Projeto de Pesquisa e Planejamento de Atividades

Aluno: Sandro de Sena Machado		**Data início curso**: 01/05/2023
Orientador: Aubert Henrik Venson		**Defesa em:** 10/2024
Curso: Data Science & Analytics	**Modalidade**: MBA	Turma: DSA231.

1. Título do projeto

Estimativas de biomassa acima do solo a partir de aprendizado de máquina e sensoriamento remoto

2. Introdução

De acordo com o Sexto Relatório de Avaliação (AR6) do Painel Intergovernamental sobre Mudanças Climáticas (IPCC), a temperatura da Terra já aumentou 1,2 °C em comparação com a média de 1850-1900, em função do aumento da concentração de gases de efeito estufa, principalmente do dióxido de carbono (CO_2), em decorrência das atividades humanas. Segundo o relatório, foram detectadas tendências de aumento da temperatura global e do nível do mar, bem como aumento da frequência e severidade de eventos climáticos extremos, entre outros efeitos (IPCC, 2023).

Neste cenário, o desmatamento e a degradação florestal são considerados responsáveis por uma parte considerável das emissões globais de gases de efeito estufa. Isto se dá pois as florestas armazenam uma quantidade significativa de carbono na forma de biomassa, e a medida que são degradadas ou desmatadas, o carbono armazenado é emitido para a atmosfera na forma de CO_2, contribuindo para o aumento das concentrações de gases de efeito estufa (Reichstein, *et al.* 2019).

Além de seu papel no armazenamento de carbono em biomassa, tanto acima do solo quanto abaixo, as florestas oferecem diversos outros serviços ecossistêmicos de provisão, regulação, cultura e suporte. Entre os quais podem-se citar proteção e conservação do solo e dos recursos hídricos, regulação do clima, conservação da biodiversidade, entre outros (Serviço Florestal Brasileiro, 2019).

Nesse sentido, a análise da biomassa pode fundamentar as tomadas de decisão acerca da valoração econômica das florestas e a promoção de práticas de manejo sustentável como programas de restauração de ecossistemas, projetos de conservação da biodiversidade, programas de monitoramento de áreas protegidas, projetos de pagamentos por serviços ambientais, entre outros (Oliveira *et al.* 2018).

Tradicionalmente, para medir a biomassa é preciso ir a campo realizar amostragem sistemática utilizando parcelas permanentes. Os métodos envolvem a instalação de parcelas de medição em locais pré-determinados na floresta e a coleta de dados repetidos ao longo

do tempo para monitorar mudanças na biomassa. Os locais de amostragem são selecionados de forma a representar uma variedade de condições florestais, como diferentes tipos de vegetação, idades das árvores, gradientes de altitude, etc (Serviço Florestal Brasileiro, 2019).

Assim, são coletadas medições da biomassa das árvores, entre outras plantas presentes. Isso pode incluir a medição do diâmetro das árvores, altura das árvores, densidade da madeira, área basal, etc. A partir disso, são utilizadas equações alométricas. A relação alométrica mais comum utilizada na estimativa da biomassa acima do solo é a Equação Alométrica de Volume. Esta equação relaciona o volume da árvore (ou de seu tronco) com seu diâmetro e altura (Ratuchne, L. 2010).

Todavia, os métodos alométricos possuem limitações para aplicações em larga escala, tendo em vista a dificuldade operacional e os custos para coletar dados em campo e manter o monitoramento contínuo. Portanto, monitorar e analisar a distribuição espacial e temporal da biomassa acima do solo de forma acurada e escalável ainda é desafiador (Belloli *et al.* 2022).

Frente a isso, a implementação de técnicas de aprendizado de máquina em sensoriamento remoto e processamento digital de imagens, têm demonstrado resultados satisfatórios em termos de acurácia e escalabilidade, ainda que seja necessário coletar amostras em campo para calibrar e validar os modelos (Ghosh & Behera, 2021).

