Predição do Incremento Médio Anual Volumétrico de Eucalyptus com Aprendizado de Máquina

Adilson Rosa Lopes[1], Jean Marcel Sousa Lira [1], Leonardo Araujo Oliveira[1], Marlon dos Santos Pereira Birindiba Garuzzo [1], Marcos Veniciu de Sá Barbalho[1], Patrick Oliveira Corrêa de Araújo[1], Gleison Augusto dos Santos [1], José Augusto Nacif[1]

[1]Universidade Federal de Viçosa (UFV) – Viçosa, MG – Brasil

{adillopes, jean.lira, leonardo.a.oliveira, marlon.garuzzo}@ufv.br
marcos.barbalho, patrick.araujo, gleison, jnacif}@ufv.br

Abstract. *This work applied machine learning algorithms to predict the future Average Annual Volume Increment (IMAVol m³/ha/year) of eucalyptus. The dataset used is composed of physiological variables and IMAVol of eucalyptus plants from a forest genetic improvement project. By applying four ML algorithms, the results were an average of 2.84 ± 0.02 and 0.83 ± 0.03 for the root mean square error (RMSE) and coefficient of determination (R^2) metrics, respectively, after performing 50 Kfold cross-validation iterations. The results are promising to support the early selection of high-volume productivity genetic materials.*

Resumo. *Este trabalho aplicou algoritmos de aprendizado de máquina (Machine Learning - ML) para predição futura do Incremento Médio Anual Volumétrico (IMAVol m³/ha/ano) de eucalipto. O conjunto de dados utilizado é composto de variáveis fisiológicas e o IMAVol de plantas de eucalipto de um projeto de melhoramento genético florestal. Aplicando quatro algoritmos de ML, os resultados foram a média de 2,84 ± 0,02 e 0,83 ± 0,03 para as métricas de raiz do erro quadrático médio (RMSE) e coeficiente de determinação (R^2), respectivamente, após a realização de 50 iterações de validação cruzada Kfold. Os resultados são promissores para apoiar a seleção precoce de materiais genéticos de alta produtividade volumétrica.*

1. Introdução

As florestas plantadas de eucalipto são uma atividade importante do agronegócio brasileiro, gerando uma diversidade de produtos como celulose, papel, painéis, compensados, carvão vegetal que gera energia para a indústria, entre outros. No Brasil, a área cultivada de eucalipto em 2021 correspondeu a 7,53 milhões de hectares, 75,8% do total de florestas plantadas [IBÁ 2021]. Isso ocorre devido às várias características silviculturais de interesse que a cultura apresenta, como boa forma do tronco, crescimento rápido, alto rendimento, tolerância a baixa fertilidade do solo, alagamento e baixa disponibilidade de água [Zaiton et al. 2020, Chen et al. 2022].

Apesar de tolerar alguns níveis de seca, a cultura do eucalipto vem sofrendo perdas de produtividade por causa do aumento da mortalidade e redução do crescimento decorrentes da maior frequência, duração e intensidade dos eventos de seca. Isto causa

prejuízos ao setor florestal brasileiro. Em função disso, as empresas que compõem o setor veem buscando desenvolver materiais genéticos de eucalipto mais resistentes à seca e altamente produtivos. Uma das formas de se obter esses materiais é através de teste de progênies em sítios sujeitos a estes eventos de seca.

Devido aos longos ciclos de reprodução, o melhoramento florestal normalmente é um processo demorado, dificultando o desenvolvimento de novos genótipos [Castro et al. 2021]. Desta maneira, a técnica de seleção precoce, que utiliza plantas jovens para obter características que serão utilizadas como preditores de crescimento e produtividade das plantas adultas, é uma maneira de acelerar o processo de melhoramento florestal, e com isso, antecipar os ganhos genéticos [Moraes et al. 2014, Corrêa et al. 2017]. Todavia, a determinação das características com maior poder de predição é um processo complexo e laborioso, de acordo com a quantidade de dados. Neste sentido, a utilização de aprendizado de máquina pode proporcionar maior rapidez ao processo.

