DOI: http://dx.doi.org/10.20435/multi.v28i69.4104
Recebido em: 18/05/2023; aprovado para publicação em: 28/06/2023

Um estudo econométrico e de Machine Learning sobre indivíduos que se tornaram pobres na pandemia a partir da PNAD-Contínua

An econometric and Machine Learning study on individuals who became poor during the pandemic based on the Continuous PNAD

Un estudio econométrico y de Machine Learning sobre personas que se empobrecieron durante la pandemia basado en la PNAD-Continua

Roberto Santolin[1]
Patrick Gomes de Oliveira[2]

[1]Doutor em Economia pelo Centro de Desenvolvimento e Planejamento Regional da Universidade Federal de Minas Gerais (CEDEPLAR/UFMG). Professor associado da Universidade Federal Rural do Rio de Janeiro (UFRRJ), *campus* Três Rios. Professor Permanente do Programa de Pós-Graduação em Economia Aplicada da Universidade Federal de Outro Preto (PPEA/UFOP). **E-mail:** rsantolin@ufrrj.br, **Orcid:** https://orcid.org/0000-0002-4997-9091

[2]Bacharel em Ciências Econômicas pela Universidade Federal Rural do Rio de Janeiro (UFRRJ), *campus* de Seropédica. **E-mail:** patrickufrrj@gmail.com, **Orcid:** https://orcid.org/0009-0005-7273-6732

Resumo: Este trabalho visa investigar a relação entre pobreza e a pandemia da covid-19, a partir de microdados da PNAD-Contínua. Para obter diferentes abordagens sobre o tema, foram utilizadas duas metodologias: 1) Econometria e 2) Aprendizado de Máquina (*Machine Learning*). O estudo tem como foco entender os principais determinantes da pobreza no período da pandemia, bem como prever a vulnerabilidade de indivíduos à pobreza utilizando *Machine Learning*. Os resultados obtidos apontam para uma maior chance de passagem para a pobreza em indivíduos não brancos, mulheres, moradores da região metropolitana, indivíduos em famílias maiores e com menor grau de instrução. Além disso, o algoritmo *XGBoost* obteve o melhor desempenho na previsão da pobreza após o balanceamento dos dados. Estes resultados podem ser utilizados para auxiliar na tomada de decisões no combate à pobreza no Brasil.

Palavras-chave: Econometria; *Machine Learning*; pobreza.

Abstract: This study aim to investigate the relationship between poverty and the COVID-19 pandemic, based on microdata from Continuous PNAD. To obtain different approaches to the topic, two methodologies were used: 1) Econometrics and 2) Machine Learning. The study focuses on understanding the main determinants of poverty during the pandemic period, as well as predicting the vulnerability of individuals to poverty using Machine Learning. The results obtained indicate a higher likelihood of transitioning into poverty for non-white individuals, women, residents of metropolitan areas, individuals in larger families, and those with lower educational attainment. Furthermore, the XGBoost algorithm performed best in predicting poverty after data balancing. These results can be used to assist in decision-making in combating poverty in Brazil.

Keywords: Econometrics; Machine Learning; poverty.

Resumen: Este estudio tiene como objetivo investigar la relación entre la pobreza y la pandemia de COVID-19, utilizando microdatos de la PNAD-Continua. Para obtener enfoques diferentes sobre el tema, se utilizaron dos metodologías: 1) Econometría y 2) Aprendizaje Automático (*Machine Learning*). El estudio se centra en comprender los principales determinantes de la pobreza durante el período de la pandemia, así como en predecir la vulnerabilidad de las personas a la pobreza utilizando el Aprendizaje Automático. Los resultados obtenidos señalan una mayor probabilidad de caer en la pobreza en personas no blancas, mujeres, residentes de áreas metropolitanas, personas en familias numerosas y con menor nivel educativo. Además, el algoritmo *XGBoost* obtuvo el mejor rendimiento en la predicción de la pobreza después del equilibrio de los datos. Estos resultados pueden ser utilizados para ayudar en la toma de decisiones en la lucha contra la pobreza en Brasil.

Palabras clave: Econometría; *Machine Learning*; pobreza.

1 INTRODUÇÃO

No campo da ciência econômica, o conjunto de métodos convencionais que trabalham com diferentes técnicas de análise de dados, seja por métodos de estimativas com objetivo de previsão, seja para realizar inferências estatísticas, recebe o nome de econometria. A ciência econômica tem evoluído em métodos mais modernos de previsão e extração de padrões em dados. É nesse sentido que surge a interação entre econometria e *Machine Learning*, também chamada de Aprendizado de Máquina. Seguindo essa proposta, este trabalho propõe-se a associar ambas as abordagens, a fim de investigar o fenômeno da pobreza no período da covid-19 entre os anos de 2019 e 2020.

A partir dos microdados da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD-Contínua), o presente estudo se divide em duas abordagens, a fim de obter diferentes óticas sobre o mesmo problema. Na primeira abordagem, a econométrica, o estudo foca em entender quais variáveis influenciaram indivíduos a passarem para a condição de pobreza em 2020, ano 1 do surto da pandemia de covid-19, a partir de um modelo *logit*. Na segunda abordagem, a *Machine Learning*, o estudo foca em utilizar algoritmos de previsão deste crescimento da taxa de pobreza. Mais especificamente, o trabalho analisa indivíduos que não eram pobres em 2019 e tornaram-se pobres a partir de 2020.

O presente trabalho está dividido em cinco seções. Além desta introdução, o item 2 apresenta a revisão de literatura, dividida entre (i) *Machine Learning* na economia aplicada e (ii) estudos sobre os determinantes da pobreza no Brasil. O tópico 3 descreve a metodologia, e o item 4 discute os resultados obtidos. Finalmente, o último item expõe as conclusões do estudo realizado.

2 REFERENCIAL TEÓRICO

2.1 *Machine Learning* na economia aplicada

O termo *Machine Learning,* ou Aprendizado de Máquina, tem sido usado para identificar diferentes técnicas pelas quais seja possível implementar

algoritmos que aprendam a realizar previsões sobre um dado conjunto de dados. Mais especificamente, o uso do termo aprendizado refere-se à programação computacional, em geral, baseada em métodos estatísticos, em que um percentual da amostra de dados, *a priori,* é fornecido ao algoritmo adotado. Este algoritmo aprende diferentes regras de associação para a previsão de uma ou mais variáveis de interesse.

A posteriori, um novo conjunto de dados, a partir da mesma amostra, é fornecido, e o algoritmo utilizado aplica a mesma regra de associação aprendida *a priori* nestes novos dados. Observa-se o percentual de acertos do algoritmo utilizado. Considera-se desejável o algoritmo que possui maior capacidade de previsão dos dados *a posteriori*. Entre os diversos algoritmos usados, há, por exemplo, Redes Neurais, Método de Bayes, Regressão Logística, Regressão Linear, *Support Vector Machine* (SVM), *K-Means*, *K-Nearest Neighbor* (KNN) ou mesmo diferentes combinações entre estes próprios algoritmos.

