Previsão de transição no mercado de trabalho com modelos de machine learning e classes desbalanceadas

Vitor Hugo Miro Couto Silva[1]*; Walter Mesquita Filho[2]

[1] Doutor em Economia, professor adjunto do Departamento de Economia Agrícola da Universidade Federal do Ceará. Av. Mister Hull, 2977 – Campus do Pici - Bloco 826; CEP: 60.440-970 Fortaleza, Ceará, Brasil.
[2] Doutor em Entomologia, ESALQ/USP. Orientador MBA data Science & Analytics. Rua Alexandre Herculano, 120 – Vila Monteiro; CEP: 13.418-445 Piracicaba, São Paulo, Brasil
*autor correspondente: vitormiro@gmail.com

Previsão de transição no mercado de trabalho com modelos de machine learning e classes desbalanceadas

Resumo

Considerando o contexto da pandemia de Covid-19, e a recessão econômica dela decorrente, o presente estudo propõe uma aplicação da abordagem de machine learning para construir modelos preditivos de manutenção ou perda de ocupação no mercado de trabalho. Utilizando um conjunto de dados de uma pesquisa por amostras de domicílios brasileiros, foram aplicados os algoritmos de Random Forest e XGBoost, baseados nos modelos de árvores de decisão. Lidando com dados em que as classes previstas se mostram desbalanceadas, tais algoritmos foram treinados em conjunto com a técnica SMOTE. Os modelos treinados apresentaram uma performance razoável, condicionada ao conjunto de dados empregado, ao tratamento destes dados e a abordagem preditiva empregada. Dessa forma, o presente trabalho apresenta discussões iniciais da aplicação de modelos de machine learning para a previsão de transição no mercado de trabalho, se propondo a contribuir com uma ainda insipiente literatura relacionada à aplicação deste tipo de modelagem ao problema.
Palavras-chave: Mercado de trabalho, dados longitudinais, machine learning, classificação.

Predicting labor market transition with machine learning models and imbalanced classes

Abstract

Considering the context of the Covid-19 pandemic, and the economic recession resulting from it, the present study proposes an application of the machine learning approach to build predictive models for the maintenance or loss of occupation in the labor market. Using a dataset from a sample survey of brazilian households, Random Forest and XGBoost algorithms were applied, based on decision tree models. Dealing with data in which the predicted classes are unbalanced, such algorithms were trained together with the SMOTE technique. The trained models presented a reasonable performance, conditioned to the dataset used, the treatment of these data and the predictive approach used. In this way, the present work presents initial discussions of the application of machine learning models for the prediction of transition in the labor market, proposing to contribute to a still incipient literature related to the application of this type of modeling to the problem.
Keywords: labor market, longitudinal data, machine learning, classification.

Introdução

A pandemia de Covid-19 foi um fenômeno sem precedentes. Para enfrentar a emergência de saúde, os países se viram obrigados a paralisar a vida econômica. Com a rápida evolução no número de casos e vítimas fatais da doença, muitos países adotaram uma série de intervenções para reduzir a transmissão do vírus e frear a rápida evolução da pandemia. A adoção de políticas de distanciamento social e restrições de atividades econômicas presenciais ganhou ainda mais força quando a Organização Mundial da Saúde [OMS] decretou o status de pandemia, em 11 de março de 2020 (OMS, 2020). Apesar das medidas de distanciamento social serem extremamente necessárias, seus impactos sobre a economia e, em específico, sobre o mercado de trabalho, foram bastante severos.

No Brasil, logo após o decreto do status de pandemia, e a tendência crescente de casos da doença, estados e municípios passaram a adotar políticas de distanciamento social. No entanto, sem uma coordenação central por parte do governo federal, medidas locais foram introduzidas em momentos diferentes e com diferentes níveis de restrição, como destacado por Moraes (2020).

No mercado de trabalho, os impactos da crise econômica foram heterogêneos. Conforme aponta a literatura que vem surgindo para relatar os impactos da pandemia de Covid-19, grupos mais vulneráveis, como as mulheres, os jovens, as minorias raciais e os indivíduos menos escolarizados foram afetados de forma mais intensa (Fairlie et al., 2020; Barbosa et al., 2020).

Considerando esse cenário no contexto da pandemia de Covid-19 e da recessão econômica decorrente dela, o presente estudo propõe a construção de modelos preditivos para a transição da situação de ocupado/empregado para desocupado/desempregado. As referências para este tipo de estimação são os modelos de transição de Markov que, no presente caso, são aplicados para mensurar a probabilidade de transição entre diferentes estados da força de trabalho[1]. Essa abordagem teve início com os trabalhos seminais de Clark e Summers (1978a e 1978b), Flinn e Heckman (1983) e Jones e Riddell (1999 e 2006).

Esse tipo de análise, considerando dados individuais, permite a identificação de como trabalhadores podem ser mais vulneráveis em momentos de recessão econômica em função de seu perfil social, econômico e demográfico. Assim, o estudo irá empregar dados longitudinais da Pesquisa Nacional por Amostra de Domicílios Contínua [PNAD Contínua], levada à campo pelo Instituto Brasileiro de Geografia e Estatística [IBGE]. Com informações da base de microdados de divulgação trimestral, a probabilidade de transição da situação de ocupado/empregado para desocupado/desempregado é estimada tendo como variáveis preditoras características demográficas, sociais e econômicas individuais.

O argumento do trabalho aqui proposto é que informações dessa natureza possibilitam melhores previsões a respeito dos movimentos do mercado de trabalho, o que permite uma vantagem por meio da antecipação, subsidiando o planejamento e a formulação de políticas mais eficientes, seja de manutenção de empregos, de qualificação ou inserção de pessoas neste mercado.

Além da construção dos modelos preditivos, uma pergunta a ser respondida pelo trabalho é: em que medida o perfil social e demográfico dos trabalhadores os torna mais vulneráveis ao desemprego em momentos de recessão econômica? A princípio, trata-se de

[1] Modelos de transição de Markov são baseados na definição de Cadeia de Markov, que descreve um processo estocástico de tempo discreto em que a transição de um estado no momento t depende somente do estado em $t-1$. Estes modelos permitem estimar uma variável dependente ao longo do tempo, condicionada ao valor imediatamente anterior e a um conjunto de covariáveis.

um problema que pode ser investigado com um instrumental estatístico/econométrico empregando modelos de variável dependente discreta. Mas dentre os objetivos específicos está o emprego de técnicas supervisionadas de machine learning visando a adoção de técnicas mais flexíveis, generalizáveis e com maior poder preditivo (Mullainathan e Spiess, 2017).

Na análise inicial dos dados, percebeu-se um forte desequilíbrio entre as classes da variável alvo, com uma razão de desequilíbrio de aproximadamente 1:7. Dados desbalanceados exigem tratamento especial na formatação de modelos preditivos, uma vez que algoritmos de classificação tradicionais costumam ser tendenciosos em favor da classe majoritária, privilegiando maximizar alguma métrica de precisão global em detrimento de previsões corretas da classe minoritária (Krawczyk, 2016; Fernández et al., 2018).

O processo de modelagem realizado no presente trabalho aplicou modelos logit para uma abordagem inicial, e os algoritmos Random Forest e XGBoost combinados com o algoritmo denominado Synthetic Minority Oversampling Technique [SMOTE] para a construção de modelos preditivos. A aplicação computacional utilizou o framework oferecido pelo tidymodels que reúne funções de diversos pacotes direcionadas para modelagem e o aprendizado de máquina.

A principal contribuição do presente trabalho é estabelecer discussões iniciais e se estabelecer como uma das primeiras referências na aplicação de modelos de machine learning para a previsão de transição no mercado de trabalho. Os resultados obtidos apontam para o potencial de construção de modelos com boa capacidade preditiva e que podem ser aplicados ao problema considerado. Foram treinados modelos com desempenho razoável, mas que podem ser aprimorados em esforços de pesquisa futuros.

Material e Métodos

Dados

Os dados utilizados são provenientes da PNAD Contínua, elaborada e levada a campo pelo IBGE. Trata-se de dados públicos, disponibilizados pelo IBGE em seu portal na internet: https://www.ibge.gov.br/estatisticas/sociais/trabalho/9171-pesquisa-nacional-por-amostra-de-domicilios-continua-mensal.html?=&t=microdados.

Atualmente, a PNAD Contínua é a principal pesquisa socioeconômica por amostragem domiciliar no Brasil. Iniciada em 2012, a pesquisa emprega plano amostral complexo e adota painéis rotativos trimestrais no esquema “1-2(5)” em que um domicílio é entrevistado em um determinado mês, passa dois meses fora da amostra e volta a ser entrevistado, repetindo

esse padrão por cinco trimestres e deixando a amostra definitivamente ao fim de cinco entrevistas (IBGE, 2014).

Com o objetivo de avaliar a transição da situação de ocupado para desocupado ou inativo entre o 1º e 2º trimestres de 2020 (doravante 2020/T1 e 2020/T2), a amostra aplicada na análise aqui realizada é composta por indivíduos que estavam ocupados no 1º trimestre. Na composição dessa amostra foram considerados indivíduos que estavam respondendo a 1ª entrevista no 1º trimestre de 2020 e a 2ª entrevista no 2º trimestre de 2020[2].

Nos microdados da PNAD Contínua, para a identificação de domicílios são fornecidos na base de dados as seguintes variáveis: Unidade Primária de Amostragem [UPA], Número de Seleção do domicílio (V1008) e Painel de Grupo da Amostra (V1014). Estas variáveis permitem a construção de uma chave de identificação do domicílio no banco de dados, o que permite a realização de análises longitudinais (por até cinco trimestres). Por sua vez, a identificação de indivíduos pode ser realizada com a construção de um identificador usando as informações de sexo (variável V2007) e a data de nascimento do morador (dia, mês e ano, dadas pelas variáveis V2008, V20081 e V20082).