As técnicas de Aprendizado de Máquina (*Machine Learning - ML*) têm sido aplicadas em várias as áreas de estudo de plantas para análises de dados de alto rendimento, realização de previsões e otimização de variáveis em sistemas biológicos complexos [Silva et al. 2019]. Neste sentido, o objetivo do trabalho foi avaliar o desempenho de quatro algoritmos de aprendizado de máquina supervisionados para a predição do Incremento Médio Anual Volumétrico (IMAVol m³/ha/ano), uma variável de produtividade, em um sítio com baixa disponibilidade hídrica. Assim, os modelos realizaram previsões do IMAVol baseado em características de plantas mais jovens, a fim de automatizar a seleção precoce de genótipos com maior potencial produtivo em ambientes com **baixa disponibilidade hídrica**.

Para apresentar o trabalho proposto, além desta seção de introdução, a Seção 2 apresenta trabalhos relacionados, a Seção 3 descreve os materiais e métodos empregados e a Seção 4 mostra os resultados alcançados e a discussão. Por fim, na Seção 5 são feitas as considerações finais e as perspectivas para trabalhos futuros.

2. Trabalhos Relacionados

Após revisão de literatura, não foram encontrados estudos que tratem do mesmo objetivo deste estudo, que é o uso de ML para prever a produtividade de eucalipto com dados de experimento de campo em **ambiente de seca**. No entanto, foram encontrados trabalhos relacionados que abordam questões semelhantes:
(i) [da Silva Tavares Júnior et al. 2020]: realizou a predição do Incremento Periódico Anual Médio em DAP (*API dbh; cm y-¹*), utilizou dados de inventário florestal de três fragmentos de floresta atlântica coletados ao longo de 4 anos, composto pelas variáveis: Diâmetro à Altura do Peito (DAP cm), Índice de Competição, Período (três categorias) e Grupo de Espécies (seis categorias de espécies agrupadas a partir de análise de *cluster*). O estudo aplicou os algoritmos de RF, MLP e SVM, comparando os resultados obtidos pela combinação de diferentes hiperparâmetros e variáveis preditoras. O trabalho avaliou o desempenho pela métrica de erro RMSE (Equação 1) e alcançou um percentual de erro (RMSE %) por volta de 12,75% com o RF.
(ii) [Li et al. 2022]: usou dados de imagens multiespectrais e de georeferenciamento de uma floresta subtropical úmida da região sul da China como conjunto de dados para predição da biomassa florestal. Foram comparados três algoritmos de aprendizado de

máquina (SVM, RF e Árvore de Decisão) com modelos de regressão linear. Os melhores resultados foram com RF, que teve um RMSE de 15,0377 (RMSE% por volta de 15%) e R^2 de 0,8670.
(iii) [Cordeiro et al. 2022]: a partir dos dados de inventário com 214 amostras de um plantio comercial de eucalipto, em um ambiente de clima equatorial úmido no estado do Amapá, foram usadas as variáveis DAP (cm), Altura total (Ht m) e Volume (m^3), comparou dois modelos estatísticos clássicos com três configurações de MLP e duas de SVM para prever o volume em povoamentos de eucalipto. O resultado foi um RMSE % variando de 9,38 a 9,81% nas técnicas estatísticas; 4,71 a 8,06% para MLP; e 7,64 a 7,81% com SVM. Os resultados da pesquisa estão próximos e até superam resultados alcançados na predição da produtividade em eucalipto pelos trabalhos acima, embora a comparação não seja totalmente equânime devido a diferença nas variáveis trabalhadas.

3. Materiais e Métodos

Esta seção apresenta a metodologia desenvolvida neste trabalho de pesquisa. Será descrito brevemente o projeto de melhoramento genético florestal que gerou os dados para a pesquisa, o conjunto de dados e a configuração experimental dos dados e algoritmos de aprendizado de máquina empregados.

3.1. Caracterização da área

O projeto de melhoramento genético florestal, que forneceu os dados para este trabalho, avaliou um teste de progênies de híbridos de irmãos completos de Eucalyptus instalada com o objetivo de selecionar genótipos com alta produtividade e potencial de tolerância à seca. O teste foi instalado em março de 2019 num sítio localizado no município de Buritizeiro, estado de Minas Gerais (17º 05' 49"S 44º 53' 09"O), o clima é tropical classificado como tipo Aw (Inverno Seco), com precipitação média anual de 1102 mm, temperatura média anual de 24,5 °C, máxima de 41 ºC e mínima igual a 5,4 ºC. O solo é do tipo argiloso (15 - 35 %) com textura média. Este sítio é sujeito a **eventos extremos de seca**, por isso possui um alto potencial para selecionar materiais genéticos tolerantes.

