ACURÁCIA DE CLASSIFICAÇÃO DE SISTEMAS INTEGRAÇÃO LAVOURA-PECUÁRIA BASEADA EM MACHINE LEARNING E BALANCEAMENTO DE CLASSES

José Galdino de Oliveira Júnior [1], Stanley Robson de Medeiros Oliveira [2], Gleyce Kelly Dantas Araújo Figueiredo [3]

[1] Doutorando em Eng. Agrícola, Laboratório de Geoprocessamento, FEAGRI/UNICAMP, Campinas - SP, dinojr95@gmail.com;
[2] Pesquisador, Embrapa Agricultura Digital, EMBRAPA, Campinas - SP;
[3] Profa. Dra., Laboratório de Geoprocessamento, FEAGRI/UNICAMP, Campinas - SP

Apresentado no
LII Congresso Brasileiro de Engenharia Agrícola - CONBEA 2023
18 a 21 de outubro de 2023 – Ribeirão Preto - SP, Brasil

RESUMO: O emprego conjunto de dados de sensoriamento remoto e tecnologias de inteligência artificial possibilitam a extração de informações importantes sobre a superfície terrestre em curto prazo. Porém, tal processo depende de uma eficiente etapa de pré-processamento digital dos dados amostrais. Logo, objetivo deste trabalho é checar a acurácia de quatro algoritmos classificadores para a classificação de áreas de sistemas Integração Lavoura-Pecuária (ILP) e avaliar a eficiência da aplicação de etapas prévias de manipulação dos dados (análise de redução de dimensionalidade das variáveis analisadas e o uso do filtro SMOTE para balanceamento de classes minoritárias). Os resultados encontrados não somente destacaram o *Random Forest* como melhor classificador entre os demais (com valores de índice Kappa, Acurácia Global, *Precision*, *Recall* e *F1-Score* iguais à 0,66, 72,69%, 0,73, 0,73 e 0,73, respectivamente), como também demonstrou que das 22 variáveis iniciais, somente 11 eram necessárias à classificação (as bandas B2, B4, B6, B7, B8A e B12 e os índices NDRE1, SAVI, VARI, NDVI e S2REP). Tal estudo também revelou a boa aplicabilidade do filtro SMOTE para o aumento da acurácia do mapeamento agropecuário.

PALAVRAS-CHAVE: Sistema ILP, sensoriamento remoto, Random Forest.

CLASSIFICATION ACCURACY OF SYSTEMS CROP-LIVESTOCK INTEGRATION BASED ON MACHINE LEARNING AND CLASS BALANCING

ABSTRACT: Remote sensing data and artificial intelligence technologies enable extracting meaningful information about the Earth's surface shortly. However, such a process depends on an efficient step of digital pre-processing of sample data. Therefore, the objective of this work is to check the accuracy of four classification algorithms for classifying areas of Integrated Crop-Livestock (ICL) systems and to evaluate the efficiency of the application of previous data manipulation steps (dimensionality reduction analysis of the analyzed variables and the use of the SMOTE filter for rebalancing minority classes). The results found not only highlighted Random Forest as the best classifier among the others (with Kappa index values, Global Accuracy, Precision, Recall, and F1-Score equal to 0.66, 72.69%, 0.73, 0.73, and 0.73, respectively) as well as demonstrating that of the 22 initial variables, only 11 were necessary for classification (bands B2, B4, B6, B7, B8A, and B12 and the NDRE1, SAVI, VARI,

NDVI, and S2REP indices). This study also revealed the good applicability of the SMOTE filter for increasing the accuracy of agricultural mapping.

KEYWORDS: ICL system, remote sensing, Random Forest.

