# Modeling of growth and production of *Eucalyptus spp* stands using supervised machine learning

**Rafael M. C. Araujo**[1], **Alvaro A. V. Soares**[1]
**Antonilmar A. L. Silva** [2], **Bruno F. Martins** [2], **Murillo G. Carneiro**[1,3]

[1] PPG em Agricultura e Informações Geoespaciais,
Instituto de Ciências Agrárias – Universidade Federal de Uberlândia (UFU).

[2]Celulose Nipo-Brasileira (CENIBRA)

[3]PPG em Ciências da Computação,
Faculdade de Computação – Universidade Federal de Uberlândia (UFU).

{rafaelmartinsdca, alvaro.soares, mgcarneiro}@ufu.br

{antonilmar.silva, bruno.martins}@cenibra.com.br

***Abstract.*** *The forest sector is one of the most important pillars of the Brazilian economy due to its high wood productivity. Several studies are being carried out seeking to develop a computational method capable of estimating production efficiently in order to reduce production costs. The objective of the study was to develop machine learning models capable of estimating present and future eucalyptus production with high precision, evaluating relevant supervised learning models, like neural networks and support vector machine (SVM), in relation to the Clutter model, widely adopted by the forestry industries. A case study conducted on real data obtained from a continuous forest inventory showed that SVM is efficient to estimate growth and production of eucalyptus.*

***Resumo.*** *O Setor florestal é um dos mais importantes pilares da economia brasileira devido a sua alta produtividade de madeira. Diversos estudos estão sendo feitos buscando desenvolver um método computacional capaz de estimar a produção de maneira eficaz a fim de diminuir os custos de produção. O objetivo do estudo foi desenvolver modelos de aprendizado de máquina capazes de estimar a produção de eucalipto presente e futura com alta precisão, avaliando modelos relevantes de aprendizado de máquina, como redes neurais e máquina de vetores de suporte (SVM), em relação ao modelo de Clutter, amplamente adotado pela indústria florestal. Um estudo de caso conduzido sobre dados reais obtidos a partir de dados de um inventário florestal contínuo mostrou que SVM é eficiente para estimar crescimento e produção de eucaliptos.*

## 1. Introdução

O setor florestal é um dos principais pilares do agronegócio brasileiro, permitindo que o brasil seja o maior produtor mundial de carvão vegetal em 2020 e o principal exportador de celulose no ano de 2021 [Ibá 2022]. As florestas plantadas no Brasil ocupam atualmente 9,93 milhões de ha (hectares), sendo 75,8% desta composta por eucalipto; o restante é divido entre plantios de pinheiros (19,4%) e outras espécies (4,8%) como seringueira, acácia, teca e paricá. Os principais estados produtores de eucalipto em área

plantada são Minas Gerais (30%), Outros (21%), Mato Grosso do Sul (14%), São Paulo (13%), Bahia (8%), Rio Grande do Sul (8%) e Paraná (6%), que juntos atingiram uma produtividade média de 38.9 $m^3ha^{-1}$ ano$^{-1}$ em 2022 [Ibá 2022].

Segundo o levantamento bibliométrico [Campos et al. 2022], o Brasil lidera o número de estudos voltados à produção e tecnologia de papel de eucalipto. Desses estudos, grande parte deles buscam melhorar as estimativas da produção afim de atender as demandas desses subprodutos, com modelos mais robustos associados a variáveis tanto do inventário, quanto fora dele. Como exemplos para a estimativa do volume de eucalipto temos [Penido et al. 2020] com avaliação de três modelos tradicionais: Modelo de Povoamento (MP) Clutter, Modelo de Distribuição Diamétrica (MDD) e Modelo de Árvore Individual (MAI); [Dai et al. 2021] com modelos que incluem índices de vegetação (IV) derivados de Sensoriamento Remoto; [Cordeiro et al. 2022] com uso de Redes Neurais Artificiais (ANN) e SVM; dentre outros dispostos na literatura, todos com o mesmo objetivo, estimar a produção de eucalipto com precisão.

Diante do potencial dessas aplicações, o presente estudo tem por objetivo desenvolver modelos de aprendizado de máquina capazes de estimar a produção de eucalipto com alta precisão, para dados de um inventário florestal contínuo de *Eucalyptus spp*. Especificamente, as contribuições alcançadas pelo presente trabalho são listadas a seguir:

- Construção de uma base de dados a partir de um inventário florestal contínuo de uma empresa multi-nacional, consistida por especialistas do setor florestal, que permita o desenvolvimento de modelos de aprendizado de máquina supervisionado para modelagem de crescimento e produção de *Eucalyptus spp*;
- Avaliação do desempenho preditivo dos modelos de aprendizado de máquina em comparação ao modelo de Clutter, amplamente adotado pela maioria das indústrias florestais, para estimar produção presente e futura de povoamentos de *Eucalyptus spp*.

Em relação ao último ponto listado, é importante destacar que nossa comparação considera apenas os poucos atributos adotados pelo modelo de Clutter para estimar a produção, diferente de outros trabalhos da literatura em que algoritmos de aprendizado de máquina consideraram um número consideravelmente maior de informações. Como a coleta dessas informações em campo representa tempo e custo adicional para as empresas, é esperado que nossa análise possa contribuir para uma melhor avaliação desses especialistas em relação à adoção dos modelos de aprendizado desenvolvidos.

