VIII Simpósio de Iniciação Científica 2025

"Do conhecimento acadêmico à transformação sustentável: inovação com validação científica"

MAPEAMENTO DA CAFEICULTURA DO SUL DE MINAS UTILIZANDO DADOS DO BRAZIL DATA CUBE E INTELIGÊNCIA ARTIFICIAL

Clara L. Moreno[1] (IC), Vanessa C. O. Souza (PQ)[1]
[1]*Universidade Federal de Itajubá*

Palavras-chave: Análise temporal. Cafeicultura. Cubo de dados. Sensoriamento remoto.

Introdução

O Brasil é o maior produtor e exportador mundial de café, com área total destinada à produção de café no país em 2025 totalizando 1,86 milhões de hectares (CONAB, 2025). Dentre os estados brasileiros, Minas Gerais destaca-se com a maior área destinada à produção cafeeira em 2025, sendo 1,08 milhões de hectares (CONAB, 2025). A macrorregião do Sul e Centro-Oeste de Minas consolida-se como o principal polo produtor do estado, respondendo por cerca de metade da produção mineira com uma colheita estimada em 12,4 milhões de sacas para 2025 (CONAB, 2025).

Dada a relevância da cafeicultura para a região, o mapeamento e monitoramento automatizado das áreas cultivadas representa uma ferramenta importante para o planejamento do setor (ALVES; VOLPATO; CAMPOS, 2021). Para solucionar as dificuldades de mapeamento apontadas pela revisão de Hunt et al. (2020), métodos que analisam as diferentes fases fenológicas do café através de séries temporais de imagens e inteligência artificial têm sido explorados, como nos trabalhos de Souza (2015) e Santos (2024).

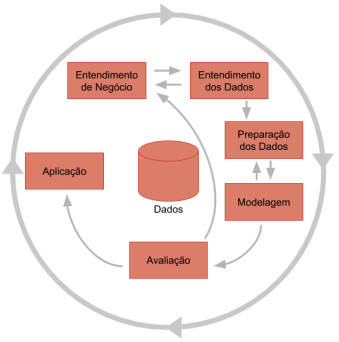
Assim, essa pesquisa tem por objetivo a identificação de áreas cafeeiras no sul de Minas Gerais a partir de dados de sensoriamento remoto, algoritmos de *Machine Learning* (ML) e *Deep Learning* (DL).

A metodologia adotada foi a CRISP-DM (*Cross Industry Standard Process for Data Mining*), uma vez que esta é referência para o desenvolvimento de projetos de mineração de dados e descoberta de conhecimento (MARTÍNEZ-PLUMED et al., 2021). O fluxo experimental partiu da obtenção de séries temporais de imagens de satélite por meio da plataforma Brazil Data Cube (FERREIRA et al., 2020). A classificação desses dados foi realizada utilizando a biblioteca open-source *sits* em R (SIMOES, 2021). E, por fim, a avaliação e comparação dos desempenhos se deu através do uso da biblioteca *scikit-learn* em Python e a inspeção visual dos mapas gerados. Cada etapa deste procedimento será detalhada adiante, na seção de Metodologia.

Metodologia

Como ilustrado na Figura 1, o processo da CRISP- DM compreende seis fases cíclicas flexíveis que permitem o retrocesso às fases anteriores quando há um novo aprendizado (WIRTH; HIPP, 2000). Neste trabalho foram aplicadas cinco fases, de Entendimento do Negócio a Avaliação, como descrito a seguir.

Figura 1 – Fases da CRISP-DM



Fonte: Adaptado de (WIRTH; HIPP, 2000).

Entendimento do Negócio

Nessa fase foram definidos o objetivo do projeto, identificação de áreas cafeeiras no sul de Minas Gerais, e a tarefa de mineração de dados a ser realizada, a classificação de áreas cafeeiras a partir de dados de sensoriamento remoto, algoritmos de ML e DL.

