Seleção de Características para Monitoramento de Variáveis Importantes para Segurança Alimentar no Estado do Ceará

Ícaro L. Rodrigues[1], Luiza C. A. Pacheco[1], Josué M. Hinrichs[1], Adilio J. Freitas[1], José Luciano M. Neto[1,2], Antonio Rafael Braga[1,3], Danielo G. Gomes[1]

[1]Grupo de Redes de Computadores, Engenharia de Software e Sistemas (GREat)
Universidade Federal do Ceará (UFC), Fortaleza - CE

[2]Ciências da Computação – Campus Iguatu,
Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), Iguatu-CE

[3]Redes de Computadores – Campus Quixadá,
Universidade Federal do Ceará (UFC), Quixadá-CE

[icarodelima, luizaclara, adiliojfreitas]@alu.ufc.br
josue2001marinho@gmail.com, luciano.neto03@aluno.ifce.edu.br
[rafaelbraga, danielo]@ufc.br

Abstract. *The Brazilian state of Ceará presented a detrimental context in its food insecurity (FI) rate in 2023, with 35% of the total population experiencing some level of FI. This study aims to identify a subset of the most indicative variables regarding FI in Ceará, with the objective of improving public policies to combat hunger in the state. For this purpose, data from the Food Security module of the Pesquisa Nacional por Amostra de Domicílios Contínua (PNADC) 2023 were used. Five feature selection techniques were applied to a set of pre-processed variables, and the 18 most frequent variables were selected, with Education and Income/Employment categories standing out.*

Resumo. *O estado do Ceará apresentou um contexto desfavorável em sua taxa de insegurança alimentar (IA) em 2023, com 35% da população total apresentando algum nível de IA. Este artigo objetiva determinar um subconjunto de variáveis de maior relevância com relação a IA no Ceará para aprimorar políticas públicas de combate à fome no Estado. Para isto, foram utilizados dados do módulo de Segurança Alimentar da Pesquisa Nacional por Amostra de Domicílios Contínua (PNADC) de 2023. Aplicamos 5 técnicas de seleção de características em um conjunto de variáveis pré-processadas e então foram selecionadas as 18 variáveis mais frequentes, nas quais se destacaram as categorias de Educação e Renda/Emprego.*

1. Introdução

No Brasil, em 2022, cerca de 67,8 milhões de pessoas viviam abaixo da linha da pobreza, enquanto 12,7 milhões estavam em situação de extrema pobreza. A região Nordeste concentrava 54,6% da população em situação de extrema pobreza [Gomes 2023]. No mesmo ano, estima-se que 41,3% dos domicílios visitados durante o II Inquérito de Insegurança Alimentar no Contexto da Pandemia da Covid-19 (II VIGISAN) se encontravam em Segurança Alimentar (SA), enquanto os outros 58,7% enfrentavam algum tipo de Insegurança Alimentar (IA) [Gandra 2022].

A Lei Orgânica de Segurança Alimentar e Nutricional (LOSAN) define a SA como a garantia do direito de que todo cidadão tenha acesso constante e permanente a alimentos de qualidade, em quantidade adequada, sem prejudicar o acesso a outras necessidades fundamentais [1]. Por outro lado, a IA ocorre quando há algum tipo de preocupação com a aquisição, redução na qualidade ou escassez dos alimentos entre os membros de uma família. No Brasil, a IA é causada principalmente pela dificuldade de acesso à alimentação, fator que está diretamente associado à renda da população e ao preço dos alimentos [Bezerra et al. 2017].