Ao conjunto de dados selecionado foram aplicados alguns filtros de forma a considerar apenas indivíduos em idade ativa no mercado de trabalho (de 15 a 65 anos) e que não apresentavam informações faltantes nas variáveis consideradas. Também foram retirados da amostra os indivíduos que declararam posição de ocupação como funcionários públicos estatutários ou militares, em razão da natureza destes vínculos.

Com a aplicação destes filtros, tem-se uma amostra composta por 23.538 observações, representando indivíduos que estavam ocupados no mercado de trabalho em 2020/T1. Destes, 20.167 permaneceram ocupados no mercado de trabalho em 2020/T2, e 3.371 passaram para a condição de desocupados ou inativos em 2020/T2.

Como mencionado, a variável resultado ou variável alvo da análise é a situação de ocupação dos indivíduos em 2020/T2. Na PNAD Contínua a identificação da situação dos indivíduos no mercado de trabalho é identificada pelas variáveis “Condição em relação à força de trabalho na semana de referência para pessoas de 14 anos ou mais de idade” (VD4001) e “Condição de ocupação na semana de referência para pessoas de 14 anos ou mais de idade” (VD4002). Dessa forma formata-se uma variável que caracteriza se o indivíduo está ocupado no mercado de trabalho ou não; neste segundo caso, podendo estar desocupado (mas ainda assim, ativo no mercado) ou inativo.

[2] Foram testadas amostras mais robustas com indivíduos identificados como respondentes das entrevistas de número 2, 3 e 4 no 1º trimestre de 2020 e entrevistas de número 3, 4 e 5 no 2º trimestre de 2020. Apesar da amostra maior, não se verificou melhora qualitativa nos resultados obtidos.

Por sua vez, o conjunto de preditores adotados nos modelos considera um conjunto de variáveis em nível individual, como características demográficas, sociais e econômicas. O conjunto de variáveis preditoras é apresentado na Tabela 1 abaixo.

Tabela 1. Descrição de variáveis preditoras.

Rendimento no mercado de trabalho.
Sexo: masculino ou feminino.
Idade. Esta variável foi categorizada em 4 grupos etários: 15-24 anos, 25-34 anos, 35-49 anos e 50-64 anos.
Cor declarada. Esta variável foi categorizada em 2 grupos: brancos (e amarelo) e preto ou pardos.
Escolaridade. Variável categoria com as seguintes classes: sem instrução, fundamental incompleto, fundamental completo, médio incompleto, médio completo, superior incompleto; superior completo.
Condição no domicílio: chefe do domicílio, cônjuge, filho ou enteado, parente ou outro.
Setor de ocupação em 2020/T1: transporte administração pública, agropecuária, alojamento e alimentação, comércio, construção cível, serviços de educação, saúde ou social, indústria, informação e comunicação, serviços domésticos, outros serviços e atividades mal definidas.
Posição de ocupação em 2020/T1: empregado no setor privado com carteira de trabalho assinada, empregado no setor privado sem carteira de trabalho assinada, empregado no setor público com carteira de trabalho assinada, empregado no setor público sem carteira de trabalho assinada, trabalhador doméstico com carteira de trabalho assinada, trabalhador doméstico sem carteira de trabalho assinada, trabalhador por conta própria, empregador e trabalhador familiar auxiliar.
Area de residência: urbana ou rural.
Região: Norte, Nordeste, Sudeste, Centro-Oeste e Sul.

Fonte: Dados originais da pesquisa

Modelos estatísticos/econométricos

A literatura que estuda as taxas de transição entre estados do mercado de trabalho foi inaugurada com o trabalho seminal de Clark e Summers (1978), e se desenvolveu com os trabalhos de Flinn e Heckman (1983) e, mais recentemente, com os trabalhos de Jones e Riddell (1999, 2006). De forma mais recente, e para o caso específico do Brasil, temos os trabalhos de Reis e Aguas (2014) e Santos, Monsueto e Varella (2021).

Considerando este arcabouço teórico, define-se $Y_{i,t}$ como uma variável aleatória discreta, que assume valores correspondentes a S estados mutuamente exclusivos e que descrevem a situação de um indivíduo i no mercado de trabalho no trimestre t.

Assim, a probabilidade de transição do estado s para o estado s' entre os trimestres $t-1$ e t é dada por eq.(1):

$$p_{i,t} = P\left(Y_{i,t} = s' | Y_{i,t-1} = s\right) \qquad s, s' = 0, 1, \cdots, S \tag{1}$$

A probabilidade representada na eq. (1) é um processo de Markov de ordem 1.

Do ponto de vista estatístico, o modelo da probabilidade de transição do estado s para o estado s' é um modelo de probabilidade condicionada, podendo ser função de um conjunto de características individuais (eq. 2), $X_{i,t-1}$. Assim, podemos reescrever:

$$p_{i,t} = P\left(Y_{i,t} = s' | Y_{i,t-1} = s\ , X_{i,t}\right) \qquad s, s' = 0, 1, \cdots, S \tag{2}$$

Nessa literatura, a estimação das probabilidades de transição no mercado de trabalho é realizada com um instrumental estatístico/econométrico empregando modelos de variável dependente discreta como os modelos da classe logit e probit.

Na presente proposta, na estimação das probabilidades de transição serão consideradas técnicas supervisionadas de machine learning, que estão ganhando notoriedade em anos recentes na análise de problemas econômicos conforme apontam Mullainathan e Spiess (2017), Athey (2017 e 2018), Athey e Imbens (2019). Obviamente, o uso de algotimos de machine learning tem como objetivo a adoção de técnicas mais flexíveis e com maior poder preditivo.

Considerando uma abordagem supervisionada de machine learning, trata-se de um problema de classificação, que pode ser modelado com o uso de diferentes algoritmos, como os algoritmos baseados em árvores de decisão, modelos de bagging (bootstrap aggregation) como o Random Forest, e algoritmos mais sofisticados de Adaptative Boosting e Gradient Boosting.

A obtenção de bons resultados preditivos com base em modelos de machine learning enfrenta alguns desafios que incluem problemas relacionados à complexidade, separabilidade, dimensionalidade e desequilíbrios entre classes. No presente estudo, talvez o principal desafio a ser considerado é dado pela presença de desequilíbrio entre as classes que representam a situação dos indivíduos no mercado de trabalho.

Algoritmos de machine learning aplicados

Os métodos de ensemble empregados em machine learning representam uma abordagem que busca melhorar desempenho preditivo ao combinar as previsões de vários modelos. Dentre os métodos de ensemble se destacam os métodos de bagging e boosting.

Assumindo como base os modelos de árvores de regressão e classificação, o bagging envolve ajustar muitas árvores em diferentes subamostras do mesmo conjunto de dados, obtidas pela aplicação do método de reamostragem de bootstrap, e calcular a média das previsões. Por sua vez, o boosting envolve a adição sequencial de membros do ensemble que corrigem as previsões feitas por modelos anteriores e gera uma média ponderada das

previsões. Dentre os métodos de boosting, os mais comuns são o AdaBoost (canonical boosting) e o Gradient Boosting.

Dos algoritmos de bagging, grande destaque é dado ao Random Forest, proposto por Breiman (2001), que é atualmente um dos mais populares em razão do seu bom desempenho preditivo exigindo relativamente pouco ajuste de hiperparâmetros. O algoritmo de Random Forest é construído a partir de um modelo de bagging de árvores de decisão. Ao agregar árvores construídas com subamostras obtidas por bootstrap, o bagging permite uma redução de variância que melhora o desempenho preditivo. No entanto, adotando os mesmos preditores, a correlação entre árvores diminui esse poder de redução de variância. O Random Forest incorpora mais aleatoriedade no processo de crescimento das árvores empregando um subconjunto de preditores aleatoriamente determinado a partir dos preditores originais.

Segundo Boehmke e Greenwell (2019), os principais hiperparâmetros a ser ajustados no algoritmo Random Forest, que possuem o maior impacto na capacidade preditiva são os seguintes: o número de árvores (“trees”), número de variáveis aleatoriamente consideradas em cada divisão (parâmetro “mtry”) e o número mínimo de pontos em um nó (parâmetro “min_n”).

Por sua vez, os algoritmos de Gradient Boosting se tornaram extremamente populares após se mostrarem bem-sucedidos ao vencer competições do Kaggle, sendo aplicados à diferentes problemas (Boehmke e Greenwell, 2019)[3]. Enquanto o Random Forest se compromete com a construção de conjuntos de árvores profundas e independentes, o Gradient Boosting se propõe a construir um conjunto de árvores rasas de forma sequencial, com cada árvore aprendendo e melhorando com os erros da árvore anterior. Embora as árvores rasas por si só sejam modelos preditivos bastante fracos, elas podem ser “impulsionadas” para produzir um poderoso conjunto que, quando ajustado adequadamente, atinge resultados difíceis de serem superados por outros algoritmos. Nessa classe de modelos, grande destaque é dado ao Extreme Gradient Boosting, ou XGBoost, desenvolvido por Chen e Guestrin (2016).

No caso do algoritmo XGBoost, os hiperparâmetros a ser ajustados são os seguintes: o número de árvores (“trees”), número de variáveis aleatoriamente consideradas em cada divisão (parâmetro “mtry”), o número mínimo de pontos em um nó (parâmetro “min_n”), a profundidade máxima de cada árvore (“tree_depth”), a taxa de aprendizado (“learn_rate”), a redução de perda mínima (“loss_reduction”) e a proporção amostrada (“sample_size”).

[3] O Kaggle é uma plataforma que hospeda competições de Data Science públicas, privadas e acadêmicas. Para saber mais: https://www.kaggle.com/ .

