Modelagem de biomassa acima do solo por meio de aprendizado de máquina e sensoriamento remoto

Sandro De Sena Machado; Aubert Henrik Venson

Resumo

A predição de biomassa acima do solo é essencial para a compreensão dos ecossistemas e o manejo sustentável dos recursos naturais. Este trabalho teve como objetivo principal desenvolver modelos preditivos utilizando dados de sensoriamento remoto e técnicas de aprendizado de máquina. A metodologia incluiu técnicas de amostragem, validação cruzada, seleção de variáveis, otimização de hiperparâmetros, explanação dos modelos com SHAP e validação dos modelos com dados do Inventário Nacional Florestal (IFN). A avaliação de acurácia foi realizada utilizando métricas obtidas através da validação cruzada dos modelos otimizados, incluindo Erro Quadrático Médio (RMSE), Erro Absoluto Médio (MAE) e Coeficiente de Determinação (R^2) ajustado. Os modelos Random Forest e XGBoost apresentaram R^2 ajustados de 62,7% e 62,0%, respectivamente, enquanto a Rede Neural Artificial (DNN) obteve R^2 ajustado de 60,4%. Ao comparar a biomassa prevista com os dados obtidos através do IFN, identificou-se uma diferença de -8,66 t/ha, indicando que as previsões do modelo Random Forest são, em média, inferiores aos valores medidos diretamente no campo. Além disso, o teste de Mann-Whitney não encontrou diferenças estatisticamente significativas nas distribuições das métricas de desempenho entre os modelos. Assim, o trabalho contribuiu para a área de gestão ambiental, oferecendo uma ferramenta que pode ser utilizada para apoiar a tomada de decisões em políticas de conservação e uso sustentável dos recursos florestais.

Palavras-chave: Florestas Aleatórias; Redes Neurais Profundas; XGBoost; Python; Google Earth Engine.

Modeling Above-Ground Biomass through Machine Learning and Remote Sensing

Abstract

The prediction of aboveground biomass is essential for understanding ecosystems and the sustainable management of natural resources. This study aimed to develop predictive models using remote sensing data and machine learning techniques. The methodology included sampling techniques, cross-validation, variable selection, hyperparameter optimization, model explanation using SHAP, and model validation with data from the National Forest Inventory (IFN). Accuracy assessment was performed using metrics obtained from cross-validation of the optimized models, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and adjusted Coefficient of Determination (R^2). The Random Forest and XGBoost models achieved adjusted R^2 values of 62,7% and 62,0%, respectively, while the Artificial Neural Network (DNN) obtained an adjusted R^2 of 60,4%. When comparing the predicted biomass with data obtained from the IFN, a difference of -8.66 t/ha was identified, indicating that the predictions of the Random Forest model are, on average, lower than the values measured directly in the field. Moreover, the Mann-Whitney test found no statistically significant differences in the performance metrics distributions among the models. Thus, this work contributed to the field of environmental management by providing a tool that can be used to support decision-making in conservation policies and the sustainable use of forest resources.

Keywords: Random Forest; Deep Neural Networks; XGBoost; Python; Google Earth Engine.

1. Introdução

De acordo com o Sexto Relatório de Avaliação (AR6) do Painel Intergovernamental sobre Mudanças Climáticas (IPCC), a temperatura global já aumentou 1,2°C em comparação à média do período de 1850-1900, principalmente devido à elevação na concentração de gases de efeito estufa, como o dióxido de carbono (CO_2), resultantes das atividades humanas. Esse aquecimento tem causado o aumento do nível do mar e a intensificação de eventos climáticos extremos, como ondas de calor, secas e tempestades (IPCC, 2023). Entre as principais fontes dessas emissões estão o desmatamento e a degradação florestal, que contribuem significativamente para o aumento de CO_2 na atmosfera, pois as florestas armazenam grandes quantidades de carbono em sua biomassa. Quando essas áreas são desmatadas ou degradadas, o carbono anteriormente retido nas árvores é liberado na forma de CO_2, exacerbando as mudanças climáticas (Reichstein et al. 2019).

Além de seu papel no armazenamento de carbono, as florestas fornecem diversos serviços ecossistêmicos, como a regulação do clima, a proteção de solos e recursos hídricos, e a conservação da biodiversidade (Serviço Florestal Brasileiro, 2019). Dessa forma, o estudo e a análise da biomassa florestal são essenciais para embasar decisões políticas e econômicas voltadas à conservação, como projetos de restauração de ecossistemas, programas de monitoramento de áreas protegidas e iniciativas de pagamentos por serviços ambientais (Oliveira et al. 2018).

Tradicionalmente, a mensuração da biomassa florestal é feita por amostragem de campo, com a instalação de parcelas permanentes, nas quais são coletados dados como o diâmetro e a altura das árvores, além da densidade da madeira. Essas informações são usadas em equações alométricas para estimar a biomassa e o volume das árvores (Ratuchne, 2010).

No entanto, a coleta de dados em campo enfrenta limitações de custo e de viabilidade quando aplicada em larga escala, tornando o mapeamento preciso da biomassa um desafio (Belloli et al. 2022). Como alternativa, o sensoriamento remoto, aliado a técnicas de aprendizado de máquina, tem se mostrado uma ferramenta eficaz para estimar a biomassa em grandes áreas, com maior precisão e menor custo. Assim, dispositivos como satélites, drones e LiDAR permitem a coleta de dados sobre a superfície terrestre à distância, os quais são processados e analisados para extrair características relevantes sobre a biomassa acima do solo (AGB).

Nesse sentido, informações acerca da biomassa podem ser extraídas como índices de vegetação, parâmetros biofísicos e medidas de textura, como o NDVI, o LAI e o GLCM

(Jensen, 2015). Além disso, a combinação de diferentes sensores, como sensores ópticos, radar e LiDAR, pode melhorar a qualidade das estimativas, uma vez que cada sensor fornece informações complementares sobre a estrutura florestal. Sensores ópticos capturam a densidade e o vigor da vegetação, enquanto os sensores radar e LiDAR conseguem penetrar no dossel das florestas, fornecendo medidas tridimensionais da altura e da estrutura da vegetação (Molisse et al. 2022).

O Google Earth Engine (GEE), plataforma em nuvem para processamento de dados de sensoriamento remoto, tem se destacado pelo seu catálogo amplo e bem documentado. Com a integração de conjuntos de dados de sensoriamento remoto a partir do GEE no Google Colab, é possível criar modelos para estimar a biomassa florestal de maneira escalável e acurada, utilizando bibliotecas de aprendizado de máquina em Python como Scikit-Learn, Keras e Tensor Flow.

Para que os modelos possam ser calibrados e validados com dados de campo, a Terra Indígena Mangueirinha, localizada no oeste do Paraná, foi escolhida como área de estudo devido à qualidade dos dados de biomassa disponíveis, coletados e consolidados pelo Inventário Florestal Nacional (IFN) em escala local (Serviço Florestal Brasileiro, 2019). Essa área situada na bacia do Rio Iguaçu, possui informações detalhadas sobre a saúde das florestas, biodiversidade e estoques de biomassa e carbono, fornecendo um valioso conjunto de dados de referência para este trabalho.

O objetivo principal deste estudo é implementar e avaliar modelos de aprendizado de máquina para estimar AGB, mapeando sua distribuição espacial e temporal. Serão testados diferentes modelos preditivos, buscando identificar aqueles que apresentam maior acurácia e comparar seus resultados com dados de campo, visando garantir confiabilidade dos modelos.

2. Metodologia

2.1. Conjuntos de dados

2.1.1. MapBiomas (Coleção 7)

A coleção 7 de cobertura da Terra do MapBiomas oferece um conjunto de dados acerca da dinâmica do uso e ocupação do solo no Brasil, permitindo análises de mudanças em diferentes classes, como agricultura, pastagem, áreas florestais e outros usos. Essa coleção foi usada para uma análise exploratória da Terra Indígena Mangueirinha, visando identificar transformações no uso e ocupação do solo ao longo dos anos, bem como pressões de atividades antrópicas na região que podem influenciar a quantidade de biomassa na área de estudo.

2.1.2. ESA CCI Global Forest Above Ground Biomass

Este conjunto de dados é produto do processamento digital de imagens derivadas de sensores radar, principalmente SAR (*Synthetic Aperture Radar*) e LiDAR, através das missões ALOS-2, Sentinel-1, ICESat e ICESat-2. Possuem escala global, resolução de ~ 100 m x 100 m no equador e estão amplamente documentados (Santoro, *et al.* 2023).

Este produto está disponível no catálogo do *Google Earth Engine*, foi validado e calibrado a partir de dados alométricos de Inventários de Florestas Nacionais, contidos nos relatórios do projeto de Avaliação dos Recursos Florestais Globais (*Global Forest Resources Assessment*) da FAO (*Food and Agriculture Organization of the United Nations*) (FAO, 2020). A produção desses dados se deu principalmente por modelos de regressão semi-empíricos, que detectam as relações entre AGB, retroespalhamento SAR e nuvens de pontos LiDAR (Santoro, *et al.* 2023).

Através da amostragem desse conjunto de dados, pode-se obter uma variável alvo (y) para o treinamento dos modelos. O programa *Climate Change Initiative* (CCI) da Agência Espacial Europeia (ESA) desenvolveu este conjunto de dados em escala global, que fornecem estimativas de AGB (unidade: tons/ha *i.e.*, Mg/ha) para os anos de 2010, 2017, 2018, 2019 e 2020, permitindo análises de mudanças ao longo do tempo e avaliações de desmatamento, degradação florestal e reflorestamento (Santoro, *et al.* 2023).

Em 2024, análises conduzidas pela Planet Labs produziram o Forest Carbon Diligence, um conjunto de dados de biomassa semelhante ao ESA CCI Biomass, o qual demonstrou forte concordância entre as estimativas de ambos os produtos, com um coeficiente de correlação de 0,87 tanto para ecossistemas de latitudes médias quanto altas (Planet Labs, 2024).

2.1.3. Harmonized Sentinel-2 MSI: MultiSpectral Instrument, Level-2A.

Sentinel-2 é uma missão de imageamento multiespectral de alta resolução e ampla cobertura, que apoia estudos de monitoramento terrestre do programa Copernicus. As imagens multiespectrais para este estudo serão obtidas através da coleção *Harmonized Sentinel-2 MSI: MultiSpectral Instrument, Level-2A* (ESA, 2023a).