Entendimento dos Dados

A área de estudo é composta pelos municípios de Bom Sucesso-MG, com área total de 705,046 km², e Oliveira-MG, com área total de 896,494 km². Arquivos vetoriais contendo a delimitação da área dos municípios e os mapas de referência das plantações de café em 2018 foram fornecidos pela Empresa de Assistência Técnica e Extensão Rural do Estado de Minas Gerais (EMATER).

Os cubos de dados de imagens de satélite são estruturas tridimensionais que integram séries temporais de imagens georreferenciadas, nas quais duas dimensões representam o espaço (latitude e longitude) e a terceira representa o tempo (FERREIRA et al., 2020). A coleção

de imagens utilizada nesse estudo foi a S2-16D-2 do satélite Sentinel-2 caracterizada por um intervalo regular de 16 dias entre observações (FERREIRA et al., 2020).

Para compor os cubos de dados, foram utilizadas imagens das bandas espectrais B02 (*Blue*), B03 (*Green*), B04 (*Red*), B06 (*Red Edge*) e B08 (*Near-infrared*). Também foram incluídos os Índices de Vegetação (IVs) *Enhanced Vegetation Index* (EVI), *Normalized Difference Vegetation Index* (NDVI), *Modified Soil Adjusted Vegetation Index* (MSAVI), *Normalized Difference Water Index* (NDWI) e *Visible Atmospherically Resistant Index* (VARI). As fórmulas para o cálculo desses índices podem ser consultadas em Zhu et al. (2024).

Preparação dos Dados

Para realizar as análises, utilizou-se um notebook VAIO FE14 com Windows 11 Home, processador AMD Ryzen 5 5500U, placa de vídeo RX Vega 7 (4000/5000), 16 GB de RAM e 256 GB de SSD. E o ambiente de desenvolvimento principal foi o RStudio 2024.12.

Como o cafeeiro possui um ciclo bienal (CONAB, 2025), foram gerados cubos de dados referentes a um período de dois anos, 01/01/2017 a 31/12/2018, com a função *sits_cube(),* totalizando 46 imagens. Os cubos criados continham bandas espectrais e IVs, provenientes do BDC (EVI e NDVI) ou calculados separadamente (NDWI, VARI, MSAVI) pela função *sits_apply()*, conforme a Tabela 1:

Tabela 1: Datasets utilizados

Dataset	Bandas espectrais e índices de vegetação
1	R, G, B, NIR
2	R, G, B, NIR, EVI, NDVI
3	R, G, B, NIR, EVI, NDVI, NDWI, VARI, MSAVI
4	R, G, B, NIR, RE
5	R, G, B, NIR, RE, EVI, NDVI, NDWI, VARI, MSAVI

No QGIS (OSGEO, 2025) foram gerados dois conjuntos amostrais com 5 mil pontos aleatórios sobre a área do município de Bom Sucesso, 2500 pontos para as classes "Não Café" e "Café". Esses conjuntos serviram de base para a criação das amostras das séries temporais por meio da função *sits_get_data()*, que posteriormente seriam utilizadas para treinamento e teste dos modelos.

Modelagem

Na fase de modelagem, foram selecionados três algoritmos de ML e dois de DL. Dentre os de ML, optou-se pelo *Support Vector Machine* (SVM), um método robusto e de bom desempenho na área (SOUZA, 2015); e pelos algoritmos baseados em árvores de decisão *Random Forest* (RF) e *Extreme Gradient Boosting* (XGBoost), dada a eficácia dessa abordagem no mapeamento do uso da terra (SANTOS, 2024; SIMOES, 2021).

Para os testes com DL foram selecionados a *Temporal Convolutional Neural Network* (TempCNN) e o *Lightweight Temporal Attention Encoder* (L-TAE). A arquitetura da TempCNN estende o conceito das redes neurais convolucionais para a análise de séries temporais. O modelo aplica filtros 1D (unidimensionais) que extraem características ao longo da dimensão temporal, analisando uma janela de observações passadas para capturar padrões (SIMOES, 2021). Já o L-TAE é uma rede neural leve que utiliza um mecanismo de atenção para focar nas observações mais relevantes de uma série temporal, alcançando desempenho superior a outras arquiteturas com menor número de parâmetros e custo computacional (GARNOT; LANDRIEU, 2020).