O restante do artigo está organizado da seguinte forma. A Seção 2 apresenta os principais trabalhos correlatos da literatura; a Seção 3 descreve a base de dados construída, o modelo de Clutter, as técnicas de aprendizado de máquina investigadas e outros detalhes do método de pesquisa adotado; e a Seção 4 apresenta os resultados obtidos para a estimativa de produção presente e futuro. Por fim, a Seção 5 traz as conclusões e trabalhos futuros relacionados à pesquisa.

## 2. Trabalhos Relacionados

Nos últimos 20 anos diversas pesquisas foram desenvolvidas utilizando métodos de aprendizagem de máquina para aprimorar a modelagem biométrica das florestas. Exemplos de uso destes modelos incluem: a estimava de variável categórica em um modelo

[Binoti 2012]; estimativa de produção [Gorgens et al. 2009]; a probabilidade de sobrevivência de uma árvore [Reis et al. 2018]; o quanto inserção de variáveis fora do inventário melhoram as previsões dos modelos [De Freitas et al. 2020].

Essa gama de possibilidades faz com que o número de estudos seja cada vez mais frequente, e isso possibilita serem divididos em duas categorias: aqueles voltados apenas para a estimativa de uma variável categórica em um modelo de produção, como exemplo a altura, e aqueles que buscam apenas aperfeiçoar o processo de estimação de produção (volume), por meio de outras abordagens. Um exemplo da primeira categoria é o estudo de [Binoti 2012], que procurou a estimativa da altura das arvores por meio de ANN, para dados de inventários florestais contínuos de povoamentos clonais de eucalipto, como diâmetro com casca, altura dominante e idade, localizados no Sul da Bahia. Como resultado, o autor concluiu que a ANN permitiu diminuir reduzir a quantidade dados coletados para calibrar os modelos e consequentemente o custo de inventário sem perda de precisão dos modelos.

Um outro exemplo é o estudo de [Were et al. 2015], que comparou a performance dos modelos Floresta Aleatória (RF), Regressão por Vetores de Suporte (SVR) e ANN para mapear a variabilidade de estoques de carbono em solo orgânico (SOC - *Soil Organic Carbon*) na reserva florestal de Mau no leste do Quênia. O resultado do estudo mostrou que os modelos superestimaram os estoques SOC com desempenho semelhante entre o SVR e a ANN.

[Reis et al. 2018] propôs estimar a probabilidade de sobrevivência e mortalidade de árvores individuais em uma floresta com colheita seletiva, usando ANN para apoiar decisões siviculturais na gestão florestal da Floresta Nacional de Tapajarós, Cuiabá-Santarém. O estudo mostrou que a ANN estima a sobrevivência individual das árvores com mais precisão, mas isso não ocorre com a mortalidade, que é um evento mais raro do que a sobrevivência.

Já para a segunda categoria, mais comum às empresas florestais e ao objetivo deste estudo, temos alguns outros exemplos, como o de [Alcântara 2012], que buscou modificações em modelos tradicionais, como o de [Clutter 1963], com o objetivo de modelar o crescimento em nível de povoamento de clones de eucalipto na região de Minas Gerais. Neste estudo foi utilizado dados de inventários florestais contínuos, como índice de local, altura dominante e idade, para estimativa da área basal e do volume através de cinco equações derivadas do modelo original de Clutter. Os resultados indicaram eficiência das equações e conclui que o índice de local, na equação de projeção da área basal pode ser substituído pela altura dominante na mesma idade sem perda de eficiência e de exatidão.

[Gorgens et al. 2009] procuraram modelos mais robustos, como a ANN, para estimar com eficiência o volume em nível de povoamento de clones de eucalipto. Para o estudo eles utilizaram dados de inventário florestal de quatro empresas localizadas no Brasil, como diâmetro na altura do peito (DAP), altura total (HT), cinco locais e duas espécies, ambos representados por variáveis binárias, onde testaram 5 tipos de normalização e equalização de dados e 4 tipos de arquiteturas para a rede. Eles concluíram que as redes que apresentaram resultados estaticamente equivalentes foram construídas com duas camadas intermediárias, com função de ativação exponencial; as redes com melhores re-

sultados apresentaram dados normalizados e equalizados pelas variações do método de interpolação linear; e em síntese, os autores encontraram que a modelagem por redes neurais foi viável.

Finalizando essa categoria, temos o estudo de [De Freitas et al. 2020], que buscaram utilizar uma ANN, para poder estimar a produtividade de eucalipto no município de Minas Gerais. Eles utilizaram dados de 507 povoamentos de inventários florestais contínuos, como o tipo de clone, espaçamento, genótipo, temperatura, dentre outros, como práticas silviculturas. Para modelar as redes, eles utilizaram a ANN com neurônios variando nas camadas ocultas, de acordo com cada configuração, algoritmos de treinamento (*Backpropagation, Resilient Propagation (Rprop +) e Quick Propagation*) e funções de ativação (logística, tangente hiperbólica e linear), totalizando 87500 redes e, para avaliação do erro, a raiz quadrada do erro-médio (RMSE). Eles concluíram que as 10 melhores redes tiveram correlação acima de 85% entre os dados observados e estimados e RMSE menor que 15% para estimar a produtividade, e que as variáveis que tiveram maior importância para estimar a produtividade foram o genótipo, espaçamento, argila, matéria orgânica, capacidade de troca catiónica (CTC), chuva, temperatura, défice de água e performance de fertilização durante a rotação.

