Área 12 - Economia Social e Demografia Econômica

MODELOS DE MACHINE LEARNING NA CLASSIFICAÇÃO DE POBREZA: UMA APLICAÇÃO PARA O ESTADO DO CEARÁ

Vitor Hugo Miro Couto Silva[1]

João Mário Santos de França[2]

Resumo

O presente artigo apresenta uma aplicação da associação entre métodos de *machine learning* e testes de elegibilidade como o *proxy means test* (PMT). Propõe-se uma discussão a respeito da aplicabilidade de modelos de *machine learning* na classificação de beneficiários de programas direcionados ao combate à pobreza, assumindo a hipótese de que esses métodos podem aprimorar mecanismos de seleção e prover melhorias na focalização de políticas. O exercício empírico realizado emprega dados da PNAD Contínua e ajusta um modelo de classificação de domicílios/famílias segundo seu status de pobreza, adotando um algoritmo *Extreme Gradient Boosting* (*XGBoost*).

Palavras-chave: Aprendizado de Máquina; Pobreza, teste de elegibilidade.

Abstract

This article presents an application of the association between machine learning methods and eligibility tests such as proxy means test (PMT). We propose a discussion about the applicability of machine learning models in classifying beneficiaries of poverty reduction programs, assuming the hypothesis that these methods can improve selection mechanisms and improve policy targeting. The empirical exercise carried out uses data from the Continuous PNAD and adjusts a household/family classification model according to their poverty status, adopting the Extreme Gradient Boosting (XGBoost) algorithm.

Keywords: Machine Learning, Poverty, proxy means test.
Códigos JEL: C14, C52, I32, I38

[1] Professor do Programa de Pós-Graduação em Economia Rural (PPGER/UFC) e pesquisador do Centro de Análise de Dados e Avaliação de Políticas Públicas (CAPP/IPECE).

[2] Professor do Programa de Pós-Graduação em Economia (CAEN/UFC) e Diretor Geral do Instituto de Pesquisa e Estratégia Econômica do Ceará (IPECE).

Os autores agradecem ao CAPP/IPECE e apoio financeiro da Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (FUNCAP).

1. INTRODUÇÃO

A focalização é uma propriedade importante no desenho de políticas públicas de combate à pobreza. Ela permite o direcionamento deliberado de recursos públicos para as pessoas que mais necessitam destes, potencializando os impactos esperados em termos da retirada de pessoas da situação de pobreza ou redução de restrições associadas à esta condição.

Uma das questões pertinentes para prover boa focalização em uma política de combate à pobreza é a identificação das famílias pobres, como beneficiárias ou potenciais beneficiárias da política. No entanto, gestores de programas sociais direcionados às famílias vulneráveis não possuem informação perfeita sobre quem se encontra em condição de pobreza. Sendo a elegibilidade do programa baseada em informação imperfeita, é inevitável a presença de algum erro de inclusão (indivíduos que não são pobres sendo beneficiados), e/ou de exclusão (indivíduos ou domicílios que são pobres não sendo beneficiados).

Existem ferramentas que permitem aprimorar mecanismos de seleção e tornam possível a redução ou eliminação dos erros de inclusão ou exclusão, baseados na coleta de dados adicionais, cruzamento de bases de dados administrativos e processamento destas informações. No entanto, esse aprimoramento envolve custos, e em um ambiente com recursos limitados, gestores devem avaliar se tais custos são justificados pela melhoria na focalização.

Em algumas situações o cadastro de beneficiários e potenciais beneficiários existe, mas os pobres são classificados com base na renda declarada e informações adicionais podem ser insuficientes para permitir uma análise mais apurada da real situação. Por sua vez, a avaliação necessária das características e bens da família pode ser cara e trabalhosa, exigindo do entrevistador maior quantidade de tempo por família.

O uso de variáveis *proxy*, como gênero, educação, tamanho da família e condições de moradia, em substituição a indicadores que não sejam facilmente manipuláveis por beneficiários reais ou potenciais, se tornaram comuns para a atribuição de pontuações e a validação de outros métodos de focalização. O esforço de desenhar ferramentas para aprimorar a identificação de beneficiários resultou nos testes de elegibilidade como o *proxy means test* (PMT). Versões do PMT baseados em regressão se tornaram populares nos anos de 1990, em particular após a contribuição de Grosh (1994, *apud* BROWN *et al*., 2016) que comparou vários programas sociais na América Latina e concluiu que essa classe de métodos produziu os melhores resultados de focalização, medidos em termos de redução de erros de inclusão (BROWN et al., 2016).

O presente estudo propõe o uso de métodos de *machine learning* associados com a aplicação de PMT. A hipótese assumida é que os modelos de *machine learning* podem melhorar o desempenho preditivo das técnicas estatísticas tradicionalmente aplicadas na estimação do PMT. Nesse sentido, realiza-se um exercício demonstrativo da aplicação de um modelo preditivo com objetivo é mostrar a viabilidade do uso destas técnicas para aprimorar decisões de políticas direcionadas ao combate à pobreza.

No exercício realizado foi utilizado o método conhecido como *Extreme Gradient Boosting*, que é um método de aprendizado supervisionado bastante elogiado por sua capacidade de gerar boas previsões. Outras técnicas também foram empregadas de forma complementar como os métodos de LASSO e de Floresta Aleatória (*Random Forest)*, utilizados em etapa de seleção de variáveis.

Os dados utilizados são provenientes da PNAD Contínua (PNADC), relativos aos microdados da 1ª visita coletados em 2019. Vale ressaltar que as pesquises domiciliares, como a PNADC, constituem uma importante ferramenta para o desenho e o acompanhamento de políticas públicas, mas o cadastro de famílias para fins administrativos é algo que está fora do propósito de tais pesquisas. Aqui os dados da PNADC foram utilizados por sua disponibilidade de variáveis e facilidade de acesso, mas se deve ter em mente que estes dados não permitem a identificação de famílias e seu uso para seleção de beneficiários de políticas.

Ainda com relação aos dados utilizados também vale mencionar que foi adotado o recorte para o estado do Ceará. Localizado na região Nordeste do Brasil, o estado do Ceará é um dos

estados com maior contingente populacional em situação de pobreza. Estimativas para o ano de 2019 mostram que mais de 40% da população do estado era classificada como pobre e 12% como extremamente pobre, segundo dados da PNADC e linhas de pobreza adotadas pelo Banco Mundial. Estes percentuais representam um contingente estimado em 3,7 milhões de pessoas pobres e 1,1 milhões em pobreza extrema.

