ANÁLISE AUTOMATIZADA DE DESMATAMENTO POR IMAGENS DE SATÉLITE UTILIZANDO MACHINE LEARNING

AUTOMATED ANALYSIS OF DEFORESTATION BY SATELLITE IMAGES USING MACHINE LEARNING

Igor Brinker Battilana[1]
[1] Universidade La Salle, Unilasalle, Canoas, Rio Grande do Sul, Brasil

Correspondência: Igor Brinker Battilana, Universidade La Salle, Unilasalle, Canoas, Estado do Rio Grande do Sul, Brasil. E-mail: igorbrinker@gmail.com

RESUMO

Este artigo apresenta uma metodologia para análise automática de desmatamento utilizando datasets de imagens obtidas por satélite e técnicas de machine learning. Com base nas imagens do satélite Sentinel-2 disponíveis no Google Earth Engine (GEE), foi aplicado o Índice de Vegetação da Diferença Normalizada (NDVI) para identificar áreas com vegetação. Posteriormente, um modelo de árvore de decisão (decision tree) foi treinado e utilizado para classificar áreas como desmatadas ou não desmatadas. A técnica proposta oferece uma abordagem eficiente e automatizada para monitorar e avaliar o desmatamento em regiões de floresta atlântica.

Palavras-chave: análise de desmatamento; imagens de satélite; inteligência artificial; NDVI; aprendizado de máquina; google earth engine, modelo de árvores de decisão.

ABSTRACT

This article presents a methodology for automatic analysis of deforestation using satellite image datasets and machine learning techniques. Based on images from the Sentinel-2 satellite available on Google Earth Engine (GEE), the Normalized Difference Vegetation Index (NDVI) was applied to identify areas with vegetation. Subsequently, a decision tree model was trained and used to classify areas as deforested or not deforested. The proposed technique offers an efficient and automated approach to monitor and assess deforestation in Atlantic forest regions.

Keywords: deforestation analysis; satellite images; artificial intelligence; NDVI; machine learning; google earth engine, decision tree model.

[1*] [*] Graduando em Ciência da Computação. Universidade La Salle. E-mail: igorbrinker@gmail.com

1. Introdução

A Amazônia, um dos maiores biomas do mundo, enfrenta um desafio crítico com o aumento contínuo do desmatamento. Segundo SPRACKLEN et al., 2012: “Amazônia Legal Brasileira compreende 60% da floresta amazônica - a maior floresta tropical contínua do planeta fornecendo importantes serviços ecossistêmicos como a provisão de água, alimentos, madeira, fibras, além de atuar como uma eficiente bomba, reciclando a água sobre toda a extensão da floresta, exercendo um papel preponderante no ciclo hidrológico que contribui para regular o clima da região.”

Este estudo busca empregar avançadas técnicas de machine learning e análises de imagens de satélite para monitorar e avaliar o desmatamento na região da Amazônia. Utilizando imagens do satélite Sentinel-2 disponíveis no Google Earth Engine (GEE), aplicamos o Índice de Vegetação da Diferença Normalizada (NDVI) para identificar áreas com vegetação e, posteriormente, desenvolvemos um modelo de árvore de decisão para classificar áreas como desmatadas ou não desmatadas. O objetivo é oferecer uma abordagem eficiente e automatizada para o monitoramento ambiental, contribuindo para a preservação deste ecossistema vital. Este artigo descreve o processo de coleta e análise dos dados, a metodologia empregada no desenvolvimento do modelo, e os resultados obtidos, discutindo suas implicações para o futuro do monitoramento ambiental na Amazônia.

2. Material e Métodos

2.1 Google Earth Engine (GEE)

O Google Earth Engine é uma plataforma onde armazena diversos dados geoespaciais providenciados por diversas instituições, como por exemplo a Administração Nacional do Espaço e da Aeronáutica (NASA) e a Agência Espacial Europeia (ESA). Baseada na nuvem, que acaba facilitando o processamento de grandes conjuntos de dados geoespaciais, como diversas frequências do espectro eletromagnético captadas por satélite. Ele fornece um ambiente para visualização, análise e compartilhamento de conjuntos de dados geoespaciais em grande escala. Adicionalmente, a ferramenta conta com uma API para desenvolvimento em Python. Vale salientar, que existem assinaturas para diversos tipos de usuários onde cada plano irá providenciar diversas novas funcionalidades.