Algoritmo para lidar com dados desbalanceados

De modo geral, algoritmos canônicos de machine learning assumem que o número de observações entre as classes de uma variável alvo é distribuído de forma uniforme, ou de forma próxima a isso. No entanto, em muitas situações reais, a distribuição de observações é bastante assimétrica ou desbalanceada, com algumas classes sendo representadas com muito mais frequência do que outras. Isso representa uma dificuldade para o aprendizado dos algoritmos, pois tendem a gerar previsões tendenciosas a favor da classe majoritária; o que constitui um problema quando o interesse é exatamente a previsão de classes minoritárias (Krawczyk, 2016).

Na análise preditiva com modelos de machine learning este é um problema relativamente comum em problemas de classificação, como a detecção de fraudes, diagnóstico de problemas de saúde, risco de crédito e a previsão de churn (perda de clientes).

Na modelagem com dados desbalanceados, é comum que as medidas de precisão se apresentem como excelentes, mas na verdade refletem apenas a distribuição da variável subjacente. Com isso, muitos analistas podem ser induzidos ao erro, uma vez que a precisão da classificação geralmente é a medida mais empregada para avaliar os modelos treinados.

Conforme Fernandez et al. (2018), as abordagens mais usadas para lidar com dados desbalanceados podem ser agrupadas em: i) Abordagens em nível de algoritmo, ii) Abordagem em nível de dados, iii) Abordagem de aprendizado sensível ao custo e, iv) Abordagem de métodos de ensemble.

As abordagens em nível de dados visam alterar o equilíbrio original entre as classes, seja replicando pontos amostrais pertencente às classes minoritárias ou criando instâncias sintéticas (oversampling), seja removendo algumas instâncias das classes majoritárias (undersampling). O objetivo é fornecer um conjunto de dados mais equilibrado para que o modelo treinado não tenha problemas para classificar instâncias de qualquer classe.

No presente trabalho adota-se o algoritmo denominado Synthetic Minority Oversampling Technique [SMOTE], que é uma das técnicas mais sofisticadas e renomadas para lidar com dados desbalanceados. A ideia chave do SMOTE é introduzir exemplos sintéticos de classes minoritárias, criados por interpolação de várias instâncias comuns nestas classes. Por esta razão, diz-se que o procedimento é focado no “espaço de características” e não no “espaço de dados” (Fernandez et al., 2018).

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.

Uma abordagem bastante popular para avaliar a performance de um modelo de classificação é baseada na matriz de confusão. Uma matriz de confusão é simplesmente uma matriz que compara as classes observadas de uma variável categórica com as classes previstas pelo modelo. Para a presente aplicação, em que a variável alvo apresenta duas classes, a matriz de confusão é uma matriz 2x2, como na Figura 1 a seguir, em que nas colunas se apresentam os valores reais/observados e nas linhas os valores preditos.

		Classe Observada	
		Positivo	Negativo
Classe Prevista	Positivo	*Verdadeiro Positivo (VP)*	*Falso Positivo (FP)*
	Negativo	*Falso Negativo (FN)*	*Verdadeiro Negativo (VN)*

Figura 1. Matriz de Confusão
Fonte: Dados originais da pesquisa

Ao prever a classe alvo (denominada como positivo) corretamente, tem-se um verdadeiro positivo [VP]. No entanto, ao prever erroneamente essa classe, tem-se um falso positivo (FP, um erro do tipo I). Alternativamente, quando se prevê a classe alternativa (denominada como negativo) e a classe alvo é a classe correta, isso é chamado de falso negativo (FN, um erro do tipo II). Quando se prevê a classe alternativa e a previsão é correta, tem-se um verdadeiro negativo.

Com base na matriz de confusão é possível calcular diversas métricas das quais destacam-se: a acurácia, a precisão, a sensibilidade e a especificidade[4].

A Acurácia (eq. 3) é uma métrica de performance geral do modelo. Ela indica as classificações corretas do modelo dentre todas as classificações realizadas ($P + VN + FP + FN$):

$$\text{acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \quad (3)$$

[4] Medidas presentes em Kuhn e Johnson (2013), Boehmke e Greenwell (2019).

A Precisão (eq. 4) é definida como a proporção de predições corretas de uma categoria em relação a todas as previsões feitas dessa categoria. Nesse caso, o número de verdadeiros positivos entre o total de positivos previstos ($VP + FP$):

$$\text{precisão} = \frac{VP}{VP + FP} \tag{4}$$

A Sensibilidade (eq. 5), também conhecida como recall, é definida como a proporção de previsões corretas da categoria alvo (positivo), ou seja, uma taxa de verdadeiros positivos.

$$\text{sensibilidade} = \frac{VP}{VP + FN} \tag{5}$$

A Especificidade (eq. 6), por sua vez, mede a proporção de previsões negativas realizadas corretamente:

$$\text{especificidade} = \frac{VN}{VN + FP} \tag{6}$$

A F1 score (eq. 7) é dada pela média harmônica entre precisão e sensibilidade, representada por:

$$F1 = \frac{2 * precisão * recall}{precisão + recall} \tag{7}$$

Além destas medidas, existem medidas compostas que são mais adequadas para situação de dados desbalanceados, como a Acurácia Balanceada, o Índice J de Youden e o Índice Kappa de Cohen [KAP]. Detalhes destas medidas são apresentados por Kuhn e Johnson (2013) e Fernandez et al. (2018).

A Acurácia Balanceada é dada por uma média aritmética simples das medidas de sensibilidade e especificidade, sendo uma média da taxa de verdadeiros positivos e verdadeiros negativos.

O Índice J de Youden é dado por: $\text{J} = \text{sensibilidade} + \text{especificidade} - 1$. Seu valor varia de 0 a 1, em que o valor zero indica que o classificar é inútil e o valor de 1 indica que não há falsos positivos ou falsos negativos, ou seja, o teste é perfeito.

O índice Kappa de Cohen [KAP] é baseado na ideia de “compensar”, da medida de acurácia (A), a porção devida ao acaso, ou seja, a porção que um classificador aleatório alcançaria. O índice é dado por: $KAP = \frac{A - A^e}{1 - A^e}$, em que A^e é uma medida de acurácia esperada. Os valores do KAP variam de -1 a 1, e valores menores que zero indicam que o desempenho do classificador é menor do que a adivinhação aleatória.

Outra forma de avaliar um modelo de classificação é por meio da curva Receiver Operating Characteristic [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 é baseada em dois parâmetros: a medida de sensibilidade e a medida de especificidade. Plotada em um plano unitário, a curva ROC resulta da representação gráfica da taxa de verdadeiros positivos pela taxa de falsos positivos (1 - especificidade).

Uma forma de simplificar a análise da curva ROC é por meio da área sob a curva ROC [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 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.

Segundo He e Ma (2013), a análise baseada na ROC, e AUC, não tem nenhum viés em relação a modelos que apresentam bom desempenho na classe majoritária em detrimento da classe majoritária – uma propriedade que é bastante atraente ao lidar com dados desequilibrados. Os autores complementam que, mesmo com críticas recentes sobre o uso da ROC, trata-se da métrica mais comum para analisar modelos com dados desequilibrados.

Resultados e Discussão

Nesta seção são apresentados resultados obtidos com a aplicação de técnicas de análise estatística de dados e modelagem preditiva com algoritmos de machine learning. A análise de dados foi realizada empregando técnicas de análise exploratória e um modelo de regressão logística. O objetivo da análise exploratória foi o de replicar indicadores oficiais publicados pelo IBGE e gerar indicadores específicos, de acordo com os interesses do trabalho, permitindo uma melhor compreensão do fenômeno estudado. Foram analisados os indicadores de taxa de participação e de taxa de desocupação discriminados de acordo com algumas das variáveis explicativas consideradas na análise.

Por sua vez, a regressão logística considerou a amostra de dados do trabalho e teve o objetivo de apresentar uma primeira análise da transição (ou não transição) da situação de ocupação entre o 1º e o 2º trimestres de 2020. Verificou-se a correlação e a significância estatística do conjunto de variáveis explicativas com a variável alvo, e testou-se a adequação deste modelo para a predição de resultados.

Por fim, testou-se a aplicação de algoritmos de machine learning para fins de construção de um modelo preditivo. Dadas as características de desbalanceamento dos

dados aplicou-se os algoritmos de Random Forest e XGBoost associados com a técnica SMOTE. A performance preditiva dos modelos construídos foi verificada com a aplicação de medidas clássicas, comentadas na seção anterior.

Todas as análises utilizaram a linguagem R e o ambiente de desenvolvimento RStudio. Na análise exploratória de dados foram empregados os microdados da PNAD Contínua, bem como os dados o Sistema IBGE de Recuperação Automática [SIDRA]. Os microdados foram baixados diretamente no ambiente do RStudio com a função get_pnadc do pacote PNADcIBGE[5]. Nesta análise foi considerado o desenho amostral da PNAD Contínua, ponderando os dados para que estes representem estimativas populacionais fidedignas do mercado de trabalho brasileiro. Por sua vez, os dados do SIDRA foram baixados diretamente no R com o uso do pacote sidrar[6].

O processo de modelagem e aplicação de algoritmos de machine learning contou com a aplicação do framework oferecido pelo tidymodels, que reúne funções de diversos pacotes direcionadas para modelagem e o aprendizado de máquina[7].

Análise de dados do mercado de trabalho nos dois primeiros trimestres de 2020

A taxa de participação na força de trabalho, indica a porcentagem de pessoas em idade de trabalhar (14 anos ou mais) que estão empregadas ou em busca de trabalho. Este indicador representa uma medida da parte ativa de toda a força de trabalho no país. A série relativa à taxa de participação, iniciada no 1º trimestre de 2012, até o último trimestre disponível (3º trimestre de 2021) está plotada no gráfico da Figura 2 abaixo.

[5] Documentação: https://cran.r-project.org/web/packages/PNADcIBGE/index.html .
[6] Documentação: https://cran.r-project.org/web/packages/sidrar/index.html .
[7] Mais sobre o *tidymodels* pode ser consultado no site: *https://www.tidymodels.org/* . Outra ótima referência é o trabalho de Kuhn e Silge (2022), disponível em: https://www.tmwr.org/ .