Esses estudos mostram a evolução das pesquisas sobre modelagem de dados florestais no Brasil, onde modelos de aprendizado de máquina e tradicionais, são utilizados frequentemente para estimativas de variáveis relacionadas ao processo de crescimento e produção do eucalipto. Entretanto existe um número menor de pesquisas com um superno aprofundamento em investigar um número maior de modelos de aprendizado de máquina para a modelagem da produção,o que abre a novos pesquisadores indagações e possibilidades a serem analisadas e testadas buscando melhorar a acurácia no processo de estimativa da produção.

## 3. Materiais e Descrição dos Modelos

Para o desenvolvimento deste estudo foram utilizados dois tipos de técnicas fortemente empregadas no meio florestal: as tradicionais, baseadas em modelos matemáticos de densidade variável, e as de aprendizado de máquina, baseadas em técnicas computacionais sofisticadas, como redes neurais e máquina de vetores de suporte. Uma breve descrição deles e da base de dados, são apresentados a seguir.

### 3.1. Conjunto de Dados

A base de dados utilizada no estudo foi concedida em parceria com uma empresa multinacional do setor florestal contendo informações referentes a 1280 unidades amostrais (parcelas) distribuídas em 507 talhões de *Eucalyptus sp* localizados no Guanhães, Minas Gerais (Figura 1). Os dados foram coletados em inventários florestais contínuos com medições anuais entre 2013 e 2019, e abrangem idades de 18 a 118 meses de idade.

### 3.2. Modelos Tradicionais

Os modelos mais utilizados de crescimento e produção em povoamentos florestais são aqueles de densidade variável [Campos and Leite 2017]. Um dos modelos mais usuais é o de [Clutter 1963], que utiliza atributos biométricos do povoamento relacionados à produção ($m^3ha^{-1}$) e produtividade ($m^3ha^{-1}$ ano$^{-1}$). Por ser amplamente adotado pelas indústrias do setor florestal, o modelo de Clutter será apresentado a seguir.

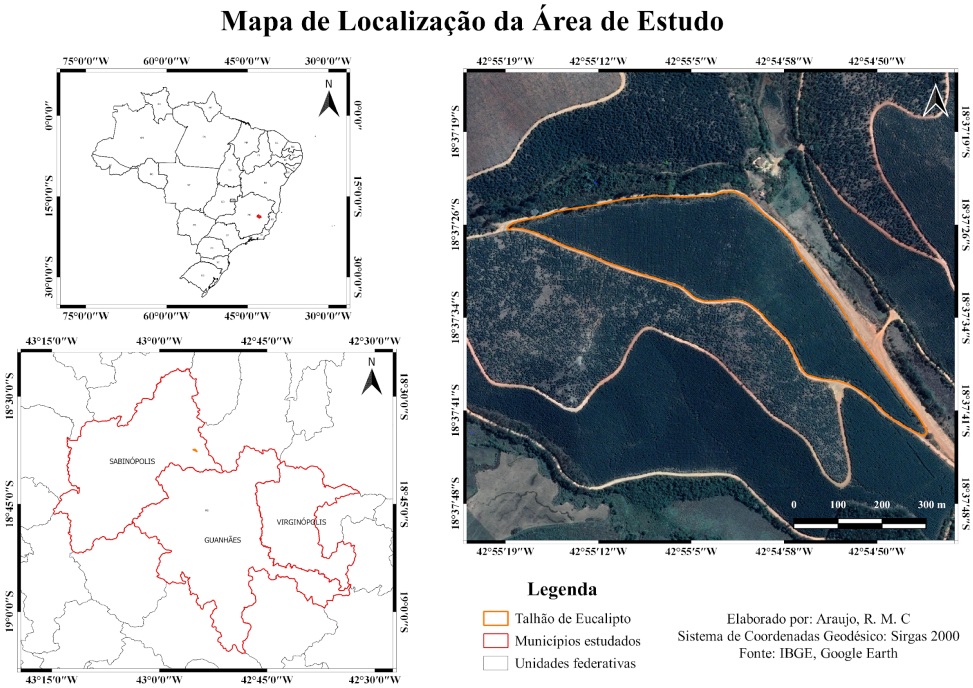


**Figura 1. Mapa de localização de um dos talhões de estudo**

### 3.2.1. Modelo de Clutter

Desenvolvido no ano de 1963, o modelo de Clutter é composto por um sistema de equações. Uma equação tem como variável resposta a área basal do povoamento e a outra cuja variável resposta é o volume. A equação de área basal permite a projeção da área basal para uma idade de interesse que por sua vez é utilizada na equação de volume para a estimação do volume nesta idade.

$$\begin{cases} LnV_2 = \beta_0 + \beta_1(\frac{1}{I_2}) + \beta_2 S + \beta_3 LnB_2 + \epsilon \\ LnB_2 = LnB_1(\frac{I_1}{I_2}) + \alpha_0(1 - \frac{I_1}{I_2}) + \alpha_1(1 - \frac{I_1}{I_2})S + \epsilon \end{cases} \tag{1}$$

em que: $I_1$ é a idade presente, em meses; $I_2$, é a idade futura, em meses; $\beta_1$, é a área basal na idade $I_1$, em $m^2ha^{-1}$; $B_2$, é a área basal na idade $I_2$, em $m^2ha^{-1}$; S, é o índice de local; $V_2$, é o volume na idade $I_2$ em $m^3ha^{-1}$; $\alpha_i$ e $\beta_i$, são os parâmetros do modelo; e $\epsilon$ é o erro aleatório, em que $\epsilon \sim$ N(0, $\sigma^2$).