O histórico da IA no Brasil nos últimos 20 anos foi registrado em grande parte pela Pesquisa Nacional por Amostra de Domicílios (PNAD), pela Pesquisa de Orçamentos Familiares (POF) e pela PNAD Contínua (PNADC) [Golgher 2024]. Esses estudos utilizam a Escala Brasileira de Insegurança Alimentar (EBIA) para classificar a IA em três níveis: leve, moderado e grave [Boas 2023]. A região Nordeste, em particular, possui resultados preocupantes de IA [2]. Entre os anos de 2013 e 2018, a região apresentou um aumento significativo nos níveis de IA moderada e grave entre famílias de baixa renda [Cherol et al. 2022]. Ademais, a pesquisa I VIGISAN de 2020 indicou que 30,8% dos domicílios nordestinos estavam em IA moderada ou grave [PENSSAN and II VIGISAN 2021]. Já na II VIGISAN, ocorrido entre 2021 e 2022, 68% dos domicílios da região enfrentavam algum nível de IA [Gandra 2022]. No Estado do Ceará, estima-se que 35% da população estava vivendo em IA em 2023 [Gomes 2023].

Além de seus levantamentos principais, a PNADC elabora também diferentes seções com extensões de outros grupos de variáveis, como o módulo de SA realizado no último trimestre de 2023. Embora esse estudo seja um suplemento muito importante, principalmente no contexto de políticas voltadas ao combate à fome (e.g. Plano Brasil Sem Fome[3], Ceará Sem Fome[4]), ele não é realizado com a mesma periodicidade do PNADC trimestral. Consequentemente, o acesso a variáveis com informações correlacionadas à SA/IA fica limitado à uma frequência anual. Contudo, o módulo de SA inclui todas as variáveis presentes nos PNADC trimestrais. Levando isso em consideração, técnicas de seleção de características ou *Feature Selection* (FS) [Chandrashekar and Sahin 2014] podem ser muito úteis. FS são técnicas ou métodos capazes de determinar as características mais relevantes de determinada variável de interesse em um conjunto de dados. Nesse caso, é possível determinar um subconjunto dessas variáveis que estão disponíveis em periodicidade trimestral e que tenham maior relação com a SA/IA no estado do Ceará.

A partir desse contexto, o presente artigo propõe a utilização de cinco técnicas diferentes de FS com os dados do módulo de SA do PNADC de 2023, monitorando as variáveis correlacionadas à SA/IA no Estado do Ceará e que estejam presentes nos PNADC trimestrais. Destarte, o objetivo deste trabalho é determinar um subconjunto de variáveis que forneça uma visão mais precisa e frequente da situação alimentar, contribuindo para a eficácia de políticas de combate à fome e alinhadas ao Objetivo do Desenvolvimento Sustentável (ODS)[5] número 2: "Fome Zero e Agricultura Sustentável".

[1]https://www.cfn.org.br/index.php/seguranca-alimentar-e-nutricional/

[2]https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/39838-seguranca-alimentar-nos-domicilios-brasileiros-volta-a-crescer-em-2023

[3]https://www.gov.br/mds/pt-br/acoes-e-programas/brasil-sem-fome

[4]https://www.cearasemfome.ce.gov.br/

[5]https://brasil.un.org/pt-br/sdgs

2. Material e Métodos

Os dados utilizados neste estudo são do módulo de SA do PNADC anual do quarto trimestre de 2023 e foram obtidos no site do IBGE[6], aqui referenciado como PNADC/SA. A variável de interesse neste trabalho é a SD17001, a qual indica o nível de SA/IA do morador de acordo com o EBIA. Além disso, utilizou-se como suporte também o dicionário de variáveis do PNADC trimestral - aqui referenciado como PNADC/T. Para acessar o conjunto de dados, foi utilizada a biblioteca PNADcIBGE, da linguagem R. Os dados foram filtrados pelo estado do Ceará e exportados em arquivos CSV (.csv). Os seguintes passos de análise e processamento foram executados em linguagem de programação *Python*.

Inicialmente, fez-se necessário manter apenas as variáveis comuns ao PNADC/T e ao PNADC/SA, reduzindo o escopo do conjunto de dados para contemplar apenas as variáveis de interseção entre esses dois conjuntos. Dentre as estendidas pelo PNADC/SA, apenas a variável de interesse SD17001 foi mantida para servir como rótulo para os dados. Esta variável classifica a situação de SA/IA do indivíduo entrevistado em quatro classes: (1) segurança alimentar, (2) insegurança alimentar leve, (3) insegurança alimentar moderada e (4) insegurança alimentar grave. A partir disso, o conjunto de dados é separado em um vetor de variáveis $\mathbf{X}$ constituído por um subconjunto de variáveis presentes tanto no PNADC/SA quanto no PNADC/T e um vetor de rótulos y correspondendo à classificação de SA/IA dos moradores entrevistados no estado do Ceará. A Figura 1 demonstra como estava a distribuição desta variável entre os moradores entrevistados.