INTRODUÇÃO: Informações sobre a mudança de uso e cobertura da terra são de vital importância para diferentes aplicações como manejo ambiental, monitoramento climático e planejamento agropecuário (YANG et al., 2021; EBRAHIMY et al., 2022). Além disso, análises contínuas e acuradas sobre as dinâmicas da superfície terrestre são uma parte essencial para o desenvolvimento de estratégias eficientes de caráter sustentável (LOUKIKA et al., 2021; YANG et al., 2021). Pois, diante do crescimento populacional mundial acelerado nas últimas décadas, a preocupação pela produção de alimentos necessária à garantia da segurança alimentar tem se tornado cada vez mais evidente (MORAES et al., 2018; SEKARAN et al., 2021). Por outro lado, particularmente em relação ao Brasil, almejando a diminuição das emissões dos gases do efeito estufa, o Governo Federal instituiu no ano de 2011 o Plano de Agricultura de Baixo Carbono (Plano ABC), que consistiu no uso de tecnologias que garantissem benefícios ecológicos como a recuperação de pastagens degradadas, maior fixação biológica do nitrogênio e maior ciclagem de nutrientes no solo (NEWTON et al., 2016; KUCHLER et al., 2022; SANTOS et al., 2022). Por meio desse projeto, os sistemas de Integração Lavoura-Pecuária-Floresta (ILPF), mais especificamente os sistemas de Integração Lavoura-Pecuária (ILP), alcançaram ótimos resultados mediante a sua aplicação em áreas com alto nível de degradação do solo presentes nos biomas brasileiros Cerrado, Pampa e Amazônico (BALBINO, BARCELLOS & STONES, 2011; MORAES et al., 2018; SOARES et al., 2020). Contudo, devido a tais sistemas apresentarem rápidas mudanças de uso da terra em escala temporal e à alta presença de nuvens em locais tropicais como o Brasil, tornam-se difíceis as etapas de mapeamento e monitoramento agrícola com boa eficiência. Neste âmbito, com os últimos avanços de tecnologias como o sensoriamento remoto e a inteligência artificial, o uso conjunto de algoritmos de aprendizado de máquina (*Machine Learning* e *Deep Learning*) e de dados de satélites orbitais (como os pertencentes às séries Landsat e Sentinel) se tornou mais eficiente. Isto possibilitou a extração de informações pertinentes em diferentes escalas espaciais e temporais ligadas, principalmente, a tarefas de classificação de uso e cobertura da terra (NABOUREH et al., 2020; LOUKIKA et al., 2021; YANG et al., 2021; EBRAHIMY et al., 2022). Todavia, a acurácia de tal mapeamento permanece atrelada a uma boa amostragem dos dados, pois, ainda são problemas recorrentes a ocorrência de alta similaridade do comportamento espectral entre determinadas classes e o desbalanceamento quantitativo de pontos amostrais (NABOUREH et al., 2020; YANG et al., 2021). Portanto, o objetivo deste estudo é verificar a acurácia de estratégias de classificação (algoritmos classificadores baseados em *Machine Learning*) e de manipulação prévia dos dados (aplicação do filtro SMOTE), visando o aumento da eficácia do mapeamento de áreas de sistema Integração Lavoura-Pecuária.

MATERIAL E MÉTODOS: O estudo foi desenvolvido em uma área pertencente à Fazenda Barbosa – Brejo/MA (3°43’12” S, 42°57’0” O), a qual está localizada em uma porção do MATOPIBA (acrônimo para as siglas dos estados de Maranhão, Tocantins, Piauí e Bahia e destaca-se atualmente no País como sendo a última fronteira agrícola e grande produtora de commodities agrícolas como soja e milho) e encontra-se representada na Figura 1. O clima dessa região é o tropical com inverno seco (Aw), com valores pluviométricos anuais que variam entre 1600 a 1900 mm/ano e temperatura média anual variável entre 20 e 26° C (ALVARES et al., 2013; MIRANDA et al., 2014).

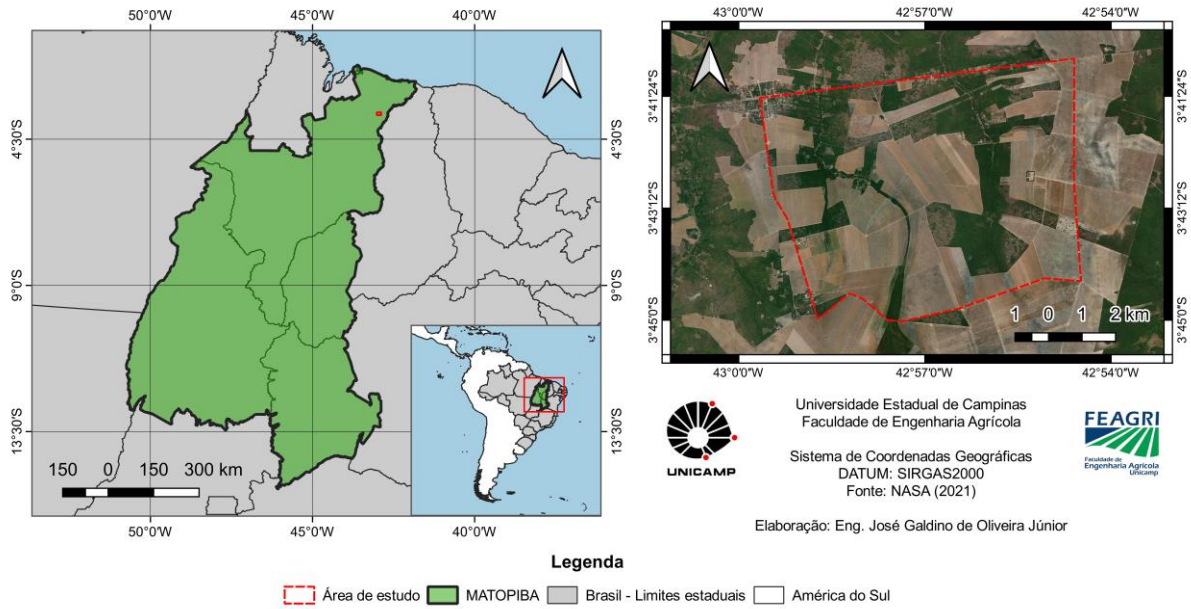


FIGURA 1. Localização espacial da área de estudo.