Por se tratar de sistema de equações simultâneas, a estimação dos parâmetros é feita pelo método de mínimos quadrados em dois estágios.

### 3.3. Modelos de aprendizado de máquina

Segundo [Brink et al. 2016], a área de Aprendizado de Máquina é considerada um ramo do setor de Inteligência Artificial (IA), que tem por finalidade o estudo e construção de sistemas capazes de aprender de forma automatizada, supervisionada, a partir de um conjunto de dados prévio.

Esses sistemas são formados pelos algoritmos, que em geral, tem a capacidade de modelar classes complexas por meio do mapeamento de uma função de entrada para saída, podendo aceitar dados preditores variados de entrada e não fazer suposições sobre a distribuição deles [Maxwell et al. 2018].

Em um sistema, caso o resultado da modelagem seja uma categoria, como exemplo o conceito de um conjunto de árvores em uma parcela, em uma das categorias A (jovem), B (adulta), C (idosa), temos uma tarefa de classificação. Porém, se for um valor numérico, como a previsão do volume de um conjunto de árvores em uma parcela, temos uma tarefa de regressão.

Dentro do setor florestal, a maior parte dos estudos são voltados para regressão, em função da estimativa de valores pontuais, como exemplo o volume. Alguns algoritmos conhecidos são utilizados, como regressão linear (LR), perceptron multi-camadas (MLP), máquina de vetores de suporte (SVM), floresta aleatória (RF) e k-vizinhos mais próximos (KNN), devido a sua alta capacidade de agregar muitas variáveis de entrada e como saída retornar uma resposta de forma precisa e pontual, a partir de um conjunto de aprendizado, treinamento e teste [Rex 2019].

### 3.3.1. Regressão Linear (LR)

A regressão linear pode ser definida como um modelo que busca estimar um número (variável $Y$ ou dependente), dados os valores de outras variáveis (variáveis $x$ ou independente), podendo ser expressa pela seguinte relação, equação 2, [Pagliarussi 2018]:

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon \tag{2}$$

em que: $Y$ é a variável resposta (dependente) a ser estimada; $\beta_0$ é o intercepto; $\beta_1$ e $x_p$ são os preditores com seus respectivos parâmetros; $\epsilon$ é o erro aleatório $\epsilon \sim$ N(0, $\sigma^2$).

### 3.3.2. Perceptron Multi-Camadas (MLP)

As redes neurais artificiais segundo [Haykin 2001] em sua forma geral, são como uma máquina projetada para modelar a forma com que o cérebro realiza uma tarefa, utilizando-se de uma rede composta por células computacionais denominadas neurônios ou unidades de processamento, que em um processo de aprendizagem tem a função de armazenar conhecimento experimental e torná-lo disponível para o uso.

O neurônio é uma unidade de processamento primordial para a operação de uma rede neural. Como ilustrado na Figura 2, ela é composta basicamente por cinco elementos: as sinapses ou elos de conexão, caracterizados por um peso ou força própria; um somador para somar os sinais de entrada; uma função de ativação para restringir a amplitude da saída de um neurônio; o bias que tem o efeito de aumentar ou diminuir a entrada líquida da função de ativação e a saída que pode ser entrada para outro neurônio ou para o resultado final [Haykin 2001].

Uma importante classe das redes neurais conhecida como MLP, tomou destaque pela sua arquitetura com uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída, com o sinal se propagando para frente, camada por camada, Figura 3.