2.2 Google Colab

O Google Colab é um ambiente de notebook Jupyter que permite a execução de código Python em blocos, podendo ser usado no navegador sem qualquer configuração. É uma

ferramenta gratuita oferecida pelo Google que fornece recursos de computação na nuvem, incluindo GPUs e TPUs, facilitando a execução de tarefas intensivas em computação. Existem limitações quanto ao uso de memória, onde é necessário diminuir fatores que aumentam o processamento de grandes quantidades de dados. Também existem planos de assinatura, onde o usuário poderá usufruir de clusters que darão um poder de processamento maior para executar tarefas.

2.3 Extração dos Dados

As imagens foram filtradas com base na ROI, intervalo de datas e percentagem máxima de cobertura de nuvens usando o conjunto de dados 'COPERNICUS/S2' no GEE.

Os dados coletados para montar o dataset de treino, foram extraídos a partir do notebook "*extract_ndvi_and_label_from_image_collection*" no qual sua exclusiva função é consumir o datasets de imagens do Google Earth Engine de uma determinada lista contendo diversas localizações da região amazônica. Onde o dataset contém a latitude e longitude da região, o NDVI, timestamps de quando as imagens foram feitas pelo satélite e label que indica a categoria de tal região. Já, no notebook "*extract_ndvi_from_image_collection*", temos os mesmos dados exceto o label de classificação, pois iremos usá-lo para aplicar o treinamento do modelo em dados reais.

2.3.1 Índice de Vegetação da Diferença Normalizada (NDVI)

Esse índice será a base da pesquisa, onde será possível julgar se a ROI definida tem algum princípio de desmatamento.

Segundo Huang et al. (2020, p. 1), "O Índice de Vegetação por Diferença Normalizada (NDVI), um dos primeiros produtos analíticos de sensoriamento remoto usados para simplificar as complexidades da imagiologia multiespectral, é agora o índice mais popular usado para avaliação da vegetação. Essa popularidade e uso generalizado estão relacionados à capacidade de calcular um NDVI com qualquer sensor multiespectral que possua uma banda visível e uma banda de infravermelho próximo (near-IR). A redução dos custos e dos pesos dos sensores multiespectrais significa que eles podem ser montados em satélites, veículos aéreos e, cada vez mais, em Sistemas Aéreos Não Tripulados (UAS)." O NDVI será o índice mais importante dessa pesquisa, pois o nosso algoritmo de machine learning irá usá-lo para treinar e, logo após, baseado nos resultados dos melhores hiperparâmetros obtidos, identificar e classificar a definição de cada valor.

Os dados serão extraído pelo cálculo feito através das bandas B4 (RED) e B8 (NIR) do satélite Sentinel-2 que faz parte do programa Copernicus, no qual é administrado pela Comissão Europeia e executado pela Agência Espacial Europeia (ESA), a Organização Europeia para a Exploração de Satélites Meteorológicos (EUMETSAT), o Centro Europeu

para as Previsões Meteorológicas a Médio Prazo (ECMWF), as agências da UE e a Mercator Océan.

A fórmula para calcular o NDVI é:

NDVI = (NIR - Red) / (NIR + Red)

Onde, segundo Jones e Vaughan (2010, p. 353), “O NDVI é o índice de vegetação por diferença normalizada. Red e NIR são medições de radiância espectral (ou refletância) registradas com sensores nas regiões vermelha (visível) e NIR, respectivamente. Radiância (watts esterradiano−1 m−2 μm−1) é a medida do fluxo de energia registrado por um sensor. Os valores de radiância são frequentemente escalados para números digitais (DN) como inteiros sem sinal de 6 bits ou 7 bits (MSS), 8 bits (TM, ETM +), ou 12 bits (Landsat 8). Reflectância é uma medida sem unidade da razão da radiação refletida por um objeto em relação à radiação incidente sobre o objeto. Os valores de NDVI variam de -1 a 1, independentemente de usar radiância, reflectância ou DN como entrada. Em geral, seus valores são negativos para corpos d'água, próximos de zero para rochas, areias ou superfícies de concreto, e positivos para vegetação, incluindo culturas, arbustos, gramíneas e florestas".