3.2. Desenho experimental

A duração do experimento será de aproximadamente sete anos, com medições periódicas em campo, iniciadas no sexto mês de plantio. O delineamento experimental é do tipo blocos ao acaso com genitores não aparentados. Apresenta 214 progênies de híbridos de Eucalyptus gerados por meio da técnica de hibridação - Protoginia Artificialmente Induzida (PAI) [Assis et al. 2005] e 6 clones testemunhas plantados para comparação dos resultados. Os tratamentos foram casualizados em 20 repetições para cada progênie e testemunha com espaçamento de 3,5 × 2,57 m (9 m^2), parcela de árvore única (Single Tree Plot - STP).

3.3. Variáveis medidas

O conjunto de dados fornecido pela equipe do projeto de melhoramento florestal é composto por (i) variáveis morfofisiológicos: Área Foliar Específica (AFE), Área Foliar Individual (AFI), Largura Foliar (LF), Comprimento Foliar (CF); Potencial Hídrico Foliar (PHF); (ii) Incremento Médio Anual Volumétrico - IMAVol (m3/ha/ano); (iii) identificação das amostras: número de identificação do indivíduo e da progênie da qual

ele faz parte. Para determinação da AFE, AFI, CF e LF, aproximadamente 20 folhas completamente expandidas do terço médio foram coletadas e digitalizadas em scanner (Hp ScanJet 200). Em seguida, as folhas foram secas em estufa de circulação forçada de ar a 70 ºC até atingir massa constante. A AFE foi determinada pela razão entre a AFI (cm²) e a massa seca (g). As imagens digitalizadas foram medidas no software Image-Pro Plus® para obtenção da AFI, CF e LF. O PHF foi determinado ao meio-dia com o auxílio de uma Bomba de Scholander em uma folha completamente expandida do terço médio da planta. O IMAvol foi calculado utilizando o volume de cada árvore do experimento produzido no espaçamento de 3,5 × 2,57 m (9 m²), extrapolado para 1 ha e dividido pela respectiva idade em meses, conforme descrito pela Equação 1.

$$IMAVol = \frac{(VOL \times 10.000)}{idade} \tag{1}$$

onde VOL: Volume da árvore em m³ no espaçamento de 9 m² e $idade$ é a idade da planta em meses. O Volume(VOL) foi calculado pela Equação 2 de [Schumacher and Hall 1933]:

$$VOL = \frac{(\pi \times DAP^2 \times Altura \times f)}{40.000} \tag{2}$$

onde π é a razão entre a circunferência e diâmetro de um círculo (3,14159); DAP é o diâmetro à altura do peito em centímetros; $Altura$ é a altura total da árvore em metros; f é o fator de forma adotado (0,45).

As coletas dos dados foram realizadas aos 6, 18, 30, 36 e 42 meses, para a última idade foi disponibilizado apenas o IMAVol. No total 83 progênies diferentes tiveram indivíduos amostrados, bem como as 6 testemunhas. As variáveis de identificação das amostras são discretas e as demais são contínuas.

3.4. Modelagem dos dados

Com o intuito de obter o melhor resultado para os modelos testados, foram experimentadas diferentes configurações e transformações dos dados. Os dados de cada coleta foram agrupados de modo a utilizar as características preditoras de plantas mais jovens como entrada dos modelos, rotulando as amostras pelo IMAVol na idade futura (alvo). As variáveis selecionadas para entrada (preditoras) foram AFE, AFI, CF, LF, PHF e o IMAVol das idades anteriores à idade alvo, aqui considerada como sendo a de 42 meses por ser a última idade coletada. Essa abordagem permitiu, por exemplo, treinar os modelos com dados de entrada das coletas de 6 e 18 meses para predizer a produtividade volumétrica aos 42 meses, no intuito de modelar a predição precoce da variável alvo. Nesse sentido, foram testadas diferentes combinações dos dados das idades coletadas, conforme a Tabela 1, visando avaliar a que obteve melhor desempenho.