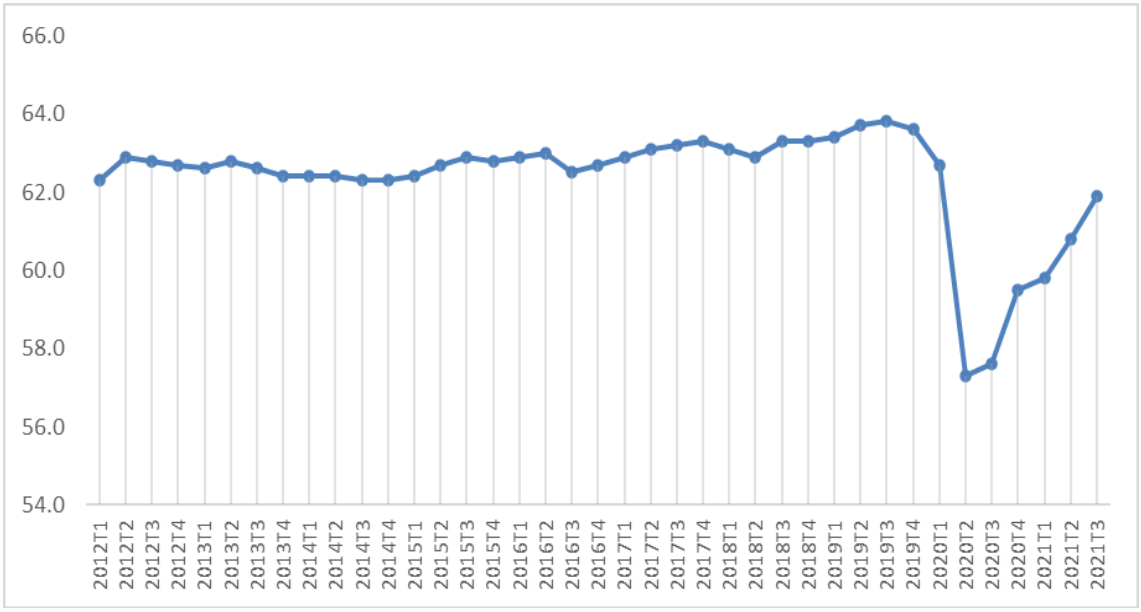

Figura 2. Taxa de participação no Brasil (2012/T1 – 2021/T3)
Fonte: Resultados originais da pesquisa

Desde o 1º trimestre de 2012 a taxa de participação esteve em um patamar médio de aproximadamente 63%. No 1º trimestre de 2020 era de 62,7%, quando apresentou uma redução brusca no segundo trimestre de 2020 para 57,3%. Nos trimestres seguintes este indicador apresentou uma tendência de crescimento, mas, até o 3º trimestre de 2021, sem atingir o patamar médio anterior à pandemia.

Por sua vez, a taxa de desemprego apresenta um comportamento mais volátil, respondendo de forma mais elástica aos movimentos da atividade econômica. Considerando a série iniciada no 1º trimestre de 2012, é possível verificar um forte aumento do desemprego entre os primeiros trimestres de 2015 e 2017, período de maior impacto da recessão econômica de 2015 e 2016. Após o período recessivo, a taxa de desocupação se estabeleceu em um patamar mais elevado. Entre o 1º trimestre de 2012 e o 4º trimestre de 2014 a taxa de desemprego estava em uma média de 7,2%. Após o período de aumento, entre os primeiros trimestres 2017 e 2020, a média foi de 12,4%, mesmo valor observado no 1º trimestre de 2020. No 2º trimestre de 2020, a taxa de desocupação avançou para o patamar de 13,6% e apresentou um forte crescimento, atingindo o maior valor da série no 3º trimestre, 14,9%.

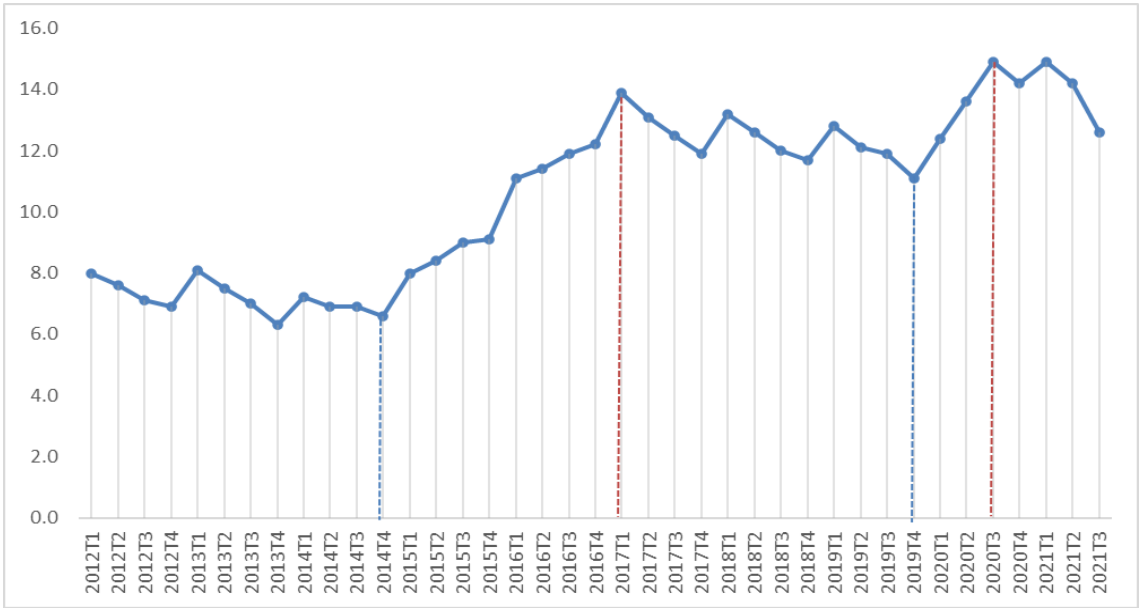

Figura 3. Taxa de desocupação no Brasil (2012/T1 – 2021/T3)
Fonte: Resultados originais da pesquisa

A inserção dos trabalhadores do mercado é reconhecidamente desigual, seja discriminada por características demográficas, regionais ou sociais. Considerando algumas destas características, as informações apresentadas nos gráficos a seguir apresentam uma breve análise exploratória do perfil dos indivíduos que estão ativos no mercado de trabalho e dos que estão desocupados.

A Figura 4 apresenta gráficos relativos às taxas de participação e de desocupação por sexo (homem e mulher). O gráfico da esquerda apresenta a taxa de participação por sexo e revela um diferencial importante quanto à inserção no mercado de trabalho de homens e mulheres. A taxa de participação dos homens é maior do que a taxa de participação das mulheres. O gráfico da direita apresenta a taxa de desocupação para homens e mulheres, revelando a maior vulnerabilidade ao desemprego por parte das mulheres.

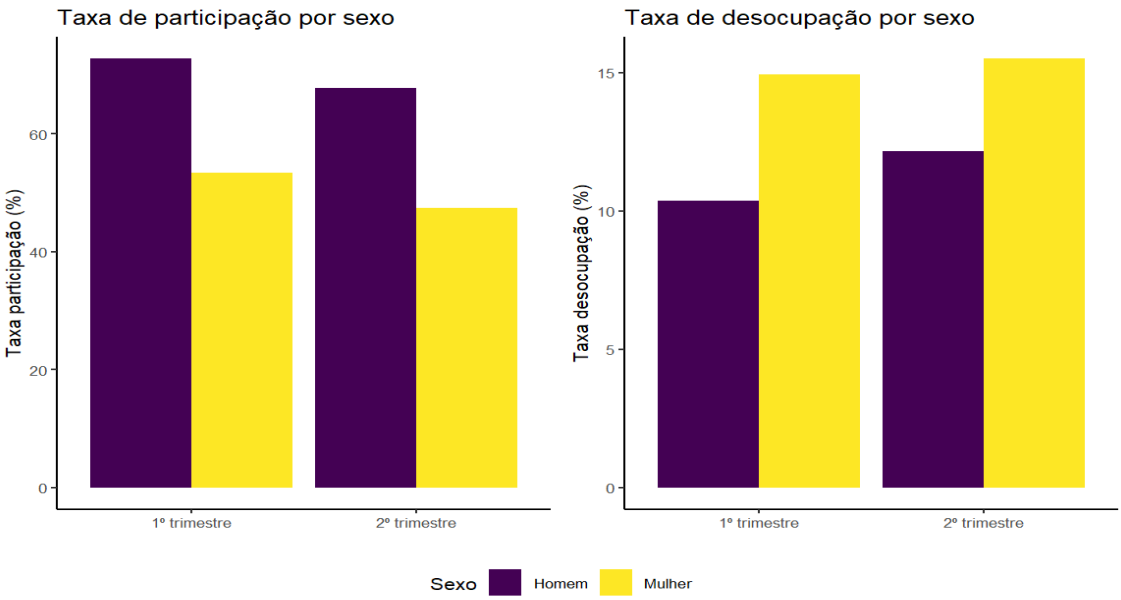

Figura 4. Taxa de participação e de desocupação por sexo (Brasil: 2020/ T1 e T2)
Fonte: Resultados originais da pesquisa

A Figura 5 apresenta informações sobre as taxas de participação e desocupação segundo a cor declarada pelos indivíduos no momento da pesquisa. O gráfico mostra que pessoas brancas possuem maior participação do que as pessoas pretas ou pardas. Por sua vez, a taxa de desocupação entre pretos e pardos também é maior, sinalizando sobre uma maior vulnerabilidade ao desemprego.