2.3.2 Árvore de Decisão (Decision Tree)

A escolha do modelo de Árvore de Decisão foi tomada a partir das características dos dados que foram extraídos e calculados pelos datasets do Earth Engine para criar o nosso próprio dataset, que apresenta algumas características onde o algoritmo de árvore de decisão pode se sair muito bem, tais como, dados onde não são normalizados, não são lineares e que devem ser categorizados. Porém, deve-se tomar o cuidado, pois o modelo tende a ser sensível a pequenas variações nos dados, o que pode levar a diferentes estruturas de árvores.

Segundo Rokach e Maimon (2005, p. 165), "Uma árvore de decisão é um classificador expresso como uma partição recursiva do espaço de instância. A árvore de decisão consiste em nós que formam uma árvore enraizada, significando que é uma árvore dirigida com um nó chamado 'raiz' que não tem arestas de entrada. Todos os outros nós têm exatamente uma aresta de entrada. Um nó com arestas de saída é chamado de nó interno ou de teste. Todos os outros nós são chamados de folhas (também conhecidos como terminais ou nós de decisão). Em uma árvore de decisão, cada nó interno divide o espaço de instância em dois ou mais subespaços de acordo com uma certa função discreta dos valores dos atributos de entrada."

2.3.3 Definição da Região de Interesse (ROI)

No estudo sobre a análise automática de desmatamento na Amazônia, abordamos a metodologia de seleção de locais específicos para o treinamento e validação do nosso modelo de machine learning. Esta seleção foi essencial para garantir a representatividade e a relevância dos dados no contexto da diversidade ambiental da Amazônia.

Para o treinamento do modelo, selecionamos áreas que oferecem uma amostra abrangente do ecossistema amazônico, incluindo regiões próximas a Manaus, áreas em Rondônia, o Parque Nacional do Pico da Neblina, e a região do Acre. Estes locais foram escolhidos por sua variedade em vegetação, níveis de impacto humano e histórico de desmatamento, fornecendo uma base de dados diversificada para o treinamento do modelo.

Além disso, para a validação do modelo, inserimos labels reais em locais selecionados estrategicamente, abrangendo diferentes graus de desmatamento. Estes locais incluíram áreas adicionais em Rondônia, regiões ao longo do Rio Amazonas e áreas fronteiriças próximas ao Peru e Colômbia. Essa seleção cuidadosa foi crucial para validar a precisão do modelo na identificação de áreas desmatadas e preservadas.

A combinação destes dados de treino e validação permitiu uma análise abrangente, aumentando a confiabilidade dos resultados do modelo. Esta estratégia garante que o modelo não apenas aprenda a partir de um conjunto de dados representativo, mas também seja validado contra uma variedade de condições ambientais, o que é fundamental para o monitoramento eficaz e a gestão sustentável do desmatamento na Amazônia.

2.4 Metodologia de Treinamento do Modelo

Destacamos que os resultados alcançados na classificação de áreas desmatadas na Amazônia são notáveis, refletindo avanços significativos no uso de tecnologias de machine learning para questões ambientais. A precisão e eficácia do modelo em identificar regiões desmatadas e não desmatadas demonstram o potencial das ferramentas de IA na análise de grandes conjuntos de dados de satélite. Esta seção detalha cada etapa do processo de treinamento, desde o carregamento e preparação dos dados até a avaliação e visualização dos resultados, ilustrando como o modelo foi desenvolvido e validado para atender aos objetivos do estudo.

2.4.3.1. Carregamento de Dados

A seleção e o uso dos dados são cruciais em qualquer modelo de machine learning. Optamos por um conjunto de dados armazenados em um arquivo CSV no Google Drive, contendo NDVI e classificações de desmatamento. Este passo foi vital para garantir que o modelo tivesse acesso a informações confiáveis e relevantes para a tarefa de classificação.