As imagens contém 13 bandas multiespectrais com resolução espaciais variadas, cobrindo diferentes regiões do espectro eletromagnético. Dessas bandas, as que compõem a luz visível (azul = ‘B2’, verde= ‘B3’; vermelho = ‘B4’) e o infravermelho-próximo (*NIR* = ‘B4’) possuem 10 m de resolução espacial; *Red-Edge* (B5; B6; B7; B8A) e o infravermelho de

ondas curtas (SWIR1 = ‘B11’; SWIR 2 = ‘B12’) possuem 20 m. As bandas ‘B1’, ‘B9’ e ‘QA60’, que possuem 60 metros de resolução, foram removidas por se tratarem de componentes desnecessários para a modelagem de biomassa (ESA, 2023a).

As imagens são ortorretificadas, georreferenciadas e radiométricamente calibradas. O pré-processamento Level-2A sobre as imagens permite remover efeitos atmosféricos e converter os valores dos pixels em reflectância. Adicionalmente, é necessário aplicar uma função filtrar as imagens com menor percentual de cobertura de nuvens, realizar o mascaramento de nuvens e aplicar o fator de escala. Ainda, reamostrar as bandas para a mesma resolução espacial é fundamental para análises posteriores (Belloli *et al.* 2022).

2.1.4. Sentinel-1 SAR GRD: C-band Synthetic Aperture Radar Ground Range Detected, log scaling

O Sentinel-1 é um radar de abertura sintética (SAR) que fornece dados sobre a estrutura da superfície da terra, e sofre menos interferência das condições climáticas ou da presença de luz solar, em comparação com sensores ópticos como o Sentinel-2. Ao contrário do Sentinel-2 que captura radiação solar refletida pela superfície, o Sentinel-1 emite pulsos de radar e registra a energia refletida pelos objetos na Terra, também chamada de retroespalhamento (*backscattering*) (ESA, 2023b).

O Sentinel-1 opera na banda C (5.405 GHz), reconhecidamente útil para mapear a biomassa florestal devido à sua alta penetração na vegetação densa em função do seu comprimento de onda entre 2,75 - 7,5 cm. O sinal de radar na banda C é capaz de penetrar na floresta e interagir com diferentes componentes da vegetação, como troncos, galhos e folhas (ESA, 2023b).

As imagens de radar serão obtidas através da coleção *Sentinel-1 SAR GRD: C-band Synthetic Aperture Radar Ground Range Detected, log scaling* que oferece quatro bandas com diferentes polarizações (VV, HH, VH e HV), com resolução espacial de 10 m e unidade de medida em escala logarítmica (dB), permitindo a detecção de mudanças sutis na biomassa. Esse produto foi pré-processado através de calibração radiométrica, remoção de ruído termal, correção do terreno com SRTM 30 (*Shuttle Radar Topography Mission*) e ASTER DEM e ortorretificação. Contudo, este produto ainda precisa de pré-processamento adicional como filtro de ruído Speckle (ESA, 2023b).

Cada tipo de polarização oferece vantagens e desvantagens para detecção de características. A polarização horizontal (HH) é sensível à estrutura horizontal da vegetação (galhos, folhas) enquanto a polarização vertical é sensível à estrutura vertical da vegetação

(troncos). Já a polarização cruzada (HV e VH) permite extrair informações sobre a orientação e propriedades dielétricas da vegetação (ESA, 2023b).

2.1.5. NASA SRTM Digital Elevation 30m

O Modelo Digital de Elevação (MDE) do SRTM (Shuttle Radar Topography Mission) é um projeto conjunto da NASA e da Agência Espacial Nacional da Alemanha (DLR) que mapeou a elevação terrestre para o ano de 2000. A topografia influencia a estrutura da vegetação e a biomassa, por isso é importante considerá-la no mapeamento. O MDE está disponível como uma imagem única global de resolução espacial de 30 m (Farr *et al.* 2007).

2.1.6. ETH Global Sentinel-2 10m Canopy Height (2020)

O conjunto de dados ETH Global Sentinel-2 10m Canopy Height (2020) fornece uma estimativa global da altura do dossel florestal em metros, com uma resolução espacial de 10 metros, derivada de dados do Sentinel-2 em conjunto com dados do GEDI. Este dado é valioso para a predição de AGB, pois a altura do dossel está diretamente relacionada ao volume de biomassa presente em uma área florestal. Ao integrar essas estimativas em modelos preditivos, é possível melhorar a precisão do mapeamento de biomassa.

2.2. Processamento dos dados

2.2.1. Pré-processamento

Para melhor aproveitamento dos dados de satélite, foi necessário realizar alguns pré-processamentos adicionais. Assim, foram aplicadas funções para mascaramento de nuvens nas imagens ópticas (Sentinel-2) e filtragem de ruído *speckle* para as imagens derivadas de SAR (Sentinel-1).

Também foi necessário realizar a reamostragem das diferentes bandas das imagens para ajustá-las para a mesma resolução espacial do conjunto de dados de referência. Assim, as bandas dos diferentes sensores foram re-amostradas para uma resolução espacial de 100 m. Isso é imprescindível para que todas as imagens tenham o mesmo tamanho de pixel e portanto sejam comparáveis quando sobrepostas (Jensen, J. 2015).

Por fim, foi preciso ajustar todos os conjuntos de dados para o mesmo sistema de referência de coordenadas. Assim, todas as imagens estão na projeção WGS 84 / EPSG: 4326, e portanto podem ser sobrepostas e comparadas geometricamente, para por exemplo, calcular área e afins (Jensen, J. 2015).

2.2.2. Amostragem em grade

Para coletar os dados das imagens de satélite, optou-se pela amostragem em grade (malha), distribuindo pontos uniformemente espaçados sobre a área de estudo com distâncias de 350 m. Em contraste, a amostragem aleatória seleciona pontos sem um padrão definido, o que pode resultar em áreas sub-amostradas ou super-amostradas.

Embora a amostragem aleatória reduza o viés, a amostragem em malha costuma ser mais eficiente para capturar padrões espaciais detalhados e garantir representatividade em áreas heterogêneas. Essa abordagem permite uma cobertura mais uniforme do terreno e facilita a captura de variações espaciais, tornando-se especialmente útil para mapeamentos de variáveis contínuas como o da biomassa.

2.2.3. Extração de características

A partir das coleções de dados de satélites, foram extraídas as características que servem como variáveis preditoras (x) para os modelos de aprendizado de máquina. Para obter as variáveis preditoras, optou-se pela integração de dados de sensores ópticos e radar (Sentinel-2 e Sentinel-1, respectivamente), considerando a complementaridade desses sensores para análises de estrutura, vigor e composição da vegetação. Essas características incluem índices de vegetação, parâmetros biofísicos, medidas texturais e topográficas (Molisse et al. 2022).

As imagens de satélite foram obtidas baseadas na correspondência temporal entre a coleção de dados ESA CCI Biomass e da localização da área de estudo para o ano de 2020, visando garantir a consistência temporal entre essas imagens e as datas dos produtos do ESA CCI Biomass. Dessa forma, as condições ambientais observadas nas imagens de satélite foram o mais semelhantes possível às condições observadas pelo produto ESA CCI Biomass. Portanto, é fundamental que haja correspondência e temporal entre as imagens de satélite e a coleção ESA CCI Biomass para poder comparar e interpretar os resultados (Hunka et al. 2023).

Utilizando a coleção de dados do Sentinel-2, foram calculados 17 índices de vegetação. Eles são: NDVI (Normalized Difference Vegetation Index), EVI (Enhanced Vegetation Index), SAVI (Soil Adjusted Vegetation Index), GNDVI (Green Normalized Difference Vegetation Index), que forneceram informações sobre a resposta espectral da vegetação, seu vigor e densidade. Além desses, também foram calculados o NBR (Normalized Burn Ratio), utilizado para detectar queimadas e áreas afetadas pelo fogo; o

NDMI (Normalized Difference Moisture Index), que ajuda a monitorar o teor de umidade da vegetação; e o NDWI (Normalized Difference Water Index), voltado para a detecção de áreas com presença de água. O MNDWI (Modified Normalized Difference Water Index) foi empregado para identificar corpos d'água em áreas urbanas.

Outros índices, como o NDVIre (NDVI de borda vermelha) e o PSRI (Plant Senescence Reflectance Index), forneceram informações adicionais sobre a saúde e senescência da vegetação. Adicionalmente, foram incluídos índices de reflectância espectral, como SRBlueRededge1, SRNIRnarrowRededge1, SRNIRnarrowRededge2, que permitem uma análise mais detalhada das diferentes bandas espectrais do infravermelho e da borda do vermelho, e os índices PSSRa (Pigment Specific Simple Ratio) e GCI (Green Chlorophyll Index), que auxiliaram no monitoramento de pigmentos específicos como a clorofila. Os índices Lcaroc e Lchloc forneceram informações sobre os carotenóides e a clorofila da vegetação. Ainda, foi calculado o parâmetro biofísico LAI (Leaf Area Index).

A partir das polarizações disponíveis na coleção de dados do Sentinel-1, foram gerados 5 índices de vegetação SAR e medidas de textura da superfície. Para sensores que operam na banda C, os índices de vegetação DPSVI (Dual Polarization SAR Vegetation Index) e DPSVIm (Modified Dual Polarization SAR Vegetation Index) se destacaram por utilizar as polarizações VV e VH para quantificar biomassa. Contudo, devido à saturação do sinal em áreas densamente florestadas, apenas o DPSVIm foi selecionado para este estudo, por sua capacidade de atenuar a saturação do sinal e melhorar o mapeamento da vegetação (dos Santos et al., 2021).

Além disso, foram incluídos outros índices relevantes, como o SARVI (SAR Vegetation Index), que aproveita a sensibilidade do radar para monitorar a vegetação em diferentes condições; a razão VV/VH, que permitiu uma análise da relação entre as duas polarizações; e o NDSI (Normalized Difference Backscattering Index), que ajudou a identificar variações no espalhamento de volta da superfície.

A partir do SRTM foi extraída a medida topográfica de elevação, fundamental para a compreensão da distribuição espacial da biomassa pelo terreno (Farr *et al.* 2007). Outro parâmetro biofísico importante é a altura da copa, extraído do conjunto de dados ETH Global Sentinel-2 10m Canopy Height (2020).

Posteriormente, foi necessário empilhar os conjuntos de dados em um único conjunto. O empilhamento visa combinar as camadas verticalmente para formar um único conjunto de dados multidimensional. Assim, cada linha no conjunto de dados representa um pixel na imagem, e cada coluna representa uma característica específica, as quais poderão ser comparadas em termos de importância.

2.2.4. Validação cruzada K-Fold e Padronização com RobustScaler

A validação cruzada é essencial para a implementação de modelos robustos à vieses. A validação cruzada acontece através da divisão dos dados em conjuntos de treino e validação, permitindo uma avaliação mais precisa da capacidade de generalização para dados novos, evitando o sobreajuste (“overfitting”) dos modelos (O’Reilly, 2019).