Em função desse cenário, o Ceará também conta com um grande conjunto de políticas sociais direcionadas aos pobres. Estratégias de combate à pobreza em nível estadual são, em grande medida, financiadas com recursos do Fundo Estadual de Combate à Pobreza, o FECOP. Os recursos do FECOP financiam um grupo heterogêneo de projetos com ações em diferentes áreas como de nutrição, habitação, educação, saúde, saneamento básico e até mesmo reforço de renda familiar. Em 2019, os recursos do FECOP foram aplicados em 74 projetos, totalizando um valor aplicado superior a R$515 milhões (CEARÁ, 2020).

Além das estratégias de combate à pobreza, o Instituto de Pesquisa e Estratégia Econômica do Ceará (IPECE) reúne esforços na construção de um sistema de informações para subsidiar as estratégias de combate à pobreza no estado. Destacam-se as iniciativas da Pesquisa Regional por Amostra de Domicílios (PRAD/CE) e a formatação de um sistema provisoriamente denominado como Sistema de Cadastramento de Beneficiários e Monitoramento de Indicadores de Resultado – SABE.

A PRAD/CE, levada a campo entre setembro e novembro de 2019, realizou o levantamento de informações socioeconômicas da população cearense, tendo representatividade estatística para o Estado, incluindo as populações das zonas urbana e rural, e suas quatorze regiões de planejamento (MEDEIROS *et al*., 2021). Por sua vez, o SABE propõe uma integração entre os dados do Cadastro Único para Programas Sociais do Governo Federal (CadÚnico) e cadastros de programas e projetos financiados pelo FECOP, como uma forma de organizar informações de beneficiários potenciais e efetivos.

Nesse sentido, o artigo apresenta uma discussão inicial a respeito da associação de modelagem preditiva, técnicas de *machine learning* e os sistemas de informação disponíveis, sob a hipótese de que essa combinação pode trazer benefícios em termos de aprimoramento das políticas de combate à pobreza, seja no Ceará, como em qualquer outra unidade federativa ou localidade.

2. REVISÃO DE LITERATURA

Em contextos com poucos recursos para coleta e processamento de informações a respeito de famílias o método de *proxy means test* (PMT) tornou-se bastante difundido, uma vez se configura em um uma alternativa rápida, que conserva o rigor científico e demanda uma menor quantidade de recursos e informações.

Um PMT pode ser desenvolvido usando dados de pesquisas por amostra de domicílios que sejam representativas e normalmente utiliza um pequeno número de perguntas, entre 10 e 30, com o objetivo de estimar a probabilidade de uma família estar em situação de pobreza (KSHIRSAGAR et al., 2017).

De forma geral, considerado um conjunto de características de domicílios e pessoas organizados de forma que se pode definir um vetor de x_{ij}, em que cada vetor representa uma característica j para um conjunto de domicílios e pessoas indexadas por i. Ponderando cada característica i por pesos γ_j, o PMT permite o cálculo de um escore de forma simples

$$PMTscore_i = \sum_j x_{ij}\gamma_j \quad [1]$$

Com base no escore é possível classificar domicílios e pessoas de acordo com sua situação de pobreza e utilizar esta informação como critério para a seleção de beneficiários de políticas sociais.

A princípio, modelos de previsão da pobreza podem ser estimados aplicando métodos simples e tradicionais como modelos lineares estimados por Mínimos Quadrados Ordinários, no caso variáveis dependentes contínuas (como renda e gastos), ou métodos de regressão logística, no caso de variável dependente categórica (identificador de pobreza). Conforme destaca Brown *et al*. (2016), métodos de regressão linear estimados por Mínimos Quadrados Ordinários representam uma das formas mais populares de obter os pesos utilizados no cálculo do escore do PMT. A aplicação de modelos de regressão linear, no entanto, sempre foi alvo de muitas críticas.

Com a crescente aplicação de métodos de *machine learning*, em especial os métodos de aprendizado supervisionado para regressão e classificação, uma nova alternativa para aprimorar mecanismos de classificação de pobreza passa a ser considerada. Dentre os modelos que mais receberam destaque na literatura destacam-se os modelos de Floresta Aleatória/*Random Forest* (RF) e LASSO (*Least Absolute Shrinkage and Selection Operator*).

Estudos de McBride e Nichols (2015 e 2018) e Sohnesen e Stender (2017) mostraram que tanto o modelo de RF quanto o LASSO podem reduzir o erro de exclusão em PMTs para selecionar famílias pobres com mais precisão.

McBride e Nichols (2016) analisaram o desempenho de métodos de RF em comparação com os resultados obtidos por modelos de regressão na aplicação de PMT para predição de pobreza com dados da Bolívia, Timor-Leste e Malawi. Com o objetivo de aprimorar o modelo criado pela *United States Agency for International Development* (USAID), os autores concluíram que os métodos de RF melhoraram significativamente o desempenho da previsão "fora da amostra".

Sohnesen e Stender (2017) aplicaram os métodos LASSO e RF para prever a pobreza usando um ano de dados para previsão dentro do mesmo ano e dois anos de dados para prever a pobreza ao longo do tempo. Os resultados obtidos apontaram que a aplicação do método de RF forneceu uma estimativa mais robusta do que os métodos de regressão linear, resultando em previsões altamente precisa em áreas urbanas e rurais. Os resultados mostraram, no entanto, que em nível nacional, o método de RF não oferece previsões melhores do que o método de LASSO e modelos de regressão linear.

Kshirsagar et al. (2017) aplicaram o método de bootstrap em conjunto com o LASSO para selecionar um subconjunto de variáveis que forneceram uma previsão precisa da taxa de pobreza.

Já o artigo de McBride e Nichols (2018) propõe o aprimoramento das ferramentas de PMT com o uso de métodos de *machine learning*. Novamente, os autores tomam como base as ferramentas empregadas pela USAID e mostram que a aplicação de técnicas de validação cruzada e métodos de *stochastic ensemble* melhoram substancialmente o desempenho preditivo das ferramentas de PMT.

Além da nascente literatura, é possível verificar que a aplicação de modelos de *machine learning* vem ganhando notoriedade prática. Exemplo que ilustra bem a atenção crescente que os modelos de *machine learning* estão recebendo nas aplicações específicas para o problema de predição de pobreza se refere a uma competição de construção de modelos preditivos de pobreza lançada em 2018 pelo Banco Mundial. A competição foi organizada na plataforma DrivenData e envolvia a previsão de pobreza com dados de pesquisas por amostras de domicílios do Malawi e da Indonésia (FITZPATRICK et al.,2018) [3].

3. ASPECTOS METODOLÓGICOS E DADOS

3.1. Dados

O conjunto de variáveis preditivas (atributos) utilizado no exercício aqui realizado é limitado pelo conjunto de variáveis presentes no questionário da PNADC. Dado o conjunto de variáveis da PNADC foram definidas a variável de interesse que represente a situação de pobreza e variáveis relacionadas a características do chefe de domicílio, além de características demográficas e

[3] https://www.drivendata.co/

estruturais dos domicílios. Após a obtenção, os dados foram submetidos a um processo de organização e limpeza.