Foram encontrados dados faltantes nas variáveis preditoras em todas as idades coletadas do conjunto de dados. Para tratar isso, foi adotada a abordagem de imputar a média da variável para os indivíduos da mesma progênie e na mesma idade da amostra. A correlação entre as variáveis foi analisada para verificar a multicolinearidade, que pode afetar o resultado da regressão. O maior coeficiente de correlação foi 0,81 entre o IMAVol aos 18 e aos 30 meses, e os demais foram menores ou iguais a 0,72. Para mitigar o

Tabela 1. Combinações de idades testadas para predição precoce do IMAVol

Idades agrupadas (entradas)	Idade futura (rótulo)	Amostras
6	42	326
6 e 18		492
6 e 18 e 30		579
6 e 18 e 30 e 36		623

problema de uma possível multicolinearidade, foi utilizada a técnica de Análise de Componentes Principais (PCA) e o número de componentes testados foi de 5. No entanto, não houve melhora nos resultados após a aplicação do PCA. Também foi testado o efeito da mudança de escala dos dados, os melhores resultados foram obtidos deixando os dados na escala original.

3.5. Configuração experimental dos algoritmos de ML

Buscando ajustar um algoritmo de ML com alta performance para a tarefa de predição precoce do IMAVol, foram escolhidos quatro algoritmos supervisionados amplamente citados na literatura para predição de produtividade de eucalipto, são eles:
i) Random Forest - RF: baseia-se no método de aprendizado de conjunto. Ele cria várias árvores de decisão (floresta) com um conjunto de dados de treinamento aleatório para cada árvore. Em problemas de regressão, a previsão do algoritmo é dada pela média das predições das árvores [Breiman 2001].
ii) Extreme Gradient Boost - XGBoost: também baseado em árvores de decisão como o RF, usa técnicas de otimização para aumentar a precisão das previsões. Funciona treinando várias árvores pequenas de maneira incremental, onde cada nova árvore tenta corrigir os erros de predição feitos pelas anteriores. O resultado final é uma combinação ponderada das previsões das várias árvores [Chen and Guestrin 2016].
iii) Multilayer Perceptron - MLP: trata-se de uma rede neural artificial alimentada para frente, com camadas de neurônios que processam informações unidirecionalmente da entrada para a saída. Cada conexão possui um peso associado e o neurônio de uma camada recebe a soma ponderada desses pesos com a saída dos neurônios da camada anterior, acrescida do viés, submetida a uma função de ativação. O resultado da predição é obtido pela camada de saída. O aprendizado ocorre pela atualização dos pesos e os vieses da rede, visando para minimizar o erro de predição, geralmente pela técnica de retropropagação de erro [Taud and Mas 2018].
iv) Support Vector Machine - SVM: fundamenta-se na teoria da aprendizagem estatística, formulado como um problema de otimização que busca minimizar o erro de regressão enquanto maximiza a margem de separação entre os vetores de suporte, representados por uma função de regressão que melhor se ajusta aos dados de entrada. [Smola and Schölkopf 2004].

3.5.1. Avaliação de desempenho e ajuste de hiperparâmetros

Os algoritmos foram avaliados pela técnica de validação cruzada *k-fold* (*Cross Validation - CV - Kfold*). Essa abordagem consiste em dividir aleatoriamente as observações em k partições de tamanho similar, usar uma partição como conjunto de validação e as outras

como conjunto de treinamento, esse procedimento é repetido até que todas tenham sido usadas como treinamento e validação ao menos uma vez. O RMSE (Equação 3) do treinamento e validação são calculados para cada partição e ao final é calculada a média do RMSE de validação de todas as partições, então esta média será o RMSE da validação cruzada. O R^2 (Equação 4) da validação cruzada foi calculado da mesma forma. No estudo, os dados foram divididos em 10 partições (k-fold=10) e o processo de validação cruzada foi repetido 50 vezes para uma avaliação mais robusta. Ao final das iterações, a média dos 50 resultados para o RMSE e R^2 foi calculada e considerada como resultado para cada métrica. Os resultados das duas avaliações estão na Tabela 2.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \tag{3}$$

onde n é o número de amostras; y é o valor da variável de saída (IMAVOl aos 42 meses) observada; $\hat{y}$ é o valor de saída predito pelos modelos.

