, Biomed central, v. 3, n. 1, p. 1–8, 2008.

CARSON, R. L.; CHASE, M. A. An examination of physical education teacher motivation from a self-determination theoretical framework. **Physical Education and Sport Pedagogy**, Taylor & Francis, v. 14, n. 4, p. 335–353, 2009.

CHEN, X.; ISHWARAN, H. Random forests for genomic data analysis. **Genomics**, Elsevier, v. 99, n. 6, p. 323–329, 2012.

COLUCCI, A. C. A.; PHILIPPI, S. T.; SLATER, B. Desenvolvimento de um questionário de freqüência alimentar para avaliação do consumo alimentar de crianças de 2 a 5 anos de idade. **Revista Brasileira de Epidemiologia**, SciELO Public Health, v. 7, n. 4, p. 393–401, 2004.

CUTLER, A.; CUTLER, D. R.; STEVENS, J. R. Random forests. In: **Ensemble machine learning**. [S.l.]: Springer, 2012. p. 157–175.

EFROYMSON, M. **Mathematical Methods for Digital Computers, chapter Multiple regression analysis**. [S.l.]: Wiley, New York, NY, 1960.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for generalized linear models via coordinate descent. **Journal of statistical software**, NIH Public Access, v. 33, n. 1, p. 1, 2010.

FUCHS, P. A. et al. Purposeful variable selection and stratification to impute missing fast data in trauma research. **The journal of trauma and acute care surgery**, NIH Public Access, v. 75, n. 1 0 1, p. S75, 2013.

GENUER, R.; POGGI, J.-M.; TULEAU-MALOT, C. Variable selection using random forests. **Pattern recognition letters**, Elsevier, v. 31, n. 14, p. 2225–2236, 2010.

GENUER, R.; POGGI, J.-M.; TULEAU-MALOT, C. **VSURF: Variable Selection Using Random Forests**. [S.l.], 2019. R package version 1.1.0. Disponível em: <https://CRAN.R-project.org/package=VSURF>.

GREENLAND, S. Modeling and variable selection in epidemiologic analysis. **American journal of public health**, American Public Health Association, v. 79, n. 3, p. 340–349, 1989.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. An introduction to statistical learning. 2009.

HOCKING, R. R. A biometrics invited paper. the analysis and selection of variables in linear regression. **Biometrics**, JSTOR, p. 1–49, 1976.

IBGE. Pesquisa de orçamentos familiares 2017 – 2018. primeiros resultados. **Instituto Brasileiro de Geografia e Estatística. Primeiros Resultados**, POF, 2019.

JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. [S.l.]: John Wiley & Sons, 2013. v. 398.

JUNIOR, G. P. d. A. Avaliação do lasso e métodos alternativos em modelos de regressão logística. Universidade Federal de São Carlos, 2021.

LAROUR, E. **Purposeful: Purposeful Selection (Hosmer, Lemeshow). R package version 0.0.0.9000.** [s.n.], 2022. Disponível em: <https://github.com/emilelatour/purposeful>.

LEDERER, D. J. et al. Control of confounding and reporting of results in causal inference studies. guidance for authors from editors of respiratory, sleep, and critical care journals. **Annals of the American Thoracic Society**, American Thoracic Society, v. 16, n. 1, p. 22–28, 2019.

LIAW, A.; WIENER, M. et al. Classification and regression by randomforest. **R news**, v. 2, n. 3, p. 18–22, 2002.

LOURENÇÃO, L. F. de P. et al. Socioeconomic factors related to food consumption and the condition of food and nutrition insecurity in preschoolers. **Nutrire**, Springer, v. 46, n. 2, p. 1–11, 2021.

LOZANO, M. et al. Procedimiento estratégico en tres fases para la selección de variables, con el fin de obtener resultados equilibrados en investigación sobre salud pública. **Cadernos de Saúde Pública**, SciELO Brasil, v. 34, 2018.

MALLEY, B.; RAMAZZOTTI, D.; WU, J. T.-y. Data pre-processing. **Secondary analysis of electronic health records**, Springer, p. 115–141, 2016.

MEINSHAUSEN, N.; BÜHLMANN, P. High-dimensional graphs and variable selection with the lasso. **The annals of statistics**, Institute of Mathematical Statistics, v. 34, n. 3, p. 1436–1462, 2006.

MOLINA, M. d. C. B. et al. Preditores socioeconômicos da qualidade da alimentação de crianças. **Revista de Saúde Pública**, SciELO Brasil, v. 44, p. 785–732, 2010.

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. **Cadernos de Saúde Pública**, SciELO Brasil, v. 23, p. 1825–1834, 2007.

NETER, J. et al. Applied linear statistical models. Irwin Chicago, 1996.

PAULAUSKAS, N.; AUSKALNIS, J. Analysis of data pre-processing influence on intrusion detection using nsl-kdd dataset. In: IEEE. **2017 open conference of electrical, electronic and information sciences (eStream)**. [S.l.], 2017. p. 1–5.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2021. Disponível em: <https://www.R-project.org/>.

RANSTAM, J.; COOK, J. Lasso regression. **Journal of British Surgery**, Oxford University Press, v. 105, n. 10, p. 1348–1348, 2018.

SANCHEZ-PINTO, L. N. et al. Comparison of variable selection methods for clinical predictive modeling. **International journal of medical informatics**, Elsevier, v. 116, p. 10–17, 2018.

SANTOS, L. P. d. et al. Comparação entre duas escalas de segurança alimentar. **Ciência & Saúde Coletiva**, SciELO Public Health, v. 19, p. 279–286, 2014.

SANTOS, P. R. **Seleção De Variáveis Para Regressão Logística Em Um Exemplo De Segurança E Frequência Alimentar**. http://repositorio.ufla.br/handle/1/39129, 2020.

SEGALL-CORRÊA, A. M. et al. Projeto: acompanhamento e avaliação da segurança alimentar de famílias brasileiras: validação de metodologia e de instrumento de coleta de informação. **Campinas: Departamento de Medicina Preventiva e Social, Universidade Estadual de Campinas/Organização Pan-Americana da Saúde/Ministério da Saúde**, 2003.

SENRA, L. F. A. d. C. et al. Estudo sobre métodos de seleção de variáveis em dea. **Pesquisa Operacional**, SciELO Brasil, v. 27, p. 191–207, 2007.

STROBL, C. et al. Conditional variable importance for random forests. **BMC bioinformatics**, Springer, v. 9, n. 1, p. 1–11, 2008.

SULLIVAN, D. H. et al. Impact of nutrition status on morbidity and mortality in a select population of geriatric rehabilitation patients. **The American journal of clinical nutrition**, Oxford University Press, v. 51, n. 5, p. 749–758, 1990.

THOMPSON, B. **Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial**. [S.l.]: Sage Publications Sage CA: Thousand Oaks, CA, 1995. 525–534 p.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. JOURNAL OF THE ROYAL STATISTICAL SOCIETY: SERIES B (STATISTICAL METHODOLOGY), 2011.

VENABLES, W. N.; RIPLEY, B. D. **Modern Applied Statistics with S**. Fourth. New York: Springer, 2002. ISBN 0-387-95457-0. Disponível em: <https://www.stats.ox.ac.uk/pub/MASS4/>.

VILOR-TEJEDOR, N. et al. Sparse multiple factor analysis to integrate genetic data, neuroimaging features, and attention-deficit/hyperactivity disorder domains. **International Journal of Methods in Psychiatric Research**, Wiley Online Library, v. 27, n. 3, p. e1738, 2018.