É válido mencionar a diferença entre econometria e *Machine Learning*, uma vez que as duas abordagens tratam os dados e modelos de maneiras diferentes. De acordo com Varian (2014), a abordagem tradicional em econometria tem o foco em especificar parâmetros de um modelo estatístico que descreve a relação entre um conjunto de variáveis. Grande parte da econometria aplicada se preocupa em detectar e resumir as relações nos dados. A ferramenta mais comum utilizada para o resumo é a análise de regressão linear. *Machine Learning*, por sua vez, oferece um conjunto de ferramentas que podem associar de forma útil vários tipos de relações não lineares nos dados, que, por vezes, não são possíveis de serem analisadas a partir de regressões lineares, usuais na econometria.

Segundo Wu *et al.* (2007), na literatura de *Machine Learning,* o foco é tipicamente no desenvolvimento de algoritmos. Athey (2019) aponta que o objetivo dos algoritmos é especificamente fazer previsões sobre algumas variáveis dadas ou classificar unidades com base em informações limitadas.

Uma das primeiras ideias de aplicações de *Machine Learning* em problemas econômicos foi proposto por Lee e Lee (1974), que descreveram um modelo para analisar o comportamento de um neurônio artificial denominado neurônio *fuzzy*. Os autores argumentaram que sistemas complexos

(como um sistema econômico, por exemplo) podem se beneficiar desta técnica, uma vez que o neurônio *fuzzy* se adapta às imprecisões resultantes do alto grau de complexidade do sistema analisado. Nesse trabalho, os autores desenvolvem os aspectos teóricos do neurônio *fuzzy* por meio de aplicações para o reconhecimento de linguagens.

Wang *et al.* (2019) utilizaram a pontuação de crédito baseada em algoritmos de Inteligência Artificial (IA), como redes neurais recorrentes, Floresta Aleatória e *XGBoost*. Os autores investigaram se os modelos de IA melhoram a inclusão financeira a partir de três métricas: taxa de aprovação, taxa de inadimplência e taxa de rejeição falsa. A partir de dados obtidos de credores on-line, os autores descobriram que os modelos de *credit scoring* baseados em IA aumentaram a taxa de aprovação e reduziram a taxa de inadimplência simultaneamente, o que resultou na inclusão financeira e na possibilidade de fornecer crédito a populações anteriormente com restrições de acesso a empréstimos.

Parray *et al.* (2020) utilizaram modelos de *Machine Learning* para prever as ações da NIFTY 50 da Bolsa Nacional Indiana com dados de ações de 1º de janeiro de 2013 a 31 de dezembro de 2018. Os modelos utilizados foram *Support Vector Machine* (SVM), *perceptron* e Regressão Logística. Como resultado, os autores observaram, a partir do método SVM, uma precisão de 89,93%; no método *perceptron,* 76,68%; e, na regressão logística, 89,93%.

Caruana e Niculescu-Mizil (2004) comparam o desempenho de sete métodos de *Machine Learning*: *SVMs*, redes neurais, árvores de decisão, k-vizinho mais próximo, *bagged trees, boosted trees* e *boosted stumps*. O trabalho tem um enfoque mais computacional e menos explicativo sobre as relações entre as variáveis e os problemas relacionados aos conjuntos de dados selecionados, sendo o foco do trabalho utilizar a métrica Área sob a Curva (*AUC, Area under the Curve*), que calcula uma taxa de verdadeiros positivos por falsos positivos classificados pelos algoritmos. Com 4.000 dados de treino e 35.222 dados de teste, as variáveis preditoras foram idade, classe de trabalho, educação, anos de educação, estado civil, ocupação, parentesco, raça, gênero, ganhos de capital, perda de capital, horas de trabalho por semana e país de origem. A variável-alvo foi se o indivíduo recebe mais do que 50 mil dólares anuais (1) ou menos do que 50 mil dólares (0) anuais.

Como resultado, os autores obtiveram um desempenho AUC superior nos modelos *Boosted Decision Tree* (0.8902), *Bagged Decision Tree* (0.9057), *Support Vector Machine* (0.8980) e *Artificial Neural Network* (0.8980).

Chakrabarty e Biswas (2018) utilizaram técnicas de M*achine Learning* e mineração de dados para analisar o problema de desigualdade de renda. Como base de dados, utilizaram a *UCI-Adult Dataset*[1] *como exemplo. O modelo classificador XGBoost* foi utilizado para prever se a renda anual de uma pessoa nos EUA é superior a 50 mil dólares ou menos. Como métricas de avaliação, os autores utilizaram a acurácia e a curva *Receiver Operating Characteristic* (ROC). Os resultados sugerem uma curva ROC acima de 0,90 e a acurácia de 88.16% para esse algoritmo.

Topiwalla (2013) também utiliza a base de dados *UCI-Adult Dataset* para avaliar problemas de classificação a partir dos algoritmos *Decision Tree, Naïve Bayes, KNN* e *Support Vector Machine,* além de algoritmos mais complexos, como *XGBoost*, *Random Forest* e modelos com *stacking,* assim como a técnica *KFold,* nos modelos baseados em árvores.

Os estudos apresentados nesta seção, embora possuam diferentes aplicações à economia, em geral, indicam o mesmo direcionamento em termos da abordagem utilizada: (i) escolha do algoritmo apropriado ao objeto de estudo; e (ii) avaliação da capacidade preditiva deste algoritmo. No próximo tópico, são apresentados modelos de visam analisar os principais determinantes da pobreza para o Brasil. Estes estudos servem como suporte para as escolhas das variáveis que compõem a análise de *Machine Learning* para os determinantes da pobreza no Brasil, que será realizada na seção de resultados.

2.2 Estudos sobre os determinantes da pobreza no Brasil

Diversos estudos têm se dedicado a avaliar quais são as principais variáveis que determinam a pobreza para o Brasil. A partir da revisão de literatura realizada no presente trabalho, a Tabela 1 resume a presença de variáveis independentes que têm sido frequentemente usadas para se avaliar

[1] Dados disponíveis em: https://archive.ics.uci.edu/ml/datasets/adult. Acesso em: maio 2023.

os condicionantes da pobreza no contexto do mercado de trabalho brasileiro.