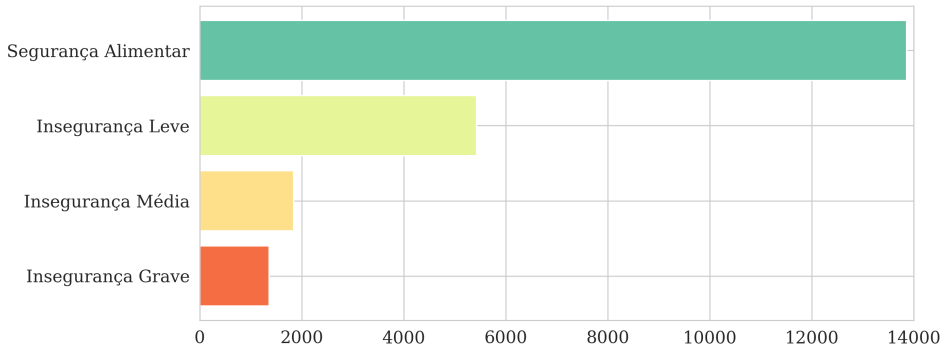

Figura 1. Histograma da SA/IA no Estado do Ceará no ano de 2023.

Posteriormente, realizou-se algumas etapas de pré-processamentos a fim de tornar os dados mais adequados para esta metodologia. O primeiro passo foi a remoção de variáveis que não agregam informações para a análise (e.g. trimestre, número do domicílio). Em seguida, foram removidas todas as variáveis que eram compostas por valores nulos acima de um limiar estabelecido em 70%.

Logo em seguida, foram aplicadas cinco técnicas diferentes de FS. Dentre estas, duas são métodos de Filtro: *Coeficiente de correlação de Pearson* e *Mutual Information*. Dois métodos de Ensemble: *minimum Redundancy - Maximum Relevance* (mRMR) e *Recursive Feature Elimination* (RFE). Além destes, também foi utilizado o algoritmo *Random Forest* (RF), que pode ser empregado para retornar a importância das *características*. Métodos de filtro se diferenciam dos demais pelo fato de realizarem análises

[6]https://www.ibge.gov.br/estatisticas/sociais/trabalho/17270-pnad-continua.html

sem necessitarem de processos de aprendizado. Ou seja, tais métodos desenvolvem a FS com base na relação entre as variáveis $\mathbf{X}$ e a variável de interesse y, enquanto os demais métodos necessitam de algum processo de aprendizado para realizar a FS.

Outra questão levantada durante a metodologia foi o desbalanceamento das classes. Como mostrado na Figura 1, cerca de 60% dos dados pertencem à classe 1. Para prevenir que a FS seja influenciada apenas por variações no comportamento da classe majoritária, foi adicionada uma etapa de *Oversampling* (OS) anterior à FS. Foram escolhidas três técnicas: *Synthetic Minority Oversampling Technique* (SMOTE) e *Adaptive Synthetic Sampling Approach* (ADASYN) [Gosain and Sardana 2017] e também um *Random Oversampling* (ROSE) [Menardi and Torelli 2014], todos disponíveis na biblioteca *imbalanced-learn* [Lemaître et al. 2017]. De formas distintas, cada técnica tem como objetivo gerar amostras sintéticas para as classes minoritárias por meio dos dados já existentes, com o fito de que fiquem balanceadas com a classe majoritária.