XIN, Y. et al. Machine learning and deep learning methods for cybersecurity. **Ieee access**, IEEE, v. 6, p. 35365–35381, 2018.

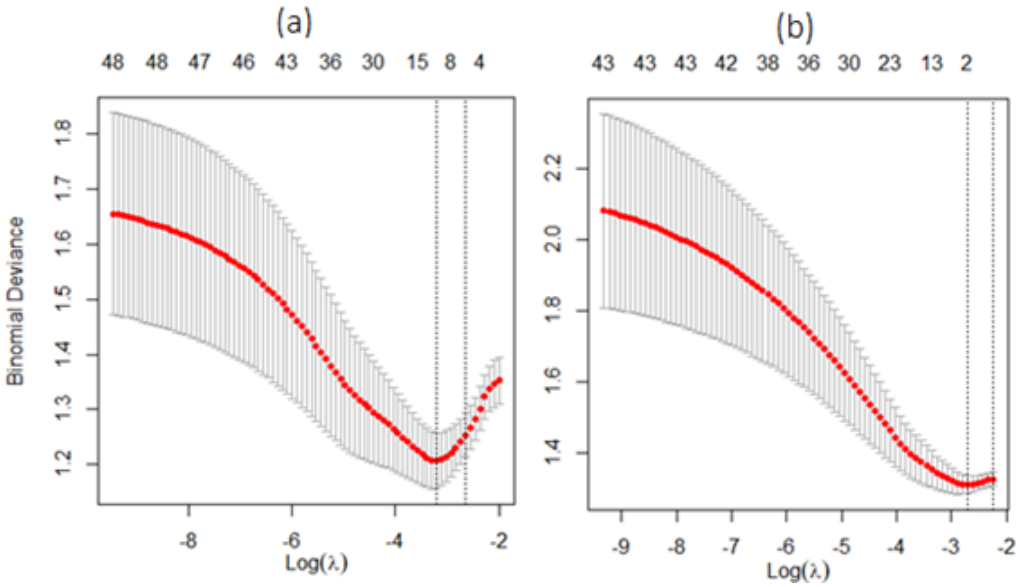
ZHANG, C.-H.; HUANG, J. The sparsity and bias of the lasso selection in high-dimensional linear regression. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 36, n. 4, p. 1567–1594, 2008.

ANEXO A - CÓDIGO, NOMES E NÍVEIS DAS VARIÁVEIS E DAS CATEGORIAS AGRUPADAS

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	0 = <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 usuário 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

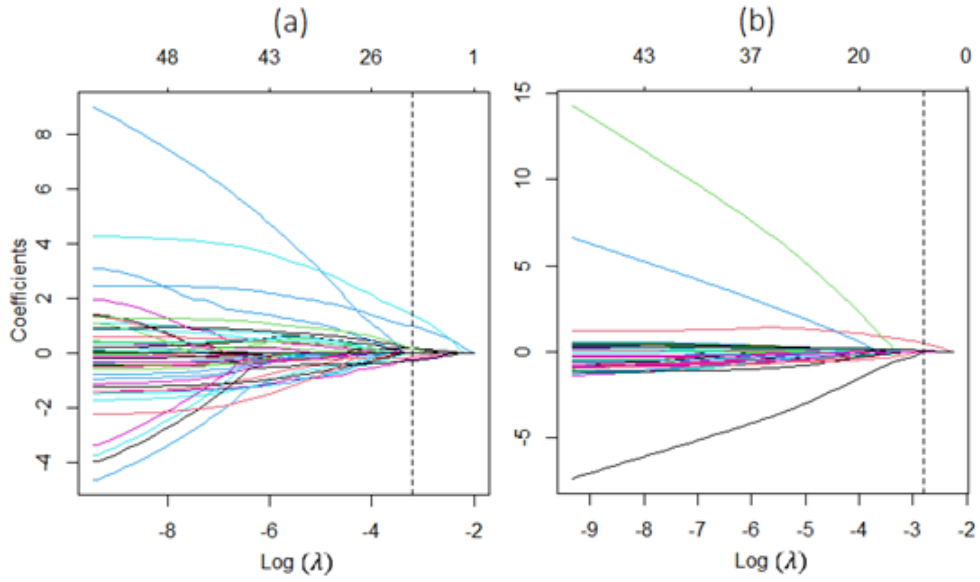
ANEXO B - ESCOLHA DO PARÂMETRO λ PARA O MÉTODO LASSO CONSIDERANDO AS VARIÁVEIS RESPOSTAS EBIA (A) E QFA (B)

Figura 2 – Validação cruzada para obtenção do parâmetro λ para as variáveis respostas da Escala Brasileira de Insegurança Alimentar – EBIA (a) e do Questionário de Frequência Alimentar – QFA (b).



Fonte: Da autora (2022)

Figura 3 – Encolhimento das variáveis para os valores de λ para as variáveis respostas da Escala Brasileira de Insegurança Alimentar – EBIA (a) e do Questionário de Frequência Alimentar – QFA (b).



Fonte: Da autora (2022)

O valor de λ foi de 0,0405 para a variável resposta segurança alimentar (EBIA) e de 0,0604 para frequência alimentar (QFA).

5 MÉTODOS DE SELEÇÃO DE VARIÁVEIS EM UM PROBLEMA DE PREDIÇÃO DE VARIÁVEIS NUTRICIONAIS DE VACAS LEITEIRAS A PARTIR DE DADOS DE ALTA DIMENSÃO

Alice Silva Duarte [1], Marina de Arruda Camargo Danes [2], Renato Ribeiro Lima[3], Izabela Regina Cardoso de Oliveira[4]

RESUMO: Modelos de regressão são técnicas utilizadas para estabelecer relação de causa/efeito entre uma variável resposta e uma ou mais variáveis explicativas. Com o avanço tecnológico, o volume e a dimensão dos dados analisados pode ser cada vez maior. Enquanto por um lado, o grande número de variáveis pode aumentar a capacidade preditiva do modelo, por outro lado, muitas destas variáveis podem contribuir pouco e gerar um alto custo computacional fazendo-se necessário a seleção de variáveis e busca por aquelas que têm maior impacto no modelo. Assim, o objetivo deste estudo foi avaliar o uso de métodos de seleção de variáveis em dados de alta dimensão obtidos com a utilização de Near-Infrared Spectroscopy (NIRS) em um problema de predição de consumo alimentar de vacas leiteiras em um rebanho. Foram avaliados os métodos Stepwise, lasso e Random Forest para seleção de variáveis, que foram comparados de maneira indireta por meio do modelo gerado pelas variáveis selecionadas por cada um deles. Após a avaliação, o modelo que selecionou as variáveis mais relevantes foi gerado utilizando regressão lasso. No entanto, esse estudo se limita na utilização dos métodos de forma independente. As contribuições deste estudo estão na comparação entre os métodos lasso e Random Forest usados separadamente para seleção de variáveis em NIRS tendo como diferencial a comparação de diferentes validações para o lasso. Estes resultados são restritos a esta base de dados, mas estudos futuros podem ser feitos considerando a união entre os métodos, outras aplicações, a validação fora da amostra e até a inclusão de outras variáveis explicativas.

Palavras-chave: NIRS. Alta dimensão. Lasso. *Random Forest*.