Neste estudo foi implementada a validação cruzada k-fold, uma técnica utilizada para avaliar a performance de modelos preditivos, dividindo o conjunto de dados em *k* subconjuntos ou *folds*. O modelo é treinado em *k-1* partes e testado na parte restante, repetindo esse processo *k* vezes, de modo que cada subconjunto seja utilizado tanto para treino quanto para teste.

Essa abordagem é importante para garantir que o modelo seja avaliado de forma robusta, minimizando o risco de sobreajuste e proporcionando uma estimativa mais confiável da sua capacidade de generalização. Também foi utilizado a função GroupKFold para garantir que as parcelas inteiras com um identificador comum permaneçam juntas em cada fold (O’Reilly, 2019).

Também foi utilizado o escalador RobustScaler do Scikit-Learn para padronizar os dados na mesma escala. Esse método foi escolhido em função da sua robustez à “outliers”, pois a distribuição de frequências dos valores de biomassa apresentaram alta variabilidade em função da heterogeneidade da área.

Contudo, somente o conjunto de treinamento foi utilizado para ajustar o escalador. Isso permitiu que os dados de validação permanecessem ocultos durante todo o processo, evitando assim o vazamento dos dados (“data leakage”). Por fim, todo o conjunto de dados é padronizado utilizando o escalador treinado. Este procedimento foi utilizado antes de cada etapa, incluindo seleção de características, ajuste dos modelos e ajuste fino dos hiperparâmetros.

2.2.5. Seleção de variáveis

Para construir um modelo confiável de estimativa de AGB, é necessário avaliar as variáveis preditoras (“features”) a serem incluídas no modelo preditivo. A escolha adequada das variáveis influencia diretamente a precisão do modelo e a interpretação dos resultados, sendo fundamental selecionar apenas aquelas que mais contribuem para a predição. Nesse contexto, dois métodos supervisionados foram avaliados para medir as importâncias das variáveis: o Decréscimo Médio em Impureza (Mean Decrease in Impurity - MDI) e o Decréscimo Médio em Acurácia (Mean Decrease in Accuracy - MDA).

O método MDI avalia a importância das variáveis com base na redução da impureza de nós em árvores de decisão, ou seja, mede o quanto uma feature contribui para a divisão dos dados dentro do modelo. Já o MDA avalia a diminuição da acurácia do modelo quando uma feature é removida, fornecendo uma indicação direta do impacto dessa variável na performance do modelo. Ambos os métodos são eficazes para identificar as variáveis mais relevantes, permitindo uma modelagem mais precisa e eficiente da biomassa (Guyon et al. 2003).

2.3. Ajuste dos modelos

2.3.1. Random Forest

Os modelos para a predição da biomassa foram ajustados utilizando algoritmos de aprendizado de máquina estatístico não-paramétricos. O primeiro modelo foi baseado no algoritmo *Random Forest* (RF), um modelo de agrupamento (*ensemble)* que utiliza a técnica de *bagging* em árvores de decisão (Bruce & Bruce, 2019).

A técnica de *bagging* consiste no ajuste de muitos modelos para amostras *bootstrapped* (amostragem com reposição) dos dados e tirando a média dos modelos. Assim, o Random Forest realiza a agregação *bootstrap* das árvores de decisão, dos registros e também nas variáveis (Bruce & Bruce, 2019).

As árvores de decisão são construídas a partir da partição recursiva dos dados a partir da minimização de critérios de impureza como a impureza de Gini e a entropia. Cada partição se refere a um valor específico de uma variável preditora e divide os dados em registros em que aquele valor preditor está acima ou abaixo do valor dividido, buscando divisões que aumentem a homogeneidade dos registros, ou seja, diminuam a impureza (Bruce & Bruce, 2019).

A biblioteca Scikit-Learn é um dos *frameworks* de aprendizado de máquina mais populares em Python, oferecendo uma ampla gama de ferramentas para tarefas de classificação, regressão e agrupamento. A implementação deste algoritmo se deu através do método *sklearn.ensemble.RandomForestRegressor* da biblioteca Scikit-Learn, utilizando os hiperparâmetros padrões.

2.3.2. Redes Neurais Artificiais

As redes neurais artificiais (ANNs), implementadas com *frameworks* como Keras e Tensor Flow, surgem como ferramentas poderosas de aprendizado de máquina, oferecendo alta acurácia e flexibilidade na modelagem (Géron, A. 2019).

Neste estudo foi implementada uma rede neural profunda (Deep Neural Network - DNN) com uma arquitetura sequencial, composta por diversas camadas densas, alinhada aos princípios fundamentais de aprendizado de máquina. Cada camada densa incluiu um número específico de neurônios, sendo 128 na primeira, 64 na segunda, 32 na terceira, 16 na quarta, 8 na última camada intermediária, todas utilizando a função de ativação ReLU, a qual introduziu não-linearidade ao modelo, permitindo que padrões mais complexos fossem aprendidos. Por fim, a camada de saída contém 1 neurônio com uma função de ativação, que indica que o modelo está tentando prever um valor numérico contínuo, o que é típico em tarefas de regressão, como estimar AGB.

Adicionalmente, técnicas de regularização, como “Dropout”, foram implementadas em duas camadas, com taxas de 30% e 10%, respectivamente, visando reduzir o risco de sobreajuste ao desativar aleatoriamente neurônios durante o treinamento. A camada de saída, composta por um único neurônio com ativação linear, foi configurada para realizar tarefas de regressão, já que o objetivo final do modelo consistiu em prever valores contínuos.

Para a compilação da rede, utilizou-se o otimizador Adam, conhecido por combinar as vantagens do método de gradiente estocástico e *momentum*, e a função de perda erro quadrático médio (MSE), adequada para problemas de regressão. O treinamento do modelo ocorreu durante 20 épocas, com o conjunto de treinamento dividido em batches, otimizando a atualização dos pesos a cada iteração.

Essa abordagem permitiu que o modelo fosse ajustado a partir de diferentes combinações de variáveis preditoras ao longo das dobras geradas pela validação cruzada. Ao final, o erro médio quadrático (RMSE) foi calculado para cada dobra, fornecendo uma métrica de desempenho robusta para avaliar a capacidade preditiva do modelo em dados não vistos.

Além disso, foi implementado o parâmetro “Early Stopping”, que interrompe o treinamento quando o modelo começa a apresentar um aumento na função de perda em relação ao conjunto de validação. Esse parâmetro garante que o treinamento será interrompido automaticamente quando não houver mais melhoria na função de perda na validação após um número definido de épocas (“patience”), preservando os melhores pesos obtidos até o momento.

A taxa de aprendizado (“learning rate”) não foi explicitamente configurada, o que implica que o otimizador Adam utilizou seu valor padrão (0.001). A taxa de aprendizado é um dos hiperparâmetros mais importantes em redes neurais, pois define o tamanho dos passos que o otimizador dá na direção de minimizar a função de perda.

2.3.3. Extreme Gradient Boosting (XGBoost)

O XGBoost (Extreme Gradient Boosting), é um algoritmo “ensemble” como o Random Forest, que combina árvores de decisão por meio da técnica de “boosting” (Chen & Guestrin, 2016). Ao contrário do “bagging”, o “boosting” ajusta sequencialmente um conjunto de modelos, onde cada modelo subsequente tenta corrigir os erros cometidos pelos anteriores. O XGBoost aprimora esse processo ao aplicar regularização para evitar sobreajuste, além de otimizações que aumentam sua eficiência computacional e desempenho preditivo (Chen & Guestrin, 2016).

Cada árvore de decisão no XGBoost é construída a partir de partições recursivas dos dados, com o objetivo de minimizar uma função de perda, como erro quadrático médio para problemas de regressão. As partições são feitas com base nos valores das variáveis preditoras, dividindo os dados em grupos mais homogêneos. O XGBoost também permite o uso de pesos e a manipulação de resíduos durante o processo de aprendizado, tornando o ajuste dos modelos mais flexível e robusto (Chen & Guestrin, 2016).

A implementação do XGBoost foi realizada através da biblioteca xgboost em Python, utilizando o método XGBRegressor, com hiperparâmetros ajustados para otimizar o desempenho do modelo. Essa biblioteca é amplamente reconhecida por sua eficiência e por oferecer diversas opções de regularização e controle do processo de boosting, permitindo o ajuste fino do modelo conforme as características dos dados.

2.4. Otimização dos hiperparâmetros

Para evitar o sobreajuste dos dados e aprimorar a acurácia dos resultados, é preciso realizar a otimização dos hiperparâmetros. Ao invés de adivinhar ou testar repetidamente os hiperparâmetros, pode-se testar programaticamente uma variedade de valores para escolher o menor valor possível que alcance a maior acurácia (Géron, A 2019).

Para os modelos Random Forest e XGBoost, optou-se pela Otimização Bayesiana, um método probabilístico que utiliza informações das avaliações anteriores para direcionar de maneira mais eficiente a busca por melhores hiperparâmetros. Ambos os métodos oferecem uma exploração mais eficaz do espaço de pesquisa, resultando em soluções mais próximas do ideal em comparação com abordagens mais simples, como a busca em grade. No entanto, é importante considerar que esses algoritmos podem ser mais complexos de implementar e, em alguns casos, podem enfrentar desafios de convergência (Géron, A 2019).

Para otimização dos hiperparâmetros da DNN, além do algoritmo Adam, amplamente utilizado por sua capacidade de ajustar as taxas de aprendizado de forma adaptativa, acelerando a convergência e melhorando o desempenho em comparação a métodos clássicos, como o SGD (Stochastic Gradient Descent), também foi utilizado o método "Random Search".

O método "Random Search" foi utilizado para otimizar os hiperparâmetros da rede neural, testando combinações aleatórias de valores dentro de intervalos pré-definidos. A vantagem do "Random Search" sobre abordagens como a "Grid Search" é a sua eficiência: em vez de testar todas as combinações possíveis de hiperparâmetros, o "Random Search "explora apenas um subconjunto aleatório, o que pode levar a boas soluções mais rapidamente, especialmente quando o número de hiperparâmetros e suas possíveis combinações é muito grande. Os hiperparâmetros otimizados na função foram o número de neurônios nas camadas ocultas, as taxas de "Dropout", e a taxa de aprendizagem.

2.5. Avaliação de acurácia

Na avaliação do desempenho de modelos de estimativa de biomassa, diversas métricas são utilizadas para quantificar sua precisão e capacidade de explicar a variabilidade nos dados. Uma das métricas mais comuns é o coeficiente de determinação (R^2), que indica a proporção da variabilidade da biomassa explicada pelo modelo.