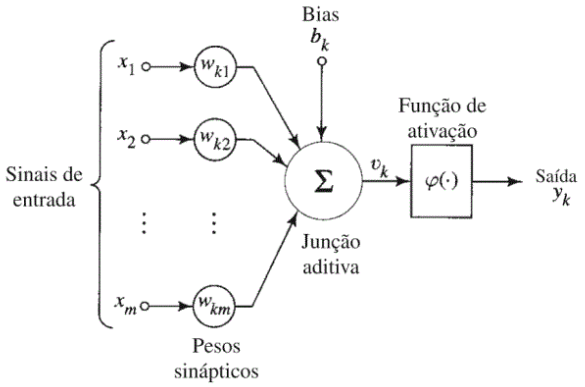


**Figura 2. Representação de um neurônio**

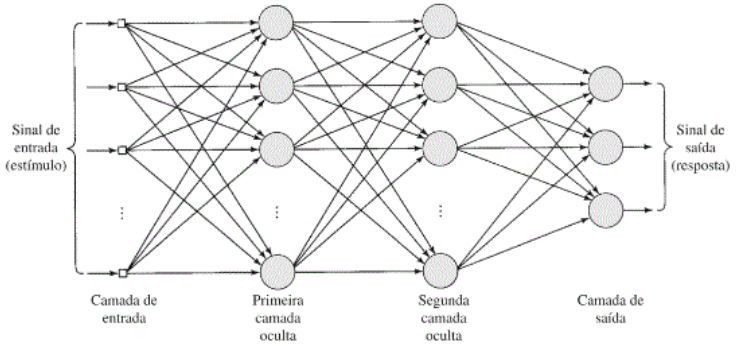


**Figura 3. Representação de uma MLP**

Ela tem sido frequentemente aplicada na resolução de problemas difíceis com a ajuda do treinamento supervisionado e do algoritmo de retro-propagação do erro. O aprendizado por retro-propagação é definido em dois passos: o primeiro para frente, onde um vetor de entrada é aplicado aos nós da rede e se propaga para frente camada por camada por meio de sinais funcionais, com os pesos sinápticos fixos, produzindo uma resposta real da rede. Durante a volta, passo para trás, os pesos sinápticos são ajustados, onde o resultado real é subtraído do desejado, a fim de produzir sinais de erro que são trazidos, a fim de corrigir os pesos sinápticos para que a reposta real da rede se mova mais perto da resposta desejada, Figura 4.

### 3.3.3. Máquina de Vetores de Suporte (SVM)

A máquina de vetores de suporte é um algoritmo da área de aprendizado de máquina popular que atribui tarefas de regressão e classificação. A regressão do vetor de suporte é semelhante a regressão linear, onde a equação da linha é denominada hiperplano. Os pontos dos lados próximos a ele são chamados de vetores de suporte, usados para plotar a linha de limite, Figura 5. A ideia por trás da SVM é tentar ajustar a melhor linha dentro de um valor limiar (distância entre o hiperplano e a linha de fronteira). Para uma regressão não linear, onde não existe hiperplano que separe instâncias positivas das negativas, a solução é mapear os dados para um espaço de maior dimensão, onde uma função Kernel

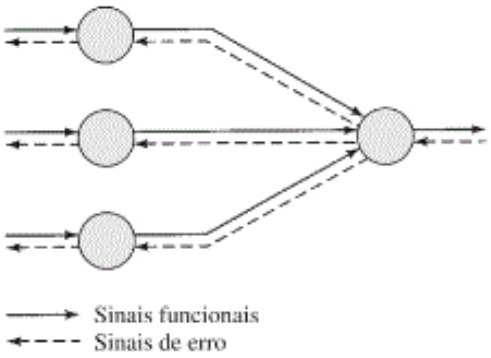


**Figura 4. Representação do princípio de retropagação**

transforma e realiza essa separação linear [Carneiro 2017].

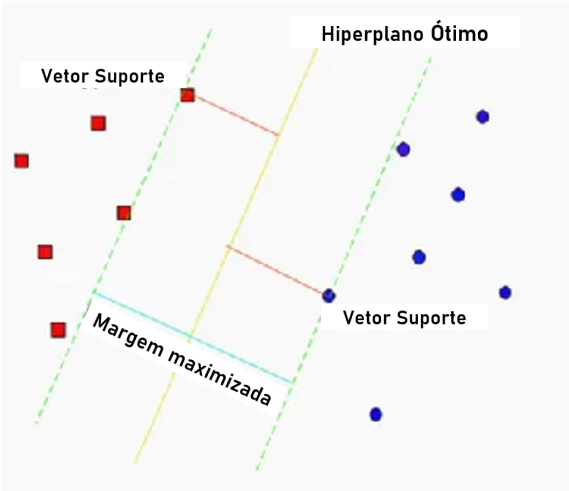


**Figura 5. Exemplo de uma SVM**

### 3.3.4. Floresta Aleatória (RF)

As florestas aleatórias conferem uma técnica de aprendizado que combina numerosos algoritmos de árvores de decisão para classificar ou prever uma determinada variável. Na regressão ao receber um vetor de entrada x composto por valores referentes às características de uma determinada área por exemplo, a RF constrói um número k de subamostras (árvores de regressão), tal como ilustrado na Figura 6, e calcula a média dos resultados, gerando a saída da floresta [Galiano et al. 2015].

Para evitar a correlação entre as diferentes árvores a RF aumenta a diversidade, fazendo a crescer a partir de diferentes dados de treinamento, pela técnica de *bagging*. Essa técnica consiste na criação de dados de treinamento por re-amostragem aleatória do conjunto de dados original, com substituição, ou seja, sem exclusão dos dados de entrada para gerar o próximo subconjunto. Dessa forma há um aumento na precisão da previsão, devido a robustez em enfrentar pequenas variações nos dados de entrada [Breiman 2001].

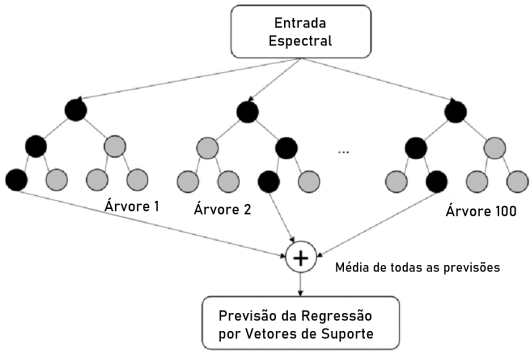


**Figura 6. Exemplo de uma floresta aleatória**

### 3.3.5. K-Vizinhos mais Próximos (KNN)

A técnica de aprendizado por vizinhos mais próximo, vem trazendo diversas funcionalidades tanto na forma não supervisionada, quanto supervisionada. Para a última ela se apresenta em duas formas de tarefa: classificação para dados discretos, e regressão para dados contínuos. Sua função trata-se basicamente em encontrar, dado uma amostra de treinamento, uma porção de dados (vizinhos) mais próxima (similar) do novo ponto, e assim prever o seu valor a partir deles. Na regressão esse valor é calculado com base na média de seus vizinhos, que pode variar em função do número k de vizinhos e dos pesos aplicados a eles, que podem ser ponderados de forma uniforme ou relativos a distância, conforme exemplificado na Figura 7, abaixo [Pedregosa et al. 2011].