2.4.3.2. Preparação dos Dados

Dividir os dados em 'features' e 'target' é uma etapa essencial na preparação para o treinamento de um modelo. Neste estudo, 'features' representam os valores de NDVI e 'target' as classificações de desmatamento. A separação subsequente em conjuntos de treino e teste permite uma avaliação justa do desempenho do modelo.

2.4.3.3. Configuração do Pipeline

A configuração do pipeline com DecisionTreeClassifier do scikit-learn estabelece a estrutura fundamental do modelo. O pipeline não só simplifica o processo de modelagem, mas também facilita a reprodução e a modificação do modelo em pesquisas futuras.

2.4.3.4. Definição de Hiperparâmetros

A escolha dos hiperparâmetros é um fator decisivo no desempenho do modelo. Definimos um conjunto extenso de hiperparâmetros para explorar uma variedade de configurações, com o objetivo de otimizar a eficácia do modelo.

O modelo foi configurado com um conjunto detalhado de hiperparâmetros. Utilizamos critérios como 'gini' e 'entropy' para a seleção de atributos, 'max_depth' variando de 300 a 5000 para definir a profundidade da árvore, e 'min_samples_split' e 'min_samples_leaf' para determinar o número mínimo de amostras necessárias em nós de divisão e folhas, respectivamente. Além disso, configuramos 'max_features' para definir o número de características a considerar na busca pela melhor divisão e 'ccp_alpha' para poda da árvore, a fim de evitar o overfitting.

2.4.3.5. Ajuste e Seleção de Hiperparâmetros

Empregamos GridSearchCV e RandomizedSearchCV, dois métodos poderosos do scikit-learn para o tuning de hiperparâmetros. O GridSearchCV examina sistematicamente todas as combinações de hiperparâmetros, enquanto o RandomizedSearchCV explora amostras aleatórias dessas combinações para encontrar a configuração ideal.

2.4.3.6. Treinamento e Avaliação

O treinamento e avaliação do nosso modelo foram realizados utilizando o DecisionTreeClassifier do scikit-learn, uma ferramenta eficaz para lidar com dados de alta dimensionalidade como os nossos. Inicialmente, treinamos o modelo com um conjunto de dados de treino, seguido de uma avaliação rigorosa usando o conjunto de testes. Durante a avaliação, geramos métricas cruciais, incluindo o relatório de classificação, matriz de confusão e acurácia, que ofereceram uma análise abrangente do desempenho do modelo. Adicionalmente, realizamos a validação cruzada, uma etapa essencial para assegurar a robustez e a confiabilidade do modelo em diferentes amostras de dados.

2.4.3.7. Visualização dos Resultados

Para uma apresentação clara e efetiva dos resultados do modelo, foram utilizadas bibliotecas como *"matplotlib"* e *"seaborn"* para gerar visualizações detalhadas, incluindo

matrizes de confusão e curvas ROC. Estas visualizações fornecem uma análise aprofundada do desempenho do modelo, destacando sua capacidade de classificar corretamente as áreas de desmatamento na Amazônia. Esta metodologia não só assegura uma abordagem rigorosa no desenvolvimento do modelo, mas também facilita a interpretação dos resultados, oferecendo gráficos claros e informativos que ilustram a eficácia do modelo na identificação de áreas desmatadas.

6. Conclusão

Concluindo, o estudo apresentado neste artigo demonstra uma abordagem significativa para o monitoramento do desmatamento na Amazônia, utilizando tecnologias avançadas de machine learning e imagens de satélite. Os resultados obtidos, embora moderados em termos de precisão e recall, oferecem uma base sólida para futuras melhorias e aplicações práticas. Este trabalho não apenas contribui para o campo do monitoramento ambiental, mas também abre caminho para pesquisas subsequentes, incentivando o aprimoramento contínuo de técnicas de análise de dados e a exploração de novas metodologias para a conservação ambiental. Os resultados detalhados e as análises complementares serão apresentados a seguir, destacando a relevância e a aplicabilidade das descobertas deste estudo no contexto global de conservação da Amazônia.