O sensoriamento remoto envolve o uso de dispositivos como satélites, drones e aeronaves para coletar dados sobre a superfície do planeta à distância. A partir destes dados, técnicas de processamento de imagem e análise espacial podem ser aplicadas para identificar e extrair características relacionadas à biomassa acima do solo (*Above Ground Biomass* - AGB). Essas características são expressas através de índices de vegetação, parâmetros biofísicos e medidas de textura como por exemplo o NDVI, LAI, GLCM, respectivamente (Jensen, J. 2015).

É recomendado combinar dados de diferentes sistemas sensores para estimar biomassa em escala, pois cada tipo de sensor oferece informações complementares sobre a estrutura e composição das florestas. Enquanto os sensores ópticos, como imagens de satélite, fornecem informações sobre a densidade, vigor da vegetação e a cobertura do dossel, os sensores radar e LiDAR (*Light Detection and Ranging*) penetram na vegetação e podem medir a altura das árvores e a estrutura tridimensional da floresta. Ao combinar essas fontes de dados, é possível obter estimativas mais precisas e abrangentes da biomassa florestal, superando limitações individuais de cada tipo de sensor e aumentando a confiabilidade das estimativas (Molisse *et al.* 2022).

Para tanto, o *Google Earth Engine* se destaca como um ambiente de desenvolvimento em nuvem para processamento digital de imagens de sensoriamento

remoto, pois permite o acesso a uma ampla gama de conjuntos de dados e ferramentas de geoprocessamento. Além disso, a plataforma está integrada ao *Google Colab*, que por sua vez, é um ambiente em nuvem que permite executar código *Python* e utilizar bibliotecas como *Scikit-Learn* e *Tensor Flow*, as quais oferecem diversas ferramentas e algoritmos que podem ser adaptados e aplicados de acordo com as necessidades específicas de cada projeto. Assim, é possível implementar e avaliar modelos de aprendizado de máquina, ajustando, calibrando e validando os modelos de forma a garantir acurácia e robustez.

3. Objetivos

3.1. Objetivo Geral

Implementar e avaliar a adequação de modelos de aprendizado de máquina para estimar biomassa acima do solo (AGB).

3.2. Objetivos específicos

I. Mapear a distribuição espacial e temporal de biomassa acima do solo.
II. Analisar a importância das variáveis preditoras.
III. Avaliar o desempenho de diferentes modelos em termos de acurácia e precisão.
IV. Discutir a validação dos modelos com dados de referência *in-situ*.

4. Material e Métodos

4.1. Área de estudo

A escolha da Terra Indígena Mangueirinha como área de estudo se deu em função da qualidade, regularidade e relevância dos dados disponíveis de AGB coletados em campo, bem como à representatividade da amostragem e às parcerias estabelecidas para a coleta e análise desses dados. Assim, obtém-se um conjunto de referência (*Ground Truth*) para o ano de 2015 para posterior calibração e validação dos modelos.

A Terra Indígena Mangueirinha, situada na região oeste do Paraná, na bacia do Rio Iguaçu, é apresentada no relatório técnico do Inventário Florestal Nacional (IFN), a partir de dados acerca da saúde e vitalidade das florestas, biodiversidade, quantitativos de biomassa e carbono em estoque, bem como dados socioambientais. Assim, o IFN serve como um conjunto de referências para formulação de políticas públicas que envolvam a conservação e recuperação de florestas (Serviço Florestal Brasileiro, 2019).

O IFN realizado pelo Serviço Florestal Brasileiro se destaca por sua abordagem direta de coleta de dados em florestas naturais e plantadas. Esse processo envolve a obtenção de amostras botânicas e de solo, medição das árvores e entrevistas com os moradores locais. Essa metodologia permite uma avaliação abrangente da qualidade e das condições das florestas, bem como sua relevância para as comunidades locais. Ainda, a

metodologia do IFN considera a regularidade da coleta dos dados, bem como a amostragem representativa em diferentes biomas, garantindo confiabilidade, atualizações periódicas e variedade dos dados (Serviço Florestal Brasileiro, 2019).