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{4}$$

onde $\bar{y}$ é a média dos valores da variável de saída observada e as demais variáveis são as mesmas da fórmula da Equação 3. Uma técnica de pesquisa em grade foi empregada para ajustar os hiperparâmetros dos algoritmos em busca de melhor desempenho, utilizando a função *HalvingGridSearchCV*, da biblioteca Sklearn [Pedregosa et al. 2011], utilizando validação cruzada k-fold (k=10).

Os algoritmos também foram avaliados quanto à importância relativa dos preditores em cada conjunto de dados (idade). A Figura 2 mostra as variáveis mais importantes para o *Random Forest*. O RF avalia a importância das variáveis preditoras pela contribuição de cada uma em reduzir a impureza nas divisões das árvores de decisão. Esta avaliação é feita calculando a média da diminuição da impureza (Gini ou entropia) resultante da divisão em cada variável e normalizando o resultado para que a soma das importâncias de todas as variáveis seja igual a 1. Desta forma o algoritmo identifica as variáveis mais importantes para a predição do modelo, permitindo uma melhor compreensão dos resultados [Pedregosa et al. 2011].

4. Resultados e Discussão

A codificação foi feita na linguagem Python, versão 3.8, e os algoritmos aplicados foram os disponíveis na biblioteca Sklearn. A execução dos códigos foi realizada em um servidor com processador AMD Ryzen 9 5900X 12-Core Processor CPU, 2.2 GHz, 64GB de memória principal, 35 TB de memória secundária e 2 GPU NVIDIA TITAN V com 12 GB de RAM.

Conforme destacado na Tabela 2, o algoritmo *Random Forest* foi o que obteve melhor resultado na predição precoce do IMAVol futuro (42 meses), uma vez que teve o menor valor da métrica de erro RMSE e o maior valor do coeficiente de determinação R^2, quando utilizadas as amostras de todas as idades como dados de entrada (6, 18, 30 e 36 meses). Até mesmo nas demais combinações dos dados o RF performou melhor, com

exceção da combinação 6, 18 e 30 meses onde o XGBoost alcançou resultado melhor. Os resultados de todos os algoritmos melhoraram à medida em que o conjunto de dados incorporou mais medições (idades), aumentando a quantidade de amostras e variáveis preditoras. Por outro lado, os piores resultados para todos os algoritmos foram com o conjunto de dados formado apenas pelas amostras coletadas aos 6 meses.
Os hiperparâmetros ajustados para o algoritmo com melhor resultado (RF aplicado aos dados de 6, 18, 30 e 36 meses) foram: *bootstrap: True; criterion: squared_error; max_depth: 20; max_features: auto; min_samples_leaf: 1; min_samples_split: 2; n_estimators: 500*. O segundo modelo a melhor performar foi o XGBoost, também

Tabela 2. Comparação do resultados dos algoritmos em cada conjunto de dados (idades), com desvio-padrão, e o % do RMSE em relação ao IMAVol médio (22.79 m³/ha/ano) na idade alvo (42 meses).

Idades	Algoritmo	RMSE	R²	RMSE(%)
6	RF	6,89±0,04	0,01±0,02	30,21
	XGBOOST	7,45±0,15	-0,16±0,05	32,66
	SVM	7,02±0,07	-0,02±0,03	30,78
	MLP	7,07±0,13	-0,05±0,02	31,00
6 e 18	RF	4,31±0,03	0,61±0,09	18,90
	XGBOOST	4,48±0,1	0,57±0,02	19,64
	SVM	5,69±0,12	0,31±0,03	24,95
	MLP	6,22±0,08	0,18±0,02	28,17
6 e 18 e 30	RF	3,35±0,03	0,76±0,06	14,69
	XGBOOST	3,32±0,09	0,76±0,01	14,56
	SVM	4,96±0,1	0,46±0,03	21,75
	MLP	5,52±0,04	0,35±0,01	24,21
6 e 18 e 30 e 36	**RF**	**2,84±0,02**	**0,83±0,03**	**12,45**
	XGBOOST	2,99±0,08	0,81±0,01	13,11
	SVM	4,53±0,1	0,56±0,02	19,86
	MLP	5,24±0,06	0,42±0,01	22,98

com dados de 6-18-30-36 meses, indicando que os algoritmos baseados em conjuntos de árvores de decisão são promissores para realizar predições usando dados da cultura do eucalipto, como também demonstrado em outros trabalhos da literatura.