Neste trabalho, a etapa de OS foi executada antes da FS no sequenciamento com o objetivo de balancear as classes e que as caraterísticas escolhidas não fossem enviesadas pela classe majoritária. Em outras palavras, que as características escolhidas não refletissem consistentemente apenas a SA e não os demais tipos de IA - que são muito importantes e determinantes uma vez que a IA grave exige ações mais imediatas que a IA leve, por exemplo. Entretanto, isto é algo que é bastante discutido na literatura. Ao passo que alguns trabalhos mostram resultados competitivos com o estado da arte utilizando OS antes da FS [Feng et al. 2023], existem estudos que mostram que a ordem de aplicação no *pipeline* vai depender muito do conjunto de dados, da combinação de cada técnica de OS e FS e também do algoritmo de classificação utilizado [Zhang et al. 2023]. Por este motivo, escolheu-se iterar entre as múltiplas combinações de técnicas de OS e FS.

Diante disso, estabeleceu-se um *pipeline* composto pela etapa de OS em série com a etapa de FS. A cada nova iteração uma nova combinação do *pipeline* era executada e seu resultado armazenado. A primeira etapa alternava entre SMOTE, ADASYN, ROSE além de uma configuração extra sem nenhum OS, ou seja, com os dados puros e desbalanceados. A segunda etapa alternava entre as cinco técnicas de FS. Ao final das iterações, as 20 combinações foram realizadas. Em seguida, foram analisadas as saídas de cada uma das combinações e então foram contabilizadas quantas vezes cada variável foi selecionada. Em outras palavras, a frequência em que as variáveis foram escolhidas entre todas as saídas do *pipeline*. Desta forma, obteve-se uma a lista ordenada das variáveis mais frequentes, considerando todas as combinações das técnicas de OS com as técnicas de FS utilizadas neste trabalho. Assim buscando alcançar o maior nível de confiabilidade e robustez nesta seleção final. Categorias foram adicionadas como forma de organizar as variáveis, contudo, embora tenham sido inferidas através do dicionário e da sua subdivisão, não representam uma categorização oficial do PNADC.

3. Resultados

Ao todo, 56 variáveis distintas foram contadas pelo menos uma vez. As 20 variáveis mais selecionadas apresentaram uma contagem maior ou igual a 9. A Tabela 1 mostra a lista destas 20 variáveis ordenadas pela frequência na qual foram selecionadas. É possível analisar que as variáveis de Renda/Emprego são as mais frequentes. Entretanto, em um recorte das cinco principais variáveis, a categoria Educação é a que mais se destaca.

Tabela 1. Lista das 20 variáveis mais frequentes entre as variáveis selecionadas.

		Informações	
Variável	Contagem	Descrição (resumida)	Categoria
VD3005	20	Anos de estudo	Educação
VD4020	18	Rendimento mensal efetivo*†	Renda/Emprego
VD3004	17	Nível de instrução mais elevado alcançado	Educação
VD3006	17	Grupos de anos de estudo	Educação
V403412	17	Rendimento bruto mensal habitual	Renda/Emprego
V2009	16	Idade do morador	Moradores
VD4017	13	Rendimento mensal efetivo⋆†	Renda/Emprego
V2001	13	Número de pessoas no domicílio	Moradores
V403411	12	Faixa do último rendimento/retirada	Renda/Emprego
V403311	12	Faixa do rendimento/retirada habitual	Renda/Emprego
V3009A	12	Curso mais elevado que frequentou	Educação
VD4019	11	Rendimento mensal habitual*†	Renda/Emprego
VD4016	11	Rendimento mensal habitual⋆†	Renda/Emprego
VD2003	11	Número de componentes do domicílio	Moradia
VD2004	11	Espécie da unidade doméstica	Moradia
V2010	11	Cor ou raça	Moradores
VD4035	10	Horas efetivamente trabalhadas*†	Renda/Emprego
V4077	9	Se conseguisse trabalho, estaria apto?	Renda/Emprego
V4039C	9	Horas trabalhadas na semana⋆	Renda/Emprego
V1023	9	Tipo de área	Moradia

* Todos os trabalhos. ⋆ Trabalho principal. † Para pessoas de 14 anos ou mais de idade.