4.2. Coleta de dados

Serão obtidas imagens de satélites baseadas na correspondência temporal com a coleção de dados ESA CCI Biomass e da localização da área de estudo. Para obter as variáveis preditoras, optou-se pela integração de dados de sensores ópticos e radar, considerando a complementaridade desses sensores para análises de estrutura, vigor e composição da vegetação (Molisse *et al.* 2022).

Nota-se a necessidade de garantir a consistência temporal entre essas imagens e as datas dos produtos do ESA CCI Biomass. Isso visa garantir que as condições ambientais durante a coleta das imagens de satélite sejam o mais semelhantes possível às condições observadas pelo produto ESA CCI Biomass (Hunka *et al.* 2023).

Isso é importante para garantir que as diferenças observadas entre as imagens de satélite e os pontos de controle sejam devidas às características da paisagem e não a mudanças temporais não controladas. Portanto, é fundamental que haja correspondência e temporal entre as imagens de satélite, a coleção ESA CCI Biomass e o IFN da área de estudo, para poder comparar e interpretar os resultados.

4.2.1. 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 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).

4.2.2. 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; enquantos as bandas 'B1', 'B9' e 'QA60' possuem 60 metros (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. Esses procedimentos serão tratados em mais detalhes na seção de processamento dos dados (Belloli *et al.* 2022).

4.2.3. 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).

4.2.4. 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).

4.3. Processamento dos dados

4.3.1. Pré-processamento

Para melhor aproveitamento dos dados de satélite, é necessário realizar pré-processamentos adicionais. Assim, serão aplicadas nas imagens, funções para mascaramento de nuvens para os sensores ópticos e filtragem de ruído *speckle* para as imagens derivadas de SAR. Também é necessário realizar a reamostragem das diferentes bandas das imagens para ajustá-las para a mesma resolução espacial. Isso é imprescindível para que todas as imagens tenham o mesmo tamanho de pixel e portanto sejam comparáveis (Jensen, J. 2015).

Por fim, é fundamental reprojetar todas as imagens para o mesmo sistema de referência de coordenadas. Assim, todas as imagens estarão na mesma projeção cartográfica e portanto poderão ser sobrepostas e comparadas geometricamente, para por exemplo, calcular área e afins (Jensen, J. 2015).

4.3.2. Extração de características

A partir das coleções de dados de satélites, serão extraídas as características que servirão como variáveis preditoras para os modelos de aprendizado de máquina. Essas

características incluem índices de vegetação, parâmetros biofísicos, medidas texturais e topográficas.

Utilizando a coleção de dados do Sentinel-2, serão calculados índices de vegetação como NDVI (Normalized Difference Vegetation Index), EVI (Enhanced Vegetation Index), SAVI (Soil Adjusted Vegetation Index), GNDVI (Green Normalized Difference Vegetation Index), que fornecem informações sobre a resposta espectral da vegetação, seu vigor e densidade.

Ainda, para a geração de parâmetros biofísicos, serão calculados o LAI (Leaf Area Index) e o FAPAR (Fraction of Absorbed Photosynthetically Active Radiation) utilizando o SNAP (Sentinel Application Platform), que oferece a ferramenta *Biophysical Processor* para o cálculos dessas características.

A partir das polarizações disponíveis na coleção de dados do Sentinel-1, pode-se gerar índices de vegetação SAR e também 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 destacam por utilizar as polarizações VV e VH para quantificar biomassa. Contudo, tendo em vista a saturação do sinal em áreas densamente florestadas, apenas o DPSVIm foi selecionado para este estudo em função de sua capacidade de atenuar a saturação do sinal e aprimorar o mapeamento da vegetação (dos Santos *et al.* 2021).

Outra medida extraível dessa coleção é a GLCM (Gray Level Co-occurrence Matrix). Essa medida de textura permite analisar a estrutura e densidade espacial da vegetação. A GLCM é uma matriz quadrada que representa a frequência com que pares de pixels com valores de cinza específicos co-ocorrem em uma imagem. Diversos recursos texturais podem ser extraídos da GLCM, como contraste, homogeneidade e entropia (Molisse *et al.* 2022).

A partir do SRTM serão extraídas as medidas topográficas, como por exemplo variáveis como declive, aspecto e altitude. Tais medidas são fundamentais para a compreensão da distribuição espacial da biomassa pelo terreno (Farr *et al.* 2007)..