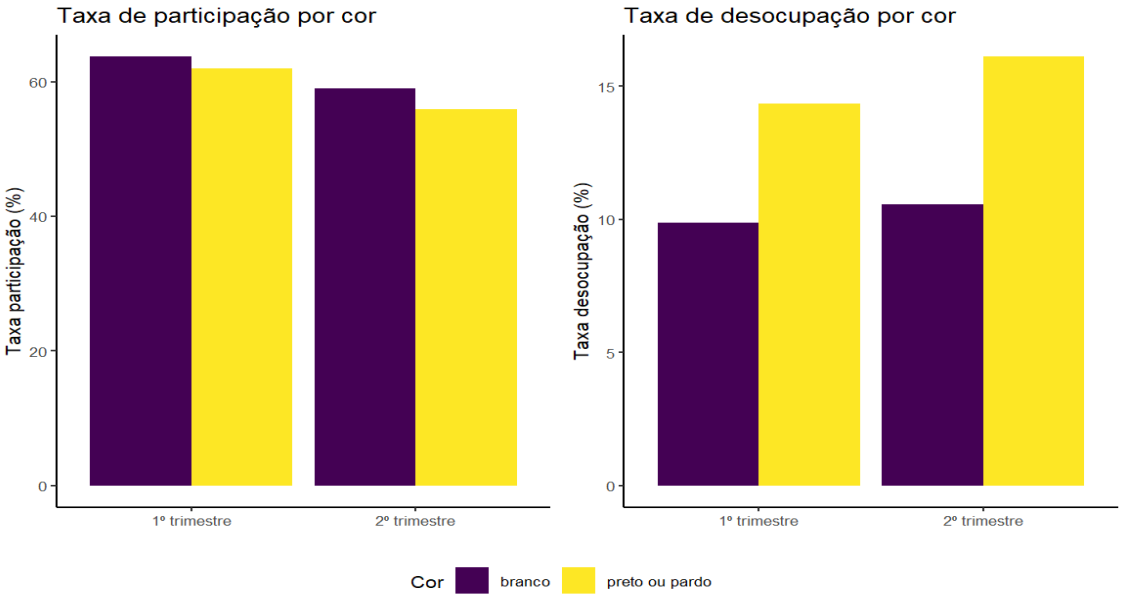

Figura 5. Taxa de participação e de desocupação por cor (Brasil: 2020/ T1 e T2)
Fonte: Resultados originais da pesquisa

Os gráficos da Figura 6 mostram as taxas de participação e desocupação por nível de escolaridade. No primeiro gráfico pode-se observar que quanto maior o nível de escolaridade,

maior é a taxa de participação. Quanto à taxa desocupação, como era de se esperar, os trabalhadores com nível superior são menos vulneráveis ao desemprego. Mas chama a atenção para os trabalhadores mais vulneráveis ao desemprego serem os que possuem escolaridade equivalente ao ensino médio incompleto.

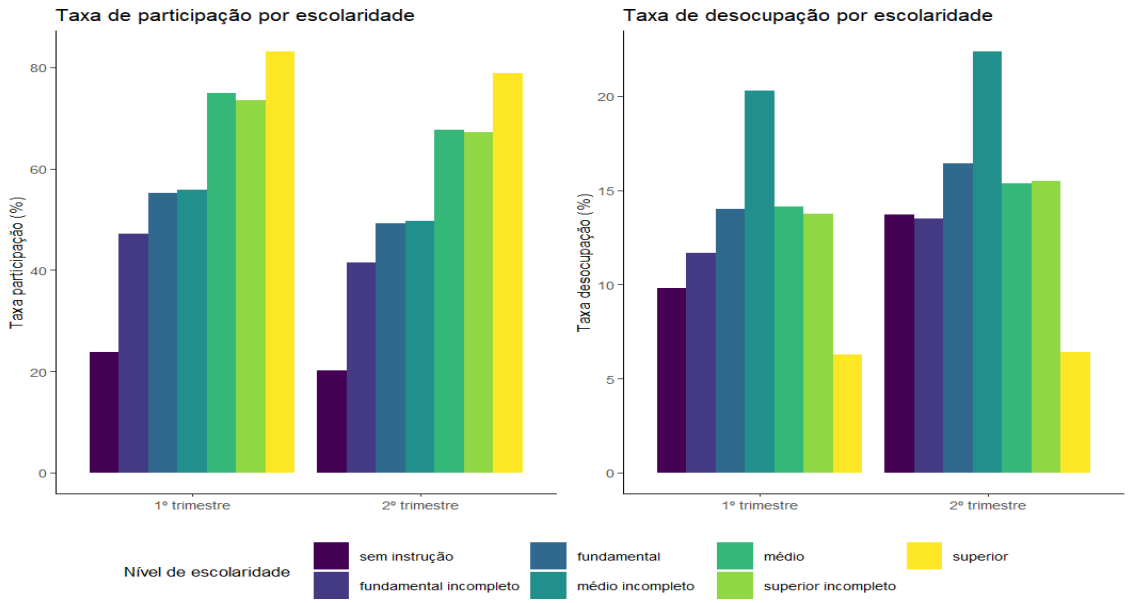

Figura 6. Taxa de participação e desocupação por escolaridade (Brasil: 2020/ T1 e T2)
Fonte: Resultados originais da pesquisa

A Figura 7 apresenta em seus gráficos as taxas de participação e desocupação segundo as grandes regiões brasileiras. Observa-se que a região Nordeste apresenta os indicadores mais frágeis em relação ao mercado de trabalho, com menor taxa de participação e maior taxa de desocupação. Por outro lado, a região Sul apresenta uma taxa de participação relativamente elevada e a menor taxa de desocupação.

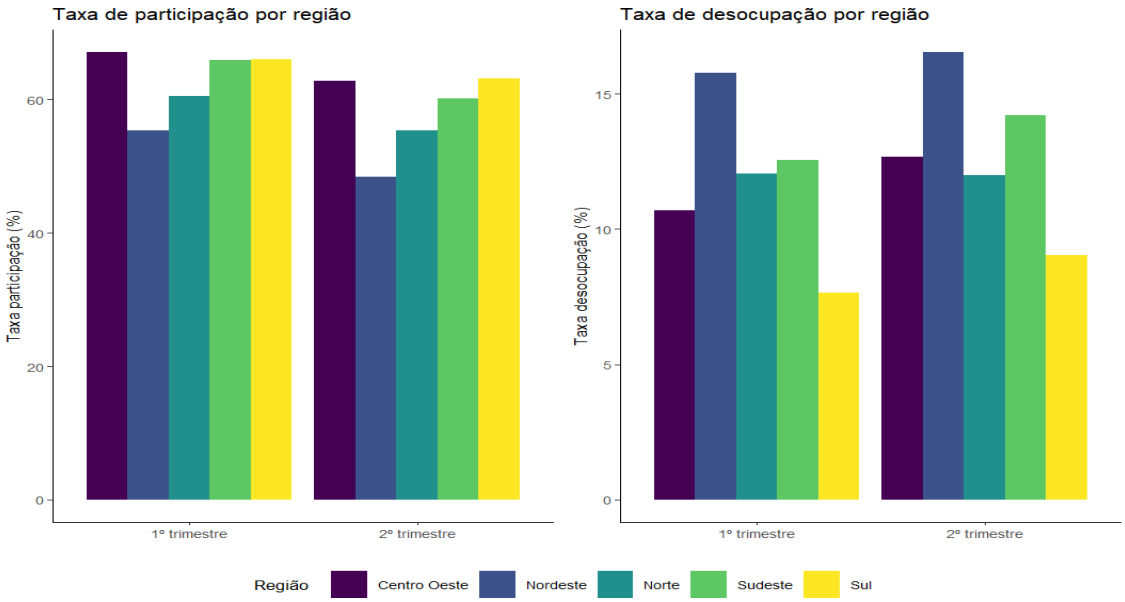

Figura 7. Taxa de participação e desocupação por região (Brasil: 2020/ T1 e T2)
Fonte: Resultados originais da pesquisa

Análise com a aplicação de regressão logística

Como relatado anteriormente, na seção Material e Métodos, a modelagem estatística empregou os microdados brutos da amostra da PNAD Contínua.

A amostra a ser utilizada nesta análise é composta por 23.538 observações, representando indivíduos identificados no 1º e 2º trimestres de 2020, que estavam ocupados no mercado de trabalho em 2020/T1. Destes, 20.167 permaneceram ocupados no mercado de trabalho em 2020/T2, e 3.371 passaram para a condição de desocupados ou inativos em 2020/T2. Essa amostra de dados foi particionada de forma que 75% dos dados compuseram o conjunto de dados de treino (17.653 observações) e 25% foram reservados para o conjunto de dados de teste (5.885 observações). A partição foi estratificada de forma a considerar a distribuição da situação ocupacional dos trabalhadores. Informações detalhadas da amostra e da partição em dados de treino e teste são apresentadas no Tabela 2 a seguir.

Tabela 2. Número de observações da amostra segundo a situação ocupacional e divisão dos dados em treino e teste.

	Amostra		Treino		Teste	
Desocupados ou inativos	3.371	(14,3%)	2.528	(14,3%)	843	(14,3%)
Ocupados	20.167	(85,7%)	15.125	(85,7%)	5.042	(85,7%)
Total	23.538	(100%)	17.653		5.885	

Fonte: Resultados originais da pesquisa

A Tabela 3, por sua vez, apresenta estatísticas descritivas da amostra utilizada.