Adicionalmente, o R^2 ajustado oferece uma visão mais precisa da qualidade do ajuste do modelo ao levar em conta o número de preditores no modelo. Diferentemente do R^2, o R^2 ajustado penaliza a inclusão de variáveis irrelevantes, sendo especialmente útil em modelos com muitos parâmetros. Um R^2 ajustado mais alto reflete a habilidade do modelo de explicar a variabilidade da biomassa sem inflar artificialmente o desempenho devido ao número de preditores (Géron, A. 2019).

Outra métrica amplamente utilizada é o erro quadrático médio (Root Mean Squared Error - RMSE), que mede a raiz quadrada da diferença média entre os valores reais e preditos da biomassa. O RMSE expressa a magnitude do erro de predição nas mesmas unidades dos dados, com valores mais baixos indicando maior precisão nas estimativas de biomassa (Géron, A. 2019).

O erro médio absoluto (Mean Absolute Error - MAE) é uma métrica complementar, que também quantifica a diferença entre os valores reais e preditos, mas utilizando a média das diferenças absolutas. Ao contrário do RMSE, o MAE não penaliza grandes erros de forma desproporcional, sendo mais robusto a outliers. Valores baixos de MAE indicam que,

em média, as predições do modelo são próximas dos valores observados, também sugerindo uma boa acurácia (Géron, A. 2019).

Essas métricas fornecem uma avaliação abrangente do desempenho do modelo, cada uma oferecendo uma perspectiva diferente sobre a qualidade das predições, permitindo um entendimento mais robusto sobre a acurácia e precisão dos modelos de biomassa.

Adicionalmente, foi implementado um teste U de Mann-Whitney para avaliar se as diferenças entre as métricas dos modelos são estatisticamente significativas.

2.6. Explanação dos modelos com pacote SHAP

À medida que a complexidade de modelos de aprendizado de máquina aumenta, torna-se mais desafiador entender como determinadas previsões são feitas e quais características têm maior impacto nessas previsões. Enquanto modelos simples possibilitam uma interpretação direta observando os pesos das características, modelos mais complexos, como ensembles e redes neurais profundas, são menos transparentes.

Nesse sentido, o pacote SHapley Additive exPlanations (SHAP) além de mostrar a importância das características, explica como cada uma contribui individualmente para as previsões finais. Isso possibilita uma interpretação mais detalhada do modelo e a identificação de como o comportamento das variáveis influenciam os modelos. Por fim, sua flexibilidade permite que seja aplicado a uma ampla variedade de modelos de aprendizado de máquina, tornando-o uma ferramenta versátil para análise de resultados (Lundberg et al. 2020).

3. Resultados e Discussão

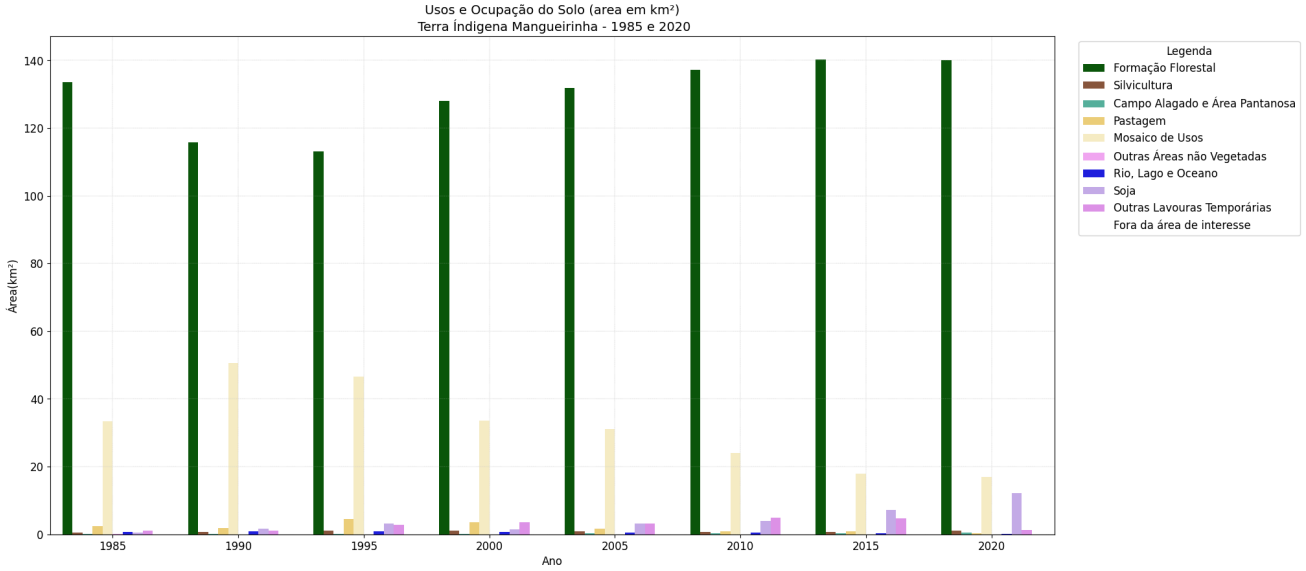
3.1. Análise Exploratória dos Dados

3.1.1. Uso e Ocupação do Solo da Área de Estudo (MapBiomas)

A partir da delimitação da área de estudo, o dado dos limites administrativos foram coletados em formato shapefile (vetorial) a partir da tabela de camadas do MapBiomas, tendo a FUNAI como fonte principal. Posteriormente, este dado foi transformado em formato GeoJSON e integrado ao Google Earth Engine.

Para analisar o uso e ocupação da terra da área de estudo, foi utilizada a Coleção 7 do MapBiomas. A partir desse conjunto de dados foi analisada a evolução temporal e espacial das classes de uso desde 1985 até 2020, conforme apresentado na Figura 1.

Figura 1. Uso e Ocupação do Solo - 1985 à 2020



Nota-se na Figura 1, um aumento da classe mosaico de usos em detrimento de formação florestal durante a década de noventa. Já nos anos 2000 em diante, a classe formação florestal volta a aumentar, bem como o cultivo de soja e áreas não vegetadas, enquanto o mosaico de usos diminui, o que pode indicar regeneração natural e também conversão para áreas agrícolas.

Na Figura 2, podemos observar os valores de área em km² (quilômetros quadrados) das classes presentes na área de estudo para o ano de 2020. Entre as classes, destacam-se “Formação Florestal", “Mosaico de Usos” e "Soja”, com valores de 140 km², 16,9 km² e 12,1 km², respectivamente.

Tabela 1: Área (km²) das classes de uso e ocupação do solo em 2020

Classe	Área (km²)
Formação Florestal	140,04
Silvicultura	1,07
Campo Alagado e Área Pantanosa	0,46
Pastagem	0,42
Mosaico de Usos	16,89
Rio, Lago e Oceano	0,14
Soja	12,11
Outras Lavouras Temporárias	1,33

Para visualizar essas transições entre as classes, foi gerado um diagrama de sankey interativo através das bibliotecas *sankee* e *plotly*, que permite avaliar o fluxo de um conjunto de valores para outro e ajudam a analisar a movimentação de dados ao longo do tempo, como conversão de uso e ocupação do Terra. Assim, constatou-se que de 1985 a 2001, 6% da classe “Mosaico de Usos” foi convertida em “Pastagem” e 7% em “outras lavouras temporárias”.

Já em relação a classe “Outras Lavouras Temporárias”, observa-se que 50% foi convertida em “Soja” até o ano de 2001. Para a classe de “Formação Florestal”, houve uma transição de 11% para “Mosaico de Usos”, enquanto 67% da classe “Pastagem” se tornou “Mosaico de Usos”. Já de 2001 a 2021, 89% da classe “Outras Lavouras Temporárias” se tornou “Soja”. Ainda, 1% da classe formação florestal e 20% da classe “Mosaico de Usos” foram convertidas na classe “Soja”. Tais resultados indicam que há uma pressão antrópica de conversão da classe “Formação Florestal”, principalmente para produção soja, ainda que apresentem níveis relativamente baixos. Para melhor visualização do diagrama de sankey, recomenda-se utilizar o código disponível no apêndice deste estudo.

3.1.2. ESA CCI Global Forest Above Ground Biomass

Inicialmente, foram calculadas algumas estatísticas descritivas para um entendimento inicial das amostras, conforme apresentado na Tabela 1:

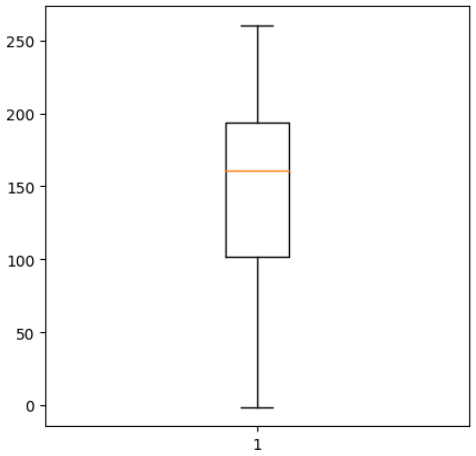
Tabela 2: Estatísticas descritivas das amostras de AGB (t/ha)

Medida	Valor
Contagem	1541
Média	144,7
Desvio padrão	63,7
Mínimo	1
25%	101
50% (mediana)	161
75%	194
Máximo	261

Observa-se que a mediana (161 t/ha) é maior que a média (144,7 t/ha), o que pode indicar uma leve assimetria, com os dados inclinados para a direita. O intervalo interquartil (IQR), que é a diferença entre o terceiro quartil (194 t/ha) e o primeiro quartil (101 t/ha), é 93 t/ha, indicando a dispersão central dos dados.

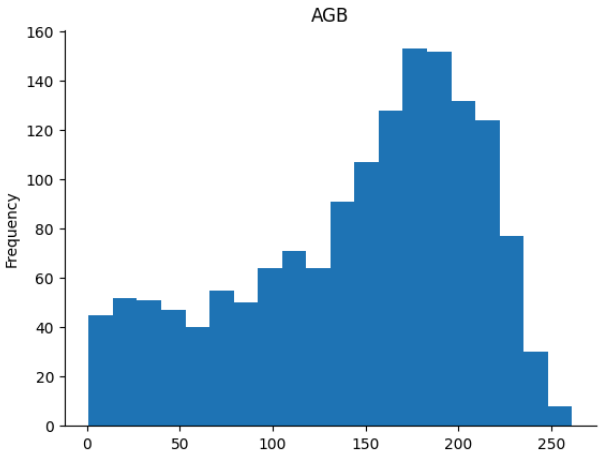
Em relação à dispersão e variabilidade, o desvio padrão (63,7 t/ha) é relativamente alto, sugerindo que os dados estão dispersos ao redor da média. Para melhor visualização dessas estatísticas descritivas de forma esquemática, é apresentado um box-plot na Figura 2.