6.1 Resultados

Os resultados do modelo de classificação para análise de desmatamento na Amazônia são significativos em várias frentes:

Precisão e Recall: O modelo apresentou precisão de 67% para a classe 0 (não desmatado) e 68% para a classe 1 (desmatado), com recall de 61% para a classe 0 e 74% para a classe 1. Isso indica um equilíbrio na identificação de ambas as classes.

F1-Score: Os f1-scores de 0.64 (classe 0) e 0.71 (classe 1) sugerem eficácia moderada em ambas as categorias.

Matriz de Confusão: O modelo classificou corretamente 417 de 680 casos como não desmatados e 570 de 775 casos como desmatados.

Acurácia Geral: A acurácia foi de aproximadamente 67.84%.

Acurácia da Validação Cruzada: A acurácia de 0.65 com desvio padrão de +/- 0.03 na validação cruzada mostra consistência do modelo.

Melhores Hiperparâmetros: Os hiperparâmetros otimizados incluem *'ccp_alpha': 0.0* (parâmetro de poda para evitar overfitting), *'class_weight': 'balanced'* (para equilibrar

classes desiguais), *'criterion': 'gini'* (medida de impureza), *'max_depth': 798* (profundidade máxima da árvore), *'max_features': 'sqrt'* (número de características a considerar ao procurar a melhor divisão), *'max_leaf_nodes': None* (número máximo de nós folha), *'min_impurity_decrease': 0.0* (limite para redução de impureza para divisão de nó), *'min_samples_leaf': 1* (número mínimo de amostras por node/folha), *'min_samples_split': 2* (número mínimo de amostras para dividir um nó interno), e *'splitter': 'random'* (estratégia para escolher divisões). Esses resultados e configurações indicam uma eficácia moderada do modelo com espaço para melhorias, particularmente na identificação de áreas não desmatadas. As escolhas de hiperparâmetros refletem um esforço para equilibrar precisão e evitar o overfitting, mantendo a árvore de decisão robusta e aplicável.

6.2 Discussão

Na discussão do estudo sobre a análise automática de desmatamento na Amazônia utilizando machine learning, destacamos a relevância dos resultados alcançados. A precisão, recall, e F1-score do modelo, embora moderados, demonstram sua capacidade de identificar áreas desmatadas, um aspecto crucial para o monitoramento ambiental. Comparando nossos achados com estudos similares, percebemos tanto avanços quanto limitações inerentes ao nosso método, especialmente relacionados à natureza do modelo de árvore de decisão e à qualidade dos dados. Estes resultados têm implicações práticas significativas, sugerindo aplicações em políticas de conservação e monitoramento. A pesquisa abre caminhos para futuras investigações, incluindo aprimoramento do modelo e exploração de novos conjuntos de dados, visando uma compreensão mais abrangente e eficaz do desmatamento na região amazônica.