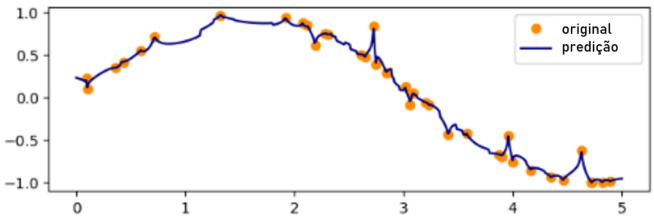


**Figura 7. Regressão KNN com (k = 5, pesos = distância)**

## 3.4. Metodologia Experimental

Para arquitetura do problema, optou-se pela criação de um código único em Python, contendo todos os algoritmos, afim de modelar os dados do inventário florestal, a partir de uma base de dados consistida. Levando em conta a modelagem tradicional, as equações de Clutter foram inseridas no código, onde o ajuste de seus coeficientes e pesos, foi feito de forma automática considerando atributos do inventário florestal, idade (meses), área basal (G), indice de sitio (S) e volume total com casca (VTCC).

No mesmo código, para parte dos modelos de aprendizado de máquina, foram inseridos os algoritmos segundo a plataforma Scikit-learn[Pedregosa et al. 2011], com a LR, MLP, RF, SVM, KNN e o algoritmo Grid Search para realizar a seleção de parâmetros dos modelos, utilizando os mesmos atributos usados para o modelo de Clutter.

- Para o algoritmo LR, foram utilizados a configuração padrão de parâmetros.

- Já no algoritmo MLP, foram avaliados os seguintes parâmetros: número de épocas igual a 400, taxa de aprendizado $\alpha \in \{0.001, 0,01, 0.05\}$, e número de neurônios para uma camada escondida $h_1 \in \{10, 50, 100, 500\}$ ou duas camadas escondidas $h_2 \in \{(10, 10), (50, 10, ), (50, 50, )\}$.
- Na SVM, foram avaliados os parâmetros de kernel Gaussiano, $\gamma \in \{2^{-10}, 2^{-8}, ..., 2^2\}$ e custo $C \in \{2^{-2}, 2^0, ..., 2^{10}\}$.
- No algoritmo RF, foi avaliado o parâmetro número de árvores $t \in \{2^1, 2^2, ..., 2^{10}\}$.
- E no KNN, foi avaliado o parâmetro número de vizinhos $k \in \{1, 3, 5, ..., 21\}$.

### 3.5. Desenho experimental

Buscando avaliar o desempenho dos algoritmos, foram testados dois experimentos buscando estimar o volume de eucalipto, onde a tarefa, os atributos e a quantidade deles, são definidas nas tabelas 1 e 2.

**Tabela 1. Experimentos e suas tarefas**

| Experimento | Tarefa | Atributos |
|---|---|---|
| 1 | Estimativa do volume presente (Clutter) | Inventário: idade (meses), área basal (G), índice de sítio (S) e volume total com casca (VTCC) |
| 2 | Estimativa do volume futuro (Clutter) | Inventário: idade atual (meses), área basal (G), índice de sitio (S), volume total com casca (VTCC), idade futura (meses) e área basal futura (G2) |

**Tabela 2. Experimentos e quantitativo de dados**

| Experimento | Tarefa | Amostras | Atributos |
|---|---|---|---|
| 1 | Estimativa do volume presente (Clutter) | 5853 | 4 |
| 2 | Estimativa do volume futuro (Clutter) | 3405 | 6 |

### 3.6. Medidas de desempenho

Como medida de desempenho para avaliação da acurácia dos modelos, foi utilizado o erro médio absoluto (MAE), (equação 3) e a raiz quadrada do erro-médio (RMSE), (equação 4), abaixo:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \ , \tag{3}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \ , \tag{4}$$

onde: $y_i$ = variável dependente obtida do inventário florestal; $\hat{y}_i$= variável dependente da estimativa dos modelos; $n$ = número de observações;

## 4. Resultados Experimentais

Foram conduzidos dois experimentos buscando estimar a produção presente e futura. A condução de cada experimento e os resultados obtidos, são apresentados abaixo.

### 4.1. Estimativa do Volume Presente

Em um primeiro experimento, buscou-se estimar o volume presente, utilizando os algoritmos das seções 3.4 e os atributos descritos na tabela 1, em um mesmo código na linguagem Python, de forma preliminar a fim de conhecer e treinar os modelos.

A estimativa do volume, foi dada pela validação cruzada, onde a base de dados foi dividida em 5 partes e realizadas 5 execuções de treinamento e teste, em cada uma delas considerando uma porção diferente dos dados para teste e as demais para treinamento dos modelos.