[1] Mestranda em Estatística e Experimentação Agropecuária (UFLA). Email: alice.sduarte15@gmail.com

[2] Professora do Departamento de Zootecnia da Universidade Federal de Lavras (UFLA). E-mail: marina.danes@ufla.br

[3] Professor do Departamento de Estatística/ICET da Universidade Federal de Lavras (UFLA).E-mail: rrlima@ufla.br

[4] Professora do Departamento de Estatística/ICET da Universidade Federal de Lavras. Email: izabela.oliveira@ufla.br

5.1 INTRODUÇÃO

Modelos de regressão são técnicas utilizadas para estabelecer relação entre uma variável resposta e uma ou mais variáveis explicativas. O objetivo destes modelos é descrever como a distribuição de uma variável resposta muda de acordo com os níveis das variáveis explicativas. Um dos problemas enfrentados na elaboração de modelos de regressão é a seleção de quais variáveis devem compor o modelo, muitas vezes referido como o problema de seleção de subconjunto (AGRESTI, 2018; ANDERSEN; BRO, 2010).

Um dos métodos mais clássicos utilizado para seleção de variáveis métodos é o Stepwise, que seleciona de forma automática as variáveis, adicionando e removendo cada uma até que nenhuma adição ou subtração melhore o desempenho do modelo (EFROYMSON, 1960; NETER et al., 1996; KUTNER et al., 2005). No entanto quando o problema a ser tratado envolve dados de alta dimensão, ou seja, dados cujo número de variáveis (p) é maior do que o número de observações (n) este método apresenta alto custo computacional (THOMPSON, 1995; NASCIMENTO, 2016; HENDERSON; DENISON, 1989) o que dificulta sua aplicação.

As bases de dados de alta dimensão estão cada vez mais presentes na realidade dos pesquisadores. Frequentemente essas bases têm muitas dimensões irrelevantes e podem mascarar ruídos nos dados (PARSONS; HAQUE; LIU, 2004). Para essas situações, o lasso (Last Absolute Shrinkage And Selection Operator), que é uma técnica de penalização proposta por Tibshirani (1996) e utilizada tanto para estimação quanto para seleção de variáveis de variáveis, tem se mostrado eficaz (MEINSHAUSEN; BÜHLMANN, 2006; LIU; LUAN; LIU, 2018; MAI; ZOU; YUAN, 2012; JANG et al., 2015; WU; YANG; LIU, 2014).

Outro tipo de abordagem muito utilizada para dados deste tipo são as técnicas de machine learning., que envolvem métodos para classificação ou predição de variáveis e, muitas vezes, são uma escolha natural para dados de alta dimensão. Alguns destes métodos podem também ser usados para detecção de variáveis importantes, como o Random Forest que fornece uma medida específica de importância de variáveis, de certa forma selecionando aquelas mais importantes. Uma implementação dessa tarefa específica foi feita em R por Genuer, Poggi e Tuleau-Malot (2015) e vem sendo utilizado na literatura com o objetivo de selecionar variáveis pela importância delas em arvores de decisão (SPEISER et al., 2019; JIANG et al., 2020; GENUER; POGGI; TULEAU-MALOT, 2015; SANCHEZ-PINTO et al., 2018).

Random Forest é um método de aprendizagem supervisionada baseado em diversas árvores de decisão e amplamente utilizado em problemas de classificação ou regressão. Este algoritmo foi proposto pela primeira vez por Ho (1995). As Random Forest (ou florestas aleatórias) combinam a simplicidade das árvores de decisão com a flexibilidade e a aleatoriedade para melhorar a precisão do modelo. A aleatoriedade está presente na seleção de atributos que serão utilizados na criação das árvores que serão treinadas com conjuntos de dados distintos, garantindo que cada arvore levará a um modelo diferente (BREIMAN, 2001).

O uso de técnicas de arvores de decisão tem sido cada vez mais usado na ciência sobretudo em dados de altas dimensões e poucas observações, conhecido como “problemas de p grande e n pequeno” (ZHANG; SINGER, 2010). Dessa forma, a fim de garantir a parcimônia de um processo de modelagem e superar outras limitações da própria estrutura simples das árvores de decisão, Zhang e Singer (2010) afirmam que o método de Random Forest surge como uma solução ideal. Os modelos Random Forest são construídos através de centenas ou milhares de árvores de decisão. Sozinhas, cada árvore não representa um bom modelo, mas ao combiná-las tem-se um ganho de valor.

Um exemplo de dispositivo que resulta em dados de alta dimensão são os Near Infraed Reflectance Spectroscopy (NIRS), que usam a região do infravermelho próximo (de 750nm a 2500 nm) para coletar informações (WILLIAMS; MANLEY; ANTONISZYN, 2019). Os dispositivos NIRS são amplamente utilizados em diagnósticos, pesquisas médicas, pesquisas em saúde, agropecuária entre outras (VILLRINGER et al., 1993; BOUSHEL et al., 2001; OBRIG, 2014; FELIX et al., 2016; MENDEZ et al., 2019).

O uso da energia do infravermelho próximo (NIRS) foi descoberto no século 19, por William Herschel, no entanto, somente em 1949, com o avanço da tecnologia que a técnica passou a ser aplicada na indústria por Karl Norris. Na agricultura e pecuária as primeiras aplicações foram com o objetivo de determinar umidade em extratos de sementes (DELPHINO, 2014; WEYER et al., 2007; BURNS; CIURCZAK, 2007).

O uso de NIRS é considerado por alguns autores como principal método de análise de valor nutritivo de cerais, variáveis relativas aos aspectos nutritivos de proteína bruta, modelos fecais entre outros. Essa técnica tem sido cada vez mais frequente por se tratar de um método não evasivo, não destrutivo e não poluente capaz de explicar boa parte da variação do consumo através das fezes dos animais tendo grande vantagem em relação aos métodos tradicionais devido ao seu baixo custo e baixo impacto ambiental (LYONS; STUTH, 1992; DELPHINO, 2014; MEINERZ et al., 2011; CHAVES, 2021)

Dentre os diversos nutrientes consumidos pelos animais destacam-se o consumo de matéria seca (CMS) como sendo parâmetro fundamental para a compreensão do comportamento nutricional de animais. A mensuração do CMS é fator importante para melhorar o manejo das pastagens otimizando assim a produção uma vez que este consumo reflete diretamente o valor nutricional e aos aspectos nutritivos do pasto (UNDI et al., 2008; UNGAR; HODGSON; ILLIUS, 1996).

Assim, o objetivo deste trabalho é avaliar o uso de métodos de seleção de variáveis em dados de alta dimensão obtidos com a utilização de NIRS em um problema de predição de consumo alimentar de vacas leiteiras em um rebanho. Esses métodos serão comparados de maneira indireta por meio do modelo gerado pelas variáveis selecionadas por cada um deles. Cada um

dos métodos será validado por validação cruzada. Avaliado e escolhido o melhor método, será apresentado um modelo para o consumo de matéria seca com as variáveis selecionadas.

Este estudo de caso traz uma aplicação de seleção de variáveis para dados de NIRS comparando os métodos Stepwise, lasso e Random Forest para selecionar os comprimentos de ondas mais importantes no sentido preditivo apresentando um modelo para o consumo de matéria seca (CMS) de vacas leiteiras. Os dados foram previamente estudados por Graças (2021) comparando métodos de correção de dispersão e derivadas espectrais. Outra contribuição do presente estudo é a incorporação das formas de validação Leave one out (LOO) para modelos selecionados por lasso. Essa validação considera a retirada de um animal ou experimento que posteriormente foram usados para testar os modelos selecionados.