4.3.3. Validação cruzada, normalização e empilhamento dos dados

Primeiramente, é necessário empilhar os conjuntos de dados. O empilhamento visa otimizar a performance dos modelos, bem como 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.

A validação cruzada e a normalização dos dados são essenciais para a implementação de modelos robustos à vieses. A validação cruzada acontece através da

divisão dos dados em conjuntos de treino e validação, a técnica permite uma avaliação mais precisa da capacidade de generalização para dados novos, evitando o sobreajuste (*overfitting*) dos modelos. Esta técnica pode ser executada através de diferentes procedimentos como por exemplo, *train-test split* e *k-fold* (O'Reilly, 2019).

A normalização garante que todas as características estejam na mesma escala, evitando que características com valores maiores tenham um impacto desproporcional nos modelos, levando a resultados enviesados. Assim, a normalização equilibra a importância das características e ainda as torna comparáveis, tendo em vista que as características em seu estado anterior possuem diferentes unidades de medidas. Além disso, os modelos têm sua performance aprimorada ao utilizar a normalização (O'Reilly, 2019).

Contudo, somente o conjunto de treinamento deve ser utilizado para ajustar o escalonador (*scaler*). Isso permite que os dados de validação permaneçam "ocultos" durante todo o processo, evitando assim o vazamento dos dados (*data leakage*). Por fim, todo o conjunto de dados é normalizado utilizando o escalonador treinado. Este procedimento deve ser utilizado antes de cada etapa, incluindo seleção de características, ajuste dos modelos e ajuste fino dos hiperparâmetros (O'Reilly, 2019).

4.4. Ajuste dos modelos

4.4.1. Random Forest

Os modelos para a predição da biomassa serão ajustados utilizando algoritmos de aprendizado de máquina estatístico não-paramétricos. O primeiro modelo será 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* de diversas árvores de classificação ou regressão, amostrando não apenas os registros, mas também as variáveis (Bruce & Bruce, 2019).

As árvores de decisão são construídas a partir da repartição recursiva dos dados a partir da minimização de critérios de impureza como a impureza de Gini e a entropia. Cada repartição 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 (*i.e.* diminuam a impureza)(Bruce & Bruce, 2019).

A potência desse algoritmo se mostra em conjuntos de dados com muitas características e registros, haja vista que este tem a habilidade de determinar quais variáveis

preditoras são mais importantes bem como as relações complexas entre si (Bruce & Bruce, 2019).

Existem duas maneiras para medir a importância das variáveis nesse contexto: pela diminuição em precisão do modelo se os valores de uma variável forem aleatoriamente permutados (*Mean Decrease in Accuracy i.e. MDA*), e pela diminuição da média na pontuação de impureza de Gini (*Mean Decrease in Impurity i.e. MDI)*. Para este estudo, a MDA foi escolhida já que esta é uma medida mais confiável, tendo em vista que a precisão é calculada dos dados fora da amostra (*out-of-bag)*, ou seja, são efetivamente uma estimativa de validação cruzada, enquanto a MDI se baseia no conjunto de treinamento (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 desse algoritmo através do método *sklearn.ensemble.RandomForestRegressor* da biblioteca Scikit-Learn é relativamente simples e intuitiva.

4.4.2. Redes Neurais Artificiais

As redes neurais artificiais (ANN's), implementadas com *frameworks* como Keras e Tensor Flow, surgem como ferramentas poderosas de aprendizado de máquina, oferecendo alta acurácia e flexibilidade na modelagem da biomassa, entre outras tarefas (Géron, A 2019).

A rede neural pode ser composta por diversas camadas de neurônios (*hidden layers*), como camadas densamente conectadas, convolucionais ou recorrentes, dependendo da natureza dos dados e da complexidade do modelo. Cada camada possui um número de neurônios que define a capacidade da rede de processar informações. Funções como ReLU e Sigmoide são utilizadas para adicionar não-linearidade ao modelo e permitir que ele aprenda relações complexas entre as variáveis (Géron, A 2019).