Figura 2. Box-plot das amostras de AGB (t/ha)



Para visualizar a distribuição das frequências dos valores das amostras, um histograma é apresentado na Figura 3. Na Figura 3, observa-se a partir do histograma que a maior frequência de valores está no intervalo de aproximadamente 140-240.

Figura 3. Histograma das amostras de AGB (t/ha)



Valores mais baixos entre 0 e 100 t/ha e valores muito altos (> 200 t/ha) têm menor frequência, indicando uma menor ocorrência desses valores no conjunto de dados. A forma da distribuição é similar a assimétrica negativa. Há uma maior concentração de valores entre 150 e 220, indicando uma inclinação dos dados para a direita.

Também foi calculado o valor de AGB para toda a área de estudo através do método *reduceRegion* utilizando o redutor de média do Google Earth Engine para comparação com os dados de referência do IFN. O cálculo mostrou uma diferença de -5,7 t/ha entre o IFN (151 t/ha) e o ESA CCI Biomass (145,3 t/ha), representando 3,77% do valor da estimativa do IFN, indicando que apesar da subestimação, este conjunto de dados está capturando a biomassa de forma razoavelmente precisa em relação ao IFN.

Nesse sentido, o conjunto de dados ESA CCI AGB se mostrou confiável para estimativas de biomassa acima do solo em escalas regionais, embora tenha limitações em áreas heterogêneas devido à sua resolução de 100 metros.

3.2. Seleção de variáveis

Foi implementada uma função que realiza uma seleção de variáveis utilizando a técnica de validação cruzada k-fold para testar a performance de diferentes combinações de variáveis na predição de biomassa tanto para o modelo Random Forest quanto para o modelo de Rede Neural Artificial. O processo começa dividindo o conjunto de dados em sete partes (folds), onde seis são usadas para treino e uma para validação, repetindo o ciclo até que todas as partes tenham sido utilizadas tanto para treino quanto para validação.

Durante cada iteração, uma lista de variáveis é testada sequencialmente, e o modelo é treinado com um número crescente de variáveis selecionadas. Para cada combinação, o modelo realiza previsões sobre o conjunto de validação, e o erro quadrático médio (RMSE) é calculado, indicando a precisão do modelo para a seleção de variáveis em questão em cada dobra.

Por fim, os resultados de todas as dobras são agregados, e a média dos RMSE é calculada, juntamente com o desvio padrão para avaliar a variabilidade dos resultados. Um gráfico é gerado ao final da função exibindo o desempenho do modelo conforme o número de variáveis selecionadas aumenta, permitindo identificar o conjunto de variáveis que minimiza o RMSE, ou seja, que resulta na melhor performance preditiva.

Esse processo foi importante para otimizar a eficiência e a precisão do modelo, evitando o uso de variáveis irrelevantes que possam aumentar o erro e tornar o modelo mais complexo sem ganhos significativos de desempenho. Os resultados desse processo podem ser observados na Figura 4, onde o melhor desempenho é obtido a partir do método MDA,

atingindo um valor mínimo de RMSE de 38,12 t/ha com 10 variáveis para o modelo Random Forest.

Considerando que o método MDA obteve o melhor resultado utilizando 10 variáveis, foram selecionadas as 10 variáveis melhor colocadas no ranking de importância. Assim, as variáveis selecionadas foram: 'b1' (Altura da copa), 'NBR', 'Sigma0_VH', 'elevation', 'DPSVIm', 'Sigma0_VV', 'Lcaroc', 'PSRI', 'MNDWI', 'SRNIRnarrowRededge2'. Na Figura 5, podemos observar que o método MDA também obteve o melhor desempenho para o modelo XGBoost com 8 variáveis.

Figura 4: Teste de seleção de variáveis com Random Forest

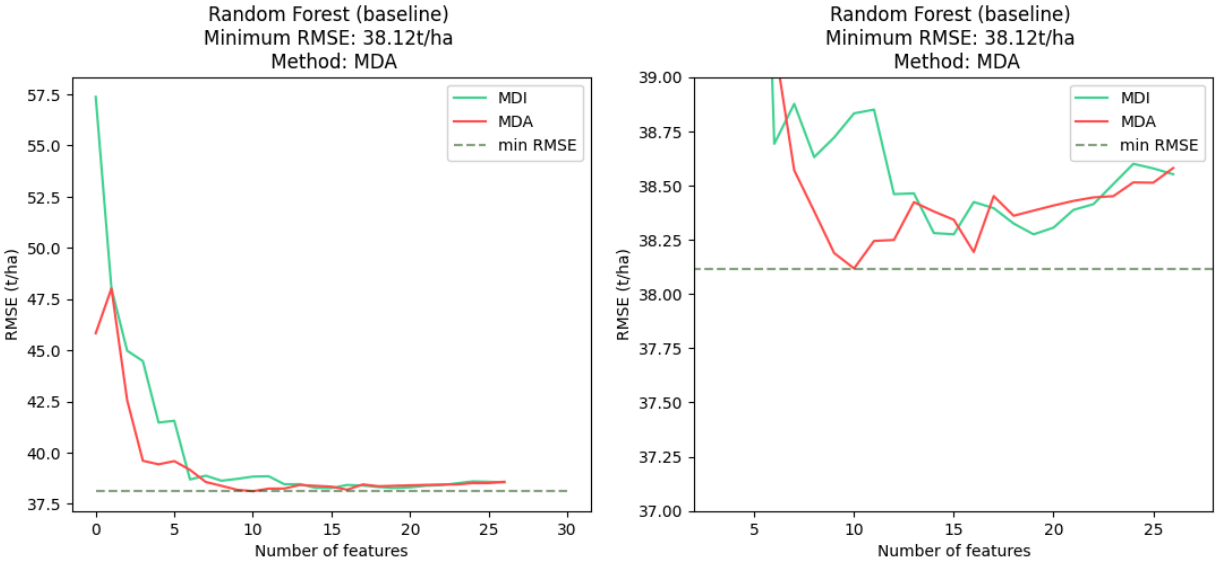
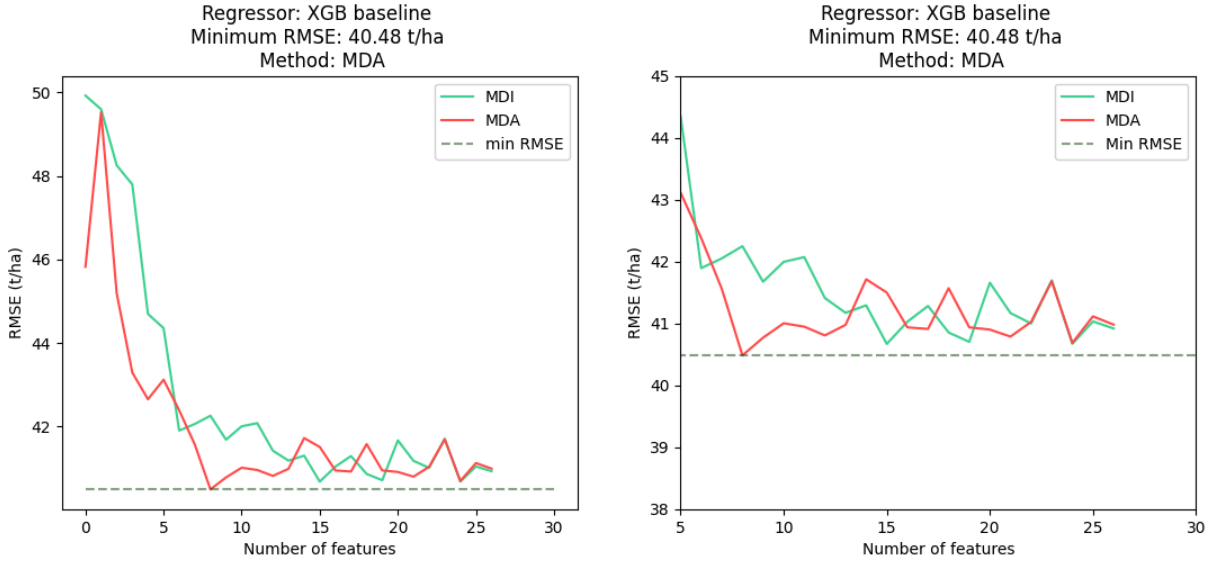
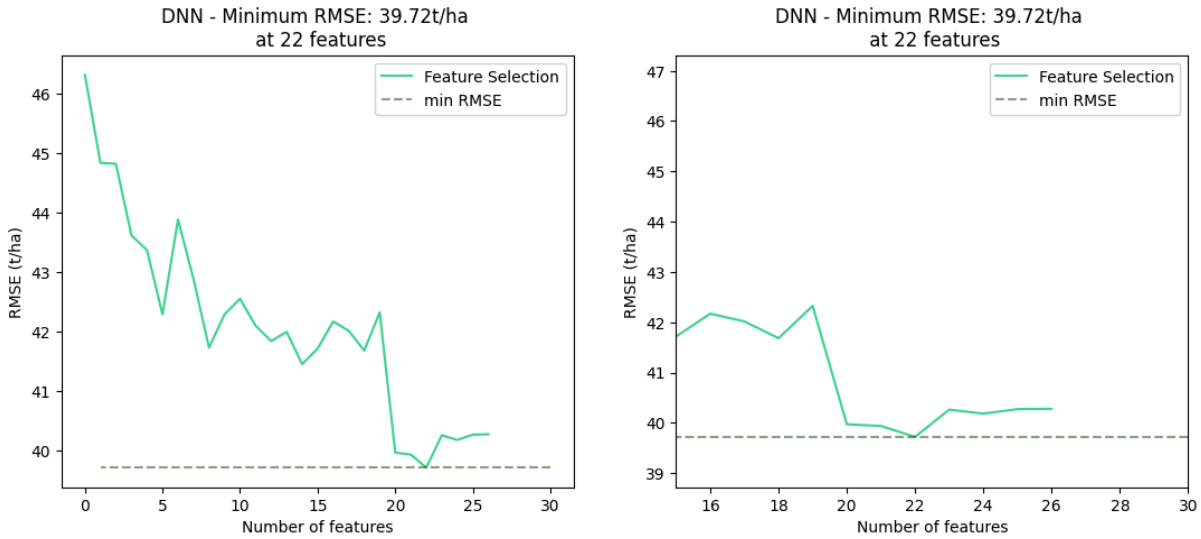


Figura 5: Teste de seleção de variáveis com XGBoost



Devido a problemas de convergência e um comportamento errático da DNN durante a tarefa de seleção de variáveis pelo mesmo método dos modelos anteriores, foi implementada uma função diferente que avalia o desempenho do modelo à medida que são adicionadas as variáveis. O gráfico do desempenho da DNN é apresentado na Figura 6, onde pode ser observado que o modelo alcançou o melhor desempenho com 22 variáveis, atingindo um RMSE de 39,72 t/ha.