A inspiração para a construção da variável de interesse está nos indicadores de pobreza propostos por Foster, Greer e Thorbecke (1984). Assim como essa classe de indicadores de pobreza, para cada indivíduo/domicílio indexado por i utiliza-se informações de renda e número de residentes no domicílio para calcular a renda domiciliar per capita, representada por r_i. Caso r_i seja inferior a linha de pobreza, dada por z, o domicílio e seus moradores são classificados como pobres; caso r_i seja igual ou maior a z, eles são classificados como não sendo pobres. Isso é representado por meio da seguinte função indicadora:

$$y_i(r_i, z) = \begin{cases} 1 & se\ r_i < z \\ 0 & se\ r_i \geq z \end{cases} \quad [2]$$

A linha de pobreza adotada é baseada no valor estabelecido pelo Banco Mundial para países de renda média-baixa, como o Brasil, o qual preconiza como pobreza as famílias com renda domiciliar per capita inferior a US$ US$ 5,50 por dia em termos de paridade de poder de compra (PPC).

No que diz respeito aos atributos, houve uma seleção prévia de variáveis com base na literatura e nas possibilidades da base dados. Os dados também são submetidos a um processo de engenharia de atributos (*feature engineering*) em que as variáveis são processadas e transformadas. Variáveis categóricas são convertidas em variáveis *dummies* e variáveis numéricas são padronizadas.

O conjunto inicial de atributos utilizados na análise está apresentado no Quadro 1 a seguir.

Quadro 1. Descrição de variáveis preditoras utilizadas na modelagem (atributos).

Descrição de variáveis
Características do chefe de domicílio/ família
Sexo do chefe de domicílio. Variável *dummy*: masculino (0) ou feminino (1).
Idade do chefe de domicílio (em anos).
Presença de cônjuge do chefe de domicílio. Variável *dummy*: não (0) ou sim (1).
Cor declarada pelo chefe de domicílio. Variável *dummy*: Branco ou amarelo (1) ou negro, pardo ou indígena (0).
Nível educacional mais elevado do chefe do domicílio. Conjunto de *dummies* para os seguintes níveis de escolaridade: sem instrução, fundamental incompleto, fundamental completo, médio incompleto, médio completo, superior incompleto; superior completo.
Situação de ocupação do chefe do domicílio no mercado de trabalho. Conjunto de *dummies* para as seguintes condições: inativo, ativo ocupado, ativo desocupado.
Chefe de domicílio recebe aposentadoria: Variável *dummy*: não (0) ou sim (1).
Características do domicílio/ família
Localização em área urbana ou rural. Variável *dummy*: urbana (0) ou rural (1).
Número de pessoas residentes.
Número de crianças (com idade menor ou igual a 14 anos) residentes.
Número de adultos residentes.
Número de cômodos.
Abastecimento de água adequado. Variável *dummy*: não (0) ou sim (1). Foi considerado adequado o abastecimento de água por rede geral ou poço profundo sendo canalizada para pelo menos um cômodo do domicílio.
Banheiro exclusivo do domicílio. Variável *dummy*: não (0) ou sim (1).
Esgotamento sanitário adequado. Variável *dummy*: não (0) ou sim (1). Foi considerado adequado o esgotamento sanitário com coleta por rede geral em área urbana ou fossa séptica em áreas rurais.

Paredes adequadas. Variável *dummy*: não (0) ou sim (1). Foram consideradas adequadas as paredes externas em alvenaria, de madeira adequada ou de taipa com revestimento.
Tipo da residência: casa ou apartamento. Variável *dummy*: urbana (0) ou rural (1).
Propriedade do domicílio: próprio quitado, próprio não pago, alugado, cedido ou outro. Conjunto de *dummies.*
Posse de aparelho de televisão. Variável *dummy*: não (0) ou sim (1).
Posse de refrigerador. Variável *dummy*: não (0) ou sim (1).
Posse de máquina de lavar roupas. Variável *dummy*: não (0) ou sim (1).
Posse de computador. Variável *dummy*: não (0) ou sim (1).
Posse de telefone celular/smartfone. Foi contabilizada de forma relativa como nº de aparelhos/ residentes.
Acesso à internet. Variável *dummy*: não (0) ou sim (1).
Posse de automóvel. Variável *dummy*: não (0) ou sim (1).
Posse de motocicleta. Variável *dummy*: não (0) ou sim (1).

Fonte: Elaboração própria com base nos dados disponíveis na PNAD Contínua.

3.2. Algoritmo de Machine Learning: XGboost

O *Extreme Gradient Boosting*, ou *XGBoost*, desenvolvido por Chen e Guestrin (2016) é um algoritmo de aprendizado de máquina, baseado em árvore de decisão e que utiliza uma estrutura de "*gradient boosting"*.

Sendo um modelo baseado no conceito de "*ensemble*" ou aprendizado por agrupamento, o algoritmo se baseia na ideia de combinar diversos modelos individuais de predição mais simples (*weak learners*), treiná-los para uma mesma tarefa de classificação, e produzir a partir destes modelos individuais um modelo agrupado mais complexo e robusto (*strong learner*). A ideia é que a combinação desses modelos fracos produzirá um desempenho preditivo melhor do que qualquer modelo básico individual, menos suscetível a problemas de viés e variância.

Considerando a aplicação de modelos de árvores de decisão, que são modelos simples e suscetíveis a alta variância, podemos agregá-los de modo a aumentar sua resistência a variações nos dados. Dessa forma, pode-se treinar várias árvores separadamente de forma a adaptar cada uma a diferentes partes da base de dados. Esse é um dos principais pilares da técnica de Floresta Aleatória/*Random Forests*, um dos modelos mais populares de *ensemble*.

Seguindo a ideia de "*ensemble*", nos algoritmos de *boosting* os modelos "fracos" são treinados de forma sequencial, construídos a partir dos modelos treinados previamente. Isso é diferente dos algoritmos de *bagging* em que os modelos são treinados de forma independente. Com base nos erros de predição dos modelos anteriores, o *boosting* estabelece ponderações maiores para os erros de classificação e de forma iterativa realiza novas predições mais resistente ao viés. Por sua vez, o "*gradient boosting"* é uma técnica de *boosting* que adota a forma de um problema de otimização numérica com o objetivo de minimizar os erros do modelo usando a técnica de gradiente descendente.

Uma descrição formal do *XGBoost*, por mais básica que seja, exige alguma notação. Assim, considere um conjunto de dados individuais indexados por i descrito da seguinte forma $\mathcal{I} \coloneqq \{1, \cdots, \mathrm{I}\}$. Para cada i define-se uma variável de interesse ou resposta y_i e um conjunto de atributos ou preditores $\boldsymbol{x}_i$. Para a análise em questão, temos que $y_i \in \{0,1\}$, sendo $y_i = 1$ se o domicílio/ família é pobre e $y_i = 0$ caso contrário. O objetivo é prever y_i com base em $\boldsymbol{x}_i$ para cada $i \in \mathcal{I}$.

