Identificação de Espécies de Eucalipto Resistentes à Seca utilizando Aprendizado de Máquina

João Arthur Gonçalves Do Vale[1], **Adilson Rosa Lopes**[1]
Patrick Oliveira Corrêa de Araújo[1], **José Augusto Miranda Nacif**[1]

[1]Instituto de Ciências Exatas e Tecnológicas (IEF) – Universidade Federal de Viçosa (UFV)
CEP 35690-000 – Florestal – MG – Brasil

{joao.vale,jnacif,adillopes,patrick.araujo}@ufv.br

Abstract. *The identification of drought-resistant eucalyptus species can bring numerous advantages to the world industry. Currently, with high climatic variations occurring globally, with emphasis on prolonged dry seasons, thousands of hectares of eucalyptus have had their production impaired. As this is a vital raw material for the industry, its loss causes damage in several segments. This work aims to use two machine learning methods: Decision Tree and Random Forest, with the purpose of identifying, based on data collected in the field, which are the most relevant variables to classify a eucalyptus species as resistant. or not to dry. In the end, it will be possible to observe that the highest accuracy of each of the models mentioned were respectively: 70.2% and 63.1%, and the most important bioidentifiers, for classification, in order of relevance were: leaf water potential, specific leaf area and leaf length.*

Resumo. *A identificação de espécies de eucalipto resistentes à seca pode trazer inúmeras vantagens para a indústria mundial. Atualmente, com as altas variações climáticas ocorrendo no âmbito global, com ênfase nas prolongadas temporadas de seca, milhares de hectares de eucalipto têm tido sua produção prejudicada. Sendo essa uma matéria prima vital para a indústria, a sua perda causa prejuízo em diversos segmentos. Este trabalho tem como objetivo a utilização de dois métodos de aprendizagem de máquina: Árvore de Decisão e Floresta Aleatória, com o propósito de identificar, com base em dados coletados no campo, quais são as variáveis mais relevantes para classificar uma espécie de eucalipto em resistente ou não à seca. Ao final será possível observar que a maior acurácia de cada um dos modelos citados foram respectivamente: 70.2% e 63.1%, sendo que os bioidentificadores mais importantes, para a classificação, em ordem de relevância foram: potencial hídrico foliar, área foliar específica e comprimento foliar.*

1. Introdução

O eucalipto é uma cultura de forte importância para o Brasil, uma vez que possui grande impacto na economia do país. Além de possuir participação significativa no PIB (produto interno bruto), gera também empregos de forma direta e indireta. No ano de 2019, segundo [IBGE 2019], o Brasil totalizou 10 milhões de hectares de áreas de florestas plantadas. Desse total, 7.63 milhões de hectares foram de eucalipto. Este está presente principalmente nos estados de Minas Gerais, Mato Grosso, São Paulo e Bahia. Vale evidenciar que, as florestas de Eucalipto no Brasil estão entre as mais produtivas no mundo.

Essa elevada produtividade é resultado do esforço de pesquisas e operações entre empresas, universidades e institutos de pesquisa ao longo dos anos.

Na atualidade a silvicultura enfrenta desafios associados a alterações climáticas. Estes desafios estão ligados à redução e má distribuição das chuvas durante o ano de forma imprevista. Por exemplo, nos anos de 2012 a 2016 o Brasil enfrentou um período de estiagem prolongada com expressiva redução das médias pluviométricas, históricas em diversas partes do país, causando um forte impacto negativo no setor florestal. Com isso, foi registrado um declínio considerável na produtividade e perda de milhares de hectares de plantio. No ano de 2014, somente na região de Minas Gerais, cerca de 150 mil hectares de plantação relacionada à silvicultura foram perdidos devido ao evento de forte seca [Comércio 2019]. Como efeito, para compensar essa perda, a plantação de eucalipto vem aumentando de forma gradativa nos últimos anos.

Os eventos de estiagem acenderam um alerta nas empresas do setor, instigando novas linhas de pesquisas voltadas à hibridação de espécies e desenvolvimento de materiais genéticos que mantenham níveis elevados de produtividade, mesmo em condições de deficiência hídrica por períodos curtos ou prolongados. Uma das novas linhas de pesquisa está relacionada ao projeto "Tolerância à Seca", desenvolvido a partir de uma parceria público-privada entre a Sociedade de Investigações Florestais (SIF), a Universidade Federal de Viçosa (UFV) e 15 grandes empresas do setor florestal brasileiro [OLIVEIRA 2021]. Um dos objetivos deste projeto é identificar bioindicadores de tolerância à seca, bem como os mecanismos adotados por esses materiais com características contrastantes (tolerantes e sensíveis), através de uma abordagem morfológica, anatômica, fisiológica, metabólica nutricional e molecular.