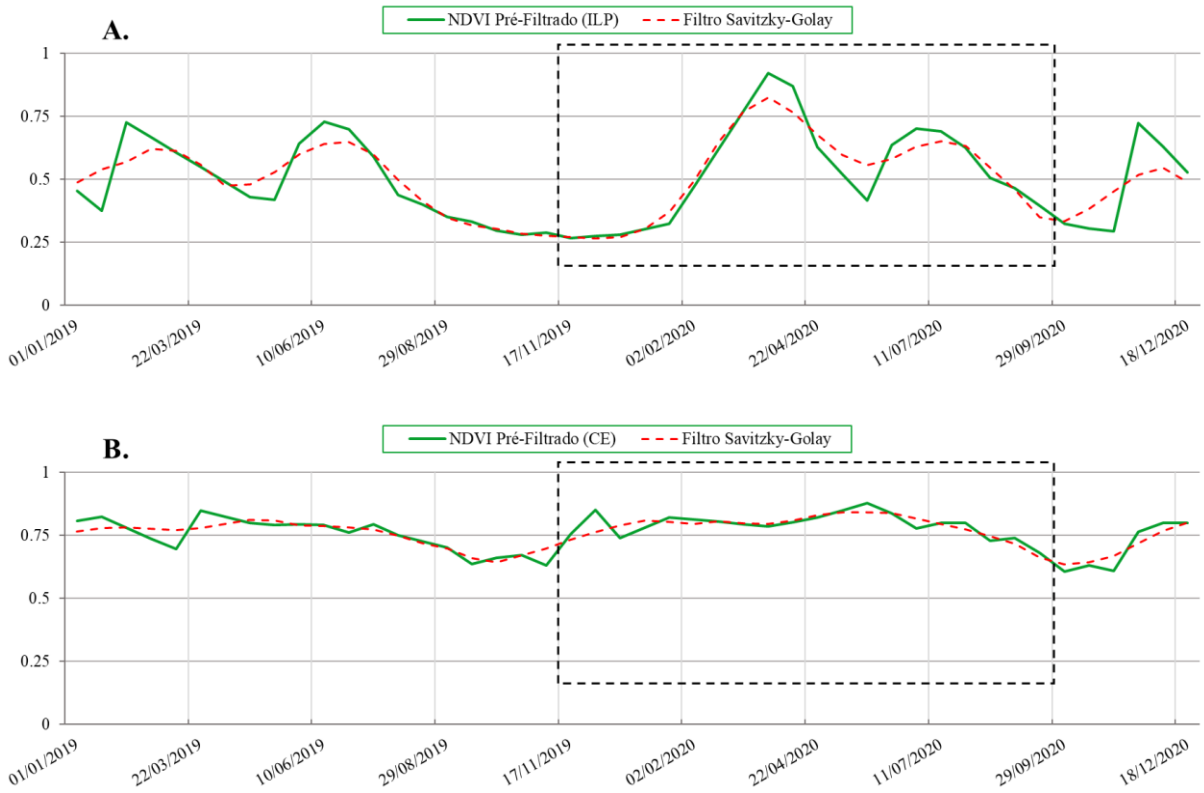
Primeiramente, todas as informações pertinentes às características do sistema ILP empregado na área de estudo foram adquiridas a partir do site da Rede ILPF (https://redeilpf.org.br/). Após isto, uma série temporal de imagens do satélite Sentinel 2, referente a safra de 2019/2020, foi manipulada na API (*Application Programming Interface*) do *Google Earth Engine* que está presente na plataforma virtual do *Google Colab*, através da biblioteca virtual "geemap" (WU, 2020). Ao todo, foram utilizadas neste trabalho 10 bandas espectrais em valores de reflectância de superfície (Level 2A) referentes às regiões do RGB (B2, B3 e B4), *red edge* (B5, B6 e B7), infravermelho próximo (B8 e B8A) e infravermelho médio (B11 e B12) e também 12 índices de vegetação: NDVI (ROUSE et al., 1974); EVI (HUETE et al., 1997); SAVI (HUETE, 1988); MSAVI (MATVIENKO et al., 2022); VARI (GITELSON et al., 2002); MCARI (DAUGHTRY et al., 1999); NDRE1, NDRE2 e NDRE3 (MATVIENKO et al., 2022); SR (BIRTH & MCVEY, 1968; JENSEN, 2011); GNDVI (GITELSON & MERZLYAK, 1998) e S2REP (GUYOT & BARET, 1988). Como etapas de pré-processamento digital dos dados orbitais, nós aplicamos inicialmente um filtro para remoção de nuvens e sombras (algoritmo FMASK) (NABOUREH et al., 2020; YANG et al., 2021; NASIRI et al., 2022). E em seguida, este conjunto de imagens foi reduzido a uma composição temporal baseada na mediana de cada pixel (NASIRI et al., 2022). Posteriormente, foi realizada a extração de um total de 918 pontos amostrais referentes aos tipos de uso e cobertura da terra existentes na área de estudo durante a safra 2019/2020, baseando-se em dados do SATVeg e MapBiomas (ESQUERDO et al., 2020; SOUZA et al., 2020). A Tabela 1 apresenta a quantidade de pontos amostrais para cada tipo de uso e cobertura da terra avaliado neste estudo.