Figura 6: Teste de seleção de variáveis com DNN



É possível considerar que essas variáveis foram selecionadas porque cada uma delas, direta ou indiretamente, está associada a fatores que influenciam a estrutura, saúde, ou densidade da vegetação. A biomassa está diretamente relacionada à estrutura e vigor da vegetação, e as variáveis escolhidas capturam informações essenciais sobre:

- Estrutura física da vegetação: ‘b1’ (altura da copa), 'Sigma0_VH', 'DPSVIm', 'Sigma0_VV'.
- Saúde e vigor da vegetação: ‘NBR’, ‘PSRI’, 'SRNIRnarrowRededge2', ‘Lcaroc’ .
- Características ambientais: ‘elevation’, ‘MNDWI’.

Portanto, é corroborado que tanto os índices de vegetação quanto os parâmetros biofísicos e medidas texturais têm uma contribuição importante na modelagem da biomassa. Essas variáveis, em conjunto, cobrem aspectos biofísicos (estrutura, altura, e densidade da vegetação), bioquímicos (teores de clorofila, senescência), e ambientais (disponibilidade de água, elevação), fornecendo ao modelo uma visão abrangente dos fatores que influenciam a biomassa. Posteriormente, aspectos do comportamento das variáveis selecionadas serão mais explorados a partir da implementação do pacote SHAP.

3.4. Otimização dos hiperparâmetros

Para otimização dos hiperparâmetros, foi implementada uma função que realiza validação cruzada k-fold dividindo os dados em diferentes dobras. Para cada dobra, o modelo é treinado em um conjunto de treino e avaliado em um conjunto de validação. O erro quadrático médio (RMSE) é calculado para cada dobra, medindo a precisão das predições feitas pelo modelo em relação aos valores reais.

Ao final, a função retorna o valor negativo do RMSE médio entre todas as dobras, pois a otimização bayesiana maximiza a função objetivo, enquanto o foco está em minimizar o erro. Esse processo passou por cinquenta iterações para ajustar os hiperparâmetros, buscando a melhor configuração que minimize o erro nas predições. É possível considerar um aumento na quantidade de iterações, apesar do alto custo computacional.

Para o modelo Random Forest, os hiperparâmetros testados incluem o número de árvores (*n_estimators*), a profundidade máxima das árvores (*max_depth*), e os critérios mínimos para a divisão dos nós (*min_samples_split*) e folhas (*min_samples_leaf*).

Com os resultados da função de otimização, um modelo Random Forest foi ajustado a partir dos hiperparâmetros otimizados, resultando num RMSE de 38,02 t/ha. Houve uma redução baixa no RMSE em relação ao modelo não-otimizado, cerca de 0,1 t/ha, o que pode indicar que o modelo pode já estar bem ajustado aos dados. Ainda que seja possível aumentar o número de iterações da função de otimização em busca de melhoria, há de se considerar o alto custo computacional desta tarefa.

Para o modelo XGBoost, os hiperparâmetros ajustados incluem a profundidade máxima das árvores *(max_depth*), o termo de regularização L1 (*alpha*), o número de árvores (*n_estimators*), a taxa de aprendizado (*learning_rate*), o parâmetro de regularização para controle de complexidade do modelo (*gamma*), e a fração de amostras utilizadas para treinar cada árvore (*subsample*).

Com os resultados da otimização, um modelo XGBoost foi ajustado utilizando os hiperparâmetros otimizados, resultando em um RMSE de 38,3 t/ha. Comparado ao modelo sem otimização, que apresentou um RMSE de 40,48 t/ha, houve uma redução de 2,18 t/ha no RMSE. Essa diminuição, embora considerável, sugere que o modelo já estava razoavelmente bem ajustado aos dados antes da otimização, embora haja algum espaço para melhorias.

Para o modelo de Redes Neurais Artificiais (DNN), os hiperparâmetros ajustados incluem o número de unidades (neurônios) em cada camada oculta, a taxa de “dropout” e a taxa de aprendizado. No modelo otimizado, a primeira camada possui 128 unidades, sem aplicação de dropout, enquanto a segunda camada tem 80 unidades e a terceira 48

unidades, com um dropout de 0,2 na terceira camada, uma quarta camada com 20 unidades e uma quinta com 8 unidades, além da camada de saída que foi mantida como antes. A taxa de aprendizado final foi de 0.001.

Após a otimização dos hiperparâmetros, o modelo DNN apresentou uma média de RMSE de 38,85 t/ha e uma média de R^2 de 61% ao longo da validação cruzada. Comparando com o modelo sem otimização, que obteve um RMSE de 39,72 t/ha, a otimização resultou em uma redução de 0,87 t/ha no RMSE. Embora essa diminuição seja modesta, ela indica que a otimização dos hiperparâmetros contribuiu para uma leve melhoria no ajuste do modelo.

Ainda que a otimização de hiperparâmetros tenha adicionado um ganho de desempenho, principalmente no XGBoost, isso pode indicar que há possibilidade para aprimorar a otimização, aumentando o número de iterações e também o alcance (range) dos parâmetros testados. Também é plausível considerar a adição de novas variáveis com maior correlação com a variável alvo e menor correlação entre as variáveis preditoras.

Além disso, deve-se considerar o teste de outras técnicas de padronização dos dados ao invés do *RobustScaler*, como o *MinMaxScaler* ou *StandardScaler* do Scikit-Learn. Outras técnicas de seleção de variáveis também podem ser implementadas e comparadas em termos de desempenho como a Regularização L1 (Lasso) e Análise de Componentes Principais (“Principal Component Analysis” - PCA).

3.5. Avaliação de acurácia

A avaliação de acurácia foi realizada utilizando as métricas obtidas através da validação cruzada dos modelos otimizados. Para cada dobra (fold), as métricas analisadas incluem o Erro Quadrático Médio (RMSE), o Erro Absoluto Médio (MAE), o Coeficiente de Determinação (R^2) e R^2 ajustado. Além disso, o teste de Mann-Whitney foi empregado para determinar se existem diferenças estatisticamente significativas nas distribuições das métricas de desempenho entre os modelos.

Para o Random Forest, o R^2 ajustado foi calculado em 0.6273, indicando que aproximadamente 62,73% da variabilidade dos dados é explicada por este modelo. Já o XGBoost apresentou R^2 ajustado de 0.6209, explicando cerca de 62,09% da variabilidade dos dados. Embora ligeiramente inferior ao R^2 ajustado do modelo Random Forest, ainda demonstra um bom nível de explicabilidade. A Rede Neural Artificial (DNN) apresentou um R^2 ajustado de 0.6043, o que significa que aproximadamente 60,43% da variabilidade dos dados é explicada. Embora o DNN também tenha um desempenho aceitável, é o modelo que apresenta a menor capacidade de explicação entre os três.

Na Tabela 3, podemos observar as métricas de validação cruzada K-fold dos modelos otimizados:

Tabela 3: Resultados da Validação Cruzada K-Fold

Modelo	RMSE (t/ha)	R^2	MAE (t/ha)
RF (Fold 1)	32,91	71,63%	26,21
RF (Fold 2)	37,76	56,28%	29,26
RF (Fold 3)	36,84	67,37%	29,40
RF (Fold 4)	39,49	66,86%	30,65
RF (Fold 5)	37,01	60,80%	30,13
RF (Fold 6)	39,29	66,28%	31,20
RF (Fold 7)	42,84	51,57%	32,97
RF - Média	**38,02**	**62,97%**	**29,97**
DNN (Fold 1)	31,97	73,22%	24,99
DNN (Fold 2)	39,59	51,94%	31,45
DNN (Fold 3)	36,77	67,49%	29,14
DNN (Fold 4)	39,49	66,86%	31,59
DNN (Fold 5)	38,49	57,58%	31,71
DNN (Fold 6)	39,69	65,58%	32,12
DNN (Fold 7)	45,98	44,23%	36,85
DNN - Média	**38,86**	**60,99%**	**31,12**
XGB (Fold 1)	32,17	7.289	25,27
XGB(Fold 2)	38,80	5.386	30,92
XGB(Fold 3)	37,39	6.640	30,21
XGB(Fold 4)	39,19	6.737	30,71
XGB (Fold 5)	38,71	5.710	31,63
RF (Fold 6)	39,15	6.652	30,98
RF (Fold 7)	42,70	5.190	32,95
XGB - Média	**38,30**	**6.229**	**30,38**

O teste de Mann-Whitney foi empregado para determinar se existem diferenças estatisticamente significativas nas distribuições das métricas de desempenho entre os

modelos. Esse teste não paramétrico é adequado para comparar duas amostras independentes e avaliar se uma das amostras tende a ter valores maiores do que a outra.

Na Tabela 4, podemos observar os resultados dos testes de Mann-Whitney:

Tabela 4: Resultados dos testes de Mann-Whitney

Comparação	Estatística U	p-valor
RF vs DNN (RMSE)	19.0	0,535
RF vs XGB (RMSE)	24.0	1,000
DNN vs XGB (RMSE)	28.0	0,710
RF vs DNN (MAE)	17.0	0,383
RF vs XGB (MAE)	18.0	0,456
DNN vs XGB (MAE)	30.0	0,535

Os resultados desse teste, que foram aplicados às métricas RMSE e MAE, indicaram que não há evidências suficientes para rejeitar a hipótese nula, sugerindo que as diferenças observadas nas métricas de desempenho entre os modelos Random Forest, XGBoost e DNN não são estatisticamente significativas.

Através dessa análise, é possível concluir que, embora haja variações nas métricas de desempenho, os modelos apresentados oferecem resultados comparáveis, permitindo que a escolha do modelo seja baseada em outros fatores, como interpretabilidade e complexidade de implementação em produção, em vez de apenas na acurácia.