Tabela 3. Estatísticas descritivas da amostra

Características (variáveis preditoras)	N = 23538
	n (%)
Rendimento do trabalho (média)	
Média	1.300
Mediana	800
Grupos de idade	
15-24	3.493 (15%)
25-34	5.661 (24%)
35-49	8.848 (38%)
50-65	5.536 (24%)
Sexo	
Homem	13.847 (59%)
Mulher	9.691 (41%)
Cor (declarada)	
Brancos (ou amarelos)	10.123 (43%)
Pretos ou pardo (ou indígenas)	13.415 (57%)
Nível de escolaridade	
Sem instrução	583 (2,5%)
Fundamental incompleto	6.232 (26%)
Fundamental	2.018 (8,6%)
Médio incompleto	1.738 (7,4%)
Médio	7.848 (33%)
Superior incompleto	1.387 (5,9%)
Superior	3.732 (16%)
Condição no domicílio	
Chefe	10.735 (46%)
Cônjuge	6.791 (29%)
Filho ou enteada	4.567 (19%)
Parente	1.334 (5,7%)
Outro	111 (0,5%)
Posição de ocupação (em 2020/T1)	
Conta própria	7.162 (30%)
Empregados no setor privado com carteira	8.298 (35%)
Empregados no setor privado sem carteira	3.310 (14%)
Empregados no setor público com carteira	297 (1,3%)
Empregados no setor público sem carteira	817 (3,5%)
Empregadores	1.148 (4,9%)
Trabalhadores domésticos com carteira	449 (1,9%)
Trabalhadores domésticos sem carteira	1.256 (5,3%)
Trabalhador familiar auxiliar	801 (3,4%)
Setor de ocupação (em 2020/T1)	
Administração pública	440 (1,9%)
Agropecuária	3.950 (17%)
Alojamento e alimentação	1.294 (5,5%)
Atividades mal definidas	10 (<0,1%)
Comércio	4.660 (20%)
construção	1.658 (7%)
Educação, saúde e serviço social	1.953 (8,3%)
indústria	2.962 (13%)
Informação e comunicação	2.411 (10%)
Outros serviços	1.250 (5,3%)
Serviços domésticos	1.718 (7,3%)
Transporte	1.232 (5,2%)
Região	
Centro Oeste	2.693 (11%)
Nordeste	5.959 (25%)
Norte	2.748 (12%)
Sudeste	7.322 (31%)
Sul	4.816 (20%)
área	
Rural	5.588 (24%)
Urbana	17.950 (76%)

Fonte: Resultados originais da pesquisa

De modo geral, a situação de um indivíduo no mercado de trabalho pode ser classificada em ativo ocupado, ativo desocupado ou inativo (fora da força de trabalho). Para o processo de modelagem a ser aplicado, estas categorias foram agrupadas em apenas duas classes: ocupado e desocupado ou inativo, referentes à situação dos indivíduos no mercado de trabalho no 2º semestre de 2020[8].

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. As variáveis categóricas foram convertidas em variáveis dummies. Entre as variáveis descritivas temos uma única variável numérica, de rendimentos do trabalho, que foi reescalonada usando o logaritmo natural (somando 1 para evitar o ln(0)).

Com o objetivo de complementar a análise de dados, e estabelecer uma primeira referência ao processo de modelagem, um primeiro modelo foi ajustado com a especificação logit, muito comum em problemas de classificação em análises estatísticas e econométricas. Os resultados desta estimação são sumarizados na Tabela 4.

Tabela 4. Coeficiente estimados e razões de chance – modelo logit

(continua)

Variável preditora	Coeficiente	Desvio-padrão	z	valor-p
Intercepto	-1.2134	0.2168	-5.5978	0.0000
Rendimento	-0.1764	0.0126	-14.0122	0.0000
Região = Sul (referência)				
Regiao = Centro.Oeste	0.3342	0.0959	3.4849	0.0005
Região = Nordeste	0.6683	0.0795	8.4047	0.0000
Região = Norte	0.3551	0.0955	3.7184	0.0002
Região = Sudeste	0.4106	0.0768	5.3454	0.0000
Area = urbano (referência)				
Area = rural	-0.2105	0.0662	-3.1786	0.0015
Sexo = Homem	-0.4361	0.0556	-7.8490	0.0000
Sexo = Mulher (referência)				
Cor = preto ou pardo (referência)				
Cor = branco	-0.2041	0.0518	-3.9385	0.0001
Idade = 50-65 (referência)				
Idade = 15-24	0.0498	0.0876	0.5691	0.5693
Idade = 25-34	-0.1717	0.0726	-2.3649	0.0180
Idade = 35-49	-0.2165	0.0614	-3.5278	0.0004
Escolaridade = superior (referência)				
Escolaridade = sem.instrucao	0.4297	0.1555	2.7634	0.0057
Escolaridade = fundamental.incompleto	0.3022	0.0955	3.1659	0.0015
Escolaridade = fundamental	0.3786	0.1090	3.4722	0.0005
Escolaridade = medio.incompleto	0.3881	0.1109	3.5007	0.0005
Escolaridade = medio	0.1333	0.0873	1.5281	0.1265
Escolaridade = superior.incompleto	0.3464	0.1173	2.9526	0.0032
Condição no domicílio = parente (referência)				
Condição no domicílio = Chefe	-0.1931	0.0977	-1.9759	0.0482

[8] Foram consideradas versões do modelo multinominal, mas sem obter um bom ajuste ao modelo com dados desbalanceados.

Tabela 4. Coeficiente estimados e razões de chance – modelo logit

(conclusão)

Variável preditora	Coeficiente	Desvio-padrão	z	valor-p
Condição no domicílio = conjuge	-0.1425	0.1007	-1.4150	0.1571
Condição no domicílio = filho.ou.enteado	0.1258	0.1020	1.2335	0.2174
Condição no domicílio = outro	-0.6577	0.3657	-1.7983	0.0721
Setor de ocupação = transporte (referência)				
Setor de ocupação = adm_publica	-0.7418	0.2845	-2.6072	0.0091
Setor de ocupação = agro	-0.7583	0.1270	-5.9714	0.0000
Setor de ocupação = alojamento_alimentacao	0.1537	0.1296	1.1861	0.2356
Setor de ocupação = ativ_mal_definidas	-11.5741	116.9169	-0.0990	0.9211
Setor de ocupação = comercio	-0.2706	0.1137	-2.3794	0.0173
Setor de ocupação = construcao	0.3111	0.1221	2.5478	0.0108
Setor de ocupação = educacao_saude_social	-0.4389	0.1532	-2.8646	0.0042
Setor de ocupação = industria	-0.2855	0.1234	-2.3130	0.0207
Setor de ocupação = info_comunicacao	-0.3711	0.1330	-2.7897	0.0053
Setor de ocupação = outros_servicos	-0.0357	0.1335	-0.2672	0.7893
Setor de ocupação = servicos_domesticos	1.5105	0.6750	2.2379	0.0252
Posição de ocupação = trabalhador_familiar_auxiliar (referência)				
Posição de ocupação = conta_propria	0.8746	0.1439	6.0800	0.0000
Posição de ocupação = empregado_setor_privado_com_carteira	0.3406	0.1545	2.2039	0.0275
Posição de ocupação = empregado_setor_privado_sem_carteira	1.3044	0.1476	8.8382	0.0000
Posição de ocupação = empregado_setor_publico_com_carteira	-0.0359	0.3716	-0.0967	0.9229
Posição de ocupação = empregado_setor_publico_sem_carteira	0.7266	0.2290	3.1736	0.0015
Posição de ocupação = empregador	0.0388	0.2148	0.1805	0.8567
Posição de ocupação = trabalhador_domestico_com_carteira	-1.4260	0.6867	-2.0768	0.0378
Posição de ocupação = trabalhador_domestico_sem_carteira	-0.2916	0.6642	-0.4390	0.6606

Fonte: Resultados originais da pesquisa

A significância estatística geral do modelo é verificada pelo teste χ^2 (qui-quadrado). O modelo estimado apresentou $\chi^2(38\ gl) = 1.468{,}83$, indicando que para o nível de significância de 5%, permite rejeitar a hipótese nula de que todos os parâmetros sejam estatisticamente iguais a zero, ou seja, pelo menos um preditor é estatisticamente significante para explicar a probabilidade de transição da situação de ocupado para a de desocupado ou inativo (Favero e Belfiore, 2017).

Nessa análise, a significância estatística de cada preditor foi avaliada com a estatística *z* de Wald (Favero e Belfiore, 2017). Ao analisar os coeficientes estimados para cada preditor, os resultados confirmam alguns efeitos já esperados.

O coeficiente estimado para a variável de rendimento se mostrou estatisticamente significante e com sinal negativo. Esse resultado aponta uma evidência de que indivíduos com menores rendimentos apresentaram maior probabilidade de transição para a situação de desemprego ou inatividade. Esse resultado é muito intuitivo, uma vez que trabalhadores que auferem menores rendimentos são também aqueles que ocupam os postos de trabalho mais vulneráveis.

Quanto à região, em comparação à região Sul, que foi a categoria base na estimação, a probabilidade de desemprego ou inatividade se mostrou significativamente maior em todas

as demais regiões. Maior probabilidade de desocupação ou inatividade também ocorre em áreas urbanas.

Considerando as características demográficas, em relação aos homens, as mulheres apresentaram maior probabilidade de migrarem para a desocupação ou inatividade. Pessoas que se declararam pretas ou pardas apresentaram maior probabilidade de ficarem desempregadas ou inativas quando comparadas com as que se declararam brancas. A posição no domicílio não se mostrou relevante do ponto de vista estatístico, com exceção do próprio chefe de família, que apresentou maior probabilidade de se manter ocupado.

Em relação ao grupo de indivíduos mais experientes, com idade entre 50 e 65 anos, grupos etários mais jovens apresentaram maior probabilidade de permanecerem ocupados. Em relação ao grupo mais jovem, com idade entre 15 e 24 anos, o resultado não se mostrou significativo do ponto de vista estatístico.

Em relação à escolaridade, trabalhadores com ensino superior completo foram considerados como categoria de referência. Apenas o parâmetro estimado referente ao nível médio não mostrou significância estatística. Quanto aos demais níveis, todos mostraram maior probabilidade de desocupação ou inatividade.