TABELA 1. Informações pertinentes aos dados amostrais utilizados no estudo.

Tipo de uso e cobertura da terra	**Abreviação**	**Quantidade de pontos (pixels) selecionados**
Sistema ILP	ILP	96
Cerrado	CE	163
Pastagem	PA	165
Vegetação nativa	VN	183
Cultura anual	CA	311

Além disso, realizamos também uma etapa de normalização dos valores destes dados amostrais (MATVIENKO et al., 2022). Através do software WEKA (*Waikato Environment for Knowledge Analysis*) versão 3.9.2, realizamos uma análise de redução da dimensionalidade das variáveis, através da técnica da aplicação de *Wrappers*, que consiste em um método conjunto que inclui o uso de algoritmos não-paramétricos como Árvores de Decisão, Redes Neurais ou *Support Vector Machines* para identificar quais variáveis dentro do conjunto de dados são mais relevantes, ou seja, apresentam maior correlação entre si, visando a otimização do processo de classificação almejado. Nesta etapa, utilizamos o algoritmo *Random Forest*, usando 500 árvores para realizar tal processo (RODRIGUEZ-GALIANO et al., 2018). Diante da escolha das variáveis que apresentaram maior importância para tal mapeamento, dividimos os dados amostrais em conjunto de teste (70%) e de validação (30%) e, posteriormente, quatro algoritmos classificadores foram escolhidos para serem avaliados neste estudo: *Random Forest* (RF), *Ada Boosting* (BO), *Bagging* (BA) e *MultiLayer Percepton* (MLP). A acurácia dos algoritmos classificadores foi avaliada sem e com a aplicação de um filtro para balanceamento de classes minoritárias (filtro SMOTE), por meio de métricas estatísticas geradas a partir de matriz de confusão, sendo elas: o Índice Kappa, Acurácia Global, *Precision*, *Recall* e *F1-Score* (DOUZAS et al., 2019; FONSECA et al., 2021; NASIRI et al., 2022).

RESULTADOS E DISCUSSÃO: Mediante a avaliação espaço-temporal dos pontos amostrais, realizada através dos dados do sistema SATVeg e do projeto MapBiomas, nós podemos detectar e diferenciar o comportamento espectral dos tipos de alvos avaliados no estudo, a partir da distinção das variações nas séries temporais do índice NDVI – *Normalized Difference Vegetation Index*, como demonstrado pelas áreas destacadas em coloração preta na Figura 2.