7. Glossário

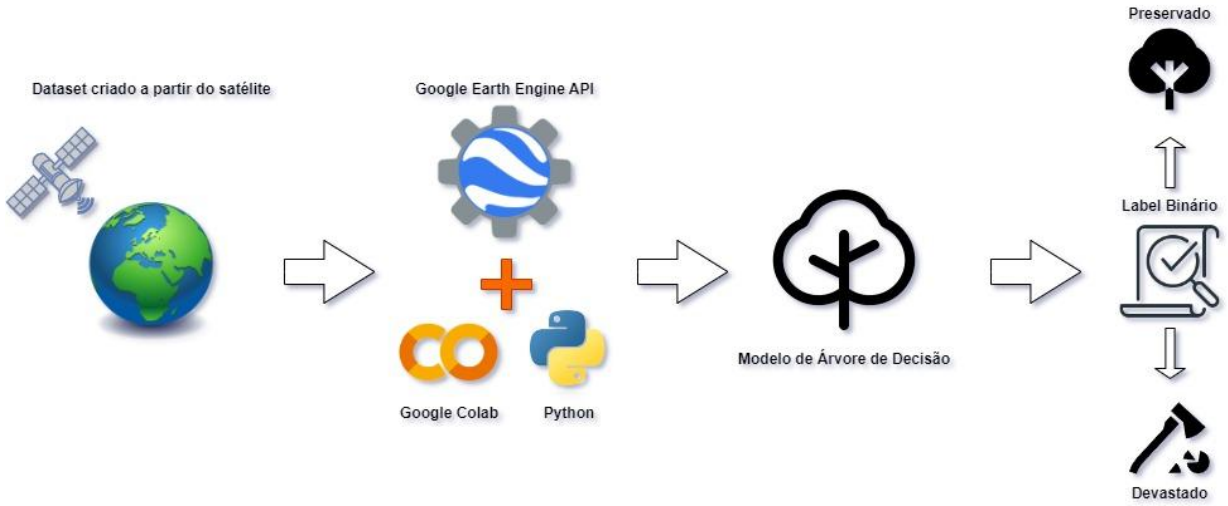


Diagrama mostrando todas as etapas do processo

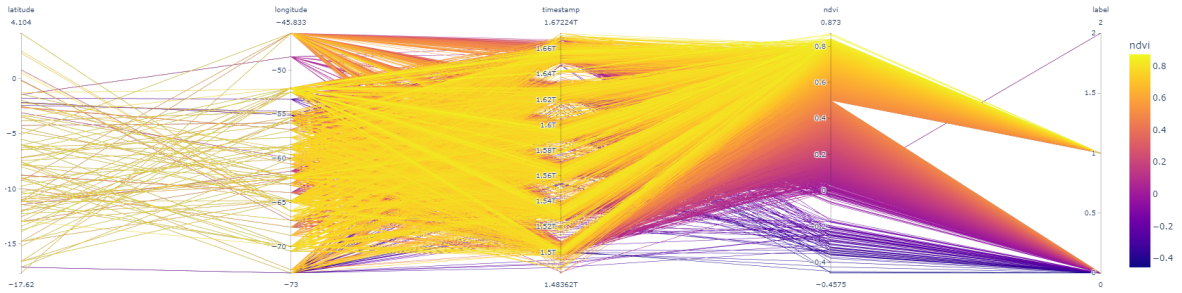


Gráfico de coordenadas paralelas para visualizar a correlação entre variáveis do dataset gerado com labels para treinar o algoritmo