A variação das médias do RMSE ao longo das repetições variou para cada algoritmo e conjunto de dados, embora tenham variado pouco considerando os modelos individualmente, mostrando que foram estáveis nas predições. As médias para RF e XGBoost estiveram sempre próximas em todos os conjuntos de dados, sempre com erros menores que SVM e MLP, com melhor performance quando avaliadas todas as idades. A única exceção foi no conjunto de dados de 6 meses, onde SVM esteve próximo ao RF e o XGBoost mais próximo a MLP. Para as demais combinações de idades, observou-se que a tendência do SVM e MLP foi de estarem próximos e com erros maiores, conforme a Figura 1. O RF foi o algoritmo com maior capacidade preditiva em todas as observações, com médias de RMSE chegando a 2,84±0,02 m³/ha/ano e R² de $0{,}83 \pm 0,03$.

A variável que mais teve importância relativa nas predições do melhor modelo (RF) foi o IMAVol (ima) em todas as idades, com uma diferença menor em relação às

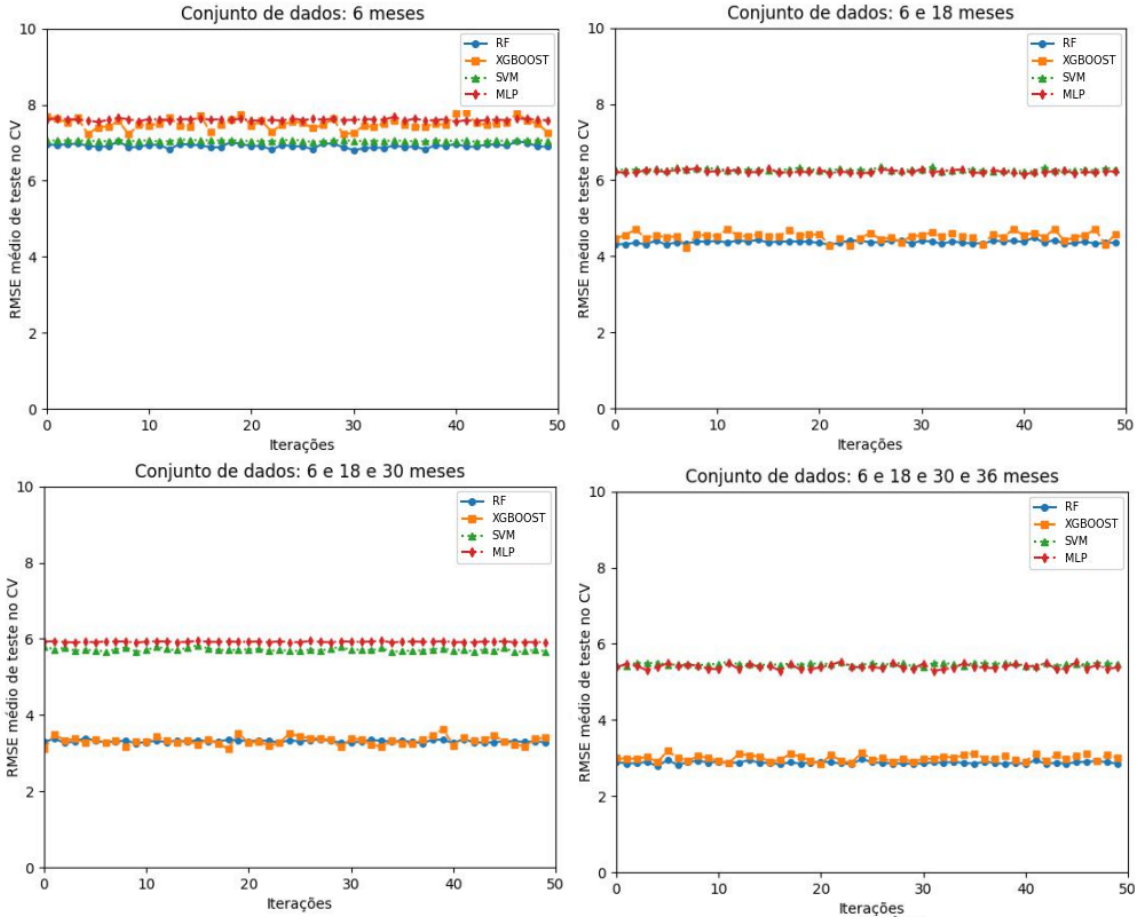

Figura 1. Variação do RMSE de validação nas iterações da validação cruzada

outras variáveis na idade de 6 meses. A importância dessa variável ficou mais acentuada na idade de 30 meses (ima_30), pois nos dois conjuntos de dados onde esteve presente, sua importância foi consideravelmente maior. De maneira oposta, a variável Área Foliar Individual (AFI) foi a que teve menor importância relativa em todas as combinações de idades, embora não tenha havido uma diferença significativa entre ela e as outras. Apesar da variável "ima" ter sido a mais importante, as demais tiveram uma importância relativa considerável, conforme mostra a Figura 2.

5. Considerações Finais

Os resultados alcançados mostram que o algoritmo *Random Forest* ajustado é promissor para apoiar programas de melhoramento florestal para seleção precoce de material genético com alto padrão de produção volumétrica, uma vez que a performance atingida se aproxima e até supera trabalhos recentes relacionados à predição da produtividade em eucalipto. 0s resultados são ainda incipientes, visto que o projeto de melhoramento florestal tem duração de 84 meses (sete anos) e os dados trabalhados são de, no máximo, 42 meses. À medida que mais dados forem incorporados à análise feita neste estudo, tanto em relação à quantidade de amostras, quanto de variáveis novas, pode haver um ganho de capacidade preditiva dos algoritmos, até mesmo pela tendência de redução do erro quando se tem mais dados. Dentre as variáveis preditoras, o IMAVol se destacou como a mais importante no resultado do melhor modelo, não obstante, todas as variáveis