A posição de ocupação também se mostrou relevante. Definindo como categoria de referência os trabalhadores familiares auxiliares, os coeficientes estimados para empregados no setor público com carteira de trabalho assinada e empregadores não mostram uma diferença estatisticamente significante na probabilidade de desocupação ou inatividade. Em relação a categoria de referência, apenas trabalhadores domésticos com carteira de trabalho assinada apresentaram maior probabilidade de se manterem ocupados. Todas as demais categorias apresentaram efeito significativo sobre a probabilidade de desocupação ou inatividade.

Quanto ao setor de ocupação, adotando pessoas ocupadas no setor de transportes como referência, se mostram mais propensos à desocupação ou inatividade os empregados nos setores de construção e de serviços domésticos.

Previsões realizadas com base nesta especificação mais simples não se mostraram satisfatórios. Mesmo com acurácia de 0,857 e precisão de 0,861, métricas como a sensibilidade de 0,994, a especificidade de 0,041 e a ROC/AUC de 0,27 indicam a inadequação de tal modelo ser aplicado para realizar previsões de transição da situação de ocupado para a de desocupado ou inativo em novos dados.

Modelagem preditiva com modelos de machine learning e a técnica SMOTE

Com a proposta de apresentar uma alternativa adequada para lidar com o problema de previsão, foram construídos modelos preditivos aplicando algoritmos de machine learning. Considerando uma amostra de indivíduos ocupados no mercado de trabalho no 1º trimestre de 2020, o modelo deve prever a condição no mercado de trabalho no trimestre seguinte, após os primeiros efeitos da pandemia de Covid-19 sobre a atividade econômica e o mercado de trabalho.

A amostra de dados empregada contou com 23.538 observações e 10 variáveis (uma variável alvo e 9 atributos). Essa amostra de dados foi particionada de forma que 75% dos dados compuseram o conjunto de dados de treino (17.653 observações) e 25% foram reservados para o conjunto de dados de teste (5.885 observações). A partição foi estratificada de forma a considerar a distribuição da situação ocupacional dos trabalhadores conforme informações apresentadas na Tabela 2.

Uma primeira característica que chama a atenção sobre os dados, e que possui implicações sobre o desempenho de modelos preditivos, diz respeito à distribuição das classes da variável dependente. Verifica-se um forte desbalanceamento entre as classes com 85,7% de indivíduos ocupados, 14,3% de desocupados ou inativos. Proporções que foram consideradas na separação de dados de treino e teste.

Deve-se destacar que esse desequilíbrio é natural do problema analisado (situação no mercado de trabalho), e não é gerado por problemas de amostragem ou erros de mensuração. Isso é importante para a definição da abordagem correta a ser aplicado na modelagem. Quando é identificado que o desequilíbrio é gerado por vieses de amostragem ou erros de mensuração, as principais abordagens costumam ser a coleta de mais informações amostrais. Nos demais casos existem algoritmos específicos como os apresentados anteriormente com base em Fernandez et al. (2018).

Dada a identificação do desbalanceamento entre as classes da variável alvo, buscou-se uma abordagem hibrida, com a combinação de modelos ensemble e a aplicação de uma técnica em nível de dados. Aplicou-se os algoritmos de Random Forest e XGBoost combinados com o algoritmo SMOTE.

Com a modelagem por meio dos algoritmos Random Forest e XGBoost, o ajuste de hiperparâmetros foi realizado com a aplicação de um procedimento de validação cruzada conhecido como k-*fold*, que separa os dados de treinamento em k partes e testa separadamente cada parte para ajustar o modelo. Aqui adotou-se $k = 10$.

Esse procedimento é projetado para selecionar um conjunto ideal de hiperparâmetros a ser adotado na versão final do modelo, utilizada para fazer previsões com base no conjunto

de dados de teste. A estratégia de ajuste de hiperparâmetros que este estudo usa é a chamada pesquisa em grade, na qual todas as combinações possíveis dos hiperparâmetros fornecidos foram testadas estabelecendo o objetivo de maximizar o valor da AUC.

Em ambos os modelos o número de árvores foi fixado em 1.000. Conforme Boehmke e Greenwell (2019), o número de árvores precisa ser suficientemente grande para estabilizar a taxa de erro; e uma boa regra é começar com 10 vezes o número de variáveis preditivas. Os autores, no entanto, enfatizam que conforme alguns hiperparâmetros forem sendo ajustados, podem ser necessárias mais ou menos árvores.

As curvas ROC obtida no procedimento de validação cruzada são apresentadas na Figura 8, a seguir.

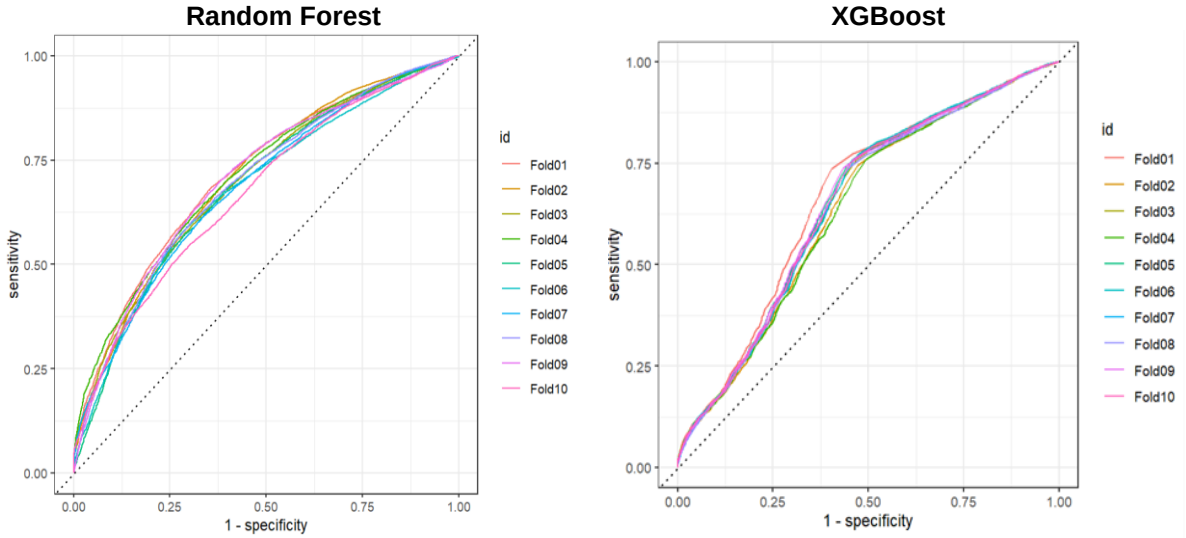

Figura 8. Curvas ROC obtidas no processo de validação cruzada k-fold ($k = 10$)
Fonte: Resultados originais da pesquisa

No caso do algoritmo Random Forest foram ajustados os hiperparâmetros de número de variáveis aleatoriamente consideradas em cada divisão da árvore (“mtry”) e número mínimo de pontos em um nó (“min_n”). Os valores otimizados foram os seguintes: mtry = 6 e min_n=37.

No caso do algoritmo XGBoost, além dos dois hiperparâmetros também presentes no modelo de Random Forest, foram ajustados valores para a profundidade máxima de cada árvore (“tree_depth”), a taxa de aprendizado (“learn_rate”), a redução de perda mínima (“loss_reduction”) e proporções amostradas (“sample_size”). Os valores ajustados com o melhor AUC foram os seguintes: “mtry” = 5, “min_n” = 8, “tree_depth”= 12, “learn_rate” = 0,0000338, “loss_reduction” = 0,771995 e “sample_size” = 0,9737.

O ajuste dos modelos nos dados de treino foi avaliado com o uso da medida de AUC. O modelo de Random Forest apresentou AUC de 0,90 no conjunto de dados de treino,

enquanto o modelo XGBoost apresentou uma AUC de 0,78. Apesar de não ser uma medida adequada para dados desbalanceados, as medidas de acurácia obtidas foram de 0,88 para o Random Forest e de 0,77 para o XGBoost.

Com os modelos ajustados, aplicou-se o classificador aos dados de teste com o objetivo de verificar a performance em dados desconhecidos ao modelo treinado. A matriz de confusão do classificador, aplicado aos dados de teste, é apresentada na Tabela 5 abaixo.

Tabela 5. Informações das matrizes de confusão para os modelos Random Forest e XGBoost

		Classe observada	
		Ocupados	Desocupados ou inativos
Classe prevista pelo modelo random forest	Ocupados	4608	644
	Desocupados ou inativos	434	199
Classe prevista pelo modelo XGBoost	Ocupados	3995	419
	Desocupados ou inativos	1047	424

Fonte: Resultados originais da pesquisa

Conforme apresentado no item Material e Métodos, com base da matriz de confusão, diferentes métricas podem ser calculadas e foram avaliadas para os dois modelos ajustados. Tais métricas são sumarizadas na Tabela 6.

Tabela 6. Sumário de métricas obtidas no conjunto de dados de teste para os modelos Random Forest e XGBoost

	Random Forest	**XGBoost**
AUC/ROC	0,717	0,727
Acurácia	0,817	0,751
Sensibilidade	0,914	0,792
Especificidade	0,236	0,503
Precisão	0,877	0,905
F	0,895	0,845
J index	0,150	0,295
KAP	0,167	0,225
Acurácia balanceada	0,575	0,648

Fonte: Resultados originais da pesquisa

Medidas usuais de acurácia e sensibilidade indicam um melhor desempenho do Random Forest, mas como discutido anteriormente, tais métricas carregam certo viés na previsão da classe majoritária. A Acurácia do Random Forest indica um percentual de quase 82% de previsões corretas entre os que seguem ocupados. Para o XGBoost esse percentual é de 75%.