Oliveira e Raiher (2021) utilizaram o modelo *logit* para analisar a inserção dos jovens brasileiros no mercado de trabalho, especialmente indivíduos pobres, avaliando qual o impacto do Programa Bolsa Família (PBF). Os dados foram obtidos da PNAD de 2014 e 2015. Os resultados do modelo *logit* calcularam a probabilidade de o indivíduo participar do Programa Bolsa Família (PBF) nos anos analisados. Os autores sugerem que a principal dificuldade para mensurar os efeitos do PBF nos diferentes grupos estudados foi em distinguir os atributos individuais (idade, raça, escolaridade, experiência profissional, renda *per capita*, gênero, número de dependentes), o que pode influenciar a variável de interesse, qual seja, a inserção do jovem no mercado formal. Como resultado, o modelo *logit* obteve um *Pseudo*-R2 de 0,37, para 2014, e 0,28, para 2015, com resultados nos coeficientes estimados estatisticamente significantes a 5%.

Tabela 1 – Artigos da revisão de literatura sobre determinantes da pobreza no Brasil

Presença das variáveis	Presente Trabalho	Ribeiro e Santolin (2021)	Scalon *et al.* (2021)	Gonçalves e Machado (2015)	Marinho e Mendes (2013)	Masri *et al.* (2021)
Educação	Sim	Sim	Sim	Sim	Sim	Sim
Raça	Sim	Sim	Sim	Sim	Sim	Sim
Gênero	Sim	Sim	Sim	Não	Sim	Sim
Urbano x rural	Sim	Sim	Sim	Não	Sim	Não
Região metropolitana	Sim	Sim	Não	Sim	Não	Não
Número de pessoas no domicílio	Sim	Não	Não	Sim	Não	Não
Ramo	Sim	Sim	Não	Não	Não	Não
Idade	Sim	Sim	Sim	Sim	Sim	Sim
Determinantes	Pobreza (ser pobre)	Pobreza (ser pobre)	Pobreza (ser pobre)	Pobreza (ser pobre)	Mercado de trabalho (estar empregado)	Mercado de trabalho (estar empregado)

Fonte: Elaboração própria.

Ribeiro e Santolin (2021) avaliaram o crescimento pró-pobre a partir de características individuais do mercado de trabalho e variáveis agregadas dos estados brasileiros. Os dados utilizados foram os microdados da PNAD, e dados agregados para Unidades Federativas (UFs), de 2004 a 2014, do Instituto de Pesquisa Econômica Aplicada (IPEA). A partir de um modelo *logit*, os autores estimaram os efeitos da educação, escolaridade, raça, gênero, trabalho informal, localidade, PIB, Índice de Gini, Benefícios do Programa Bolsa Família e do Benefício de Progressão Continuada sobre a pobreza. Como resultado, obtiveram um *Pseudo*-R2 de 29,23%, para indivíduos na condição de extrema pobreza[2].

Ribeiro e Marinho (2012) estudaram a alocação de horas trabalhadas para adultos e crianças no nível individual, medindo a pobreza de tempo para o Brasil com dados da PNAD de 2009. Os autores selecionaram as seguintes variáveis: renda mensal de todo o trabalho realizado para pessoas com mais ou 10 anos; número de pessoas residentes no domicílio (incluindo pessoas que residem lá como inquilinos, funcionários da casa e seus parentes); anos de educação; idade dos ocupantes na data de referência. Os autores concluem que o perfil de pobre de tempo é a mulher adulta, de cor negra, com baixa escolaridade e residente na área urbana da Região Nordeste, morando em domicílio com poucas pessoas e mãe de filhos com menos de 14 anos.

Scalon *et al.* (2021), com dados da PNAD de 2001, 2008 e 2015, desenvolveram um estudo para analisar a distribuição de renda e as características de grupos, sendo a renda como proporções da renda familiar média *per capita*, escolaridade, raça, localidade, zona urbana ou zona rural e gênero. A partir de um modelo *logit* multinomial, os autores avaliaram os efeitos dessas variáveis. Eles estimaram os coeficientes e a razão de chances para verificar a influência das variáveis socioeconômicas na classe social que se encontravam (extrema pobreza, vulneráveis, classe média-baixa, classe média-alta, ricos) ao longo dos anos. Como resultado, as variáveis com maiores efeitos foram a raça (indivíduos pretos e brancos) e regiões (Nordeste e Sudeste) para os grupos de indivíduos extremamente pobres e vulneráveis.

[2] A variável de extrema pobreza assumia valor "1" se a renda familiar mensal *per capita* fosse menor que R$ 138 mês, e "0", caso contrário.

Gonçalves e Machado (2015) utilizaram os microdados da Pesquisa Mensal de Emprego (PME) para o período 2002-2011 nas seis regiões metropolitanas brasileiras abrangidas pela PME. Estimando um modelo multinomial e calculando os *odds ratio*, os autores investigaram quais características da família se relacionaram com uma maior ou menor chance de pertencer a cada uma das categorias de pobreza: *Always Poor, Usually Poor, Churning or Occasionally Poor, and Never Poor*. Como resultado, obtiveram que o tamanho da família, a proporção de crianças e o desemprego entre familiares aumentam a probabilidade de pertencerem às classes pobres. Pessoas residentes nas regiões metropolitanas do Nordeste também têm uma alta probabilidade de pertencerem a essas classes. Em contrapartida, famílias cujos membros concluíram o ensino médio e/ou superior têm menos chances de pertencerem às classes pobres.

Masri, Flamini e Toscani (2021), a partir de dados da PNAD de 2012 a 2020, estudaram o impacto de curto prazo da pandemia da covid-19 no mercado de trabalho brasileiro. Os autores selecionaram variáveis como gênero, raça, área, idade, educação formal e situação de emprego. Os resultados apontaram que a pobreza e o índice de Gini teriam aumentado acentuadamente como consequência, para cerca de 13,9% e 0,58, respectivamente, em comparação com cerca de 4,7% e 0,53 pré-covid, em maio de 2020. Os autores observaram que a política de transferência de renda realizada amorteceu a crise nos meios de subsistência dos brasileiros, reduzindo a pobreza e a desigualdade de renda para 4,4% e 0,51, respectivamente, abaixo dos níveis pré-covid.

3 METODOLOGIA

3.1 Análise econométrica

Na etapa 1, o objetivo é entender como as variáveis se relacionam com a pobreza na pandemia. Para isso, será estimado um modelo *logit*, obtendo seus coeficientes e calculando o antilogaritmo e o inverso do antilogaritmo para obter as interpretações dos coeficientes. Pode-se definir o modelo a partir da equação:

$$P(pobre) = \frac{1}{1 + e^{-(b0+b1npessdom+b2genero+b3idade+b4regiao+b5raça+b6ramo+b7instrução+u)}}$$

Em que *P*(*pobre*) é o cálculo da probabilidade que capta os indivíduos em domicílios que ficaram abaixo da linha da pobreza no período da pandemia em 2020, mas que não eram pobres em 2019; *npessdom* é o número de pessoas em cada domicílio; *gênero*, se masculino ou feminino; *idade* se refere à idade da pessoa na amostra; *região* é o local onde a pessoa reside; *raça*, a etnia da pessoa; *ramo* é o ramo de atividade que a pessoa trabalha; *instrução* são os anos de estudo; e *u* é o erro idiossincrático do modelo. A estimação será realizada por meio do modelo *logit*.