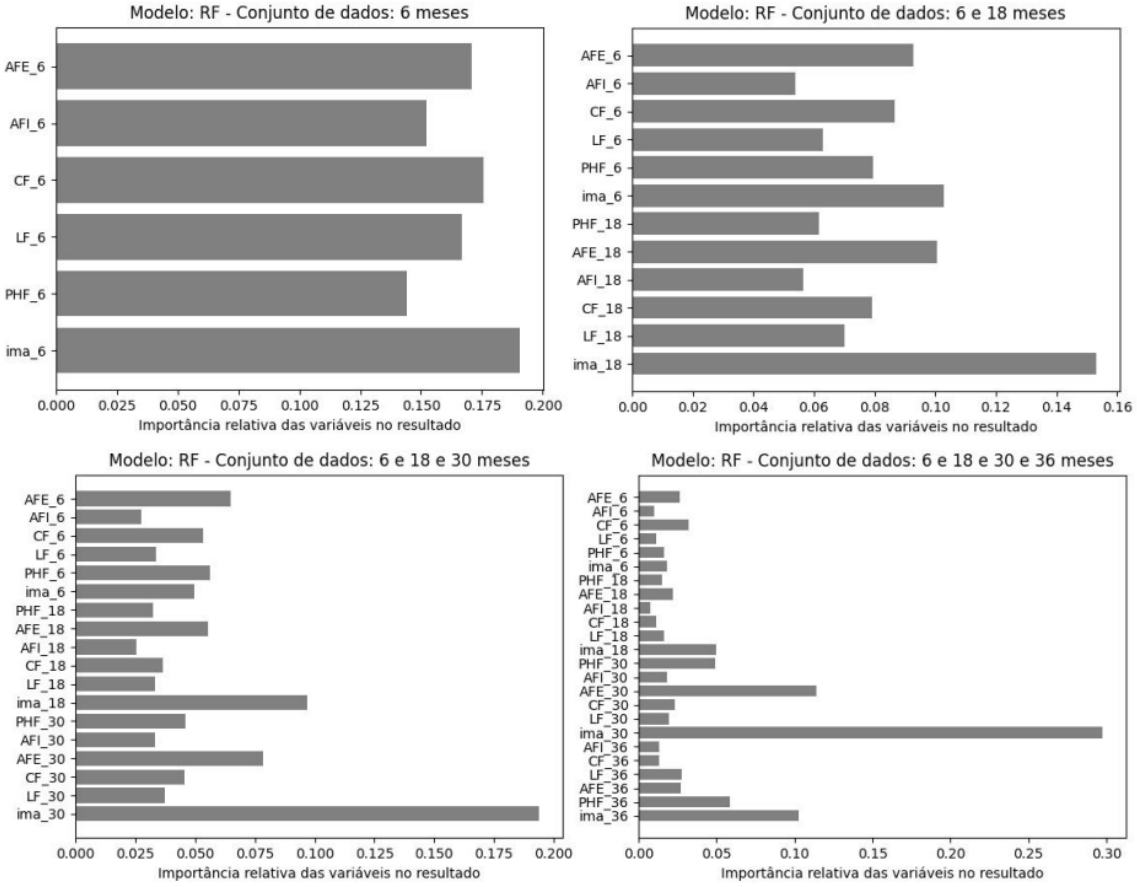

Figura 2. Importância relativa das variáveis nos resultados

fisiológicas e o status hídrico tiveram importância considerável nos resultados. Para trabalhos futuros, propõe-se testar outras combinações de idades, aplicar outros algoritmos de aprendizado de máquina e adquirir uma gama mais ampla de dados como parâmetros genéticos, análises bioquímicas e nutricionais das plantas, temperatura e precipitação, a fim de modelar de forma mais abrangente as relações entre características genotípicas e ambientais, buscando as que mais influenciam na predição precoce de produtividade volumétrica.

Agradecimentos

Gostaríamos de agradecer CAPES, CNPq, Fapemig (Projeto #APQ-02062-21) e SIF (Sociedade de Investigações Florestais) pelo financiamento a este trabalho.

Referências

Assis, T. F., Warburton, P., and Harwood, C. (2005). Artificially induced protogyny: an advance in the controlled pollination of eucalyptus. *Australian Forestry*.

Breiman, L. (2001). Random forests. *Machine Learning*.

Castro, C. A. O., dos Santos, G. A., Takahashi, E. K., Nunes, A. C. P., Souza, G. A., and de Resende, M. D. (2021). Accelerating eucalyptus breeding strategies through top grafting applied to young seedlings. *Industrial Crops and Products*.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*.

Chen, W., Zou, Y., Dang, Y., and Sakai, T. (2022). Spatial distribution and dynamic change monitoring of eucalyptus plantations in china during 1994–2013. *Trees*.

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*.