Os modelos foram comparados em termos de RMSE (*Root Mean Squared Error*) e MAE, sendo o RMSE a raiz quadrada do erro médio entre os valores preditos e os valores da amostra usada para validação (amostra de teste) e o MAE (*Mean absolut error*) o erro médio absoluto entre os valores observados na amostra de teste e os valores preditos usando o modelo desenvolvido na parcela de treino.

5.2 MATERIAL: A BASE DE DADOS

O estudo foi realizado utilizando dados das curvas de calibração provenientes do Centro de Pesquisa Better Nature, localizado em Ijaci, Minas Gerais, Brasil. Todos os protocolos foram aprovados pelo Comitê de Ética em Pesquisa com Animais da Universidade Federal de Lavras. A base contém 2208 variáveis das quais 2203 são informações sensoriais vindas do NIRS, a produção de leite, e três variáveis para identificação respectivamente da amostra, do experimento e da vaca.

Foram utilizadas 234 amostras de fezes, de 64 vacas leiteiras em lactação organizadas em 5 experimentos, não equilibrados, realizados em condições semelhantes. As fezes foram coletadas em baldes (coleta total) e pesadas concomitantemente à defecação durante três períodos de amostragem de 8 horas. As alíquotas de fezes foram congeladas imediatamente ao período da coleta e formada amostra composta por vaca. Os consumos foram avaliados registrando a quantidade de alimento fornecido e as sobras sendo o consumo de matéria seca determinado por secagem a 105ºC (AOAC, 1990).

Estes dados foram analisados anteriormente por Graças (2021) em um estudo utilizando calibrações do Software Unscrambler com o objetivo de desenvolver curvas de calibração de NIRS para o consumo e composição da dieta consumida avaliando diferentes técnicas de regressão e de validação dos modelos preditivos.

A variável resposta de interesse é o consumo de matéria seca (CMS), obtida pela divisão da quantidade de nutrientes excretados nas fezes pela quantidade ingerida e subtraindo o resultado de 100%.

5.2.1 Métodos de seleção de variáveis

Ao longo deste trabalho foram utilizadas três técnicas para selecionar as variáveis mais importantes para a variável resposta em questão, o método clássico Stepwise, penalização lasso e Random Forest. Uma vez que as variáveis foram selecionadas por cada método, foi ajustado um modelo linear múltiplo para a variável CMS.

O método Stepwise é um dos métodos mais usados em regressão e é composto pela união do método forward e backward. Ele atua nos dois sentidos, adiciona e remove variáveis e, a cada adição, é verificado se alguma das variáveis existentes pode ser excluída. A seleção é finalizada quando nenhuma adição ou exclusão melhora o desempenho do modelo por meio de validação cruzada do erro predito, Critério de AIC, Critério de BIC ou R^2 Ajustado (JAMES et al., 2013; NETER et al., 1996; EFROYMSON, 1960; KUTNER et al., 2005). Neste trabalho o critério utilizado para adição e exclusão de variáveis foi o critério de AIC através do pacote MASS (RIPLEY et al., 2015) do Software R (R Core Team, 2021).

O segundo método utilizado para selecionar variáveis neste estudo foi a regressão ou penalização lasso, que tem como vantagem o balanço entre viés e variância. De modo simplificado o viés mede o quão próximas as estimativas do modelo ajustado estão das verdadeiras respostas enquanto a variância é uma medida de dispersão que mede a distância entre os valores preditos e a média sintetizando o quanto o modelo consegue se adaptar às mudanças de amostra (JUNIOR, 2021).

Assim, além de diminuir a variância do modelo, ainda se tem a vantagem em casos em que há variáveis altamente correlacionadas, de selecionar apenas uma delas, zerando os coeficientes das outras de forma a minimizar a penalização facilitando a interpretação do modelo (SILVA, 2018). O estimador do lasso é definido, na sua forma lagrangeana, conforme Hastie, Tibshirani e Friedman (2009), por:

$$\hat{\beta}^{lasso} = argmin_{\beta} \left\{ \frac{1}{2} \sum_{j=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{5.1}$$

sendo $\sum_{j=1}^{p} |\beta_j|$ a penalização lasso. Devido à natureza desta restrição, tornar λ suficientemente grande faz com que alguns coeficientes sejam exatamente zero fazendo uma seleção contínua de subconjuntos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A seleção de variáveis por lasso foi feita utilizando o pacote glmnet (FRIEDMAN; HASTIE; TIBSHIRANI, 2010) do R (R Core Team, 2021). O parâmetro de complexidade λ foi definido por validação cruzada e as variáveis com menor erro na validação cruzada por meio do λ mínimo foram selecionadas.

O outro método utilizado para selecionar as variáveis mais relevantes do NIRS fecal foi o método de *random forest*, que cria diversas árvores de decisão para predizer melhor o resultado. A utilização de *random forest* com o objetivo de selecionar variáveis executa, internamente,

validação cruzada k-fold parametrizada. Neste estudo, para que 1/5 dos dados seja deixado de fora para treino e o restante usado para treinamento do modelo. Foram usadas 2000 arvores com o objetivo de verificar a importância das variáveis independentes e prever o comportamento da variável resposta.

A seleção de variáveis por Random Forest foi realizada utilizando o pacote VSURF (GENUER; POGGI; TULEAU-MALOT, 2015) em um algoritmo que seleciona variáveis em três etapas por meio de uma medida de importância dada a cada árvore. A primeira etapa chamada de "thresholding step" elimina variáveis irrelevantes do conjunto, a segunda etapa "interpretation step" seleciona as variáveis relacionadas ao modelo para fins de interpretação e ultima etapa chamada de "prediction step" refina essa seleção eliminando a redundância no conjunto de variáveis selecionadas anteriormente para fins preditivos.

Os parâmetros utilizados para o argorítimo de Random Forest estão Tabela 5.1:

Tabela 5.1 – Parametrização da seleção de variáveis por Random Forest

Parâmetro	Valor	Justificativa	Significado
mtry	47	$\sqrt{p} = \sqrt{2204}$ = 46,95	Variáveis amostradas aleatoriamente como candidatas em cada divisão.
ntree	2000	Default	Número de árvores em cada floresta cultivada. Parâmetro padrão de Random Forest.
parallel	TRUE	-	Indicação lógica sobre a execução em paralelo em vários núcleos
ncores	12	O máximo da maquina	Número de núcleos a serem usados.

Fonte: Da autora (2022).

O tempo de processamento de cada algoritmo foi medido pela biblioteca tictac (IZRAILEV, 2014).

5.2.2 Processo de validação

A qualidade de predição dos modelos desenvolvidos foi medida pelo desempenho quando implementados em um conjunto de dados de validação, sendo o ideal que esse conjunto de dados seja externo e não utilizado na calibração. No entanto, nem sempre é possível obter uma amostra externa sendo uma alternativa a validação interna utilizando parte dos dados da própria amostra, o que é chamado de validação cruzada (REFAEILZADEH; TANG; LIU, 2009).

O processo de *data splitting* consiste na separação dos dados em duas partes, uma delas para o treinamento (train dataset) e outro para teste (test dataset). Assim, os dados foram primeiramente separados, a amostra de treino usada para calibração dos modelos e, em seguida, o modelo resultante foi testado e comparado com a amostra de teste. O objetivo é verificar o erro de predição frente a uma parte da amostra fora da base de aprendizado (STONE, 1978).