3.2 Análise em *Machine Learning*

Na etapa 2, o objetivo é, a partir do conjunto de dados, prever a que classe o indivíduo pertence, com foco na classe minoritária: não pobre. Para isso, será seguida uma série de passos sugeridos por Fayyad, Piatetsky-Shapiro e Smyth (1996), a qual eles nomeiam como *Knowledge Discovery in Databases* (KDD). Esse método é dividido em: seleção dos dados, pré-processamento, transformação, mineração, interpretação e geração de conhecimento.

Como avaliação do resultado do estudo, o primeiro passo foi a execução automatizada de diversos algoritmos discutidos na revisão de literatura sobre *Machine Learning* na economia aplicada, e então os algoritmos foram comparados com métricas de ajustamento.

Particularmente, os resultados que apresentaram os melhores ajustes foram aqueles relativos ao método *XGBoost*. O *XGBoost* é um algoritmo de *Machine Learning* que utiliza árvores de decisão para prever valores de uma variável alvo. Esse algoritmo adiciona gradualmente novas árvores de decisão ao modelo para melhorar a precisão. O modelo utiliza gradientes para minimizar a função objetivo do modelo. Esse processo é repetido até que o modelo atinja um ponto de convergência (onde não é possível melhorar o ajuste do modelo) ou um número máximo de árvores seja alcançado. Ao final, o *XGBoost* retorna um modelo que combina todas as árvores construídas durante o treinamento e previsões para novos exemplos. O modelo também utiliza técnicas de regularização para evitar o sobreajuste.

A função objetivo que o algoritmo *XGBoost* minimiza é composta por dois termos: a função de perda e a função de regularização. Expandindo essas funções:

$$l^{(t)} = \sum_{i=1}^{n} l\left(y_i, y_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

(1)

Em que $l^{(t)}$ é a função objetivo no estágio t; n é o número de exemplos n conjunto de treinamento; y_i é o valor verdadeiro da variável alvo para o i-ésimo exemplo; $y_i^{(t-1)}$ é o valor previsto pela árvore de decisão adicionada ao modelo no estágio $t-1$ para o i-ésimo exemplo; $l\left(y_i, y_i^{(t-1)} + f_t(xi)\right)$ é a função de perda que mede o erro entre a previsão do modelo e o valor verdadeiro da variável alvo para o i-ésimo exemplo; t é o número máximo de árvores a serem adicionadas ao modelo; e $\Omega(f_i)$ é a função de regularização da j-ésima árvore adicionada ao modelo.

A função de perda pode ser qualquer função diferenciável, dependendo do tipo de problema que está sendo resolvido (classificação, regressão e clusterização). No presente trabalho, o problema é de classificação.

A função de regularização ômega é usada para evitar o sobreajuste e pode ser escrita como:

$$\Omega(f_j) = \gamma T + \frac{1}{2}\lambda \sum_{i=1}^{T} w_i^2$$

(2)

Em que y e λ são hiperparâmetros que controlam a importância da regularização; T é o número de folhas das árvores; e w_i é o peso associado à i-ésima folha da árvore. Por fim, a árvore de decisão adicionada ao modelo no estágio t é definida por:

$$f_t(x) = w_{q(x)}$$

Em que x é o vetor de características do exemplo a ser previsto; $q(x)$ é a função que mapeia o vetor de características do exemplo para uma

das folhas da árvore; e $w_{q(x)}$ é o peso associado à folha da árvore mapeada pelo exemplo. Dessa forma, é estimado o modelo que minimiza a função objetivo, evitando sobreajuste.

Para avaliar o desempenho geral dos modelos, utilizam-se algumas métricas de problemas de classificação em *Machine Learning*. São elas:

$$precisão = \frac{VerdadeiroPositivo}{VerdadeiroPositivo + FalsoPositivo}$$

(4)

A precisão é uma medida que indica o quão precisas são as previsões do modelo. Em outras palavras, é a proporção de previsões corretas que o modelo faz.

$$recall = \frac{VerdadeiroPositivo}{VerdadeiroPositivo + FalsoNegativo}$$

(5)

O *recall* é uma medida de quão completas são as previsões do modelo. Em outras palavras, é a proporção de exemplos da classe verdadeira que o modelo consegue prever corretamente.

Por sua vez, o *F1-score* é uma métrica que combina a precisão e o *recall* em um único número. Ela é calculada como a média harmônica da precisão e do *recall*.

$$f1 = \frac{2 * precisão * sensibilidade}{precisão + sensibilidade}$$

(6)

A curva *Receiver Operating Characteristic* (ROC) é um gráfico que mostra a relação entre a taxa de verdadeiros positivos (TPR) e a taxa de falsos positivos (FPR) para vários níveis de corte do modelo. É utilizada para avaliar modelos de classificação binária e por ser mais robusta do que a taxa de acerto, ao considerar o *trade*-off entre a TPR e a FPR.

Já a *Area Under the Curve* (AUC) é uma medida de desempenho para avaliar modelos de classificação binária. Ela é calculada como a área sob a curva ROC. A métrica varia de 0 a 1, sendo valores próximos de 1 o

indicativo de um bom desempenho do modelo e valores próximos de 0 um desempenho ruim.

O objetivo é observar os resultados a partir dessas métricas, comparar os algoritmos e verificar qual se saiu melhor para prever quem se tornou pobre durante a pandemia. Isso pode ser importante para entender qual algoritmo é mais adequado ao problema e tomar decisões sobre qual deles deve ser utilizado.

Contudo, um problema encontrado no conjunto de dados foi o desbalanceamento de dados. O desbalanceamento ocorre quando um conjunto de dados apresenta uma distribuição desigual entre as classes. Na presente amostra utilizada, 7,63% das pessoas passaram para a pobreza entre 2019 e 2020. Esse desequilíbrio pode levar a uma série de problemas na aplicação de algoritmos de aprendizado de máquina, uma vez que os modelos são treinados, principalmente, com base na classe majoritária e tendem a ser enviesados em relação a ela. Como consequência, o modelo pode não aprender adequadamente os padrões e as características distintivas da classe minoritária, dificultando a generalização e identificação correta dos indivíduos que se tornaram pobres.