Diante deste cenário, no presente trabalho objetiva-se estudar a aplicação de duas técnicas de aprendizado de máquina para definir um modelo que seja capaz de classificar clones de eucalipto resistentes à seca. As técnicas que serão abordadas são: Árvore de Decisão e Floresta Aleatória, por serem métodos capazes de entregar um resultado possível de ser mapeado de forma regressiva. Os bioidentificadores morfológicos e fisiológicos mensurados no projeto "Tolerância à Seca" [OLIVEIRA 2021], que foram coletados no período de 18 meses após a semeação, serão utilizados como entrada para o modelo. Ao final da execução de cada um dos métodos, pretende-se chegar à uma conclusão de quais são as variáveis que apresentam um melhor comportamento para realizar as classificações dos indivíduos, tolerantes ou sensíveis à seca. Isso se dará a partir da análise do resultado dos modelos construídos.

Portanto, neste trabalho será abordado inicialmente o referencial teórico, com intuito de contextualizar aos leitores dos conceitos tratados, tanto aqueles referentes à área de engenharia florestal, quanto os referentes à área de computação. Ademais, serão abordados os materiais e métodos utilizados para a execução do trabalho. Em sequência, serão apresentados os resultados dos modelos a nível de acurácia e relevância de variáveis. Por fim, serão tratadas as conclusões e recomendações para trabalhos futuros.

2. Referencial teórico

Esta seção está divida em três tópicos, sendo eles: melhoramento genético de plantas, árvore de decisão e floresta aleatória. Em cada um deles será abordada uma breve introdução sobre os conceitos que os representa com intuito de contextualizar os leitores

do artigo sobre o tema tratado no trabalho.

2.1. Melhoramento genético de plantas

O melhoramento genético de plantas pode ser definido como uma produção de vegetais realizada por cruzamento seletivo, hibridação ou ferramentas de biotecnologia [CropLife 2020]. Destaca-se que, a hibridação é o cruzamento de diferentes tipos de espécies, com intuito de formar um indivíduo (planta) modificado geneticamente com as características das espécies utilizadas. A partir do melhoramento genético é possível destacar não só a criação de populações mais eficientes em termos produtivos, mas também altamente capazes de adaptarem-se a diversas condições edafoclimáticas que antes seria menos favorável ao indivíduo cultivado [e Silva F.; Neves I.; Paiva V.; Santiago A.; Ribeiro D. 2005]. Além disso, pode-se destacar um grande aliado ao melhoramento genético: a identificação de bioindicadores, características da planta, que podem prever um comportamento de uma planta ao longo dos anos logo nos primeiros meses de plantação [OLIVEIRA 2021].

A seguir as figuras abaixo representam um exemplo de seleção de indivíduos a partir de sua hibridação. A Figura 1 apresenta duas espécies de eucalipto que são definidas respectivamente como espécies: A e B. A espécie "A" representa uma árvore com alta capacidade produtiva, alta qualidade de madeira, porém, possui sensibilidade à doença e à seca. Em contrapartida, a espécie "B" representa uma árvore com uma capacidade produtiva mediana, uma qualidade de madeira intermediária, porém uma alta resistência à doença e à seca. Ao realizar o cruzamento dessas duas espécies, elas passam a ser uma progênie, que poderá ser clonada e semeada em campo. O resultado do cruzamento pode ser visto na Figura 2. Vale realçar que, a árvore do meio atendeu as expectativas, pois apresenta todas as características positivas tanto da espécie "A", como da espécie "B". Desse modo, ela é selecionada e sua genética é clonada e semeada para uma produção em escala comercial, que pode ser visto na Figura 3.