Ao longo deste estudo foram utilizadas duas técnicas de validação cruzada. Para a seleção por Random Forest, o primeiro método utilizado foi a validação cruzada k-fold. Essa técnica divide os dados em k subconjuntos do mesmo tamanho, um destes subconjuntos é separado para teste o os k-1 subconjuntos restantes usados para treino do modelo. Esse processo é realizado k vezes alternando o subconjunto de teste (WONG, 2015). Esta validação foi utilizada por defaut para o modelo de *random forest* sendo 1/5 dos dados separados para teste.

Para a seleção de modelos por lasso, além da validação k-fold foi utilizado também o método leave-one-out. Esse método é um caso específico de validação amplamente utilizado em estudos experimentais por apresentar uma investigação do modelo em relação as unidades experimentais (CAWLEY; TALBOT, 2003). Considera-se neste método a retirada de uma amostra que pode ser um animal, um rebanho, um experimento ou uma unidade experimental. A divisão é repetida N vezes até que todas as observações ou unidades experimentais tenham sido utilizadas para validação (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Na realidade deste estudo, para a penalização lasso foi utilizada tanto a validação leave-one-animal-out (LOAO) quanto a validação leave-one-experiment-out (LOEO). A validação LOAO foi realizada em uma sequência de 64 modelos, ou seja, em cada um dos 64 modelos uma das 64 vacas foi deixada de fora e posteriormente utilizada para validar os resultados. Já a validação LOEO, foi realizada em um ciclo de 5 experimentos deixados um a um de fora para posterior validação. Validar os modelos deixando uma vaca/experimento de fora é importante para que seja garantido uma relação mínima entre a parcela utilizada para treino e os dados de teste para que o modelo não seja, por exemplo, treinado com uma vaca ou um experimento que também compõe a amostra de teste (KOCAGUNELI; MENZIES, 2013; MAGNUSSON et al., 2019).

5.3 RESULTADOS E DISCUSSÃO

As variáveis analisadas neste estudo são, na grande maioria, valores obtidos pela espectroscopia de reflectância no infravermelho próximo (NIRS) com exceção apenas da produção de leite. Estas variáveis são geradas pelas cores refletidas captadas pelo sensor e são caracterizadas com coordenadas. Neste contexto, para cada unidade experimental, 2203 variáveis são geradas correspondentes a parcela de cores refletidas pelo sensor.

Assim, selecionar quais os comprimentos de ondas são mais importantes e qual o melhor método para selecioná-los se torna tarefa essencial para modelagem de dados. Desta forma os resultados deste estudo buscam, num primeiro momento, identificar qual o método de seleção mais eficaz e em seguida aplicá-lo na base de dados avaliando o método de seleção quanto à capacidade preditiva do modelo.

5.3.1 Seleção de Variáveis

Um dos métodos mais clássicos para selecionar variáveis é o método *stepwise*. No entanto esse método demonstra problemas quando aplicados a base de dados de alta dimensão ($p > n$), uma vez que este método busca diretamente todas as variáveis relevantes deixando desafiadora a tarefa de eliminar variáveis podendo levar o método a falha (AN et al., 2008; WASSERMAN; ROEDER, 2009; HWANG; HU, 2015). Assim, por se tratar de dados de NIRS cujo número de variáveis é aproximadamente 10 vezes maior do que o número de observações foi inviável o uso de Stepwise de maneira direta devido ao tempo de processamento fora da viabilidade do estudo.

A regressão lasso, por outro lado, tem desempenho reconhecido em dados de alta dimensão, de maneira geral, e também tem sido utilizada na literatura em vários contextos de dados provenientes de NIRS (CAI et al., 2020; LUAN; LIU; LIU, 2020; LIU; LUAN; LIU, 2018; HUO et al., 2020).

Utilizando a penalização lasso para predição da variável o consumo de matéria seca (CMS), foram selecionadas 10 variáveis em mais de 70% das amostras considerando o critério de validação leave one animal out para validação. A seleção de variáveis foi repetida 64 vezes, cada uma delas deixando uma das vacas de fora, registrando as variáveis selecionadas em cada repetição.

Na validação leave one experimente out o ciclo abrange 5 repetições em que, em cada repetição, um dos experimentos era deixado de fora para ser usado de teste do modelo. Por esta técnica de validação foram selecionadas 11 variáveis em mais de 60% dos modelos.

Abordagens de Machine Learning também vêm sendo amplamente utilizadas para dados de alta dimensão, inclusive em dados de NIRS (BAATH et al., 2020; LIU; YANG; DENG, 2015; RUSSELL-BUCKLAND et al., 2018). Random Forest (ou florestas aleatórias) é um destes métodos. Neste trabalho o uso de Random Forest foi usado, principalmente, para selecionar variáveis utilizando a importância de variáveis no sentido preditivo. Por este método foram selecionadas aproximadamente 6 variáveis

Após a aplicação dos três métodos citados anteriormente, sendo o lasso considerado com três validações diferentes, os resultados são mostrados na Tabela 5.2.

As métricas apontadas na tabela 5.2 se referem ao resultado médio das bases de teste na validação cruzada.

Tabela 5.2 – Comparação entre os métodos de seleção de variáveis lasso utilizando método de validação LOAO, LOEO, k-fold, Random Forest e Stepwise

	RMSE	MAE	Tempo de execução da etapa de seleção (em segundos)
CMS: LASSO LOAO	2,6359	1,8867	0,66
CMS: LASSO LOEO	2,829	2,0305	0,48
CMS: LASSO (k-fold)	2,3496	1,9257	0,51
CMS: Random Forest	2,8015	1,8619	617261,64
CMS: Stepwise	-	-	-

Fonte: Da autora (2022).

Considerando como critério de escolha do modelo o menor valor de RMSE[1] e MAE[2] o método de seleção mais eficiente foi o lasso. Cabe destacar que, o método *k-fold* foi aplicado na seleção por lasso apenas para fins comparativos com Random Forest uma vez que este método pode apresentar viés por possibilitar que a parcela de treino e de teste contenha o mesmo animal/experimento.

O tempo de execução apontado pela tabela foi medido por meio da biblioteca tictoc (IZRAILEV, 2014) e equivale em todos os métodos à execução utilizando um núcleo de processamento. No entanto, como o tempo da execução do algoritmo de Random Forest foi demasiado alto (aproximadamente 171 horas) foi utilizado, apenas neste caso, o uso de 12 núcleos permitindo que fosse possível executá-lo em aproximadamente 13 horas. Esse tempo de processamento equivale apenas à etapa de seleção de cada modelo. Uma vez selecionadas as variáveis o modelo obtido foi executado, em todas as situações, em menos de 0,87 segundos.

Os métodos de seleção lasso apresentaram menor raiz quadrática média dos erros (RMSE) entre os valores da amostra de teste e os valores preditos pelo modelo linear treinado com as variáveis selecionadas além de ser reproduzido de forma mais eficiente, considerando como critério de eficiência o tempo de execução da etapa de seleção. Resultados similares foram encontrados na literatura comprovando a eficácia da seleção por lasso em dados de NIRS (ZHAO; OGDEN; REISS, 2012; BRICKLEMYER et al., 2018; WANG et al., 2021)

No entanto, esse estudo se limita na utilização dos métodos de forma independente. Outros autores obtiveram bons resultados aplicando mais de um método simultaneamente, sendo o lasso destaque na etapa inicial de redução dos preditores (CAI et al., 2020; SHUKLA; BHATT; PANI, 2020). O uso de lasso apresenta vantagens devido à sua flexibilidade possibilitando a utilização do método para outros tipos de variáveis respostas além da normal, como dados de

[1] **RMSE (Root Mean Squared Error)**: medida que calcula a raiz quadrática média dos erros entre os valores da amostra de teste e os valores preditos pelo modelo treinado com a parcela de treino

[2] **MAE (Mean Absolute Error):** Erro absoluto médio entre os valores da amostra de teste e os valores preditos pelo modelo treinado com a parcela de treino.