Como apresentado anteriormente, considera-se um modelo simples dado por uma árvore de decisão, f_k, tal que $f_k(\boldsymbol{x}_i)$ é um escore ou peso que a árvore f_k atribui para i, dados os atributos

$\boldsymbol{x}_i$. Se temos um conjunto de K árvores, dado por $\theta = \{f_1, \cdots, f_K\}$, o escore total é dado por $\hat{y}_i \coloneqq \sum_{k=1}^{K} f_k(\boldsymbol{x}_i)$, que constitui uma predição para y_i baseada em $\boldsymbol{x}_i$, usando θ [4].

A principal questão é selecionar cada elemento de θ a partir de um espaço funcional $\mathcal{F}$ de *step functions*. O algoritmo *XGBoost* seleciona árvores de $\mathcal{F}$ com o objetivo de minimizar uma função de perda. O procedimento de otimiação é recursivo, tal que em cada iteração uma nova árvore entra em θ. A função de perda é então definida por:

$$L(\theta) = \sum_{i \in \mathcal{I}} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad [3]$$

, onde l é uma função de perda logística e Ω é um termo de regularização que penaliza a complexidade de cada árvore em θ.

O *XGBoost* ajusta a complexidade de cada $f_k \in \theta$ com o seguinte termo de regularização:

$$\Omega(f_k) = \gamma T_k + \frac{1}{2}\lambda \|\omega_k\|^2$$

, em que $\gamma \in (0, \infty)$ e $\lambda \in (0, 1]$ são parametros de regularização, T_k é o número de nós terminais ou ramos de cada árvore, ω_k é um vetor de escores, um para cada ramo, em f_k. Mais detalhes sobre o algoritmo *XGBoost* podem ser consultados em Chen e Guestrin (2016).

3.3. Algoritmos adicionais utilizados

Uma etapa importante, seja na aplicação do PMT, mas também na aplicação das técnicas de *machine learning*, é a seleção de variáveis. Adicionalmente ao *XGBoost*, proposto no presente exercício, foram empregados os métodos de LASSO e um modelo de Floresta Aleatória/*Random Forest* para definir a importância de atributos e realizar mais uma seleção de variáveis a serem incorporadas na versão final do modelo.

Métodos *stepwise* aplicados em conjunto com estimações de modelos lineares por MQO são bastante aplicados. De forma alternativa, e por um processo mais elaborado, pode-se aplicar o método de *Least Absolute Shrinkage and Selection Operator* (LASSO), apresentado inicialmente por Tibshirani (1996).

A técnica de LASSO incorpora um termo para regularizar para zero os coeficientes de variáveis que fornecem menos correlação com uma variável dependente. Assim, o LASSO permite reduzir o número de variáveis preditoras, de forma a manter apenas um conjunto de variáveis que seja bem correlacionado com a variável dependente. Trabalhos como Belloni e Chernozhukov (2013) e Hastie *et al*. (2015) sugeriram um método de estimação por MQO pós LASSO, propondo uma primeira etapa em que a técnica de LASSO é aplicada para selecionar variáveis. Em uma segunda etapa estima-se um modelo por MQO empregando as variáveis selecionadas por LASSO.

Por sua vez, o algoritmo de Floresta Aleatória, que é por si só um algoritmo de aprendizado supervisionado, permite derivar a importância de cada variável para a previsão. Utilizando medidas de "impureza" como o Gini e entropia é possível treinar uma árvore calculando o quanto cada atributo diminui a impureza. Quanto mais um atributo reduz a impureza, mais importante ele é. Em um algoritmo de Floresta Aleatória, a redução da impureza de cada atributo pode ser calculada tomando um média entre as árvores do agrupamento.

3.4. Avaliação do modelo de preditivo

Após treinar um modelo de *machine learning* é importante testá-lo para definir se o modelo é capaz de generalizar bem para novos dados e cumprir com o seu propósito. Se o modelo é capaz de prever muito bem os dados de treino, mas é ruim ao prever dados de teste, temos um problema de *overfitting*.

[4] Os escores são escalares e no caso de uma probabilidade, considerando que y_i é binário, temos que $\hat{y}_i = \left(1 + exp(-\sum_{k=1}^{K} f_k(\boldsymbol{x}_i))\right)^{-1}$.

Uma outra forma de visualizar a performance de um modelo de classificação é formatar a matriz de confusão (*confusion matrix*) do modelo. A matriz de confusão é uma matriz 2x2, onde as linhas representaram os valores reais e as colunas os valores preditos, assim ela mostra o número de casos em que o nosso modelo acertou ou errou em cada categoria.

Quadro 2. Matriz de Confusão.

		Previsto	
		Não pobre ($y = 0$)	**Pobre ($y = 1$)**
Observado	**Não pobre ($y = 0$)**	*Verdadeiro Negativo (VN)*	*Falso Positivo (FP)*
	Pobre ($y = 1$)	*Falso Negativo (FN)*	*Verdadeiro Positivo (VP)*

Na diagonal principal temos os acertos na forma de **verdadeiros positivos** e **verdadeiros negativos**. Por sua vez, os dois valores fora da diagonal principal nos mostram o número de vezes que o modelo errou em sua previsão são os **falsos positivos** e falsos **negativos**. De forma sistematizada temos:

- Verdadeiros Positivos: classificação correta da classe Positivo;
- Falsos Negativos (Erro Tipo II): erro em que o modelo previu a classe Negativo quando o valor real era classe Positivo;
- Falsos Positivos (Erro Tipo I): erro em que o modelo previu a classe Positivo quando o valor real era classe Negativo;
- Verdadeiros Negativos: classificação correta da classe Negativo.

Por esses valores nós conseguimos calcular a acurácia, precisão, sensibilidade e a medida de F1-score.

A Acurácia indica uma performance geral do modelo. Dentre todas as classificações, quantas o modelo classificou corretamente;

$$acurácia = \frac{VP + VN}{Total}$$

A Precisão é definida como a proporção de predições corretas de uma categoria em relação a todas as previsões feitas dessa categoria.

$$precisão = \frac{VP}{VP + FP}$$

A medida conhecida como Recall ou sensibilidade é definida como a proporção de previsões corretas da categoria alvo, verdadeiros positivos em relação a soma dos verdadeiros positivos com os falsos negativos.

$$recall = \frac{VP}{VP + FN}$$

F1-Score é dado pela média harmônica entre precisão e sensibilidade, sendo uma métrica que representa em um número único a qualidade geral do nosso modelo.

$$F1 = \frac{2 * precisão * recall}{precisão + recall}$$

Outra forma de avaliar o modelo é por meio da curva ROC (do inglês, *Receiver Operating Characteristic Curve*) que é uma representação gráfica que ilustra o desempenho (ou performance) de um sistema classificador binário à medida que o seu limiar de discriminação varia.