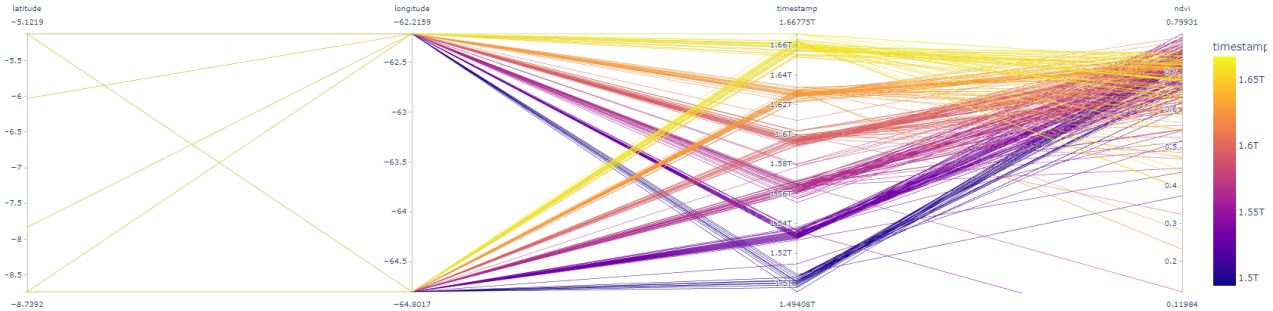
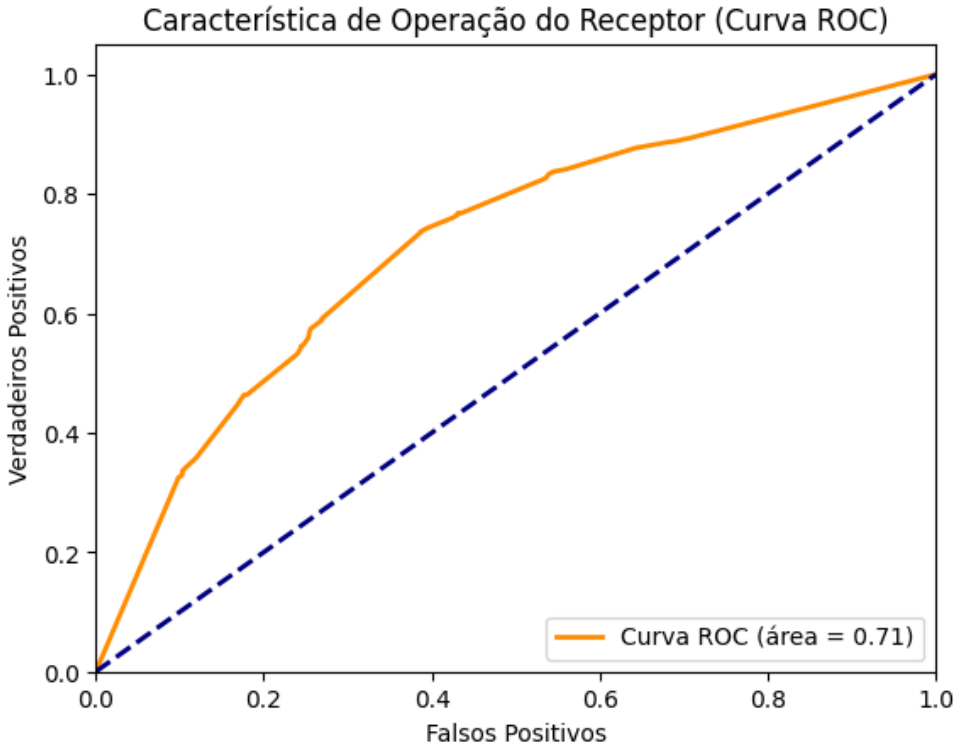
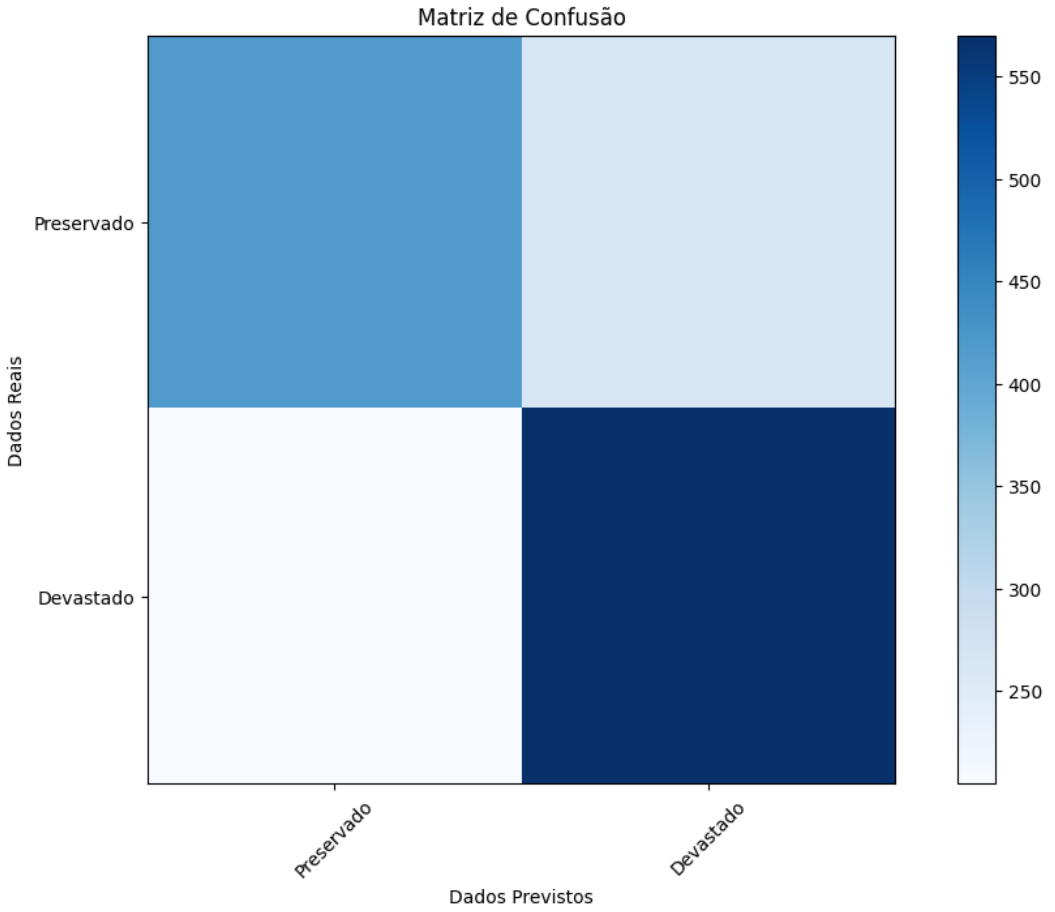