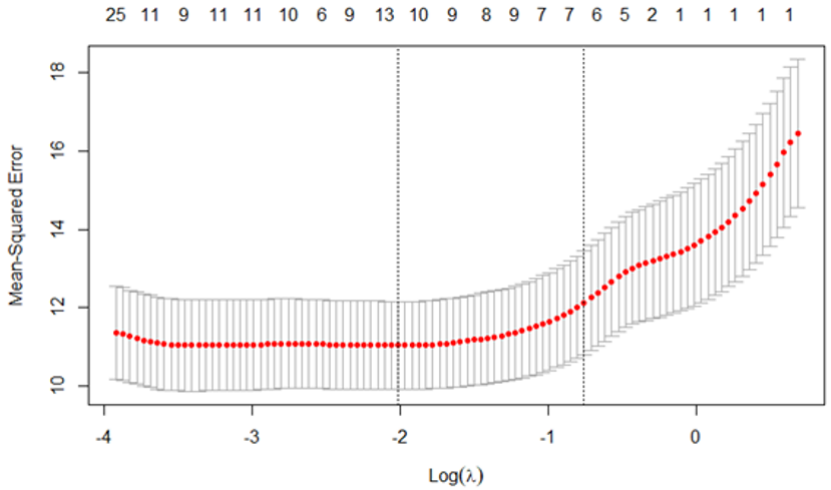
contagens, dados de proporção e outros tipos de dados assumindo uma distribuição na família dos modelos lineares generalizados.

Outros métodos abordados na literatura que também apresentam bons resultados quando aplicados a dados de alta dimensão de sensor NIRS foram os métodos Sparce partial least Square (SPLS) proposto por Chun e Keleş (2010) e o método PSL desenvolvido por (S SJÖSTRÖM M, 2001). Ambos os métodos reduzem a dimensão do conjunto de variáveis resposta por meio de seleção retirando variáveis altamente correlacionadas, reduzindo assim a dimensionalidade dos dados. Além destes, vários outros métodos são utilizando para esse tipo de dados, tais como Principal Components Analysis (PCA), BLASSO (Bayesian Lasso Regression), Algoritmo Genérico (AG), Algoritmo das Projeções Sucessivas (APS) entre outros (FERREIRA, 2018; GOMES et al., 2012; SOARES et al., 2010).

5.3.2 Modelo para predição de Consumo de Matéria Seca

Escolhida a técnica de seleção de variáveis o modelo foi obtido através da aplicação de lasso na base de dados completa. A estimativa do parâmetro λ foi feita por validação cruzada e o valor de lambda que minimiza o erro de validação cruzada foi de $0,1336145$ $log(0,1336145 \approx -2,012797)$

Figura 5.1 – Valores de λ para o erro quadrático médio de validação cruzada.



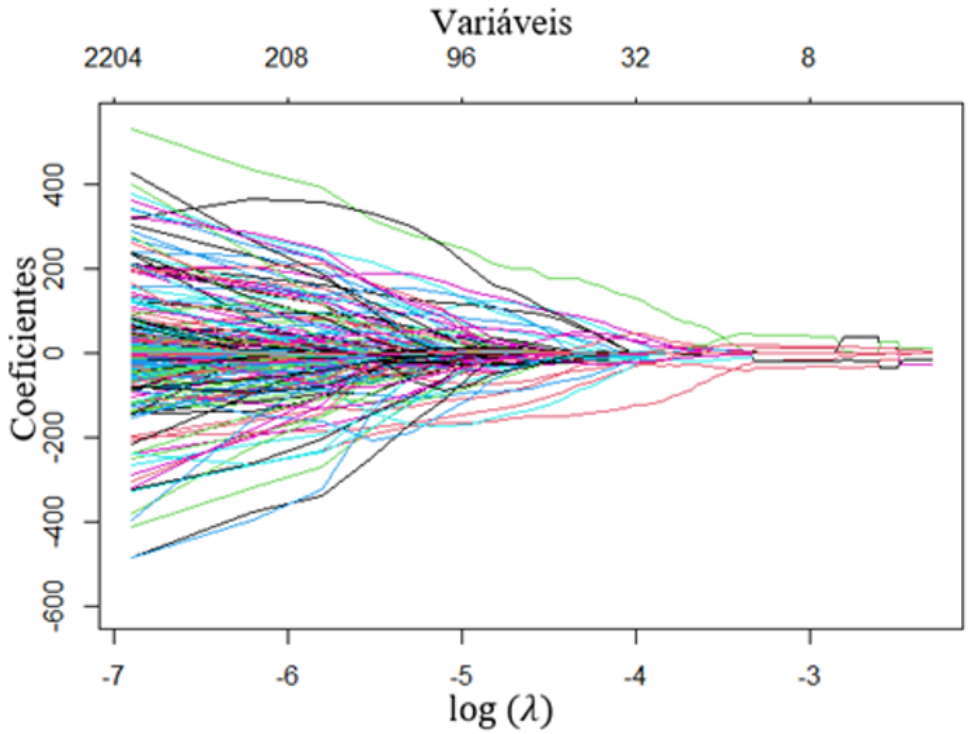
Fonte: Da autora (2022).

Assim, utilizando $\lambda = 0,1336145$, as variáveis são reduzidas de 2203 para apenas 10, equivalente a menos de 1% das observações. Isso se deve a característica do lasso penalizar os coeficientes levando vários deles a 0 (RANSTAM; COOK, 2018; TIBSHIRANI, 2011). Essa redução nos preditores soluciona boa parte dos problemas de alta dimensão (QIAN et al., 2019; BERMINGHAM et al., 2015). De acordo com Cai et al. (2020), apesar do ótimo desempenho do lasso em cenário de NIRS, a dispersão do sinal fluorescente pode ignorar informações de

estrutura com informações importantes de borda devido ao excesso de dispersão, o autor ainda sugere o uso de lasso combinado com outros métodos para reconstrução desta estrutura.

Shukla, Bhatt e Pani (2020) em um estudo utilizando dados de NIRS para previsão da qualidade do diesel encontrou bom desempenho na utilização do lasso seguido pela modelagem por *Random Forest*.

Figura 5.2 – Redução de variáveis por penalização lasso em função dos valores de λ



Fonte: Da autora (2022).

Tabela 5.3 – Estimativa dos coeficientes do modelo

Variáveis	Estimativa	Valor-p
Intercepto	46,8269	0,00411
X1 - Produção de Leite	0,4073	0,000043
X1903 - X5.160.821	4064,7289	0,83149
X1902 - X5.164.678	3063,7218	0,74955
X1904 - X5.156.964	-7123,479	0,47945
X2120 - X4.323.827	-69,7117	0,10062
X641 - X10.028.502	-2017,535	0,05823
X38 - X12.354.343	-139,0129	0,00034
X632 - X10.063.216	1924,1759	0,46385
X633 - X10.059.359	278,3659	0,91767

Fonte: Da autora (2022).