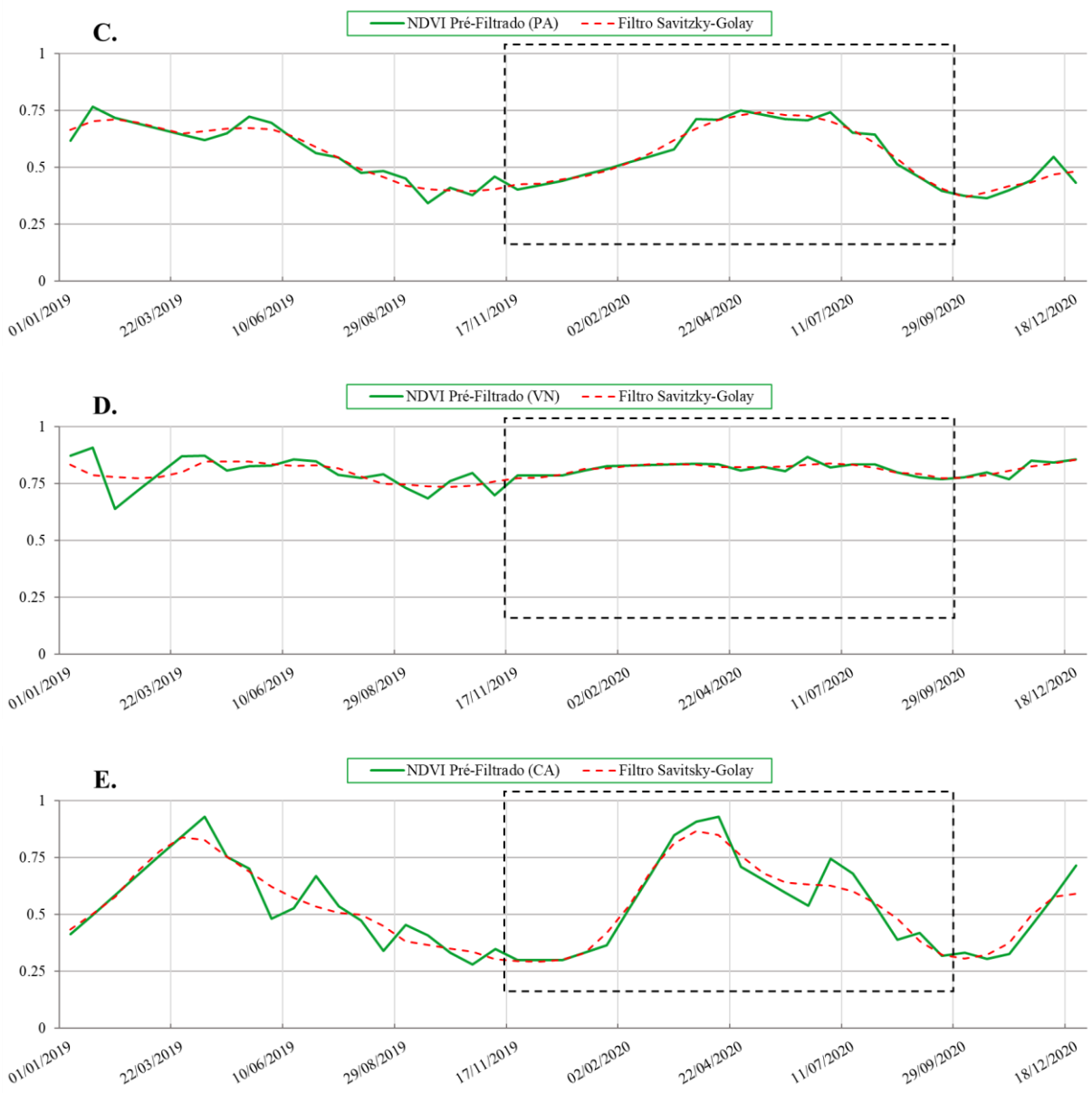


FIGURA 2. Representação temporal do comportamento espectral dos alvos analisados no estudo (ILP, CE, PA, VN e CA), através do índice NDVI obtido através do sistema SATVeg (itens A, B, C, D e E, respectivamente).

A partir da análise de redução de dimensionalidade da base de dados, os resultados demonstraram que as variáveis mais importantes foram as bandas B2, B4, B6, B7, B8A e B12 e os índices NDRE1, SAVI, VARI, NDVI e S2REP. Tais variáveis foram, posteriormente, utilizadas na etapa de classificação de uso e cobertura da terra mediante a aplicação dos quatro algoritmos: *Random Forest*, *Bagging*, *Ada Boosting* e MLP. As matrizes de confusão normalizadas e geradas após a aplicação do filtro SMOTE estão representadas nas Figuras 3A, 3B, 3C e 3D, respectivamente.