Nos três modelos de ML foram utilizados os valores padrão de seus hiperparâmetros. E nos modelos de DL foi realizada a otimização dos seus hiperparâmetros por meio de *sits_tuning()*, cujos resultados estão nas Tabelas 2 e 3.

Tabela 2: Hiperparâmetros da TempCNN

Dataset	1	2	3	4	5
cnn_layers	128	64	128	64	64
cnn_kernels	5	5	7	5	3
cnn_dropout_rates	0.3	0.15	0.3	0.3	0.4
lr	0,000608	0,000669	0,00408	0,00165	0,00743
weight_decay	0,00528	0,000767	0,00387	0,00012	0,0000000229

Tabela 3: Hiperparâmetros do L-TAE

Dataset	1	2	3	4	5
batch_size	32	64	128	32	32
otimizador	adamw	adamw	adam	adam	adamw
lr	0,007416	0,000436	0,000264	0,000306	0,000380
weight_decay	$1{,}095e^{-08}$	$2{,}121e^{-07}$	0,003725	$3{,}253e^{-05}$	$6.867e^{-06}$
eps	$1e^{-08}$	$1e^{-07}$	$1e^{-08}$	$1e^{-07}$	$1e^{-07}$

Avaliação

A avaliação do desempenho dos modelos foi realizada em duas etapas. Primeiramente, calculou-se a acurácia geral sobre o conjunto de teste utilizando a

função *sits_accuracy()*. Em um segundo momento, para uma análise mais detalhada dos mapas de classificação, além da análise visual, foram calculadas as métricas de *Intersection over Union* (IoU) e o coeficiente de Dice (*F1-score)* a partir de um mapa de referência.

Por fim, para verificar a capacidade de generalização dos melhores modelos, estes foram aplicados ao município de Oliveira-MG.

Resultados e discussão

Os dois melhores resultados quantitativos por dataset podem ser observados na Tabela 4 e os resultados completos encontram-se em Métricas de avaliação dos mapas. O XGBoost, algoritmo de ML, destaca-se em todos os *datasets*, suas métricas variam pouco entre seu pior (*dataset 2*) e melhor resultado (*dataset 5*). A TempCNN foi o modelo de DL com as melhores métricas nos conjuntos de 1 a 3, porém, seu desempenho equipara-se ao XGBoost apenas no conjunto 3. O L-TAE passa a desempenhar melhor nos dois últimos *datasets* em que há a presença da banda espectral *Red Edge,* superando o XGBoost. O RF e o SVM não apresentaram resultados satisfatórios em nenhum conjunto.

Tabela 4 - Melhores resultados por dataset

Dataset	Modelo	Acurácia	IoU	Dice
1	XGBoost	0,92	0,58	0,73
	TempCNN	0,92	0,50	0,66
2	XGBoost	0,93	0,55	0,71
	TempCNN	0,94	0,47	0,64
3	XGBoost	0,94	0,59	0,74
	TempCNN	0,93	0,59	0,7453
4	XGBoost	0,92	0,58	0,73
	L-TAE	0,91	0,59	0,74
5	XGBoost	0,94	0,60	0,75
	L-TAE	0,87	0,67	0,80

É possível observar na Figura 2 que os modelos baseados em árvores de decisão, RF e XGBoost, sofrem poucas variações no seu desempenho com os diferentes datasets. Junto ao TempCNN, esses modelos tiveram uma queda do Coeficiente de Dice e, consequentemente, da IoU no dataset 4, o que seria coerente pela diminuição dos dados presentes no cubo. Porém, essa queda de desempenho também é observada no dataset 2, que acrescenta os IVs NDVI e EVI às bandas espectrais, uma tendência não esperada já que estes índices tendem a auxiliar no estudo de vegetações (SOUZA, 2015).