O modelo produzido com as variáveis selecionadas acima na amostra de validação possui $R^2 = 0,598$ com $R^2_{ajustado} = 0,491$. Assim o modelo selecionado por lasso foi capaz de explicar aproximadamente 59,8% da variação do Consumo de Matéria Seca na amostra de validação utilizando 10 variáveis, sendo 9 delas comprimentos de ondas proveniente do sensor NIRS e a variável produção de leite. Destaca-se que apenas duas delas são significantes a 5%, a produção de leite e uma das coordenadas do NIRS.

5.4 CONSIDERAÇÕES FINAIS

O conjunto de dados é proveniente de um mesmo rebanho, assim apresenta como desvantagem a possibilidade do modelo ser menos preciso do que quando usado em rebanhos e ambientes diferentes.

Foi possível executar a seleção de variáveis por lasso considerando três tipos de validação, são elas: validação LOAO (*leave-one-animal-out*), LOEO (*leave-one-experiment-out*) e a validação *k-fold*. Destaca-se que, a validação *k-fold* foi usada apenas para fins comparativos com o modelo *Random Forest* que utiliza esta mesma validação por defaut na implementação aqui utilizada. No entanto, ela produz viés uma vez que os dados são compostos por várias observações de um mesmo animal podendo, o animal, estar tanto na parcela de treino como na amostra de teste.

Utilizando como critério de seleção de modelos o RMSE o método selecionado neste contexto foi a regressão por lasso utilizando 10 preditores, cujo modelo reduziu a dimensão dos dados em 99,54%.

As contribuições deste estudo estão na comparação entre os métodos lasso e Random Forest usados separadamente para seleção de variáveis em NIRS, tendo como diferencial a comparação entre diferentes validações para o lasso. Estes resultados são restritos a esta base de dados, estudos futuros podem ser feitos considerando a união entre os métodos, outras aplicações, validação fora da amostra e a comparação entre lasso e Random Forest com os métodos tradicionalmente utilizados, como Análise de componentes principais (*Principal Componentes Analysis*) e Regressão por mínimos quadrados parciais (*Partial Least Squares*).

REFERÊNCIAS

AGRESTI, A. **An introduction to categorical data analysis**. [S.l.]: John Wiley & Sons, 2018.

AN, H. et al. Stepwise searching for feature variables in high-dimensional linear regression. The London School of Economics and Political Science, 2008.

ANDERSEN, C. M.; BRO, R. Variable selection in regression—a tutorial. **Journal of chemometrics**, Wiley Online Library, v. 24, n. 11-12, p. 728–737, 2010.

BAATH, G. S. et al. Predicting forage quality of warm-season legumes by near infrared spectroscopy coupled with machine learning techniques. **Sensors**, MDPI, v. 20, n. 3, p. 867, 2020.

BERMINGHAM, M. L. et al. Application of high-dimensional feature selection: evaluation for genomic prediction in man. **Scientific reports**, Nature Publishing Group, v. 5, n. 1, p. 1–12, 2015.

BOUSHEL, R. et al. Monitoring tissue oxygen availability with near infrared spectroscopy (nirs) in health and disease. **Scandinavian journal of medicine & science in sports**, Wiley Online Library, v. 11, n. 4, p. 213–222, 2001.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BRICKLEMYER, R. S. et al. Comparing vis–nirs, libs, and combined vis–nirs-libs for intact soil core soil carbon measurement. **Soil Science Society of America Journal**, Wiley Online Library, v. 82, n. 6, p. 1482–1496, 2018.

BURNS, D. A.; CIURCZAK, E. W. **Handbook of near-infrared analysis**. [S.l.]: CRC press, 2007.

CAI, M. et al. Nir-ii/nir-i fluorescence molecular tomography of heterogeneous mice based on gaussian weighted neighborhood fused lasso method. **IEEE Transactions on Medical Imaging**, IEEE, v. 39, n. 6, p. 2213–2222, 2020.

CAWLEY, G. C.; TALBOT, N. L. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. **Pattern Recognition**, Elsevier, v. 36, n. 11, p. 2585–2592, 2003.

CHAVES, A. d. L. Uso da espectroscopia de reflectância no infravermelho próximo (nirs) para monitorar o status nutricional de ovinos em sistemas integrados na caatinga. 2021., 2021.

CHUN, H.; KELEŞ, S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 72, n. 1, p. 3–25, 2010.

DELPHINO, T. R. A espectroscopia de reflectância no infravermelho próximo pode predizer consumo, digestibilidade e balanço do nitrogênio de dietas alto concentrado de ovinos. Universidade Estadual Paulista (Unesp), 2014.

EFROYMSON, M. **Mathematical Methods for Digital Computers, chapter Multiple regression analysis**. [S.l.]: Wiley, New York, NY, 1960.

FELIX, J. C. et al. Predição de fósforo, carbono e nitrogênio em solos de basalto, por meio de espectroscopia nir. **Pesquisa Agropecuária Brasileira**, SciELO Brasil, v. 51, p. 1405–1416, 2016.

FERREIRA, R. d. A. Comparação de métodos de seleção de variáveis em regressão aplicados a dados genômicos e de espectroscopia nir. Universidade Federal de Viçosa, 2018.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for generalized linear models via coordinate descent. **Journal of statistical software**, NIH Public Access, v. 33, n. 1, p. 1, 2010.

GENUER, R.; POGGI, J.-M.; TULEAU-MALOT, C. Vsurf: an r package for variable selection using random forests. **The R Journal**, v. 7, n. 2, p. 19–33, 2015.

GOMES, A. d. A. et al. Algoritmo das projeções sucessivas aplicado à seleção de variáveis em regressão pls. Universidade Federal da Paraíba, 2012.

GRAÇAS, L. E. C. D. spectroscopia de reflectância no infravermelho próximo de fezes para predizer variáveis nutricionais de vacas leiteiras confinadas. Orientadora: Marina de Arruda Camargo Danes, 2021. Dissertação (Mestrado) – Universidade Federal de Lavras (UFLA)., 2021.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. An introduction to statistical learning. 2009.

HENDERSON, D. A.; DENISON, D. R. Stepwise regression in social and psychological research. **Psychological Reports**, SAGE Publications Sage CA: Los Angeles, CA, v. 64, n. 1, p. 251–257, 1989.

HO, T. K. Random decision forests. In: IEEE. **Proceedings of 3rd international conference on document analysis and recognition**. [S.l.], 1995. v. 1, p. 278–282.

HUO, J. et al. Lasso based similarity learning of near-infrared spectra for quality control. In: IEEE. **2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS)**. [S.l.], 2020. p. 424–427.

HWANG, J.-S.; HU, T.-H. A stepwise regression algorithm for high-dimensional variable selection. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 85, n. 9, p. 1793–1806, 2015.

IZRAILEV, S. tictoc: Functions for timing r scripts, as well as implementations of stack and list structures. **R package version**, v. 1, 2014.

JAMES, G. et al. **Springer texts in statistics: an introduction to statistical learning**. [S.l.]: Springer New York: Imprint: Springer New York, NY, 2013.

JANG, W. et al. Some properties of generalized fused lasso and its applications to high dimensional data. **Journal of the Korean Statistical Society**, Springer, v. 44, n. 3, p. 352–365, 2015.

JIANG, F. et al. Estimating the growing stem volume of coniferous plantations based on random forest using an optimized variable selection method. **Sensors**, MDPI, v. 20, n. 24, p. 7248, 2020.

JUNIOR, G. P. d. A. Avaliação do lasso e métodos alternativos em modelos de regressão logística. Universidade Federal de São Carlos, 2021.