A medida de sensibilidade indica que, do total de ocupados, 91% foram previstos corretamente pelo algoritmo de Random Forest, enquanto para o XGBoost esse percentual foi de 79%. Já a medida de precisão indica que o XGBoost obteve um desempenho levemente melhor, cometendo menos erros na previsão de ocupados. Esse resultado sinaliza que 90%

do total de ocupados previstos pelo modelo se mostraram corretas, contra 87% do Random Forest. Por sua vez, a medida F, que é dada pela média harmônica de precisão e sensibilidade, indica melhor desempenho do Random Forest, mas viesado pelo valor de sensibilidade, que é bastante influenciado pelo desequilíbrio entre as classes.

Na presente aplicação, a métrica de especificidade assume uma importância diferenciada, por representar o quão bem a transição de ocupado para desocupado/inativo foi prevista. Neste caso, o modelo XGBoost apresenta um desempenho melhor, com especificidade de 0,503, o que sinaliza que o algoritmo previu corretamente 50% dos que se tornaram desocupados/inativos. Com o Random Forest apenas 23% dos desocupados/inativos foram previstos corretamente. Esse resultado também permite destacar que mesmo adotando modelos considerados adequados para lidar com dados desbalanceados, ainda permanecem algumas dificuldades e limitações para a predição da classe minoritária.

Acurácia balanceada, Índice J e KAP mostram resultados de desempenho modesto, mas na comparação entre os algoritmos, indicam uma melhor performance do XGBoost.

Com foco no valor da AUC/ROC a performance dos dois modelos foi muito semelhante, com valores de 0,717 e 0,727 para o Random Forest e o XGBoost, respectivamente. As curvas ROC obtidas com aplicação dos modelos treinados aos dados de teste são apresentadas na Figura 9.

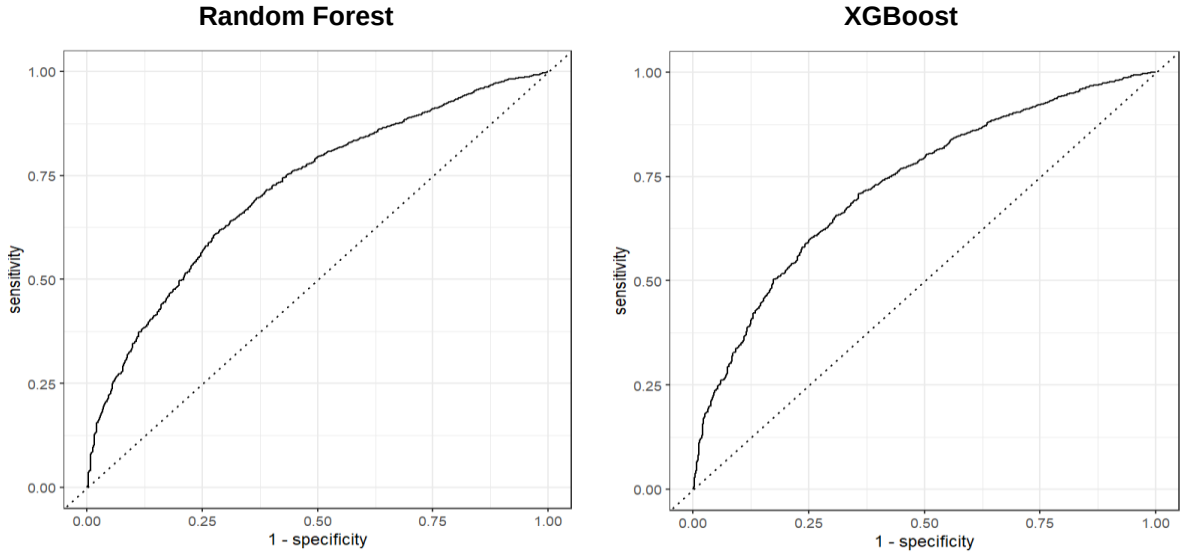

Figura 9. Curvas ROC obtidas com a aplicação do modelo aos dados de teste
Fonte: Resultados originais da pesquisa

Considerações Finais

Construir um classificador preciso a partir de um conjunto de dados desequilibrados é realmente uma tarefa desafiadora. Isso porque algoritmos tradicionais tendem a maximizar a métrica de acurácia geral da previsão, e ao fazer isso tendem a se tornar tendenciosos para a classe majoritária. O presente estudo se propôs a tarefa de treinar modelos de classificação baseados em dois dos principais algoritmos de machine learning: Random Forest e XGBoost. A abordagem aplicada combinou ambos os algoritmos com a técnica SMOTE.

O objetivo da aplicação de tais modelos é prever a situação de indivíduos no mercado de trabalho em momentos de recessão, como a que foi vivenciada no início da pandemia de Covid-19 em 2020. A aplicação da modelagem preditiva por algoritmos de machine learning ao problema em questão possui o potencial de prover informações relevantes para a análise de mercado de trabalho e até mesmo para subsidiar a tomada de decisão política.

No contexto de uma literatura praticamente inexistente, no que diz respeito ao uso de modelos de machine learning para a previsão de transição no mercado de trabalho, o presente trabalho se propõe a estabelecer discussões iniciais e as primeiras referências para este tipo de análise.

Os classificadores treinados apresentaram uma performance razoável, condicionada ao conjunto de dados empregado, ao tratamento destes dados e a abordagem preditiva empregada. Isso foi evidenciado pelo desempenho limitado dos modelos ajustados na previsão da classe minoritária. Além dos métodos empregados, existem diferentes técnicas que podem ser testadas, com bom potencial de aprimorar os resultados obtidos na presente aplicação. Abordagens como o “aprendizado sensível ao custo” ou “modificações de algoritmos” não foram testadas no presente trabalho e podem se mostrar mais adequadas para lidar com o desbalanceamento das classes a serem preditas no problema em questão.

Referências

Aquino. E. M. L. et al. 2020. Medidas de distanciamento social no controle da pandemia de Covid-19: potenciais impactos e desafios no Brasil. Ciência & Saúde Coletiva 25: 2423-2446.

Athey. S. 2017. Beyond prediction: Using big data for policy problems. Science 355(6324): 483-485.

Athey. S. 2018. The impact of machine learning on economics. In: The economics of artificial intelligence: An agenda. University of Chicago Press.

Athey. S; Imbens. G.. 2019. Machine learning methods economists should know about. Arxiv. Disponível em: https://arxiv.org/abs/1903.10075 . Acesso em: 20 out. 2021.

Barbosa. A. L. N. H.; Costa. J. S.; Hecksher. M. 2020. Mercado de trabalho e pandemia da covid-19: Ampliação de desigualdades já existentes? In: Mercado de Trabalho: conjuntura e análise – IPEA 69: 55-63.

Boehmke, B.; Greenwell, B. 2019. Hands-on machine learning with R. Chapman and Hall/CRC, New York, NY, USA.

Breiman, L. 2001. Random forests. Machine learning 45(1): 5-32.

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

Clark. K. B.; Summers. L. H. 1978. Labor force transitions and unemployment. NBER Working Paper 277.

Clark. K. B.; Summers. L. H. 1978. The Dynamics of Youth Unemployment. NBER Working Paper 274.

Fairlie. R. W.; Couch. K.; Xu. H. 2020. The impacts of covid-19 on minority unemployment: first evidence from April 2020 CPS microdata. NBER Working Paper 27.246.

Fávero. L. P.; Belfiore. P. 2017. Manual de análise de dados. Elsevier, Rio de Janeiro, RJ, Brasil.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., Herrera, F. 2018. Learning from imbalanced data sets. Springer, Berlin, Alemanha.

Flinn. C. J.; Heckman. J. J. 1983. Are unemployment and out of the labor force behaviorally distinct labor force states? Journal of labor economics 1(1): 28-42.

Hastie. T.. Tibshirani. R.; Friedman. J.. 2009. The elements of statistical learning: data mining, inference and prediction. Springer, New York, NY, USA.

He, H. e MA, Y. 2013. Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons, Hoboken, New Jersey, USA.

IBGE. Instituo Brasileiro de Geografia e Estatística. 2014. Pesquisa Nacional por Amostra de Domicílios Contínua: Notas Metodológicas. Volume 1. Rio de Janeiro. Disponível em: https://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_continua/Notas_metodologicas/notas_metodologicas.pdf Acesso em: 10 out. 2021.

James, G.; Witten, D.; Hastie, T.; Tibshirani, R. 2021. An introduction to statistical learning. Springer, New York, NY, USA.

Jones. S. R. G.; Riddell. W. C.. 2006. Unemployment and nonemployment: heterogeneities in labor market states. The Review of Economics and Statistics 88 (2): 314-323.

Jones. S. R. G; Riddell. W. C.. 1999. The measurement of unemployment: An empirical approach. Econometrica 1999: 147-161.

Krawczyk, Bartosz. 2016. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence 5 (4): 221-232.

Kuhn, M., e Johnson, K. 2013. Applied predictive modeling. Springer, New York, NY, USA.

Moraes. R. F. 2020. Medidas legais de incentivo ao distanciamento social: comparação das políticas de governos estaduais e prefeituras das capitais no Brasil. Ipea (Nota Técnica. n. 16). Disponível em: https://is.gd/JIxLs6 . Acesso em: 15 out. 2021.

Mullainathan. S.; Spiess. J.. 2017. Machine learning: an applied econometric approach. Journal of Economic Perspectives 31(2): 87-106.

Organização Mundial da Saúde [OMS]. 2020. WHO Director-General's opening remarks at the media briefing on COVID-19-11 March 2020. Disponível em: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. Acesso em: 15 out. 2021.

Reis. M.; Aguas. M.. 2014. Duração do desemprego e transições para o emprego formal, a inatividade e a informalidade. Economia Aplicada 18: 35-50.

Santos. R. T.; Monsueto. S. E.; Varella. A. C. do N. 2021. Quem fica desempregado primeiro? Uma análise de transição. Economia e Sociedade 30: 447-466.