A ROC possui dois parâmetros: a medida de sensibilidade, que representa a "taxa de verdadeiro positivo", e a medida de especificidade, que representa a "taxa de falso positivo". Plotada em um plano unitário, a curva ROC resulta da representação gráfica dos índices de sensibilidade e especificidade (na verdade 1- especificidade).

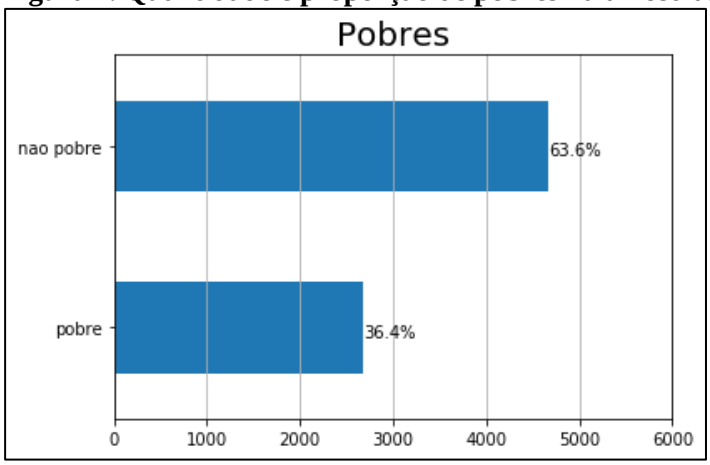
Uma forma de simplificar a análise da curva ROC é por meio da AUC (*area under the ROC curve*), que nada mais é que uma maneira de resumir a curva ROC em um único valor, agregando todos os limiares da ROC, calculando a "área sob a curva".

O valor do AUC varia entre 0 (zero) e 1 e o limiar entre a classe é 0,5. Ou seja, acima desse limite, o algoritmo classifica em uma classe e abaixo na outra classe. Um modelo cujas previsões estão 100% erradas tem uma AUC de 0, enquanto um modelo cujas previsões são 100% corretas tem uma AUC de 1.

4. PROCESSO DE MODELAGEM E RESULTADOS

O processo de *data wrangling* envolve a obtenção, limpeza e tratamento dos dados brutos. A partir dos microdados da PNAD Contínua filtrados para o estado do Ceará, dados individuais forem agregados por domicílios utilizando variáveis que permitem tal identificação. O processo de engenharia de atributos foi iniciado com o cálculo do indicador de pobreza, realizado com a transformação e cálculo de variáveis de renda (rendimentos de diferentes fontes) e renda domiciliar *per capita*.

Figura 1. Quantidade e proporção de pobres na amostra.



Fonte: Elaboração própria. Microdados da PNAD contínua (2019).
Estimação com Python e biblioteca Pandas e Matplotlib.

Algumas características domiciliares também foram transformadas de forma a obter variáveis como abastecimento de água adequado, esgotamento sanitário adequado e paredes adequadas.

A base de dados inicial contou com 7.515 observações e 26 variáveis (uma variável alvo e 25 atributos). Foram identificados valores ausentes na variável de esgotamento sanitário adequado e optou-se por excluir tais observações (linhas). Com isso a base de dados final passou a contar com 7.348 observações. Essa amostra de dados foi particionada de forma que 75% dos dados compuseram o conjunto de treino (5.511 observações) e 25% foram reservados para o conjunto de teste (1.837 observações). A divisão considerou a proporção de pobres da amostra de forma a manter a estratificação nos conjuntos de treino e de teste.

Variáveis categóricas foram codificadas de modo que cada categoria passasse a ser representada por uma variável *dummy*. Por sua vez, variáveis numéricas foram padronizadas de forma que todos os valores estejam em um mesmo intervalo, entre 0 e 1[5].

[5] Existem diferentes procedimentos de reescalonamento dos quais se destacam a normalização e a padronização. A normalização assume que os dados estão normalmente distribuídos e os redimensiona de modo que a distribuição se centre em torno de 0 com um desvio padrão de 1. No entanto, a normalização é bastante sensível a presença de outliers e não pode garantir escalas balanceadas. Por outro lado, a padronização é menos afetada por outliers, mas comprime todos os valores em uma faixa estreita.

4.1. Seleção de atributos

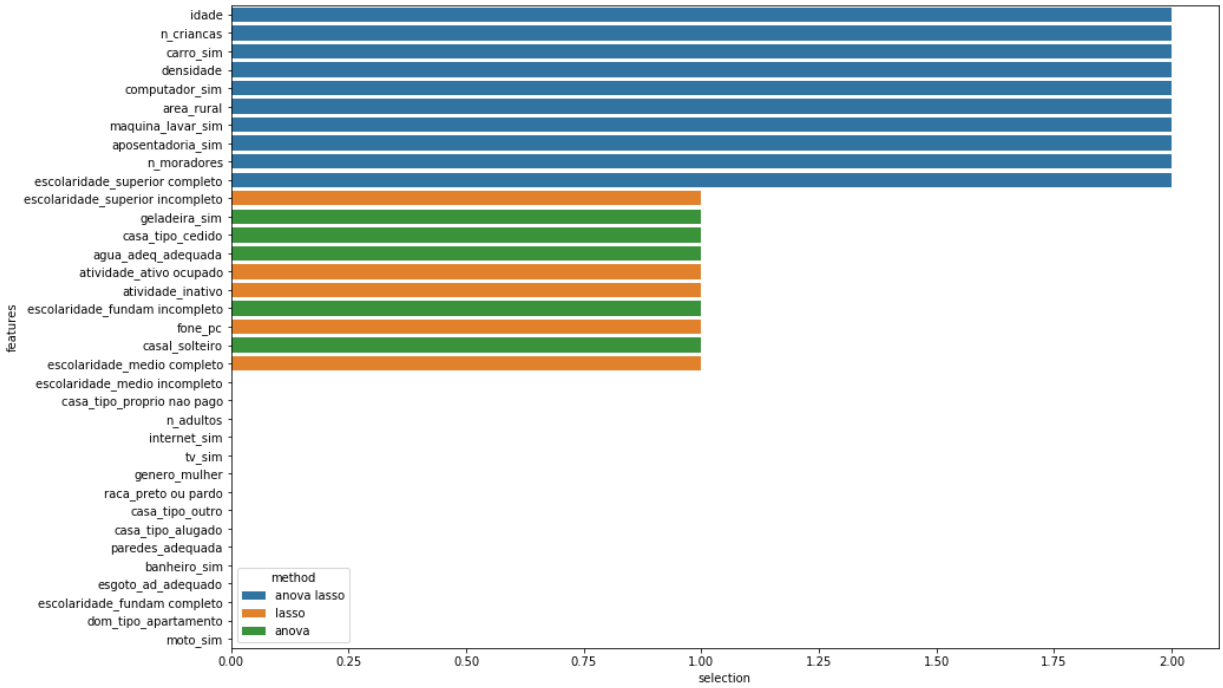
A seleção de atributos compreende o processo de seleção de um subconjunto de variáveis preditoras para construir o modelo de aprendizado de máquina. As vantagens de treinar um modelo em um conjunto mais restrito de variáveis incluem um maior poder preditivo com redução do sobreajuste (*overfitting*) dos dados, além de tornar o modelo mais fácil de interpretar.