KOCAGUNELI, E.; MENZIES, T. Software effort models should be assessed via leave-one-out validation. **Journal of Systems and Software**, Elsevier, v. 86, n. 7, p. 1879–1890, 2013.

KUTNER, M. H. et al. Applied linear statistical models, 2005. **McGraw Hill Irwin, New York. NY**, p. 409, 2005.

LIU, C.; YANG, S. X.; DENG, L. Determination of internal qualities of newhall navel oranges based on nir spectroscopy using machine learning. **Journal of Food Engineering**, Elsevier, v. 161, p. 16–23, 2015.

LIU, J.; LUAN, X.; LIU, F. Adaptive jit-lasso modeling for online application of near infrared spectroscopy. **Chemometrics and Intelligent Laboratory Systems**, Elsevier, v. 183, p. 90–95, 2018.

LUAN, X.; LIU, J.; LIU, F. Multilevel lasso-based nir temperature-correction modeling for viscosity measurement of bisphenol-a. **ISA transactions**, Elsevier, v. 107, p. 206–213, 2020.

LYONS, R. K.; STUTH, J. W. Fecal nirs equations for predicting diet quality of free-ranging cattle. **Rangeland Ecology & Management/Journal of Range Management Archives**, v. 45, n. 3, p. 238–244, 1992.

MAGNUSSON, M. et al. Bayesian leave-one-out cross-validation for large data. In: PMLR. **International Conference on Machine Learning**. [S.l.], 2019. p. 4244–4253.

MAI, Q.; ZOU, H.; YUAN, M. A direct approach to sparse discriminant analysis in ultra-high dimensions. **Biometrika**, Oxford University Press, v. 99, n. 1, p. 29–42, 2012.

MEINERZ, G. R. et al. Valor nutritivo da forragem de genótipos de cereais de inverno de duplo propósito. **Revista Brasileira de Zootecnia**, SciELO Brasil, v. 40, p. 1173–1180, 2011.

MEINSHAUSEN, N.; BÜHLMANN, P. High-dimensional graphs and variable selection with the lasso. **The annals of statistics**, Institute of Mathematical Statistics, v. 34, n. 3, p. 1436–1462, 2006.

MENDEZ, J. et al. Trends in application of nir and hyperspectral imaging for food authentication. **Scientia Agropecuaria**, Universidad Nacional de Trujillo. Facultad de Ciencias Agropecuarias, v. 10, n. 1, p. 143–161, 2019.

NASCIMENTO, A. R. d. Aplicações em quimiometria. 2016.

NETER, J. et al. Applied linear statistical models. Irwin Chicago, 1996.

OBRIG, H. Nirs in clinical neurology—a 'promising'tool? **Neuroimage**, Elsevier, v. 85, p. 535–546, 2014.

PARSONS, L.; HAQUE, E.; LIU, H. Subspace clustering for high dimensional data: a review. **Acm sigkdd explorations newsletter**, ACM New York, NY, USA, v. 6, n. 1, p. 90–105, 2004.

QIAN, J. et al. A fast and flexible algorithm for solving the lasso in large-scale and ultrahigh-dimensional problems. **BioRxiv**, Cold Spring Harbor Laboratory, p. 630079, 2019.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2021. Disponível em: <https://www.R-project.org/>.

RANSTAM, J.; COOK, J. Lasso regression. **Journal of British Surgery**, Oxford University Press, v. 105, n. 10, p. 1348–1348, 2018.

REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. **Encyclopedia of database systems**, Springer, v. 5, p. 532–538, 2009.

RIPLEY, B. et al. **MASS: Support functions and datasets for Venables and Ripley's MASS. R package version 7.3-40**. 2015.

RUSSELL-BUCKLAND, J. et al. Abroad: a machine learning based approach to detect broadband nirs artefacts. In: **Oxygen Transport to Tissue XL**. [S.l.]: Springer, 2018. p. 319–324.

S SJÖSTRÖM M, E. L. W. Pls-regression: a basic tool of chemometrics. chemometrics and intelligent laboratory systems 58: 109-130. 2001.

SANCHEZ-PINTO, L. N. et al. Comparison of variable selection methods for clinical predictive modeling. **International journal of medical informatics**, Elsevier, v. 116, p. 10–17, 2018.

SHUKLA, A.; BHATT, H.; PANI, A. K. Variable selection and modeling from nir spectra data: A case study of diesel quality prediction using lasso and regression tree. In: IEEE. **2nd International Conference on Data, Engineering and Applications (IDEA)**. [S.l.], 2020. p. 1–6.

SILVA, C. B. P. d. A técnica lasso e suas potencialidades na seleção de variáveis para modelos lineares. 2018.

SOARES, S. F. C. et al. Um novo critério para seleção de variáveis usando o algoritmo das projeções sucessivas. Universidade Federal da Paraíba, 2010.

SPEISER, J. L. et al. A comparison of random forest variable selection methods for classification prediction modeling. **Expert systems with applications**, Elsevier, v. 134, p. 93–101, 2019.

STONE, M. Cross-validation: A review. **Statistics: A Journal of Theoretical and Applied Statistics**, Taylor & Francis, v. 9, n. 1, p. 127–139, 1978.

THOMPSON, B. **Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial**. [S.l.]: Sage Publications Sage CA: Thousand Oaks, CA, 1995. 525–534 p.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. JOURNAL OF THE ROYAL STATISTICAL SOCIETY: SERIES B (STATISTICAL METHODOLOGY), v. 1, p. 267–288, 1996.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of The Royal Statistical Society: Series B(Statistical Methodology)**, 2011.

UNDI, M. et al. Comparison of techniques for estimation offorage dry matter intake bygrazing beef cattle. **Canadian Journal of Animal Science**, NRC Research Press Ottawa, Canada, v. 88, n. 4, p. 693–701, 2008.

UNGAR, E.; HODGSON, J.; ILLIUS, A. The ecology and management of grazing systems. **Chapter Ingestive Behaviour**, p. 185–218, 1996.

VILLRINGER, A. et al. Near infrared spectroscopy (nirs): a new tool to study hemodynamic changes during activation of brain function in human adults. **Neuroscience letters**, Elsevier, v. 154, n. 1-2, p. 101–104, 1993.

WANG, K. et al. A new ensemble modeling method for multivariate calibration of near infrared spectra. **Analytical Methods**, Royal Society of Chemistry, v. 13, n. 11, p. 1374–1380, 2021.

WASSERMAN, L.; ROEDER, K. High dimensional variable selection. **Annals of statistics**, NIH Public Access, v. 37, n. 5A, p. 2178, 2009.

WEYER, L. et al. **Practical guide to interpretive near-infrared spectroscopy**. [S.l.]: CRC press, 2007.

WILLIAMS, P.; MANLEY, M.; ANTONISZYN, J. **Near infrared technology: getting the best out of light**. [S.l.]: African Sun Media, 2019.

WONG, T.-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. **Pattern Recognition**, Elsevier, v. 48, n. 9, p. 2839–2846, 2015.

WU, L.; YANG, Y.; LIU, H. Nonnegative-lasso and application in index tracking. **Computational Statistics & Data Analysis**, Elsevier, v. 70, p. 116–126, 2014.

ZHANG, H.; SINGER, B. H. **Recursive partitioning and applications**. [S.l.]: Springer Science & Business Media, 2010.

ZHAO, Y.; OGDEN, R. T.; REISS, P. T. Wavelet-based lasso in functional linear regression. **Journal of computational and graphical statistics**, Taylor & Francis, v. 21, n. 3, p. 600–617, 2012.