Ao equilibrar as classes, é possível proporcionar uma distribuição mais homogênea dos dados, permitindo que o algoritmo de aprendizado de máquina identifique e aprenda com mais eficácia os padrões associados à classe minoritária. Isso melhora a capacidade do modelo em prever corretamente aqueles que se tornaram pobres, garantindo uma representação mais precisa da realidade e evitando o viés em favor da classe majoritária.

Para tratar o desbalanceamento, foi utilizado o método *Instance Hardness Threshold* (IHT), desenvolvido por Smith, Martinez e Giraud-Carrier (2014). De acordo com Verdikha, *et al.* (2018), esse método identifica dados com uma alta probabilidade de eles serem classificados incorretamente pelo algoritmo de *Machine Learning*. Ao se aplicar o IHT, é reduzida a probabilidade de erro nos resultados, a partir da redução do viés no algoritmo causado pelo desbalanceamento entre as classes "0" e "1". Sem aplicar esse método, o modelo aprenderia mais sobre a classe dominante do que a menos representada, tornando-se viesado. O IHT utiliza a probabilidade de um elemento pertencer à classe que será reduzida. Desse modo, é

selecionado a que possuir maior probabilidade, ou seja, os elementos que parecem menos com a classe minoritária (no caso desse estudo, a *pobre* =1). A função utilizada é dada por:

$$IH_h(\langle x_i, y_i \rangle) = 1 - p(y_i, h)$$

(7)

em que, x_i é o vetor de variáveis utilizadas para prever a variável pobreza, y_i é o rótulo correto (se ficou pobre ou não); *h* representa o modelo de classificação utilizado (no presente trabalho, o *XGBoost*); e *p*(*yi*, *h*) é a probabilidade atribuída pelo classificador *h*.

Na função *p*(*yi*, *h*) foi utilizado o classificador *Random Forest* de que a instância x_i pertença à classe correta y_i. Quanto maior essa probabilidade mais fácil é para o classificador *h* prever a classe correta da instância. Subtraindo essa probabilidade de 1, obtêm-se a dificuldade da instância, ou seja, a incerteza na previsão.

3.3 Fonte de Dados

Os dados foram extraídos da Pesquisa Nacional de Amostra a Domicílios (PNAD) para os anos de 2019 e 2020 com uma amostragem de 20.752 indivíduos. Os períodos de pesquisa foram os quartos trimestres de cada ano para manter a rotatividade do IBGE e selecionar os mesmos indivíduos.

Na estimativa econométrica, no modelo *logit*, da equação (1), as seguinte variáveis foram utilizadas: a variável dependente *pobre* recebe o valor "1" se a renda familiar mensal *per capita* for de até R$ 451 mensais e "0" caso contrário[3].

As variáveis independentes são: *n_pess_dom* significa número de pessoas no domicílio; *mulher* é uma variável *dummy* que indica se o indivíduo é

[3] Este valor considera a metodologia do Banco Mundial, que atribui a situação de pobreza para indivíduos que sobrevivem com menos de US$ 3,3 por dia. Os valores utilizados entre 2019 e 2020 foram deflacionados a partir do IPCA para o ano de 2020. O valor do dólar considerado foi de R$/US$ 4,55, que foi a média da taxa de câmbio entre os anos de 2019 e 2020.

mulher (1) ou não (0); *idade* representa a idade do indivíduo; *rm* indica se o indivíduo mora em região metropolitana (1) ou não (0); *nao_branco* indica a etnia do indivíduo, recebendo 1 para negro, pardo, amarelo e indígenas e recebendo 0 para branco; a variável *urbano* indica se o indivíduo vive em área urbana (1) ou no meio rural (0); a variável *ramo* indica em qual ramo o indivíduo trabalha: Indústria Geral (1), Construção (2), Comércio (3), Reparação de Veículos Automotores e Motocicletas (4), Transporte, Armazenagem e Correio (5), Alojamento e Alimentação (6), Informação, Comunicação e Atividades Financeiras, Imobiliárias, Profissionais e Administrativas (7), Administração Pública, Defesa e Seguridade Social (8), Educação, Saúde Humana e Serviços Sociais (9), Outros Serviços (10), Serviços Domésticos (11), Atividades Mal Definidas (12). A variável *instrução* indica o grau de instrução do indivíduo, sendo "Sem instrução e menos de 1 ano de estudo" (1), "Fundamental incompleto ou equivalente" (2), "Fundamental completo ou equivalente" (3), "Médio incompleto ou equivalente" (4), "Médio completo ou equivalente (5)", "Superior incompleto ou equivalente" (6), "Superior completo" (7).

4 RESULTADOS E DISCUSSÃO

A Tabela 2 mostra os resultados do modelo *logit* para os dados selecionados, que obteve um pseudo-R2 de 0,6277. O resultado do *pseudo-R2* significa que as variáveis do modelo explicam 62,77% do fenômeno analisado (passagem para a pobreza na pandemia). Para uma melhor interpretação dos resultados obtidos, calcula-se o antilogaritmo dos coeficientes, a fim de estimar a razão de chance para cada variável. O cálculo do antilogaritmo é realizado da seguinte forma: se o coeficiente *a* for positivo, utiliza-se *exp(a)* para calcular as razões de chances da ocorrência do evento, dada a ocorrência da variável. Se o coeficiente *a* for negativo, o cálculo da razão de chance é feito a partir da seguinte manipulação: 1/*exp*(*a*).

Tabela 2 – Resultados estimados no modelo *logit*

Variáveis independentes	Coeficientes	Razão de chance
n_pess_dom	0,106***	1,11
mulher	0,194***	1,21
idade	-0,001	1,00
rm	0,185***	1,20
nao_branco	0,385***	1,47
urbano	-0,080	1,08
ramo_1	-3,838***	46,43
ramo_2	-4,031***	56,32
ramo_3	-3,519***	33,75
ramo_4	-3,876***	48,23
ramo_5	-3,892***	49,01
ramo_6	-3,627***	37,60
ramo_7	-4,266***	71,24
ramo_8	-4,674***	107,13
ramo_9	-4,172***	64,85
ramo_10	-3,688***	39,96
ramo_11	-4,030***	56,26
instrucao_1	1,152***	3,16
instrucao_2	1,156***	3,18
instrucao_3	0,964***	2,62
instrucao_4	0,957***	2,60
instrucao_5	0,893***	2,44
instrucao_6	0,796***	2,22

Nota: Valores significativos em p-valores (p): * $p < 0,1$; ** $p < 0,05$; e *** $p < 0,01$

Fonte: Elaboração própria, a partir dos dados da pesquisa.

A partir dos resultados obtidos na Tabela 2, observa-se que o aumento de uma unidade de pessoa no domicílio eleva em 1,11 a chance de o indivíduo dessa família estar abaixo da linha da pobreza durante a pandemia. Esse resultado vai de acordo com o estudo de Gonçalves e Machado (2015), em que encontraram o resultado de que, para o aumento do tamanho da família em uma unidade, essa chance se eleva em 1,60.