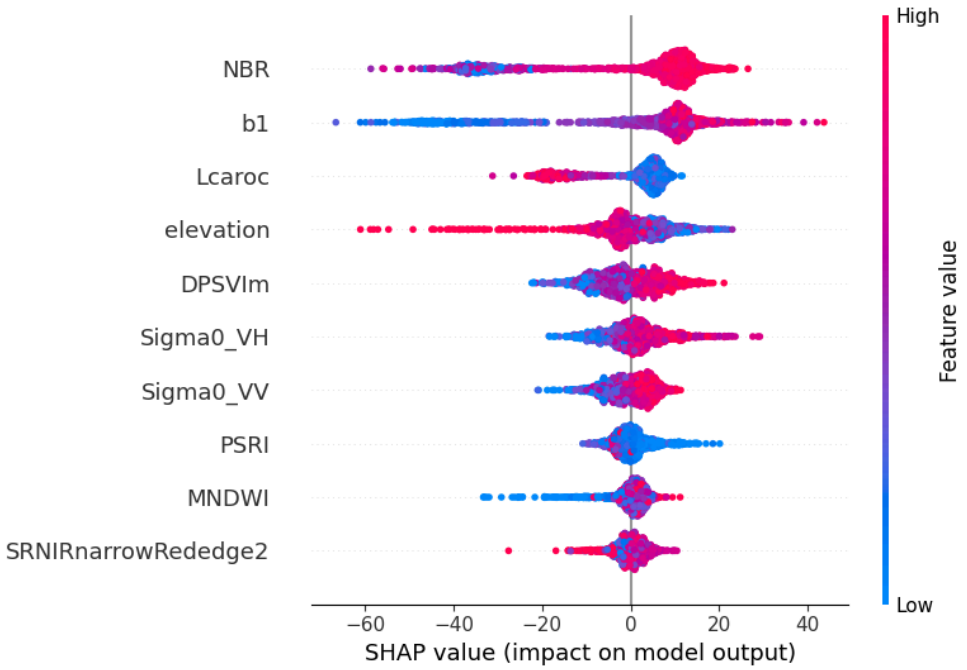
3.6. Explanação dos modelos com SHAP

Para tentar entender em mais detalhes o comportamento das variáveis nos diferentes modelos, foi utilizado o pacote SHAP. A seguir, são apresentados os gráficos resultantes da implementação desse pacote em relação às variáveis selecionadas e seus respectivos modelos.

A Figura 7 ilustra a importância das variáveis nas predições do modelo Random Forest. A variável ‘NBR’ destaca-se como a mais influente, apresentando grande variabilidade em seu impacto sobre as estimativas de biomassa.

Altos valores de ‘NBR’ (em vermelho) aumentam a predição, enquanto baixos valores (em azul) a diminuem, indicando que áreas com vegetação saudável estão correlacionadas a uma maior biomassa.

Figura 7: Gráfico de comportamento das variáveis no Random Forest



A variável ‘b1’ também é significativa, mostrando uma dispersão elevada nos valores SHAP, o que sugere efeitos variáveis nas predições. O mesmo se aplica a ‘Lcaroc’, cujos níveis de carotenóides impactam substancialmente a estimativa de biomassa.

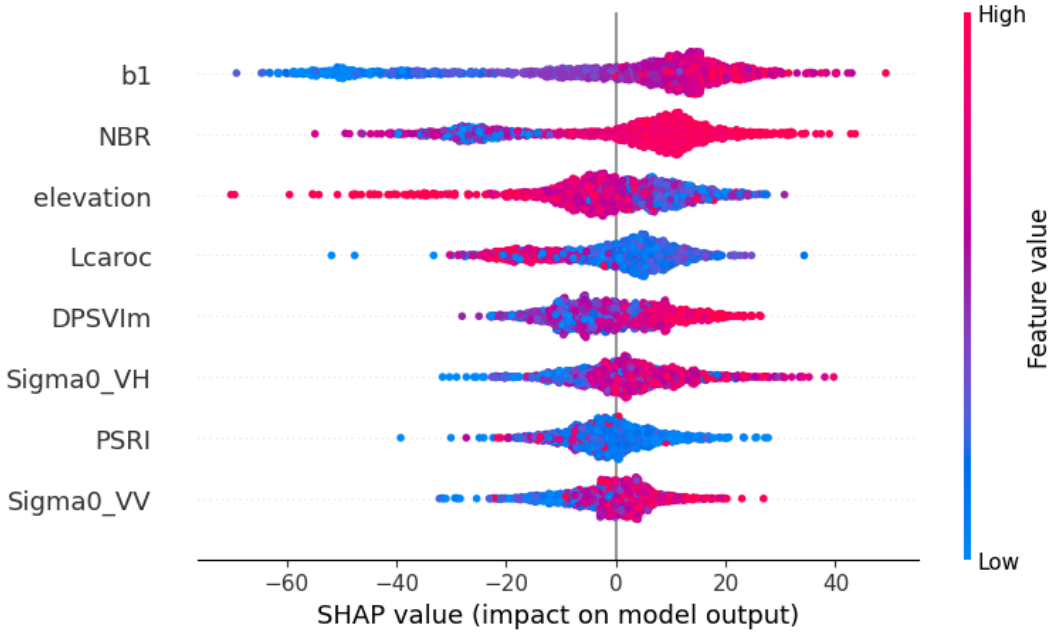
Em contraste, a variável ‘elevation’ apresenta uma relação mais clara, com elevações altas (em vermelho) associadas a reduções na biomassa, refletindo condições ambientais de menor densidade vegetacional.

As variáveis de radar, ‘Sigma0_VH’ e ‘Sigma0_VV’, demonstram um impacto mais consistente, com menor dispersão nos valores SHAP, indicando que o sinal de retorno do radar afeta as estimativas de forma previsível. Valores baixos dessas variáveis estão relacionados a uma diminuição da biomassa, enquanto valores altos indicam vegetação densa.

Outras variáveis, como ‘DPSVIm’ e ‘MNDWI’, apresentam variabilidade, mas com efeitos menos significativos. Os resultados enfatizam a sensibilidade do modelo a indicadores biofísicos e estruturais, especialmente os relacionados à saúde da vegetação. Assim, a modelagem da biomassa requer uma consideração cuidadosa das múltiplas variáveis e suas interações, dado que o impacto de cada uma depende de seu valor específico no contexto do modelo.

Em relação ao modelo XGBoost, podemos notar na Figura 8, algumas diferenças no comportamento das variáveis em relação ao modelo Random Forest.

Figura 8: Gráfico de comportamento das variáveis no XGBoost



No XGBoost, a variável ‘b1’ (altura da copa) emerge como a mais relevante, desempenhando um papel central nas predições, com uma ampla variação em seus valores SHAP, o que sugere um impacto considerável, tanto positivo quanto negativo. Isso contrasta com o modelo anterior, onde o NBR era a variável mais influente, indicando que cada algoritmo dá prioridade a diferentes variáveis na modelagem da biomassa.

Além disso, no XGBoost, o impacto das variáveis parece ser mais extremo, com ‘b1’ e NBR apresentando uma amplitude maior de valores SHAP, o que reflete uma sensibilidade maior do modelo a essas variáveis, potencialmente capturando interações complexas que o modelo anterior não detectou. Variáveis como ‘elevation’ e ‘Sigma0_VH’ também mostram um impacto mais equilibrado e simétrico no XGBoost, sugerindo que esse modelo é mais eficaz em capturar variações sutis nos dados.

Outra conclusão relevante é que o XGBoost tende a gerar uma maior dispersão nos impactos das variáveis, especialmente em ‘b1’ e ‘NBR’, o que pode indicar que o modelo lida melhor com variáveis que têm efeitos mais dinâmicos e não-lineares sobre a biomassa.

De forma geral, o XGBoost demonstra uma capacidade de ajustar as interações entre as variáveis e suas predições, tornando-o mais eficiente em explorar a variabilidade presente nos dados e capturar os efeitos complexos que influenciam a biomassa. Devido a dificuldades operacionais, o pacote SHAP não foi implementado para o modelo de Rede Neural Artificial (DNN).

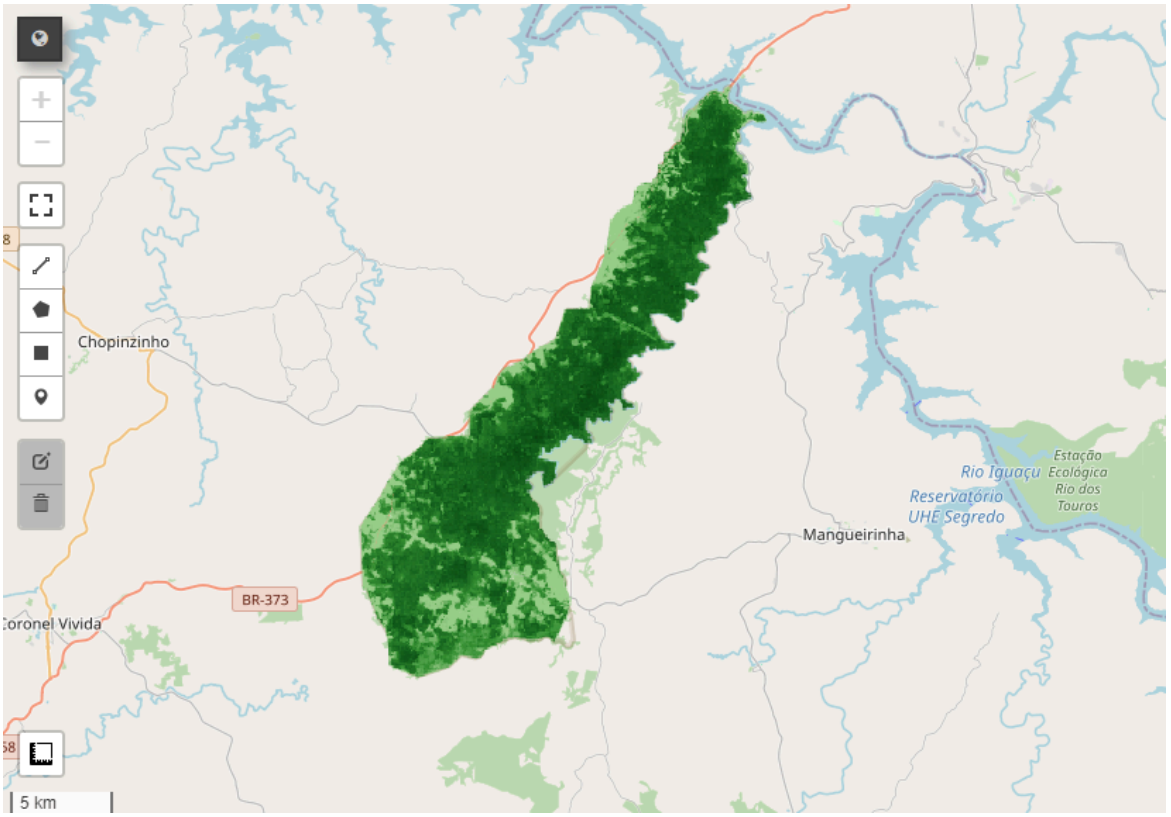
3.7. Mapeamento de Biomassa Acima do Solo (AGB)

Para o mapeamento de AGB, foi utilizado um modelo Random Forest em função dos resultados apresentados anteriormente. As amostras foram divididas em treino em teste numa proporção de 70% para treino e 30% para teste. Para o treinamento do modelo foram utilizadas as 10 variáveis melhor ranqueadas pelo método MDA conforme a etapa de seleção das variáveis, e também os hiperparâmetros encontrados na função de otimização Bayesiana.