Para o treinamento de redes neurais, é necessário definir a função custo, critério que o modelo busca minimizar durante o treinamento. Métricas como erro quadrático médio (MSE) e erro absoluto médio (MAE) são comumente utilizados como funções de perda em redes neurais (Géron, A 2019).

O treinamento das redes neurais também necessita de otimizadores, como Adam ou SGD (*Stochastic Gradient Descent*), para atualizar os pesos da rede neural e minimizar o erro de predição. Ainda, considera-se o número de épocas (*epochs*) define quantas vezes o conjunto de treino é utilizado para treinar a rede neural. Ainda, considera-se que a curva de aprendizado do modelo ajuda a visualizar como o desempenho do modelo melhora com o aumento do número de épocas (Géron, A 2019).

Recomenda-se utilizar a técnica *Early Stopping* que interrompe o treinamento do modelo quando o desempenho na validação começa a diminuir, evitando o sobreajuste, bem como o *Batch Size* que define o número de exemplos que são utilizados para atualizar os pesos da rede durante o treinamento. Ainda, pode-se aplicar técnicas como regularização L1 ou L2 que penaliza pesos grandes na rede, e também o *Dropout* que desativa aleatoriamente neurônios durante o treinamento (Géron, A 2019).

Também é necessário se atentar para a otimização da taxa de aprendizado (*learning rate*). Valores muito altos de taxa de aprendizado podem impedir que o modelo atinja a solução ideal, enquanto valores muito baixos podem tornar o treinamento extremamente lento (Géron, A 2019).

4.5. 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).

Uma abordagem comum para essa otimização é a busca em grade (*Grid Search*), que é robusta e relativamente fácil de implementar. Nesse método, define-se um espaço de pesquisa, onde são especificados intervalos de valores para cada hiperparâmetro a ser otimizado. A busca em grade então gera todas as combinações possíveis desses valores e treina o modelo para cada uma delas, avaliando seu desempenho em um conjunto de validação (Géron, A 2019).

A melhor combinação de valores, aquela que resulta no melhor desempenho na validação, é selecionada como a configuração ideal dos hiperparâmetros. Apesar de sua simplicidade e robustez, a busca em grade pode ser computacionalmente cara, especialmente para problemas com muitos hiperparâmetros, e não garante a solução ótima, já que a busca é limitada aos valores predefinidos no espaço de pesquisa (Géron, A 2019).

Para a otimização dos hiperparâmetros em redes neurais, algoritmos de otimização mais sofisticados são mais adequados. Alguns exemplos populares incluem o Adam e o SGD (*Stochastic Gradient Descent*), um algoritmo clássico que utiliza a descida do gradiente, e a Otimização Bayesiana, um método probabilístico que utiliza informações de avaliações anteriores para direcionar a busca por melhores soluções. Esses algoritmos exploram o espaço de pesquisa de forma mais eficiente e podem encontrar soluções mais próximas do ideal. Embora estes algoritmos de otimização sejam mais eficientes e possam encontrar soluções mais precisas do que a busca em grade, eles também podem ser mais complexos de implementar e entender, além de apresentar problemas de convergência (Géron, A 2019).

4.6. Avaliação de acurácia

Na avaliação do desempenho de modelos de estimativa de biomassa, várias 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. Valores próximos a 1 indicam que o modelo explica a maior parte da variabilidade, enquanto valores próximos a 0 indicam uma explicação insuficiente (Géron, A 2019).

Outra métrica importante é o erro quadrático médio (RMSE), que mede a diferença média entre os valores reais e preditos da biomassa. Um RMSE baixo sugere uma alta acurácia na predição da biomassa pelo modelo (Géron, A 2019).

O erro médio absoluto (MAE) é uma métrica semelhante ao RMSE, medindo a diferença média absoluta entre os valores reais e preditos da biomassa. Um MAE baixo também indica alta acurácia na predição da biomassa. (Géron, A 2019).

Além dessas métricas, outras são relevantes, como o erro percentual médio absoluto (MAPE), que avalia a precisão do modelo em relação à magnitude da biomassa, e o R^2 ajustado, que considera o número de variáveis no modelo, penalizando aqueles com muitas variáveis (Géron, A 2019).