Em relação a acurácia dos modelos, foi utilizado o MAE (equação 3) e o RMSE (equação 4), com seus respectivos desvios padrões. Os resultados para a estimativa do volume presente são apresentados na Tabela 3.

**Tabela 3. Erro da estimação do volume de produção presente**

| Alg. | MAE | RMSE |
|---|---|---|
| Clutter | $7.38 \pm 0.07$ | $10.36 \pm 0.65$ |
| Linear | $11.15 \pm 0.20$ | $15.40 \pm 0.28$ |
| KNN | $8.43 \pm 0.23$ | $12.13 \pm 0.73$ |
| RF | $7.80 \pm 0.10$ | $10.83 \pm 0.62$ |
| MLP | $9.05 \pm 1.15$ | $11.89 \pm 1.35$ |
| SVM | $\mathbf{7.13 \pm 0.06}$ | $\mathbf{9.95 \pm 0.65}$ |

Observa-se nesse primeiro experimento que o modelo tradicional de Clutter apresentou valores de estimativa de erro melhores que quase todos os algoritmos de aprendizado de máquina, mostrando a robustez de seu sistema de equações, ficando atrás apenas do algoritmo SVM, que obteve os melhores resultados para a estimativa de volume presente, com MAE, 7.13 $m^3ha^{-1}$, e o RMSE 9.95 $m^3ha^{-1}$.

Isso corrobora o porquê desse modelo tradicional ser a categoria mais utilizada pelas empresas florestais no Brasil, o que é explicado pelo seu menor grau de complexidade e por atender aos requisitos de planejamento de produção de madeira [Castro et al. 2013a]. Contudo, é importante destacar que mesmo em um cenário com poucos atributos a serem considerados, o algoritmo de aprendizado de máquina SVM conseguiu obter melhor desempenho preditivo do que o modelo referência (Clutter).

### 4.2. Estimativa do Volume Futuro

Embora a análise da estimativa do volume presente seja interessante para analisarmos o desempenho dos modelos para estimar a produção atual a partir de atributos coletados naquele mesmo momento, ela deve ser tratada como uma análise preliminar. A análise da estimativa do volume futuro é o grande interesse das indústrias do setor florestal. Partimos assim para o segundo experimento a fim de estimar a produção futura considerando o conjunto de dados obtido a partir do inventário contínuo disponibilizado pela empresa.

Foram utilizados os mesmos modelos e atributos da seção 4.1, destacando-se que no formato futuro, duas novas variáveis são acrescentadas ao modelo de Clutter, a idade futura, que marca o intervalo de tempo para a estimativa do volume futuro prevista pelo modelo a partir do espaço de tempo entre as medições e a área basal futura, que é calculada no sistema de equações. Estas também foram usadas como entrada nos modelos de aprendizado de máquina, a fim de obter, um critério igual de comparação entre eles.

À validação cruzada foi novamente usada, com validação da acurácia dos modelos por meio do MAE e RMSE. Os resultados para a estimação do volume futuro, são apresentadas na Tabela 4. Neste segundo experimento, observa-se que os algoritmos de aprendizado, exceto o KNN, conseguiram se sobressair ao modelo tradicional de Clutter, nas estimativas de erro para o volume futuro. A SVM mais uma vez apresentou bom desempenho preditivo e se mostrou a melhor técnica em termos de MAE, com erro de 24.78 $m^3ha^{-1}$, enquanto a MLP obteve o melhor resultado de RMSE, com erro de 39.27 $m^3ha^{-1}$.

Notasse a pouca discrepância de estimativa de um modelo para o outro, mostrando assim, a eficiência da modelagem ao conjunto de dados. Entretanto, vemos que para a estimativa futura os modelos de aprendizado conseguem ser mais eficazes. Vale ressaltar que outros estudos da literatura, tal como [Castro et al. 2013b], apresenta um comparativo favorável à MLP frente ao modelo de Clutter. Contudo, o nosso estudo difere dos demais por realizar essa análise considerando o cenário em que o mínimo de atributos estão disponíveis (modelo de Clutter), o que é uma contribuição importante tanto para a literatura quanto para as indústrias do setor florestal.

**Tabela 4. Erro da estimação do volume de produção futuro**

| Alg. | MAE | RMSE |
|---|---|---|
| Clutter | 28.22 ± 2.19 | 42.52 ± 3.76 |
| Linear | 25.44 ± 1.33 | 39.75 ± 2.73 |
| KNN | 29.71 ± 1.19 | 44.82 ± 3.39 |
| RF | 27.92 ± 1.97 | 42.49 ± 2.98 |
| MLP | 25.30 ± 1.28 | **39.27** ± **2.51** |
| SVM | **24.78** ± 1.70 | 39.36 ± 3.94 |

## 5. Conclusões

Um estudo foi realizado para um conjunto de dados de 507 talhões da cultura *Eucalyptus spp*, por meio da modelagem tradicional de Clutter e a de aprendizado de máquina, com os algoritmos LR, SVM, MLP, RF, KNN para estimativa de volume, nesses locais. Os modelos aqui apresentados demostram forte precisão, o que pode ser avaliado pela estimativa da acurácia dos erros, através do MAE e RMSE. Os modelos de aprendizados conseguem estimativas mais precisas tanto na forma futura, quanto no presente, quando comparados ao tradicional de Clutter, com destaque para a SVM nos experimentos realizados.