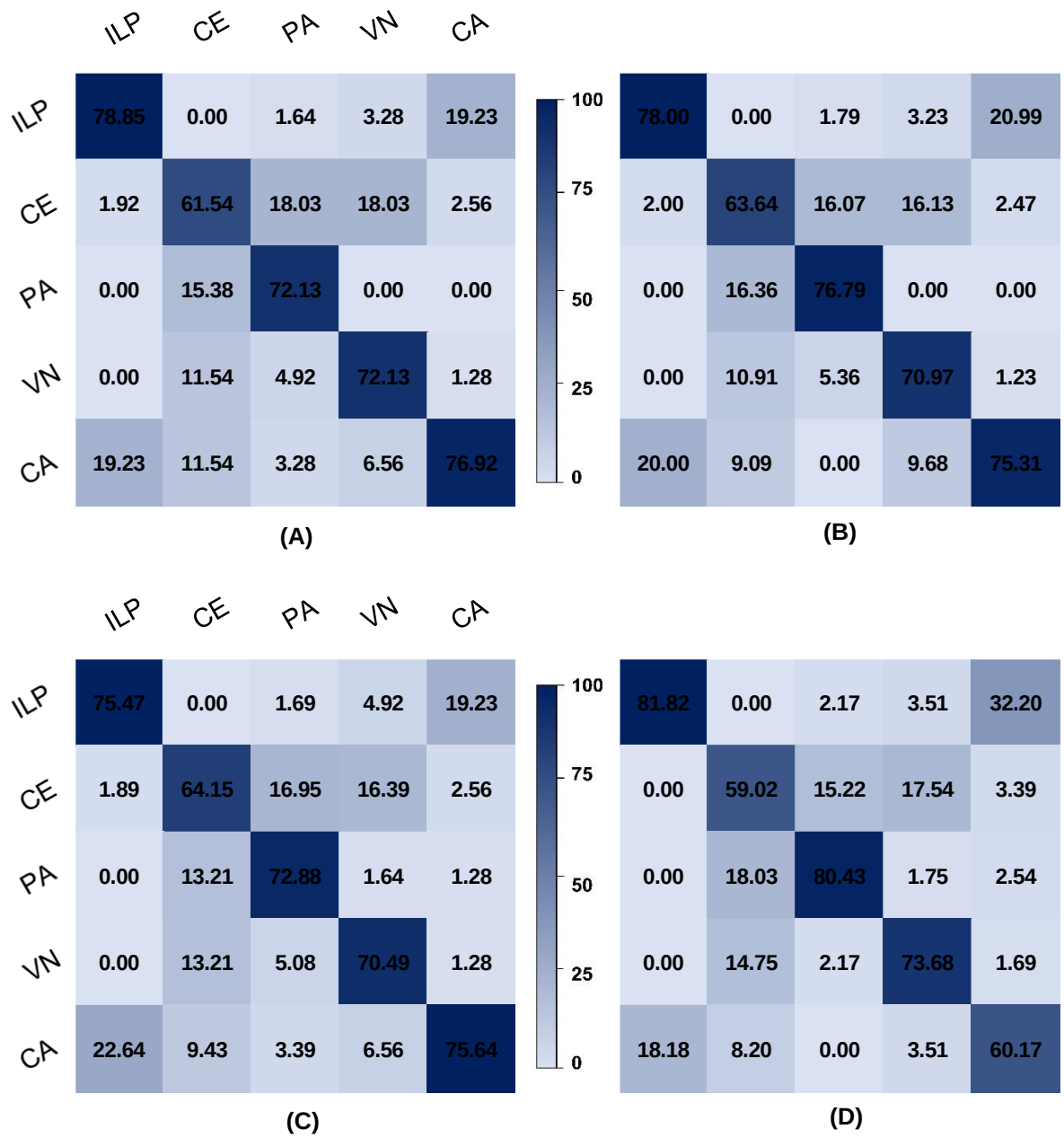


FIGURA 3. Matrizes de confusão geradas a partir dos quatros algoritmos *Random Forest* (A), *Bagging* (B), *Ada Boosting* (C) e MLP (D).

Através desta Figura, percebemos que o algoritmo MLP apresentou a maior acurácia de classificação da classe ILP entre os demais classificadores (81,82%, Figura 3D), porém, em relação às outras classes este algoritmo apresentou maiores erros de comissão. Por outro lado, os algoritmos *Random Forest* e *Bagging* apresentaram os melhores desempenhos de classificação de modo geral, pois, apresentaram menores erros de omissão e maior acerto de classificação. Todavia, em todos os classificadores, houve dois casos de confusão mútua entre as classes mapeadas (Figuras 3A, 3B, 3C e 3D). No primeiro caso, a confusão foi entre as classes de ILP e CA, que pode ser explicado pelo fato de que este sistema de produção integrado somente se difere de um ciclo anual de cultura agrícola pela inserção de pastagem animal, em consórcio ou rotação, durante o período estiagem das chuvas (BALBINO, BARCELLOS & STONES, 2011; MANABE, MELO & ROCHA, 2018). Ou seja, têm-se aí uma alta similaridade de assinaturas espectrais entre tais sistemas de produção que só se altera

minimamente na escala temporal (Figura 2A e 2E), devido aos ciclos mais curtos empregados no sistema de integração Lavoura-Pecuária (MANABE, MELO & ROCHA, 2018; KUCHLER et al., 2022). No segundo caso, houve uma confusão entre a classe de CE e as de PA e VN, demonstrando que a vegetação heterogênea do Cerrado dificultou a classificação de outros tipos de vegetação existentes na área a ser analisada. Na Tabela 2 estão representadas as métricas de avaliação da acurácia de classificação. Onde, confirmamos que mesmo com os erros de classificação, os algoritmos RF e BA alcançaram resultados semelhantes estatisticamente entre si. Apesar disso, *Random Forest* deteve o melhor desempenho de classificação, após aplicação do filtro SMOTE, com valores de índice Kappa, Acurácia Global, *Precision*, *Recall* e *F1-Score* iguais à 0,66, 72,69%, 0,73, 0,73 e 0,73, respectivamente (MAXWELL et al., 2018).