Um procedimento ainda bastante corriqueiro é a seleção manual de atributos. De certa forma, o conjunto inicial de dados brutos já contou com uma pré-seleção baseada na disponibilidade de dados e de variáveis apontadas na literatura como bastante correlacionadas com a condição de pobreza em nível individual e domiciliar (citar alguns textos). No entanto, sobre os dados do conjunto inicial, foram aplicados dois métodos para seleção "automatizada" de atributos: a análise de variância e a regularização LASSO.

O método de Análise de Variância (ANOVA) é baseada na ideia de que a variância é uma medida de propagação e pode ser interpretada como uma medida de impacto. Assim, a análise permite identificar as variáveis que explicam grande medida da variância na variável alvo. Já a regularização LASSO penaliza coeficientes de variáveis que apresentam correlação mais fraca com a variável explicada.

Na Figura 2, a seguir, as barras azuis indicam variáveis selecionadas (definindo um parâmetro de seleção de 15 atributos) por ambos os métodos, ANOVA e LASSO, os outros são selecionados por apenas um dos dois métodos, barras laranja para ANOVA e verdes para o LASSO.

Figura 2. Importância de atributos por meio de ANOVA e LASSO.

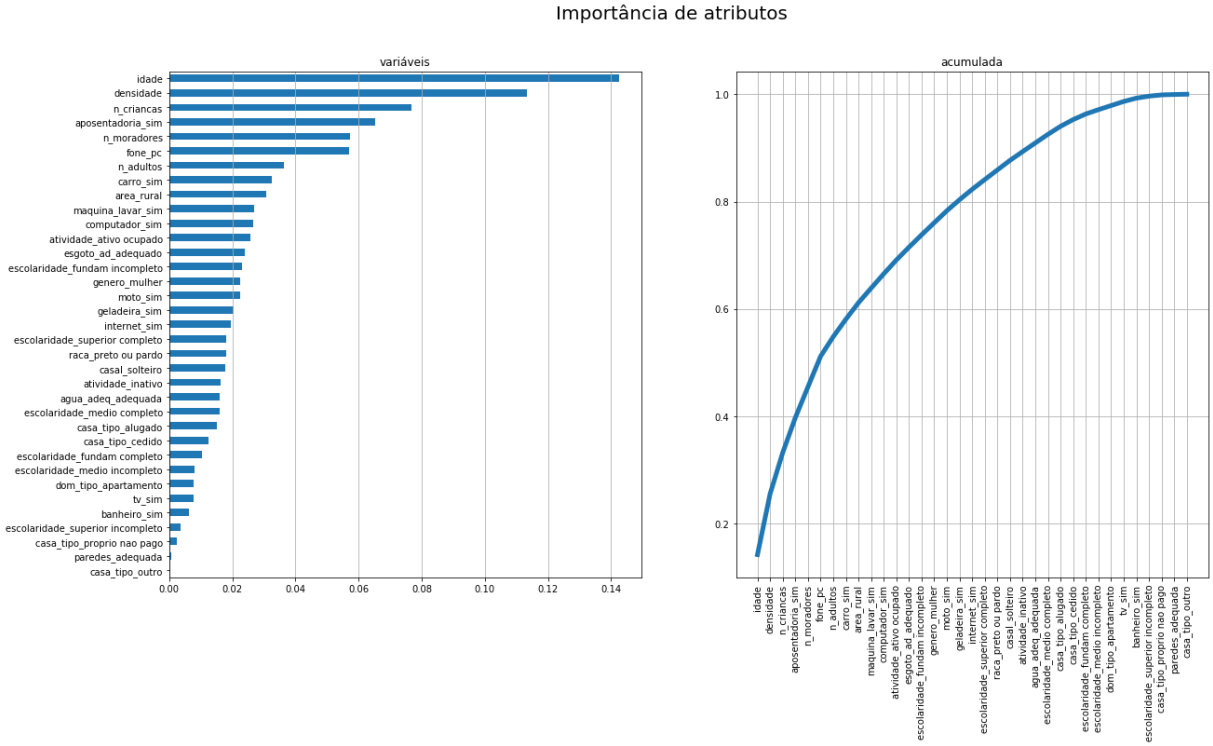


Fonte: Elaboração própria. Microdados da PNAD contínua (2019).
Estimação com Python e biblioteca pandas e matplotlib.

A importância dos recursos é calculada a partir do quanto cada recurso diminui a entropia em uma árvore.

Uma vez que as variáveis consideradas apresentavam distribuições bastante assimétricas optou-se pelo processo de padronização MinMax-Scaler.

Figura 3. Importância de atributos por meio de um modelo de Floresta Aleatória.



Fonte: Elaboração própria. Microdados da PNAD contínua (2019).
Estimação com Python e biblioteca pandas e matplotlib.

4.2. O processo de modelagem e resultados

O modelo construído com o algoritmo *XGBoost* é um classificador que prevê probabilidades de uma família/domicílio, e seus respectivos membros, estarem em situação de pobreza, dependendo das características individuais e do domicílio. A probabilidade prevista de pobreza é denotada por $\hat{y}_i$ para um domicílio i.

Os dados devem nos permitir determinar as melhores configurações para os parâmetros do modelo que geram a previsão mais precisa. Uma boa combinação de parâmetros é necessária para garantir que o modelo possa ser generalizado e evitar o sobreajuste. O controle desses parâmetros é particularmente importante para o XGBoost porque seu ajuste de modelo sequencial permite que as árvores sejam adicionadas continuamente, o que pode levar ao sobreajuste.

Usamos 75% das observações para treinar o modelo XGBoost. Os 25% restantes foram usados como dados de teste para validar o desempenho do modelo mais bem ajustado (a amostra de validação).