A Figura 2 apresenta gráficos da distribuição das classes de SA/IA para as variáveis mais relevantes relacionadas a educação. Em 2(a) temos a variável que foi selecionada com mais frequência e que corresponde aos anos de estudo do morador (VD3005). É possível notar pequenas variações nas proporções de SA/IA até os 12 anos de estudo. Acima de 12 anos de estudo, observa-se uma considerável redução, especialmente na IA grave, e um aumento igualmente considerável na SA. Além disso, nas Figuras 2(b) e 2(c) temos, respectivamente, as variáveis VD3004 e VD3006. O padrão observado nelas é análogo: pequenas variações até perceber a extremidade mais a direita, onde não há quase nenhum registro de IA grave e um aumento considerável dos moradores em SA. De acordo com a Figura 2 a educação pode ter um impacto positivo sobre a SA, mas apenas quando acima de um certo limiar (e.g. ensino médio completo, ou mais de 12 anos de estudo).

Acerca das variáveis relacionadas a Renda/Emprego, destacam-se as variáveis VD4020 (Figura 3) e V403411 (Figura 4). A Figura 3 mostra que a maioria dos moradores em SA possuíam rendimentos concentrados entre mil e 10 mil reais, enquanto a renda de pessoas em IA grave estava em maior parte abaixo dos mil reais de rendimento efetivo mensal. Ademais, a Figura 4 mostra de forma visualmente clara o aumento na proporção de moradores em SA conforme a faixa de rendimento aumenta. É possível ainda observar uma tendência de queda dos graus de IA.

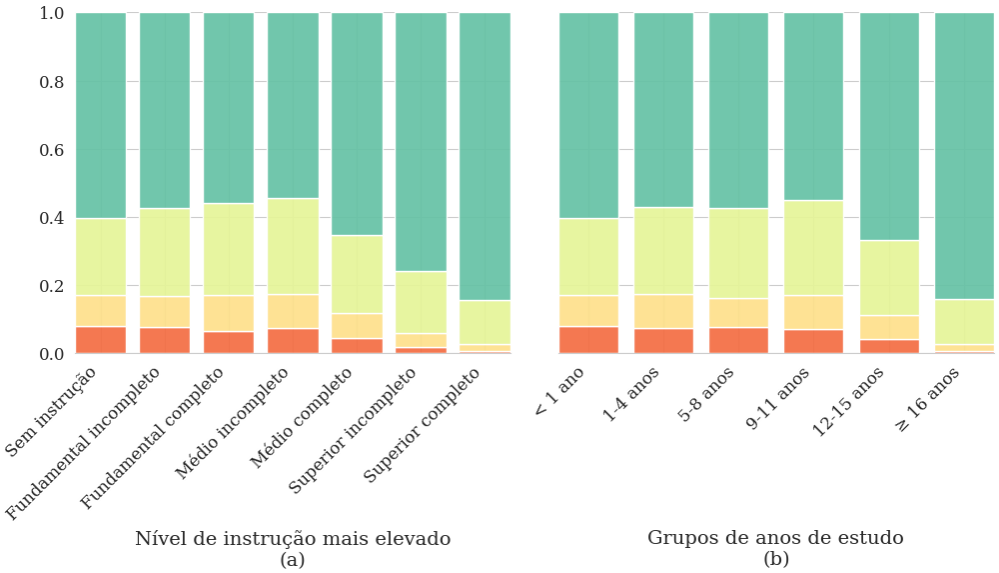

Figura 2. Proporção da SA/IA para as principais variáveis relacionadas à Educação (pessoas de 5 anos ou mais e padronizado para o Ensino fundamental com duração de 9 anos). (a) VD3004 - Nível de instrução mais elevado alcançado e (b) VD3006 - Grupos de anos de estudo.