Para a variável *raça*, foi encontrado o aumento de 1,47 na chance de ter se tornado pobre durante a pandemia para a população não branca, indo

de acordo com os trabalhos de Ribeiro e Santolin (2021), que apontou 1,32 de chance de indivíduos não brancos se encontrarem na pobreza; Scalon *et al.* (2021), que indicou um aumento de 3,14 na mesma direção; e Gonçalves e Machado (2015), que mostraram uma redução de 1,14 na chance de ser pobre caso o indivíduo seja branco. Esses resultados apontam para a disparidade social entre raças no Brasil, que se manteve durante a pandemia.

Para a variável *gênero*, os achados foram o aumento em 1,21 para mulheres na chance de ter se tornado pobres na pandemia. Os trabalhos da revisão têm resultados similares, sendo 1,06 para Ribeiro e Santolin (2021), e 1,98 para Scalon *et al.* (2021), para o aumento de chances de mulheres estarem na condição de pobreza. Assim como a desigualdade de raças anteriormente vista, a desigualdade de gênero também se manteve na pandemia.

Para a variável *região*, os resultados apontam que morar na região metropolitana aumentou cerca de 1,20 a chance de estar abaixo da linha da pobreza. Para Ribeiro e Santolin (2021), morar na região metropolitana diminui as chances de ser pobre em 1,22.

Nos ramos de trabalho, destaca-se o *ramo_8* (Administração pública, defesa e seguridade social), que possui o mais alto valor de redução de chance de passar para a pobreza durante a pandemia, enquanto o *ramo_3* (Construção) possui o mais baixo valor. Outros ramos de trabalho podem ser verificados na Tabela 3.

Para os níveis de instrução, quanto maior, menor é o aumento da chance de o indivíduo se encontrar abaixo da linha de pobreza. Por exemplo, sair da *instrução_1* (Sem instrução ou menos de 1 ano de estudo) para a *instrução_7* (Ensino superior completo) diminui de 3,16 para 2,22. Isso mostra que, quanto maior o nível de instrução, menor a chance de ter se tornado pobre durante a pandemia. Esses resultados estão conforme os trabalhos já citados, sendo uma redução de 1,15 para cada ano de estudo, segundo Ribeiro e Santolin (2021), e 1,69 para cada ano de estudo, segundo Scalon *et al*. (2021).

Após esta primeira abordagem realizada e constatando que todas as variáveis utilizadas na análise foram importantes para explicar a mudança na condição de pobreza dos indivíduos da amostra, passa-se para análise

Machine Learning, com o objetivo de melhorar a capacidade preditiva do modelo.

A matriz de confusão da regressão logística revela que o modelo foi capaz de identificar corretamente, aproximadamente, 54,6% das pessoas que não se tornaram pobres durante a pandemia e, aproximadamente, 67,2% das pessoas que se tornaram pobres. No entanto, houve um alto número de falsos positivos, representando cerca de 45,4% das previsões da classe *pobre* = 0, o que indica que o modelo erroneamente classificou muitas pessoas como pobres, quando estas não se tornaram. Além disso, houve falsos negativos, correspondendo a aproximadamente 32,8% das previsões da classe *pobre* = 1, ou seja, pessoas que se tornaram pobres, mas o modelo não conseguiu identificá-las. Isso sugere que a regressão logística, nesse caso, pode ter limitações em prever adequadamente as pessoas que se tornam pobres durante a pandemia, possivelmente devido à complexidade do problema ou à necessidade de ajustar melhor os parâmetros do modelo. O alto percentual de falsos positivos e a quantidade moderada de falsos negativos apontam para uma precisão limitada do modelo, o que pode ter implicações negativas na tomada de decisões com base nessas previsões.

Ainda, sem fazer o balanceamento, a Tabela 3 mostra as previsões dos algoritmos utilizados no presente estudo. Como pode ser observado, o algoritmo *XGBoost* foi o mais eficiente na classificação da situação socioeconômica dos indivíduos, obtendo o maior percentual de acerto (26,67%), e um AUC de 0,6221, indicando um desempenho médio na distinção entre as duas classes. O resultado obtido pode ser comparado com o resultado da *Logistic Regression – logit*, que foi o mesmo algoritmo utilizado na análise econométrica do modelo. Embora a *Logistic Regression – logit* tenha obtido um AUC maior, ele não foi capaz de aprender de forma satisfatória sobre a classe "1", em razão do desbalanceamento dos dados.

Tabela 3 – Resultado dos algoritmos de classificação

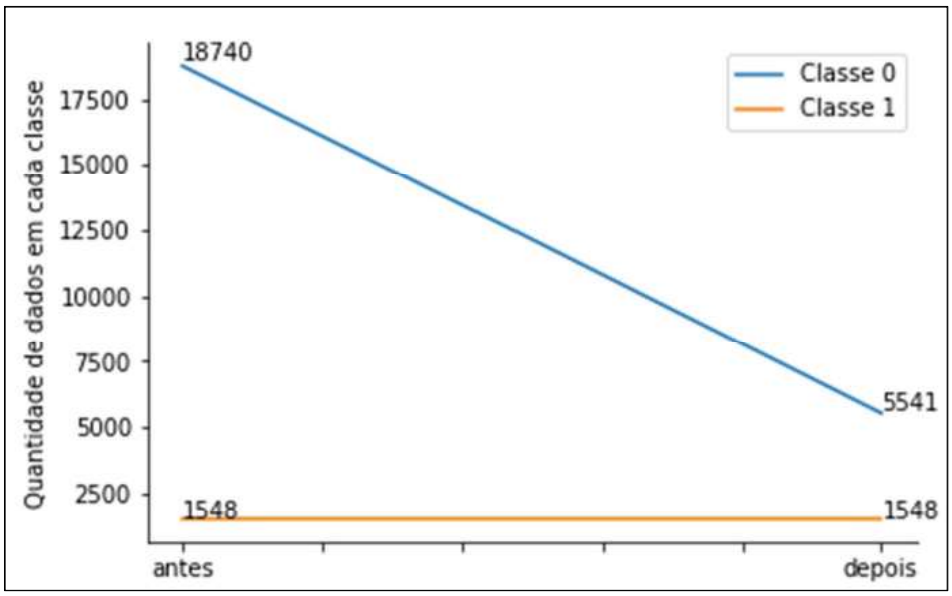
Algoritmos	AUC	Total Previsto Classe 1	% Acertos Classe 1
XGBoost	0,6221	15	26,67%
K Neighbors	0,5358	152	9,87%
Naive Bayes	0,6295	3671	9,59%
Extra Trees	0,5495	299	8,36%
Random Forest	0,5763	182	7,14%
Ada Boost	0,6544	0	0,00%
Linear Discriminant Analysis	0,6522	0	0,00%
Logistic Regression	0,6515	0	0,00%
Light Gradient Boosting Machine	0,6565	1	0,00%

Fonte: elaboração própria, a partir dos dados da pesquisa.