Figura 2 - Gráfico do Coeficiente de Dice por dataset

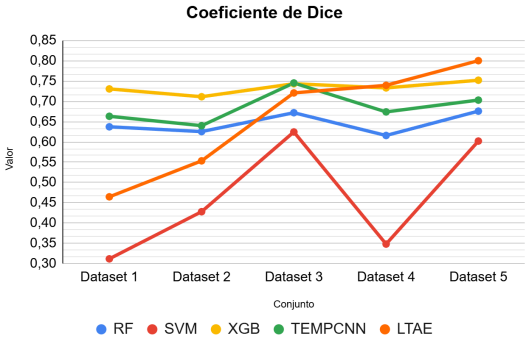
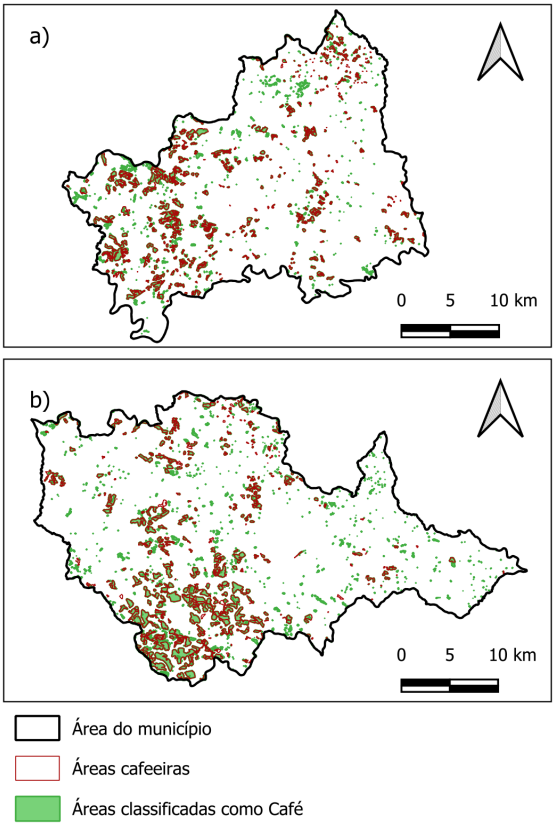


Figura 3 - a) Mapa classificado do L-TAE para Bom Sucesso-MG. b) Mapa classificado do L-TAE para Oliveira -MG.



Apesar da acurácia sobre o conjunto de teste do L-TAE ter sido a menor dentre os melhores modelos, como verifica-se na Tabela 4, ele obteve as maiores métricas de semelhança entre o mapa gerado e o de referência. Na Figura 3 a) que contém o mapa de café do L-TAE para Bom Sucesso, pode-se notar que a classificação de áreas cafeeiras se deu de forma correta

em sua maioria. O mesmo verificou-se em Oliveira, Figura 3 b), alcançando 0,70 de IoU e 0,82 de Dice, valores superiores aos de Bom Sucesso, demonstrando uma boa generalização do modelo.

Além disso, após uma inspeção dos falsos positivos do modelo nas duas cidades, verificou-se que parte deles eram, na verdade, lavouras de café. Assim, a existência de omissões no mapa de referência utilizado para a validação, sugere que as métricas de desempenho do modelo estão, possivelmente, subestimadas.

Conclusões

Os resultados demonstraram que o modelo L-TAE com o *dataset* 5 foi a abordagem mais eficaz para o mapeamento automatizado de café, alcançando um Coeficiente de Dice de 0,80 em Bom Sucesso e 0,82 em Oliveira. Isto ressalta a relevância de se combinar bandas espectrais e índices de vegetação para a tarefa. Também pode-se destacar o desempenho consistente do XGBoost que, embora superado pelo L-TAE no melhor cenário, se apresenta como uma alternativa viável por ter um custo computacional menor que os modelos de *Deep Learning*.