4.7. Explanação dos modelos com pacote SHAP

Um modelo ideal de aprendizado de máquina é aquele que não só oferece alta precisão, mas também é facilmente interpretável, permitindo uma compreensão intuitiva de seu desempenho, similar ao que é proporcionado pelas árvores de decisão. No entanto, à medida que a complexidade do modelo 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. É nesse contexto que um explicador de modelo pode ser útil para fornecer insights sobre os resultados.

Nesse sentido, o pacote SHapley Additive exPlanations (SHAP) se destaca como um explicador de modelos que são considerados “caixas-pretas”. Desenvolvido por Lundberg et al. (2020), o SHAP tem como propósito aumentar a confiança do usuário no modelo, fornecer insights para sua melhoria e auxiliar na compreensão do problema modelado.

O SHAP vai além de simplesmente mostrar a importância das características, explicando como cada uma contribui individualmente para as previsões finais. Isso possibilita uma interpretação mais detalhada do modelo e a identificação de possíveis vieses.

Entre as vantagens de utilizar o SHAP, destacam-se:

- Interpretabilidade local: O SHAP explica as previsões individuais, permitindo entender por que um modelo específico fez uma determinada previsão.
- Transparência: Fornece insights sobre o comportamento interno do modelo, auxiliando na identificação de vieses e problemas potenciais.
- Flexibilidade: Pode ser usado com uma ampla variedade de modelos de aprendizado de máquina.

5. Resultados Esperados

O Trabalho de Conclusão de Curso (TCC) proposto visa abordar uma série de objetivos fundamentais para a compreensão e análise da distribuição da biomassa acima do solo. Inicialmente, espera-se que o estudo mapeie tanto a distribuição espacial quanto temporal da biomassa na área de estudo. Este mapeamento será essencial para posterior comparação com o IFN da área de estudo.

Uma parte crucial do TCC será a análise da importância das variáveis preditoras, identificando quais dessas devem ser selecionadas para otimizar a performance de modelos de aprendizado de máquina.

Para atingir esses objetivos, o estudo empregará modelos de aprendizado de máquina, incluindo Random Forest e Redes Neurais Profundas. Esses modelos serão ajustados e otimizados através de técnicas como pré-processamento de dados de satélite, extração de características das imagens de satélite e normalização dos dados e validação cruzada.

Posteriormente será discutido a comparação dos resultados dos modelos com dados de referência coletados in-situ, permitindo uma avaliação precisa do desempenho e da acurácia de cada modelo.

Além disso, uma etapa crucial do estudo será a explanação dos modelos utilizando o pacote SHAP. Essa técnica fornecerá *insights* sobre como cada variável contribui individualmente para as previsões dos modelos, permitindo uma interpretação mais profunda e uma compreensão mais completa dos resultados.

Ao final do TCC, espera-se não apenas uma descrição detalhada da distribuição da biomassa acima do solo, mas também uma análise aprofundada dos fatores que a influenciam, bem como uma avaliação rigorosa do desempenho dos modelos utilizados. Esses resultados poderão contribuir para o avanço do conhecimento sobre esse importante aspecto dos ecossistemas terrestres.

6. Cronograma de Atividades

Atividades planejadas	Mês									
	1	2	3	4	5	6	7	8	9	10
Revisão Bibliográfica	x									
Introdução e Objetivos		x								
Metodologia e Resultados Esperados			x							
Coleta e Processamento dos Dados				x						
Validação Cruzada, normalização e empilhamento				x						
Ajuste dos modelos					x					
Otimização dos hiperparâmetros						x				
Explanação do modelo com SHAP						x				
Avaliação da precisão							x			
Resultados							x			
Discussão								x		
Conclusões									x	
Correções										x
Sintetização em artigo										x
Elaboração da apresentação										x

Projeto de Pesquisa; Resultados Preliminares; Entrega do Trabalho de Conclusão de Curso; Entrega da Apresentação da Defesa

7. 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

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 >