O próximo passo foi a redução do desbalanceamento entre as classes. A Figura 1 representa esse balanceamento de classes ao aplicar o método *Instance Hardness Threshold,* mantendo os valores da classe minoritária (1), isto é, *pobre* = 1, e reduzindo os da classe majoritária (0), isto é, *pobre* = 0. Como dito na seção de metodologia, esse processo serve para reduzir o desbalanceamento das classes, melhorando, assim, a capacidade preditiva do modelo.

Com isso, obteve-se uma redução da classe majoritária de 18740 para 5541 (uma redução de, aproximadamente, 70% da amostra), garantindo que o modelo treinado não fosse enviesado em favor da classe majoritária (*pobre* = 0), o que poderia levar a resultados incorretos. Com a redução da classe majoritária, o algoritmo de *Machine Learning* foi capaz de se concentrar nas características específicas da classe minoritária, aumentando a precisão e a capacidade de generalização do modelo.

Figura 1 – Resultado do balanceamento dos dados utilizando o método *Instance Hardness Threshold*



Fonte: Elaboração própria, a partir dos dados da pesquisa.

A Tabela 4 realiza um comparativo entre o algoritmo *XGBoost* sem e com o desbalanceamento dos dados.

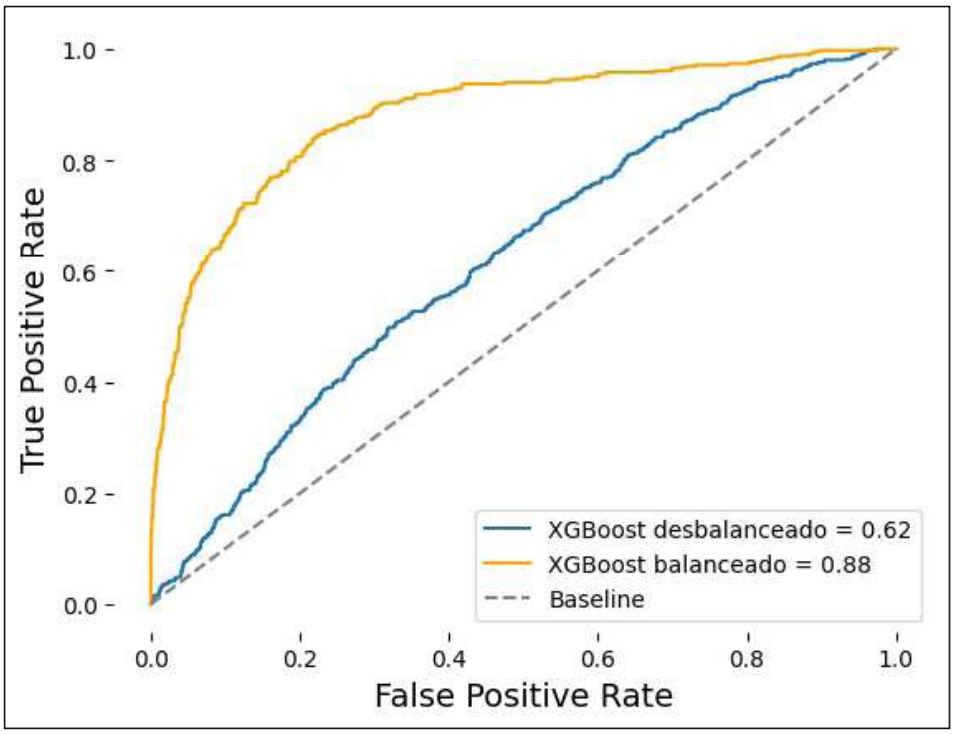
Tabela 4 – Resultado dos modelos de classificação

Algoritmo	AUC
XGBoost sem balanceamento	0,62
XGBoost com balanceamento	0,88

Fonte: Elaboração própria, a partir dos dados da pesquisa.

A Figura 2 apresenta a AUC, como já discutido na metodologia, e calcula o gráfico *True Positive Rate* (TPR) contra *False Positive Rate* (FPR). O TPR é a fração de exemplos da classe positiva que são corretamente classificados como positivos pelo modelo, enquanto o FPR é a fração de exemplos da classe negativa que são incorretamente classificados como positivos.

Figura 2 – Comparativo da curva ROC para o algoritmo aplicado em dados sem balanceamento X com balanceamento dos dados



Fonte: Elaboração própria, a partir dos dados da pesquisa.

De acordo com os resultados das tabelas 4 e 5, é possível observar que o modelo com o balanceamento dos dados apresentou um resultado significativamente melhor da métrica AUC, o que indica que a técnica de balanceamento dos dados afetou positivamente o desempenho do modelo.

O modelo com desbalanceamento apresentou precisão média de 0,72, o que significa que fez previsões corretas 72% das vezes, enquanto o modelo sem balanceamento possui precisão média de 46,5%. Ainda, o modelo com balanceamento tem um *recall* médio de 0,805, o que significa que ele está detectando corretamente 80,5% dos exemplos da classe, e *f1-score* médio de 0,74; por sua vez, o modelo sem balanceamento possui *recall* médio de 0,50 e *f1-score* médio de 0,48. Estes indicadores mostram que o modelo com balanceamento é mais preciso do que o modelo sem balanceamento.

Tabela 5 – Resultado do algoritmo *XGBoost* sem e com balanceamento dos dados

		precisão	*recall*	*f1-score*
XGBoost sem balanceamento	Não pobres	0,93	1,00	0,96
	Pobres	0,00	0,00	0,00
	Média	0,465	0,5	0,48
XGBoost com balanceamento	Não pobres	0,94	0,80	0,86
	Pobres	0,50	0,81	0,62
	Média	0,72	0,805	0,74

Fonte: Elaboração própria, a partir dos dados da pesquisa.

Os resultados aqui obtidos vão de acordo com Chakrabarty e Biswas (2018), que obtiveram AUC acima de 0,90 para um conjunto de dados sobre renda a partir da implementação do algoritmo *XGBoost*. Topiwalla (2013) também obteve para o *XGBoost* os melhores resultados em comparação com os demais algoritmos AUC de 0,9252 para 10 *folds*, 0,9241 para 100 *folds* e 0,9275 para 10 *folds,* com otimização de hiperparâmetros. Chen (2021) obteve, para o mesmo conjunto de dados citado, resultados comparativos entre modelos, sendo o melhor a Floresta Aleatória, seguida do *XGBoost*. Para este último algoritmo, foram obtidas as métricas com precisão de 76,69%, *recall* de 62,31% e *f1-score* de 68,76%.