Para o ajuste de hiperparâmetros do modelo não há uma regra geral sobre qual procedimento é melhor. Esse ajuste é algo empírico baseado no teste de diferentes combinações possíveis. No presente exercício foi definido um espaço de busca para diferentes hiperparâmetros e foi adotado um método de busca aleatória com um número de iterações fixo associado com um procedimento de validação cruzada. O método utilizado foi baseado no algoritmo *RandomizedSearchCV* da biblioteca *Scikit-Learn* com um número de 100 iterações, validação cruzada com 5 separações e o objetivo de maximizar a medida de acurácia.

O melhor conjunto de hiperparâmetros foi definido com a seguinte configuração:

- subamostra para modelos de arvore individuais = 0,9;
- nº de árvores do *boosting* = 1500;
- nº mínimo de amostras para divisão de nós (internos) = 8;
- nº mínimo de amostras para divisão de ramos = 3;
- nº máximo de atributos para uma divisão = 4;

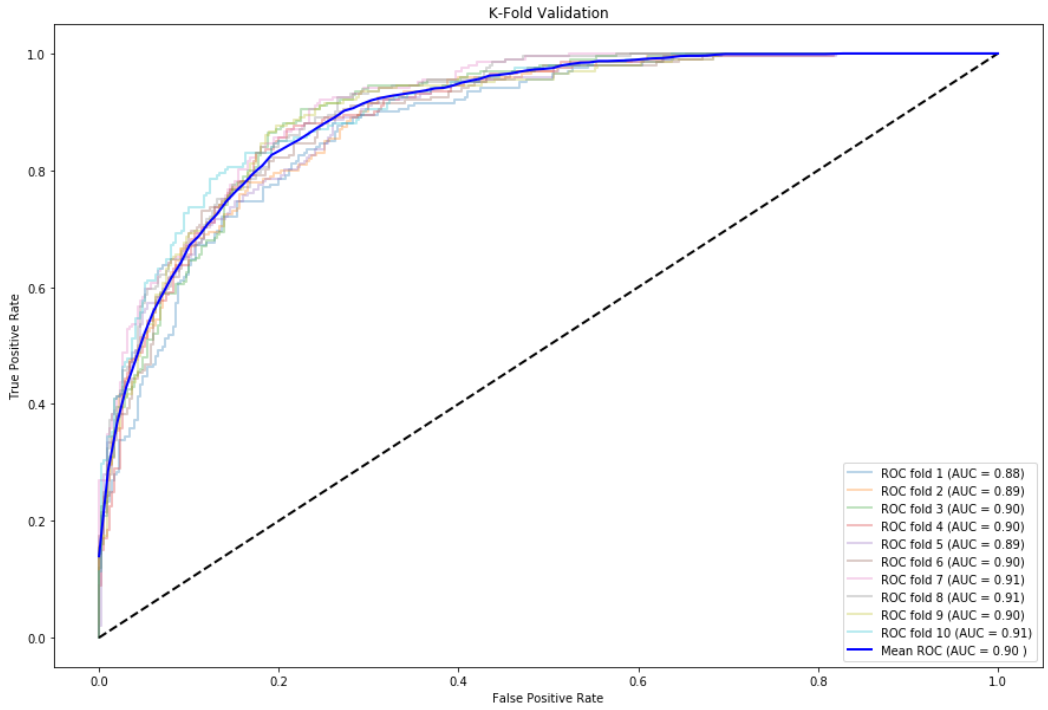
- profundidade máxima da árvore = 2;
- taxa de aprendizagem = 0,1;
- acurácia média no “melhor modelo” de 0,82.

Para a validação do modelo foi realizado um processo de validação cruzado (*k-fold cross-validation*). Trata-se de um procedimento que consiste em dividir os dados de treino k vezes em conjuntos de treinamento e validação, e para cada divisão o modelo é treinado e testado. O procedimento é usado para verificar o quão bem o modelo é capaz de ser treinado e empregado para prever com dados inéditos.

Para cada divisão (*fold*) foi plotada uma curva ROC. Também foi calculado o indicador de área sob a curva ROC (AUC) indicando a probabilidade de que o modelo classifique uma observação escolhida aleatoriamente como positiva ($Y = 1$) mais alta do que uma negativa ($Y = 0$), também escolhida aleatoriamente. As curvas ROC obtidas com o procedimento de validação cruzada com 10 *folds* ($k = 10$) podem ser visualizadas no gráfico da Figura 4.

De acordo com os resultados do processo de validação obteve-se um AUC médio de 0,90. Considerando ser este um bom valor, seguiu-se adiante com o ajuste do modelo preditivo e a aplicação no conjunto de teste.

Figura 4. Curvas ROC.

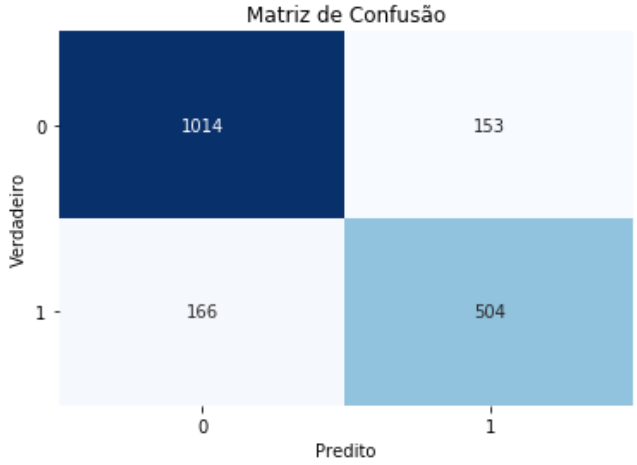


Fonte: Elaboração própria. Microdados da PNAD contínua (2019).
Estimação com Python e biblioteca pandas e matplotlib.

A aplicação do modelo ao conjunto de dados de teste simula uma situação em que o modelo é posto em produção, ou seja, o momento em que ele passaria a receber informações e realizar a classificação, no presente contexto, classificar famílias em situação de pobreza ou não.

Com a aplicação do modelo aos dados de teste, a matriz de confusão obtida está apresentada na Figura 5 a seguir.

Figura 5. Matriz de Confusão.



Fonte: Elaboração própria. Microdados da PNAD contínua (2019).
Estimação com Python e biblioteca pandas, seaborn e matplotlib.

Com base nas informações da matriz podemos obter medidas de desempenho preditivo do modelo. Temos que o modelo previu 657 domicílios pobres, dos quais 504 são predições corretas e 153 são falsos positivos. Isso corresponde a uma precisão de 77%. Por sua vez, considerado os 670 domicílios pobres no conjunto de teste, a medida de sensibilidade (recall) mostra que 78% dos domicílios pobres foram previstos corretamente.

A medida de acurácia total aponta para uma precisão geral do modelo é de cerca de 83%. Erros de exclusão e inclusão ficaram em 14% e 13%, respectivamente. Por sua vez, a AUC de 0,91 indica um modelo com bom desempenho preditivo. Tais medidas são apresentadas na Tabela 1 a seguir.