Gráfico de coordenadas paralelas para visualizar a correlação entre variáveis do dataset gerado sem labels para o algoritmo categorizar após o treinamento do mesmo



A curva ROC apresentada ilustra a capacidade do modelo de diferenciar entre as classes positivas e negativas. A área sob a curva (AUC) é de 0.71, o que indica uma boa capacidade preditiva, pois uma AUC de 1.0 representaria uma capacidade perfeita de classificação, e uma AUC de 0.5 indicaria uma capacidade não melhor que a aleatória. A curva está significativamente acima da linha tracejada, que representa a classificação aleatória, o que sugere que o modelo tem uma taxa satisfatória de verdadeiros positivos em relação à taxa de falsos positivos.



A matriz de confusão apresentada mostra a performance do modelo na classificação de áreas como 'Preservadas' ou 'Devastadas'. As células da diagonal representam as classificações corretas, enquanto as células fora da diagonal representam as classificações incorretas. A matriz indica um número substancial de verdadeiros positivos e verdadeiros negativos, o que é positivo. No entanto, também há uma quantidade considerável de falsos negativos e falsos positivos, sugerindo que há margem para melhorar a precisão do modelo na classificação das áreas de interesse.

REFERÊNCIAS

SPRACKLEN, D.V.; ARNOLD, S.R.; TAYLOR, C.M. Observations of increased tropical rainfall preceded by air passage over forests. Nature, London, v. 489, p. 282-285, 2012.

GOOGLE. Introduction to Google Earth Engine. Disponível em: https://newsinitiative.withgoogle.com/resources/trainings/introduction-to-google-earth-engine/. Acesso em: 17 Out. 2023.

GOOGLE. Earth Engine access. Atualizado em: 16 Mai. 2023. Disponível em: https://developers.google.com/earth-engine/guides/access. Acesso em: 17 Out. 2023.

GOOGLE. An Intro to the Earth Engine Python API. Atualizado em: 23 Out. 2023. Disponível em: https://developers.google.com/earth-engine/tutorials/community/intro-to-python-api. Acesso em: 18 Out. 2023.

HUANG, S.; TANG, L.; HUPY, J. P.; WANG, Y.; SHAO, G. A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing. 2020. Disponível em: https://link.springer.com/article/10.1007/s11676-020-01155-1. Acesso em: 09 Nov. 2023.

JONES, H. G.; VAUGHAN, R. A. Remote sensing of vegetation: principles, techniques and applications. Resenha por KULAWARDHANA, Ranjani Wasantha. Publicado em: 22 de junho de 2011. Disponível em: https://doi.org/10.1111/j.1654-1103.2011.01319.x. Acesso em: 10 Nov. 2023.

ROKACH, Lior; MAIMON, Oded. Decision Trees. In: The Data Mining and Knowledge Discovery Handbook. Cap. 9. Janeiro de 2005. p. 165-192. DOI:10.1007/0-387-25465-X_9.