5 CONCLUSÃO

Este trabalho investigou a relação entre pobreza e a pandemia da covid-19, a partir de microdados da PNAD Covid-19. A análise se inicia definindo os determinantes da pobreza no Brasil a partir da literatura econômica recente e traz a relação entre técnicas econométricas e de *Machine Learning*. Os resultados aqui expostos e comparados com os observados na literatura acadêmica expõem uma fragilidade social que faz com que alguns indivíduos possam se encontrar na pobreza mais facilmente do que outros. Atributos individuais, como diferenças raciais, de gênero e de capital humano, contribuem fortemente para a manutenção da desigualdade de renda e da pobreza no país.

Os resultados estimados com *Machine Learning* apontam para um aumento da eficácia preditiva ao utilizar algoritmos mais robustos que a Regressão Logística, como o *XGBoost*. Essa eficácia aumenta ao utilizar outras técnicas, como o balanceamento de classes.

Ao utilizar algoritmos de M*achine Learning* no conjunto de dados selecionado, ganha-se uma alta capacidade preditiva, como confirmado pelas métricas avaliadas (AUC, precisão, *recall* e *f1-score*). Porém, para modelos mais robustos, como *XGBoost,* a interpretação é reduzida por conta da complexidade do algoritmo. Esse *trade-off* deve ser escolhido com base no enfoque da análise. Em resumo, o debate entre predição e explicação é sobre a escolha entre um modelo que é preciso, mas pode ser difícil de interpretar, ou um modelo que é fácil de interpretar, mas pode ter menor precisão na previsão. A escolha depende das necessidades do problema em questão e da importância da interpretação dos resultados.

REFERÊNCIAS

ATHEY, Susan; IMBENS, Guido. Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, [*s.l.*], v. 11, n. 1, p. 685–725, 2019.

CARUANA, Rich; NICULESCU-MIZIL, Alexandru. An Empirical Evaluation of Supervised Learning for ROC Area. *In*: INTERNATIONAL WORKSHOP, 1., Valencia, 2004. Valencia: ROCAI, 2004.

CHAKRABARTY, Navoneel; BISWAS, Sanket. A Statistical Approach to Adult Census Income Level Prediction. *In*: INTERNATIONAL CONFERENCE ON ADVANCES IN COMPUTING, 1., [*s.l.*], 2018. *Proceedings* [...]. [*s.l.*]: IEEE, 2018.

CHEN, Li-Pang. Supervised Learning for Binary Classification on US Adult Income. *Journal of Modeling and Optimization*, [*s.l.*], v. 13, n. 2, p. 80-91, 2021.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. *AI Magazine*, Washington, v. 17, n. 3, p. 37-54, 1996.

GONÇALVES, Solange Ledi; MACHADO, Ana Flávia. Poverty dynamics in Brazilian metropolitan areas: An analysis based on Hulme and Shepherd's categorization (2002–2011). *EconomiA*, Niterói, v. 16, n. 3, p. 376-94, 2015.

LEE, Samuel; LEE, Edward. Fuzzy Sets and Neural Networks. *Journal of Cybernetics*, [*s.l.*], v. 4, n. 2, p. 83-103, 1974.

MARINHO, Emerson; MENDES, Sérgio. The impact of government income transfers on the Brazilian job market. *Estudos Econômicos*, São Paulo, v. 43, n. 1, p. 29-50, Jan./Mar. 2013

MASRI, Diala; FLAMINI Valentina; TOSCANI, Frederik. The Short-Term Impact of COVID-19 on Labor Markets, Poverty and Inequality in Brazil. *International Moneraty Fund Working Paper* [online], [*s.l.*], 2021.

OLIVEIRA, Gilson de; RAIHER, Augusta Pelinski. The inclusion of poor youth in the Brazilian labour market and the impact of the Bolsa Família programme. *CEPAL Review*, [*s.l.*], n. 135, 2021.

PARRAY, Irfan Ramzan; KHURANA, Surinder Singh; KUMAR, Munish; ALTALBE, Ali. Time series data analysis of stock price movement using machine learning techniques. *Soft Computing*, [*s.l.*], v. 24, p. 16509-517, 2020.

RIBEIRO, Jouse; SANTOLIN, Roberto. An evaluation of the structure of the labour market, assistance policies and sectoral productivity on the pro-poor growth for Brazil from 2004 to 2014: a dynamic panel analysis. *Journal of International Development*, [*s.l.*], v. 33, n. 5, p. 927-44, 2021.

RIBEIRO, Lilian Lopes; MARINHO, Emerson. Time poverty in Brazil: measurement and analysis of its determinants**.** *Estudos Econômicos*, São Paulo, v. 42, n. 2, p. 285–306, 2012.

SCALON, Celi; CAETANO, André Junqueira; CHAVES, Hugo; COSTA, Luana. Back to the past: gains and losses in Brazilian society. *The Journal of Chinese Sociology*, [*s.l.*], v. 8, n. 3, 2021.

SMITH, Michael; MARTINEZ, Tony; GIRAUD-CARRIER, Christophe. An instance level analysis of data complexity. *Machine learning*, [*s.l.*], v. 95, n. 2, p. 225-56, 2014.

TOPIWALLA, Mohammed. *Machine learning on UCI adult data set using various classifier algorithms and scaling up the accuracy using extreme gradient boosting*. 2013. (Dissertation for Big data and Analytics) **-** University of SP Jain School of Global Management, Portland, 2013.

VARIAN, Hal. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, [*s.l.*], v. 28, n. 2, p. 3-28, 2014.

VERDIKHA, Naufal Azmi; ADJI, Teguh Bharata; PERMANASARI, Adhistya Erna. Study of undersampling method: instance hardness threshold with various estimators for hate speech classification. *IJITEE*, Yogyakarta, v. 2, n. 2, p. 39-44, 2018.

WANG, Hongchang; LI, Chunxiao; GU, Bin; MIN, Wei. Does AI-based credit scoring improve financial inclusion? Evidence from online payday lending. *In*: INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS, 40., 2019, Munich. *Proceedings* [...]. ICIS: Munich, 2019.

WU, Xindong; KUMAR, Vipin; QUINLAN, Ross; GHOSH, Joydeep; YANG, Qiang; MOTODA, Hiroshi; MCLACHLAN, Geoffrey; NG, Angus; LIU, Bing; YU, Philip; ZHOU, Zhi-Hua; STEINBACH, Michael; HAND, David; STEINBERG, Dan. Top 10 algorithms in data mining. *Knowledge and Information Systems*, v. 14, n. 1, p. 1-37, 2007.