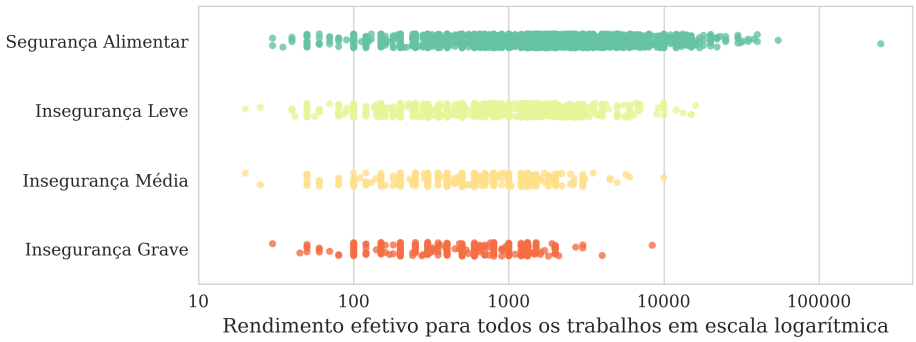

Figura 3. Distribuição da SA em relação a variável VD4020 - Rendimento mensal efetivo de todos os trabalhos (pessoas acima de 14 anos) em escala logarítmica.

Dentre as demais categorias, destaca-se a variável V2010 - Cor ou raça. Dela é possível visualizar que a maior parte da população cearense é de pardos (67%), seguido de brancos (25%) e pretos (6%), enquanto amarelos e indígenas representam menos de 1% dos residentes do estado. Ainda assim, a população branca é a que mais vive em SA (67,7%) e também a que menos está em IA (5,3%). Menos de 60% dos pretos e pardos estão em SA e mais de 6% vivem em IA. Enquanto isso, a variável V1023 - Tipo de área mostrou que as áreas fora da capital e região metropolitana tiveram o maior número de pessoas em IA grave. Já a maioria das pessoas residentes da região metropolitana de Fortaleza encontravam-se em SA.

4. Conclusão

Sabe-se que o PNADC/SA é uma extensão muito importante para a compreensão de questões referentes à fome no Brasil. Entretanto, trata-se de um estudo de periodicidade anual e, consequentemente, o acesso às variáveis diretamente ligadas à SA fica limitada a

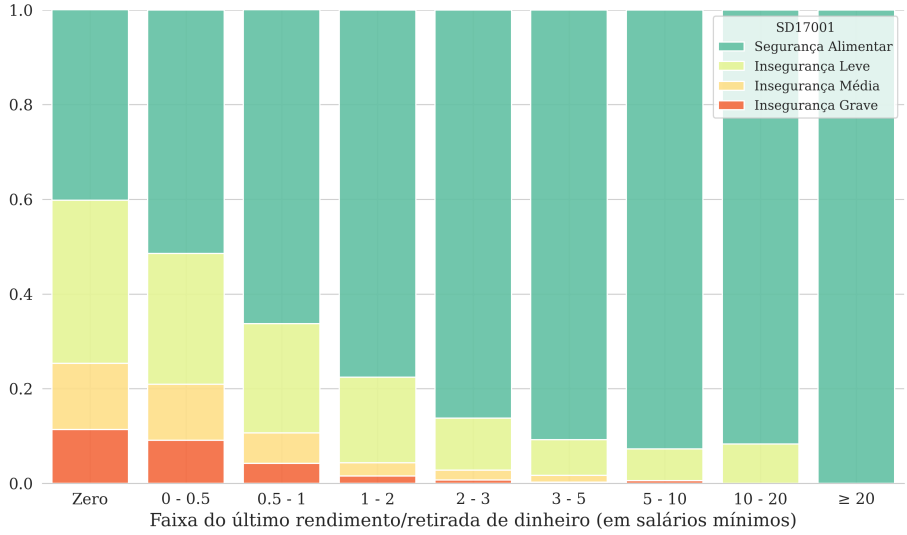

Figura 4. Distribuição da SA em relação a variável V403411, correspondente ao número da faixa do rendimento/retirada em dinheiro no mês de referência

este período. Por meio da análise de dados e da aplicação de um sumário de técnicas de FS, obteve-se as 20 variáveis (Tabela 1) que possuem maior importância para a variável que categoriza os níveis de SA/IA, e que estão presentes nos PNADC trimestrais. Esta contribuição permite um direcionamento melhor de variáveis-chave acerca da fome no estado do Ceará e que podem ser monitoradas com maior frequência. Desta maneira, possibilitando tomadas de decisão mais ágeis e políticas públicas mais assertivas e focadas para combater a fome. Além disso, a metodologia apresentada fornece possibilidades de aplicações para outros estados e regiões do Brasil.