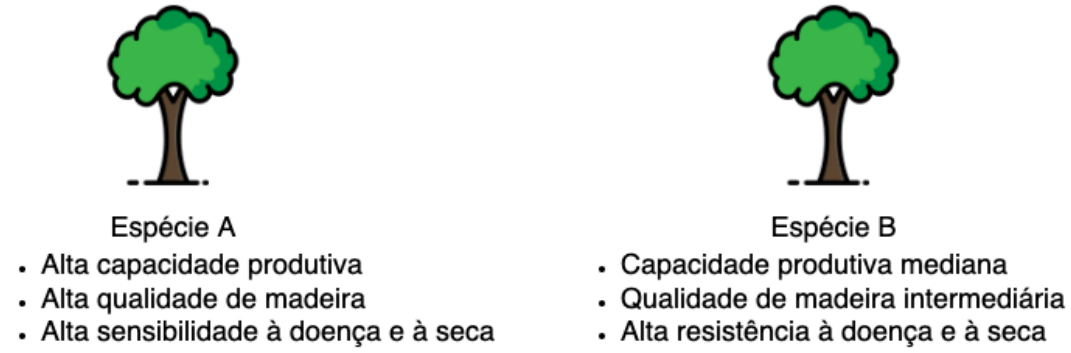

Figura 1. Espécies de eucaliptos com suas respectivas características

Figura 2. População Híbrida: Progênie [Espécie A + Espécie B]

Figura 3. Clones gerados para o plantio comercial

Cabe ressaltar que o melhoramento genético de plantas é um fator determinante para o sucesso das grandes plantações em todo o mundo. Caso não tivesse ocorrido o melhoramento genético no último século, os cultivares disponíveis suportariam somente uma população global de alguns milhões de pessoas. Portanto, a população mundial não teria jamais atingido a marca de bilhões e a teoria Malthusiana, que diz: "O crescimento populacional superará a oferta de alimentos, gerando fome e miséria no mundo todo", seria atendida [Aluízio Borém 2021].

2.2. Aprendizado de máquina

Aprendizado de Máquina (AM) é uma área de pesquisa da Inteligência Artificial que gira em torno do desenvolvimento de programas de computador (software) com a capacidade de aprender a executar uma dada tarefa com sua própria experiência [Lorena et al. 2021]. Por conseguinte, tem-se a criação de softwares capazes de aprenderem por si sós, com base em um conjunto de dados pré estabelecidos [Cerri 2017]. É possível destacar dois exemplos de utilizações do AM sendo eles: problemas que envolvem uma classificação e agrupamento de dados, e problemas que envolvem uma previsão de séries temporais [Cerri 2017]. Neste presente artigo será abordado estritamente o seu uso para problemas de classificação.

2.3. Árvore de decisão

Árvore de decisão (*Decision Tree* - DT) pode ser definida como um modelo que particiona um conjunto de dados em subconjuntos, até o momento que os conjuntos obtidos contenham apenas um tipo ou uma maior parte dele [Cerri 2017]. O elemento base da árvore

é chamado de nó-raiz, nele está contido o atributo mais relevante para a classificação do conjunto de dados, visto que, ele será responsável pelo primeiro particionamento dos mesmos. Ademais, os outros elementos da árvore de decisão são os nós-internos e nós-folhas, como pode ser visto na Figura 4. Dentro do campo do AM a árvore pode ser utilizada tanto para a classificação quanto para a regressão dos dados. Neste presente artigo o seu uso se dará com foco na criação de um modelo para a classificação.

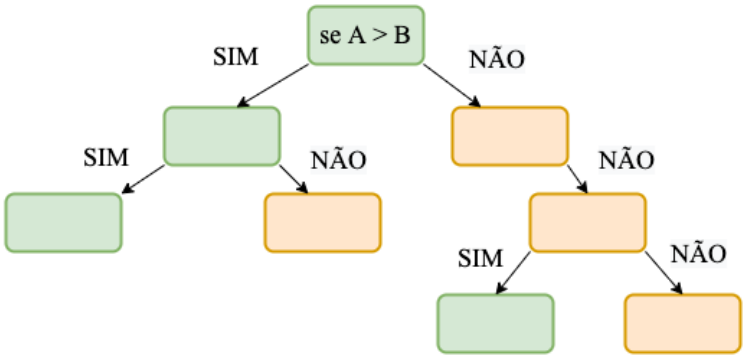

Figura 4. Estrutura árvore de decisão

2.4. Floresta Aleatória

Floresta aleatória (*Random Forest* - RF) é um método de AM que se baseia na geração de uma floresta composta por diversas árvores de decisão [Ponte et al. 2020], como pode ser visto na Figura 5. Desse modo, invés de um resultado final baseado apenas em uma única árvore, nós teremos uma saída com base em dezenas, centenas e até milhares de árvores. A partir disso as predições se tornam melhores se comparadas com aquelas feitas apenas por uma única árvore.