Ademais, trabalhos futuros devem focar na validação do método em outras regiões e períodos, utilizando séries temporais de anos subsequentes, além do uso de mapas de referência aprimorados para uma avaliação mais precisa dos classificadores.

Agradecimentos

Agradeço à UNIFEI pelo incentivo à pesquisa e pela concessão da bolsa de Iniciação Científica através do Programa Institucional de Bolsa de Iniciação Científica (PIBIC-UNIFEI).

Referências

ALVES, Helena Maria Ramos; VOLPATO, Margarete Marin Lordelo; CAMPOS, Beatriz Fonseca Dominik. **Mapeamento automatizado de áreas de café em Minas Gerais**. Brasília, DF: Embrapa Café, 2021. PDF (26 p.). (Documentos / Embrapa Café, ISSN 1678-1694; 13).

CONAB - COMPANHIA NACIONAL DE ABASTECIMENTO. **Acompanhamento da safra brasileira de café**, Brasília, DF, v.12, n. 2, segundo levantamento, maio 2025.

FERREIRA, Karine R. et al. Earth observation data cubes for Brazil: Requirements, methodology and products. **Remote Sensing**, v. 12, n. 24, p. 4033, 2020.

GARNOT, V. S. F.; LANDRIEU, L. Lightweight Temporal Self-attention for Classifying Satellite Images Time Series. In: LEMAIRE, V. et al. (org.). **Advanced Analytics and Learning on Temporal Data**. AALTD 2020. Cham: Springer, 2020. (Lecture Notes in Computer Science, v. 12588). cap. 12. DOI: 10.1007/978-3-030-65742-0_12.

HUNT, David A. et al. Review of remote sensing methods to map coffee production systems. **Remote Sensing**, v. 12, n. 12, p. 2041, 2020.

MARTÍNEZ-PLUMED, Fernando *et al*. **CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories**. IEEE Transactions on Knowledge and Data Engineering, [S. l.], v. 33, n. 8, p. 3048-3061, 2021. DOI: 10.1109/TKDE.2019.2962680. Disponível em: https://doi.org/10.1109/TKDE.2019.2962680. Acesso em: 30 jul. 2025.

OSGEO (OPEN SOURCE GEOSPATIAL FOUNDATION). **QGIS**. Versão 3.40. [S. l.]: OSGEO (Open Source Geospatial Foundation), 2025. Programa de computador. Disponível em: https://qgis.org/. Acesso em: 22 ago. 2025.

SANTOS, Pietra Brandão Ramos dos. **Cubo de Dados Geográfico aplicado à classificação automática de áreas cafeeiras no Sul de Minas Gerais**. Orientador: Vanessa Cristina Oliveira de Souza; Coorientador: Flávio Belizário da Silva Mota. 2024. 15 p. Trabalho de Conclusão de Curso (Sistemas de Informação) - Universidade Federal de Itajubá, Itajubá, 2024.

SIMOES, Rolf et al. Satellite image time series analysis for big earth observation data. **Remote Sensing**, v. 13, n. 13, p. 2428, 2021.

SOUZA, Carolina Gusmão. **Uso de séries temporais para o mapeamento da cafeicultura**. 2015. 162 p. Tese (Doutorado em Engenharia Florestal) – Universidade Federal de Lavras, Lavras, 2015.

WIRTH, Rüdiger; HIPP, Jochen. **CRISP-DM: Towards a Standard Process Model for Data Mining**. In: INTERNATIONAL CONFERENCE ON THE PRACTICAL APPLICATIONS OF KNOWLEDGE DISCOVERY AND DATA MINING, 4., 2000, Manchester, Reino Unido. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining. [S. l.]: [s. n.], 2000. p. 29-39.

ZHU, Hongyan et al. Intelligent agriculture: Deep learning in UAV-based remote sensing imagery for crop diseases and pests detection. **Frontiers in Plant Science**, v. 15, p. 1435016, 2024.