O coeficiente de determinação (R^2) foi 65,38%, enquanto o erro quadrático médio (RMSE) foi de 37,99 t/ha e o erro absoluto médio (MAE) foi de 30,21 t/ha. Além disso, o cálculo média da biomassa prevista pelo modelo Random Forest na área de estudo foi de 142,34 t/ha. Ao comparar a biomassa prevista com os dados obtidos através do IFN, identificou-se uma diferença de -8,66 t/ha. Isso significa que as previsões do modelo Random Forest são, em média, inferiores aos valores medidos diretamente no campo. Já em relação ao conjunto de dados ESA CCI Biomass, a diferença foi ainda menor, com um resultado de -2,96 t/ha.

Para ilustrar os resultados do modelo Random Forest, a Figura 9 mostra o mapa de densidade de AGB na área de estudo:

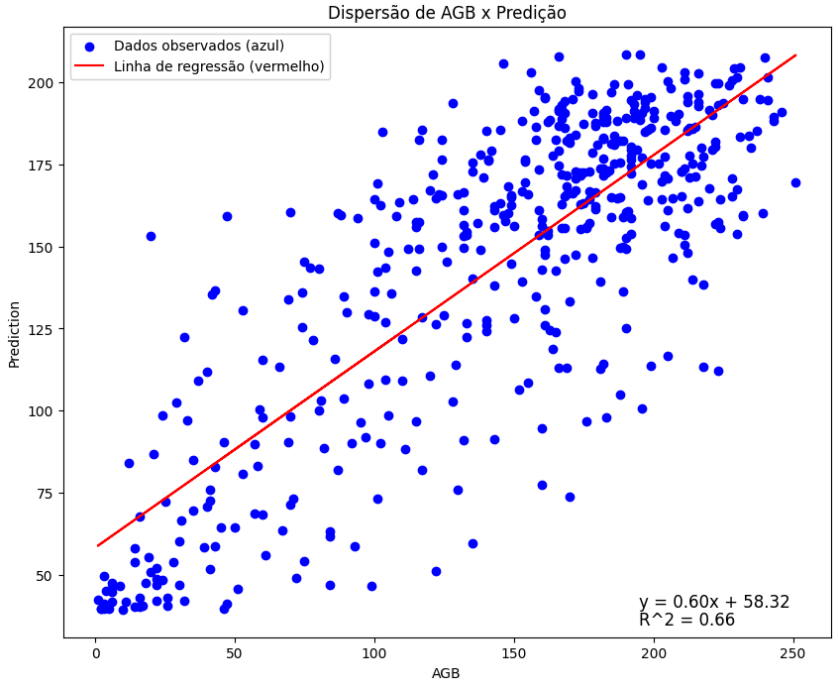
Figura 9: Biomassa Acima do Solo (AGB) na Terra Indígena Mangueirinha



Posteriormente foi realizada uma análise de regressão linear entre os valores observados de biomassa aérea (AGB) e os valores previstos pelo modelo de floresta aleatória. A variável dependente (AGB) e a predição do modelo foram extraídas e ajustadas a uma regressão linear simples, utilizando o método dos mínimos quadrados ordinários (OLS).

A Figura 10 apresenta um gráfico de dispersão que mostra a relação entre os valores das amostras de AGB e as predições do modelo, onde os pontos azuis representam os dados observados e a linha vermelha representa a reta de regressão linear ajustada, permitindo a visualização dos resíduos.

Figura 10: Valores reais x Predição (Random Forest no Google Earth Engine)



4. Considerações Finais

Este trabalho destaca a importância da implementação de técnicas de ciência de dados na modelagem de biomassa a partir de aprendizado de máquina e sensoriamento remoto para a obtenção de estimativas precisas e confiáveis. Tais técnicas abrem novas possibilidades para o monitoramento e gestão da biomassa, especialmente em contextos de conservação ambiental e planejamento sustentável.

É notável que o modelo Random Forest implementado no Google Earth Engine tenha obtido resultados satisfatórios mesmo com a subestimação dos valores de AGB tanto em relação ao conjunto de dados ESA CCI Biomass quanto do IFN.

Contudo, é preciso ressaltar que o modelo não está capturando completamente a variabilidade da biomassa. Essa diferença pode ter implicações para a aplicação dos resultados do modelo em práticas de gestão e planejamento ambiental. Portanto, é essencial entender as causas dessa subestimação, que podem incluir limitações nos dados utilizados para treinar o modelo, a escolha das variáveis preditoras ou as características específicas do ecossistema em questão. Avaliações adicionais e ajustes no modelo podem ser necessários para melhorar a precisão das previsões de biomassa, o que é fundamental para garantir uma gestão sustentável dos recursos naturais.

Referências

Belloli, T. F., Guasselli, L. A., Kuplich, T. M., Ruiz, L. F. C., Arruda, D. C. de, Etchelar, C. B., & Simioni, J. D. (2022). Estimation of aboveground biomass and carbon in palustrine wetland using bands and multispectral indices derived from optical satellite imageries PlanetScope and Sentinel-2A. Journal of Applied Remote Sensing, 16(3), 034516. DOI: 10.1117/1.JRS.16.034516: doi:10.1117/1.JRS.16.034516

Bruce, P. & Bruce, A. O'Reilly. Estatística Prática para Cientistas de Dados - 50 conceitos essenciais. 2019. Starlin Alta Editora e Consultoria Eireli. ISBN: 978-85-508-0603-7

DEFRIES, R. S., HANSEN, M. C., TOWNSHEND, J. R. G., & JANETOS, A. C. Classificação global da cobertura da terra em resolução espacial de 8 km: o produto de cobertura da terra MODIS. Remote Sensing of Environment, v. 89, n. 1, p. 112-124, 2004.

ESA, 2023a. Sentinel-2 Mission. [S.l.]: European Space Agency, 2023. Disponível em: https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2. Acesso em: 09/04/2024.

ESA, 2023b. Sentinel-1 Mission. European Space Agency. https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-1. Acesso em: 09/04/2024.

Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D.E., 2007, The shuttle radar topography mission: Reviews of Geophysics, v. 45, no. 2, RG2004, at https://doi.org/10.1029/2005RG000183.

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd ed.). O'Reilly Media, Inc.

Ghosh, S.M. Behera, M.D. Aboveground biomass estimates of tropical mangrove forest using Sentinel-1 SAR coherence data-the superiority of deep learning over a semi-empirical model. Comput. Geosci., 150 (2021), Article 104737

GUYON, Isabelle; ELISSEEFF, André. An introduction to variable and feature selection. *Journal of Machine Learning Research*, v. 3, n. 6, p. 1157-1182, 2003.

Hunka, N., Santoro, M., Armston, J., Dubayah, R., McRoberts, R. E., Næsset, E., Quegan, S., Urbazaev, M., Pascual, A., May, P. B., Minor, D., Leitold, V., Basak, P., Liang, M., Melo, J.,

Herold, M., Málaga, N., Wilson, S., Durán Montesinos, P., ... Duncanson, L. (2023). On the NASA GEDI and ESA CCI biomass maps: aligning for uptake in the UNFCCC global stocktake. Environmental Research Letters, 18(12), 124042. DOI: 10.1088/1748-9326/ad0b60: doi:10.1088/1748-9326/ad0b60

IPCC, 2023: Summary for Policymakers. In: Climate Change 2023: Synthesis Report.Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland, pp. 1-34, doi: 10.59327/IPCC/AR6-9789291691647.001.

JENSEN, John R. Sensoriamento remoto do ambiente: Uma perspectiva em recursos terrestres. 3. ed. São Paulo: Pearson Education do Brasil, 2015.

Oliveira, G. A. de, Silva, L. F. da, Nascimento, J. S., Agostinho, P. R., & Padovan, M. P. (2018). Valoração Econômica de Serviços Ambientais em Sistemas Agroflorestais Biodiversos: um Estudo de Caso no Assentamento Lagoa Grande, em Dourados/MS. Anais do AGROECOL 2018; 11 a 14 de novembro de 2018, Campo Grande/MS, 13(2), AGROECOL - Sistemas agroflorestais em bases agroecológicas.

LUNDBERG, S. M., LEE, S.-I., EILERTSEN, G., & SHAPLEY, R. SHAP: Explaining Black Box Machine Learning Models. arXiv preprint arXiv:2003.13377, 2020.

Molisse, G. Emin, D. and Costa, H. Implementation of a Sentinel-2 Based Exploratory Workflow for the Estimation of Above Ground Biomass. 2022. IEEE Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), Istanbul, Turkey, 2022, pp. 74-77, doi: 10.1109/M2GARSS52314.2022.9839897.

Ratuchne, L. C. Equações alométricas para estimative de biomassa, carbono e nutrientes em uma Floresta Ombrófila Mista. Universidade Estadual do Centrooeste, Paraná, 2010.

Reichstein, M., Carvalhais, N. Aspects of Forest Biomass in the Earth System: Its Role and Major Unknowns. Surv Geophys 40, 693–707 (2019). https://doi.org/10.1007/s10712-019-09551-x

Santoro, M.; Cartus, O. (2023): ESA Biomass Climate Change Initiative (Biomass_cci): Global datasets of forest above-ground biomass for the years 2010, 2017, 2018, 2019 and 2020, v4. NERC EDS Centre for Environmental Data Analysis, 21 April 2023. doi:10.5285/af60720c1e404a9e9d2c145d2b2ead4e.
https://dx.doi.org/10.5285/af60720c1e404a9e9d2c145d2b2ead4e

Serviço Florestal Brasileiro. Inventário Florestal Nacional: principais resultados: Terra Indígena Mangueirinha. Brasília, DF: MAPA, 2019. 76p. (Série Relatórios Técnicos - IFN). Disponível em: < https://snif.florestal.gov.br/images/pdf/publicacoes/periodo_eleitoral/publicacoes_ifn/relatorios/IFN_TI_mangueirinha_2019_periodo_eleitoral.pdf >

Apêndice

O código utilizado para a elaboração deste trabalho está disponível no GitHub do autor:

< https://github.com/sdesena/AGB-Modeling-TCC >