Corrêa, T. R., de Toledo Picoli, E. A., de Souza, G. A., Conde, S. A., Silva, N. M., Lopes-Mattos, K. L. B., ..., and Oda, S. (2017). Phenotypic markers in early selection for tolerance to dieback in eucalyptus. *Industrial Crops and Products*.

da Silva Tavares Júnior, I., Torres, C. M. M. E., Leite, H. G., de Castro, N. L. M., Soares, C. P. B., Castro, R. V. O., and Farias, A. A. (2020). Machine learning: Modeling increment in diameter of individual trees on atlantic forest fragments. *Ecological Indicators*.

IBÁ (2021). *Relatório Anual IBÁ - Indústria Brasileira de Árvores, 2021*.

Li, Y., Wang, R., Shi, W., Yu, Q., Li, X., and Chen, X. (2022). Research on accurate estimation method of eucalyptus biomass based on airborne lidar data and aerial images. *Sustainability*.

Moraes, C. B. D., de Freitas, M., Casella, T., Pieroni, G. B., Vilela de Resende, M. D., Zimbacks, L., and Mori, E. S. (2014). Estimativas de parâmetros genéticos para seleção precoce de clones de eucalyptus para região com ocorrência de geadas. *Scientia Forestalis*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

Schumacher, F. X. and Hall, F. S. (1933). Logarithmic expression of timber-tree volume. *Journal of Agricultural Research*, 47(9):719–734.

Silva, J. C. F., Teixeira, R. M., Silva, F. F., Brommonschenkel, S. H., and Fontes, E. P. (2019). Machine learning approaches and their current application in plant molecular biology: A systematic review. *Plant Science*.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*.

Taud, H. and Mas, J. (2018). *Multilayer Perceptron (MLP)*.

Zaiton, S., Sheriza, M. R., Ainishifaa, R., Alfred, K., and Norfaryanti, K. (2020). Eucalyptus in malaysia: Review on environmental impacts. *Journal of Landscape Ecology*.