TABELA 2. Métricas de avaliação da acurácia aplicadas no estudo (com o filtro SMOTE).

	Índice Kappa	**Acurácia Global (%)**	***Precision***	***Recall***	***F1 - Score***
RF	**0,61 (0,66)**	**70,55 (72,69)**	**0,73 (0,73)**	**0,70 (0,73)**	**0,68 (0,73)**
BA	0,62 (0,66)	71,27 (73,03)	0,74 (0,73)	0,71 (0,73)	0,69 (0,73)
BO	0,64 (0,65)	72,36 (72,04)	0,75 (0,72)	0,72 (0,72)	0,71 (0,72)
MLP	0,60 (0,58)	70,18(67,10)	0,75 (0,70)	0,70 (0,67)	0,67 (0,66)

Resultados semelhantes a estes também foram encontrados por Naboureh et al. (2020) em seus estudos, demonstrando assim, as vantagens de se aplicar o filtro SMOTE para balancear classes minoritárias. Entretanto, Douzas et al. (2019) reforçaram que o uso deste filtro também pode gerar ruídos no conjunto de dados amostrais, diminuindo a acurácia do mapeamento a depender do algoritmo classificador utilizado, como exemplificado no caso do MLP (Tabela 2). Exigindo-se, portanto, cautela do analista em tal processo.

CONCLUSÕES: Os procedimentos de manipulação prévia dos dados (a técnica de *Wrappers* e a aplicação do Filtro SMOTE) proporcionaram um aumento considerável na eficiência de classificação dos algoritmos *Bagging* e *Random Forest*, destacando-se este último como sendo o que alcançou os melhores resultados após tal ação. Contudo, sugerimos o emprego de algoritmos de *Deep Learning* à classificação de áreas com sistemas ILP em trabalhos futuros, visto que, todos os classificadores de *Machine Learning* alcançaram uma acurácia inferior a 80% neste estudo.

AGRADECIMENTOS: Os autores agradecem ao Programa de Pós-graduação em Engenharia Agrícola da UNICAMP e à CAPES pelos incentivos técnicos e institucionais que viabilizaram o referido trabalho.

REFERÊNCIAS:

ALVARES, C. A. et al. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711-728, 2013.

BALBINO, L. C.; BARCELLOS, A. D. O. & STONES, L. F. **Marco referencial: integração lavoura-pecuária-floresta**. Embrapa: Brasília, 2011.

BIRTH, G. S. & MCVEY G. Measuring the color of growing turf with a reflectance Spectrophotometer. **Agronomy Journal**, v. 60, p. 640-643, 1968.

DAUGHTRY, C. S. et al. Estimating Corn Leaf Chlorophyll Concentration from Leaf and Canopy Reflectance. **Remote Sensing of Environment**, v. 74, n. 2, p. 229-239, 1999.

DOUZAS, G. et al. Imbalanced learning in land cover classification: Improving minority classes' prediction accuracy using the geometric SMOTE algorithm. **Remote Sensing**, v. 11, n. 24, 3040, 2019.

EBRAHIMY, H. et al. Effectiveness of the integration of data balancing techniques and tree-based ensemble machine learning algorithms for spatially-explicit land cover accuracy prediction. **Remote Sensing Applications: Society and Environment**, v. 27, 100785, 2022.

ESQUERDO, J. C. D. M. et al. SATVeg: A web-based tool for visualization of MODIS vegetation indices in South America. **Computers and Electronics in Agriculture**, v. 175, 105516, 2020.

FONSECA, J. et al. Increasing the effectiveness of active learning: Introducing artificial data generation in active learning for land use/land cover classification. **Remote Sensing**, v. 13, n. 13, 2619, 2021.

GITELSON, A. A. et al. Novel algoritms for remote estimation of vegetation fraction. **Remote Sensing of Environment**, v. 80, p. 76-87, 2002.

GITELSON, A. A., & MERZLYAK, M. N. Remote sensing of chlorophyll concentration in higher plant leaves. **Advances in Space Research**, v. 22, n. 5, p. 689-692, 1998.