Espera-se futuramente buscar melhorias nos modelos apresentados, aplicando técnicas mais robustas de aprendizado de máquina como as Redes Neurais Convolucionais, em que estudos preliminares, tal como [Santos 2021], demonstraram grande potencial na estimativa de atributos do inventário, como diâmetro e altura no modelo florestal.

## Agradecimentos

Os autores agradecem o apoio da empresa Celulose Nipo Brasileira (CENIBRA).

## Referências

Alcântara, A. E. M. (2012). *Alternativas de modelagem para projeção do crescimento de eucalipto em nível de povoamento. 2012. 52 p.* PhD thesis, Dissertação (Mestrado)–Universidade Federal de Viçosa, Viçosa.

Binoti, M. (2012). *Emprego de redes neurais artificiais em mensuração e manejo florestal. 2012. 130 f.* PhD thesis, Tese (Doutorado em Ciência Florestal)–Universidade Federal de Viçosa, Viçosa, MG.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Brink, H., Richards, J., and Fetherolf, M. (2016). *Real-world machine learning*. Simon and Schuster.

Campos, A. S., Silva, D. S., Lienard, R. B. D., Freitas, M. C. S., Moura, J. N. M., and Trazzi, P. A. (2022). Análises bibliométricas sobre a produção e tecnologia de celulose e papel de eucalipto. In *9° Congresso Florestal Brasileiro*, volume 1, pages 365–368.

Campos, J. and Leite, H. (2017). Mensuração florestal: perguntas e respostas (636 p.). *Viçosa: Editora UFV*.

Carneiro, M. G. (2017). *Redes complexas para classificação de dados via conformidade de padrão, caracterização de importância e otimização estrutural*. PhD thesis, Universidade de São Paulo.

Castro, R., Boechat Soares, C. P., Leite, H. G., Lopes de Souza, A., Saraiva Nogueira, G., and Bolzan Martins, F. (2013a). Individual growth model for eucalyptus stands in brazil using artificial neural network. *International Scholarly Research Notices*, 2013.

Castro, R. V. O., Soares, C. P. B., Martins, F. B., and Leite, H. G. (2013b). Crescimento e produção de plantios comerciais de eucalipto estimados por duas categorias de modelos. *Pesquisa Agropecuária Brasileira*, 48:287–295.

Clutter, J. L. (1963). Compatible growth and yield models for loblolly pine. *Forest science*, 9(3):354–371.

Cordeiro, M. A., Arce, J. E., Guimarães, F. A. R., Bonete, I. P., Silva, A. V. d. S., Abreu, J. C. d., and Binoti, D. H. B. (2022). Estimativas volumétricas em povoamentos de eucalipto utilizando máquinas de vetores de suporte e redes neurais artificiais. *Madera y bosques*, 28(1).

Dai, P. V. S., Baio, F. H. R., AZEVEDO, G. d., Fagundes, L. A., and Trento, A. C. S. (2021). Estimativa de volume de madeira baseada em índices de vegetação. *Scientia Forestalis*, 49(129):e3301.

De Freitas, E. C. S., de Paiva, H. N., Neves, J. C. L., Marcatti, G. E., and Leite, H. G. (2020). Modeling of eucalyptus productivity with artificial neural networks. *Industrial Crops and Products*, 146:112149.

Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., and Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural

networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71:804–818.

Gorgens, E. B., Leite, H. G., Santos, H. d. N., and Gleriani, J. M. (2009). Estimação do volume de árvores utilizando redes neurais artificiais. *Revista Árvore*, 33:1141–1147.

Haykin, S. (2001). *Redes neurais: princípios e prática*. Bookman Editora.

Ibá (2022). *Relatório Anual*. Industria Brasileira de Árvores.

Maxwell, A. E., Warner, T. A., and Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, 39(9):2784–2817.

Pagliarussi, M. S. (2018). O ensino do modelo clássico de regressão linear por meio de simulação de monte carlo. 12:e152100.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Penido, T. M. A., Lafetá, B. O., Nogueira, G. S., Alves, P. H., Gorgens, E. B., and Oliveira, M. d. (2020). Modelos de crescimento e produção para a estimativa volumétrica em povoamentos comerciais de eucalipto. *Scientia Forestalis*, 48(128):e3340.

Reis, L. P., de Souza, A. L., dos Reis, P. C. M., Mazzei, L., Soares, C. P. B., Torres, C. M. M. E., da Silva, L. F., Ruschel, A. R., Rêgo, L. J. S., and Leite, H. G. (2018). Estimation of mortality and survival of individual trees after harvesting wood using artificial neural networks in the amazon rain forest. *Ecological Engineering*, 112:140–147.

Rex, F, E. (2019). *Estimativas do estoque e dinâmica de biomassa acima do solo utilizando diferentes abordagens estatìstcas e dados Lidar em floresta tropical. 2019. 119 f.* PhD thesis, Dissertação (Mestrado)–Universidade Federal do Parana, Curitiba, PR.

Santos, D. E. C. (2021). Regressão e deep learning aplicados em modelagem de árvore individual para eucalipto.

Were, K., Bui, D. T., Dick, Ø. B., and Singh, B. R. (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an afromontane landscape. *Ecological Indicators*, 52:394–403.