Trabalhos futuros podem avaliar o desempenho individual de cada técnica de FS utilizada neste artigo. Por exemplo, o mRMR, à priori, aparenta fornecer uma seleção promissora, na medida em que já busca eliminar variáveis correlacionadas entre si. Além disso, é imprescindível ressaltar a possibilidade de aplicação de técnicas e modelos de *machine learning* para monitorar essas variáveis-chave com o objetivo de alertar os poderes públicos para possíveis situações de risco na questão da fome. Também cabe destacar a carência de dados de outras edições do PNADC/SA, uma vez que o primeiro foi realizado em 2023. Espera-se que, a partir de uma maior quantidade de dados de anos subsequentes, seja possível desenvolver análises consideravelmente mais robustas e confiáveis. Ademais, o IBGE não divulga os dados a nível municipal, o que possibilitaria visualizações adicionais e tomadas de decisão mais assertivas pelos gestores públicos.

Agradecimentos

Danielo G. Gomes agradece ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de produtividade (processo 311845/2022-3) e à FUNCAP pelo apoio financeiro na execução do projeto Cientista Chefe da Transformação Digital do Estado do Ceará (processo 06681109/2023). Ícaro de Lima Rodrigues também agradece o apoio do CNPq (processo 140696/2023-7).

Referências

Bezerra, T. A., Olinda, R. A. d., and Pedraza, D. F. (2017). Insegurança alimentar no brasil segundo diferentes cenários sociodemográficos. *Ciência & Saúde Coletiva*, 22:637.

Boas, L. G. V. (2023). A escala brasileira de insegurança alimentar (ebia) e as principais condicionantes da (in) segurança alimentar no brasil. *Geoconexões*, 1(15):114–134.

Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. *Computers & electrical engineering*, 40(1):16–28.

Cherol, C. C. d. S., Ferreira, A. A., Lignani, J. d. B., and Salles-Costa, R. (2022). Regional and social inequalities in food insecurity in brazil, 2013-2018. *Cadernos de Saúde Pública*, 38(12).

Feng, F., Li, K.-C., Yang, E., Zhou, Q., Han, L., Hussain, A., and Cai, M. (2023). A novel oversampling and feature selection hybrid algorithm for imbalanced data classification. *Multimedia Tools and Applications*, 82(3):3231–3267.

Gandra, A. (2022). 2º inquérito nacional sobre insegurança alimentar no contexto da pandemia da covid-19 no brasil: Pesquisa aponta que fome atinge 33, 1 milhões de pessoas no país. *Agência Brasil, Rio de Janeiro*, 8:2022–06.

Golgher, A. B. (2024). Food insecurity in brazil by household arrangements and characteristics between 2004 and 2022. *Cadernos de Saúde Pública*, 40(5).

Gomes, I. (2023). Pobreza cai para 31,6% da população em 2022, após alcançar 36,7% em 2021. https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/38545-pobreza-cai-para-31-6-da-populacao-em-2022-apos-alcancar-36-7-em-2021. Acesso em: 22 de julho 2024.

Gosain, A. and Sardana, S. (2017). Handling class imbalance problem using oversampling techniques: A review. In *2017 international conference on advances in computing, communications and informatics (ICACCI)*, pages 79–85. IEEE.

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

Menardi, G. and Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*, 28:92–122.

PENSSAN and II VIGISAN (2021). Insegurança alimentar e covid-19 no brasil: inquérito nacional sobre insegurança alimentar no contexto da pandemia da covid-19 no brasil. *Belo Horizonte: Instituto Vox Populi*.

Zhang, C., Soda, P., Bi, J., Fan, G., Almpanidis, G., García, S., and Ding, W. (2023). An empirical study on the joint impact of feature selection and data resampling on imbalance classification. *Applied Intelligence*, 53(5):5449–5461.