Tabela 1. Medidas de desempenho.

Acurácia Total	0,83
Precisão	0,77
Sensibilidade (recall)	0,78
Erro de exclusão	0,14
Erro de inclusão	0,13
F1-score	0,76
AUC	0,91

Fonte: Elaboração própria. Microdados da PNAD contínua (2019).
Estimação com Python e biblioteca scikit-learn.

Estes resultados estão bastante compatíveis com os resultados apresentados na literatura, embora diversas ressalvas devam ser consideradas a respeito da diferenciação entre os processos de tratamento de variáveis, fluxo de trabalho e dos próprios algoritmos de predição.

Os resultados obtidos no exercício realizado mostram uma acurácia total acima de 80%. Ainda é um percentual relativamente baixo, se comparado aos vencedores da competição promovida pela DrivenData e apresentados por FITZPATRICK et al. (2018), mas em um nível bastante aceitável para o exercício aqui apresentado. Isso sinaliza a oportunidade de aprimoramentos no modelo, realizando ajustes no fluxo de trabalho e adotando algoritmos mais avançados.

De um ponto de vista mais prático, a substituição da base de dados da PNAD Contínua por dados do CadÚnico é adequada aos propósitos aqui apresentados. Realizando exercício semelhante com a base do CadÚnico é possível realizar simulações e avaliações da focalização de políticas

vigentes, além de verificar a adequação deste tipo de ferramenta de forma mais próxima da realidade e estimar melhor o benefício efetivo de sua aplicação.

5. CONSIDERAÇÕES FINAIS

O objetivo da pesquisa apresentada neste artigo é apresentar a possibilidade de emprego das técnicas de *machine learning*, aliadas ao método de *proxy means test*, no aprimoramento da seleção de beneficiários de políticas de combate à pobreza e políticas sociais direcionadas à população mais vulnerável economicamente.

Em situações de informação imperfeita, em que a renda não é observada ou reportada de forma correta, a proposta de critérios baseados em *proxy mean test* podem prover melhorias na seleção de beneficiários melhorando a focalização de estratégias de combate à pobreza. E como apresentado no exercício apresentado neste relatório, a associação com modelos de *machine lerarning* podem constituir uma alternativa viável para aprimorar os critérios baseados em PMT melhorando o desempenho preditivo destes aos reduzir erros de inclusão e exclusão.

Como abordado na introdução do presente texto, um dos objetivos do exercício realizado e da discussão proposta é o aprimoramento das estratégias de combate à pobreza no Ceará. Iniciativas do Governo do Estado do Ceará podem conduzir para a formatação de um sistema de cadastro de beneficiários de forma a alimentar um sistema inteligente de classificação. Algumas iniciativas em desenvolvimento, como o SABE, que utiliza a base do CadÚnico e cadastros de beneficiários dos programas financiados pelo FECOP, podem cumprir com este papel e representam um passo importante nesse sentido.

Outra questão pertinente para discussões futuras diz respeito à forma como políticas e programas podem, de fato, adotar esse tipo de ferramenta. Discussões sobre a forma como novas tecnologias podem ser incorporadas para melhorar a tomada de decisões em políticas públicas devem ser desenvolvidas considerando diversos aspectos que incluem os pontos de vista econômico, ético, político e até mesmo o operacional.

REFERÊNCIAS BIBLIOGRÁFICAS

ATHEY, S. The impact of machine learning on economics. In: **The economics of artificial intelligence: An agenda**. University of Chicago Press. p. 507-547; 2018.

ATHEY, S; IMBENS, G. W. Machine learning methods that economists should know about. **Annual Review of Economics**, v. 11, p. 685-725, 2019.

BELLONI, A.; CHERNOZHUKOV, V.. Least squares after model selection in high-dimensional sparse models. **Bernoulli**, v. 19, n. 2, p. 521-547, 2013.

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

BROWN, C.; RAVALLION, M.; VAN DE WALLE, D.. **A poor means test? Econometric targeting in Africa**. The World Bank, 2016.

CHEN, T.; GUESTRIN, C.. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. p. 785-794, 2016.

CEARÁ. Secretaria do Planejamento. Relatórios de Gestão do Fundo Estadual de Combate à Pobreza. Fortaleza, 2020.

FITZPATRICK, C.; BULL, P.; DUPRIEZ, O.. **Machine Learning for Poverty Predictions: A comparative Assessment of Classification Algorithms**. A project of the World Bank Knowledge for Change (KCP) Program, 2018.

FOSTER, J.; GREER, J.; THORBECKE, E. A class of decomposable poverty measures. **Econometrica**, v. 52, n. 3, p. 761-768, 1984.

GÉRON, A.. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems**. O'Reilly Media, 2019.

GROSH, M.; BAKER, J. Proxy Means Tests for Targeting Social Programs. **LSMS Working Paper** No. 118. The World Bank,Washington, DC, 2015.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: Data mining, inference, and prediction**. Springer New York, 2013.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning: With applications in R**. Springer New York, 2013.

KAMBUYA, P.. Better Model Selection for Poverty Targeting through Machine Learning: A Case Study in Thailand. **Thailand and The World Economy**, v. 38, n. 1, p. 91-116, 2020.

MCBRIDE, L.; NICHOLS, A.. Improved poverty targeting through machine learning: An application to the USAID Poverty Assessment Tools. **Unpublished manuscript.** Disponível em: http://www. econthatmatters. com/wp-content/uploads/2015/01/improvedtargeting_21jan2015. pdf, 2015.

MCBRIDE, L.; NICHOLS, A.. Retooling poverty targeting using out-of-sample validation and machine learning. **The World Bank Economic Review**, v. 32, n. 3, p. 531-550, 2018.

MEDEIROS, C. N.; OLIVEIRA, J. L.; SILVA, V. H. O. Pesquisa Regional Por Amostra De Domicílios Do Estado Do Ceará (Prad/Ce): Notas Metodológicas. **PRAD Informe N.1**, Instituto de Pesquisa e Estratégia Econômica do Ceará, Fortaleza/CE, 2021.

MULLAINATHAN, S.; SPIESS, J..Machine learning: An applied econometric approach. Journal of Economic Perspectives, 31(2): 87-106, 2017.

SOHNESEN, T. P.; STENDER, N.. Is random forest a superior methodology for predicting poverty? An empirical assessment. **Poverty & Public Policy**, v. 9, n. 1, p. 118-133, 2017.

STORM, H.; BAYLIS, K.; HECKELEI, T.. Machine learning in agricultural and applied economics. **European Review of Agricultural Economics**, v. 47, n. 3, p. 849-892, 2020.

VARIAN, H. R. Big data: New tricks for econometrics. **Journal of Economic Perspectives**, v. 28, n. 2, p. 3-28, 2014.