GUYOT, G. & BARET, F. Utilisation de la haute resolution spectrale pour suivre l'etat des couverts vegetaux. **Spectral Signatures of Objects in Remote Sensing**, p. 279-287, 1988.

HUETE, A. R. A soil-adjusted vegetation index (SAVI). **Remote Sensing of Environment**, v. 25, n. 3, p. 295-309, 1988.

HUETE, A. R. et al. A comparison of vegetation indices over a global set of TM images for EOS-MODIS. **Remote Sensing of Environment**, v. 59, n. 3, p. 440-451, 1997.

JENSEN, J. R. Sensoriamento remoto da vegetação. In: JENSEN, J. R. **Sensoriamento remoto do ambiente**: uma perspectiva em recursos terrestres. Parêntese: São José dos Campos, 2011.

KUCHLER, P. C. et al. Monitoring Complex Integrated Crop–Livestock Systems at Regional Scale in Brazil: A Big Earth Observation Data Approach. **Remote Sensing**, v. 14, n. 7, 1648, 2022.

LOUKIKA, K. N. et al. Analysis of land use and land cover using machine learning algorithms on google earth engine for Munneru river basin, India. **Sustainability**, v. 13, n. 24, 13758, 2021.

MANABE, V. D. MELO, M. R. S., & ROCHA, J. V. Framework for mapping integrated crop-livestock systems in Mato Grosso, Brazil. **Remote Sensing**, v. 10, n. 9, 2018.

MATVIENKO, I. et al. Bayesian Aggregation Improves Traditional Single-Image Crop Classification Approaches. **Sensors**, v. 22, n. 22, 8600, 2022.

MAXWELL, A. E. et al. Implementation of machine learning classification in remote sensing: An applied review. **International Journal of Remote Sensing**, v. 39, n. 9, p. 2784-2817, 2018.

MIRANDA, E. E. et al. **Proposta de delimitação territorial do MATOPIBA**. Nota técnica 1. EMBRAPA. Grupo de Inteligência Territorial Estratégica (GITE). 2014. Disponível em: https://www.embrapa.br/gite/publicacoes/NT1_DelimitacaoMatopiba.pdf. Acesso em: 03 maio 2023.

MORAES, A. et al. **Integrated crop-livestock systems as a solution facing the destruction of Pampa and Cerrado biomes in South America by intensive monoculture systems**. In Agroecosystem Diversity: Reconciling Contemporary Agriculture and Environmental Quality, p. 257-273, 2018.

NABOUREH, A. et al. RUESVMs: An ensemble method to handle the class imbalance problem in land cover mapping using google earth engine. **Remote Sensing**, v. 12, n. 21, 3484, 2020.

NASIRI, V. et al. Land Use and Land Cover Mapping Using Sentinel-2, Landsat-8 Satellite Images, and Google Earth Engine: A Comparison of Two Composition Methods. **Remote Sensing**, v. 14, n. 9, 1977, 2022.

NEWTON, P., et al. Overcoming barriers to low carbon agriculture and forest restoration in Brazil: The Rural Sustentável project. **World Development Perspectives**, v. 4, p. 5-7, 2016.

RODRIGUEZ-GALIANO, V. F. et al. Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. **Science of the Total Environment**, v. 624, p. 661-672, 2018.

ROUSE, J. W. et al. **Monitoring vegetation systems in the great plains with ERTS**, Proceedings: Third Earth Resources Technology Satellite-1 Symposium, Greenbelt: NASA SP-351, p. 3010-3017, 1974.

SANTOS, C. O. et al. Assessing the Wall-to-Wall Spatial and Qualitative Dynamics of the Brazilian Pasturelands 2010–2018, Based on the Analysis of the Landsat Data Archive. **Remote Sensing**, v. 14, n. 4, 1024, 2022.

SEKARAN, U. et al. Role of integrated crop-livestock systems in improving agriculture production and addressing food security – A review. **Journal of Agriculture and Food Research**, v. 5, 100190, 2021.

SOARES, M. B. et al. Integrated production systems: An alternative to soil chemical quality restoration in the Cerrado-Amazon ecotone. **Catena**, v. 185, 104279, 2020.

SOUZA, C. M. et al. Reconstructing three decades of land use and land cover changes in Brazilian biomes with Landsat archive and earth engine. **Remote Sensing**, v. 12, n. 17, 2735, 2020.

WU, Q. geemap: A Python package for interactive mapping with Google Earth Engine. **Journal of Open Source Software**, v. 5, n. 51, 2305, 2020.

YANG, Y. et al. Testing accuracy of land cover classification algorithms in the Qilian mountains based on gee cloud platform. **Remote Sensing**, v. 13, n. 24, 5064, 2021.