A RF, assim como a árvore de decisão, pode ser utilizada tanto para a classificação quanto para a regressão de dados. Sendo que, quando se trata de classificação, a predição final é decidida por voto majoritário, ou seja, é levado em consideração o resultado de cada uma das árvores e ao final será considerado aquele que aparecer em maior número. Para a regressão, a decisão final é uma média realizada entre as decisões individuais de cada árvore [Ponte et al. 2020]. Neste trabalho o seu uso se dará com foco no modelo de classificação.

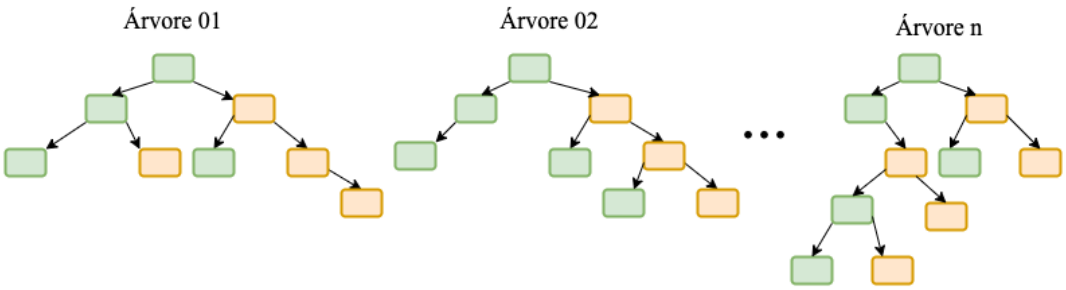

Figura 5. Estrutura floresta aleatória

3. Materiais e Métodos

Os materiais utilizados neste artigo correspondem aos dados coletados no projeto "Tolerância à seca" [Opiniões 2020]. Nesse projeto foram criados a partir da hibridação de espécies cerca de 28 progênies distintas. A partir destas foram gerados clones que resultaram em cerca de 335 indivíduos (plantas) que foram plantados em campo. O local escolhido para a plantação foi a região de Buritizeiro - MG que é conhecida por possuir uma baixa precipitação anual, logo sofrendo demasiadamente com diversos períodos de estiagem durante o ano [OLIVEIRA 2021]. Salienta-se que, os 335 indivíduos foram utilizados como entrada para os algoritmos abordados neste artigo.

Ao longo do tempo foram captadas amostras dos clones plantados em Buritizeiro, em média foram coletadas 15 folhas de cada clone. O período de captação de amostras ocorreu respectivamente em: 6 meses, 18 meses e 30 meses, nos seguintes períodos: Setembro de 2019, 2020 e 2021 [OLIVEIRA 2021], como pode ser visto na Tabela 1. Doravante, a partir das amostras foram identificados as seguintes características da planta (bioidentificadores): área foliar específica, área foliar individual, comprimento e largura foliar, como também o potencial hídrico foliar, que foram devidamente inseridos em uma base dados [OLIVEIRA 2021]. Em seguida, os dados foram analisados de forma que fez-se possível determinar correlações entre seus valores e o seus potenciais como um bioidentificador eficaz de tolerância.

A área foliar específica foi obtida a partir da mensuração das áreas de cada folha dividido por sua massa. A área foliar individual se deu através da média da área de todas as folhas coletadas. Em seguida o comprimento e largura foliar foi obtido a partir da média de ambas as medidas de cada indivíduo. Por fim, o potencial hídrico foliar foi obtido através da bomba de Scholander, que é um instrumento que pode medir o potencial aproximado de água dos tecidos vegetais, seguido de fotos de cada folha a partir de um dispositivo infravermelho. Vale ressaltar que, o atributo "indicador de vida da planta" foi utilizado com o objetivo de definir um limiar entre um indivíduo resistente ou não à seca. Este dado foi coletado no período de 30 meses, pode ser visto na Tabela 1, e identifica os indivíduos, dos meses anteriores, que sobreviveram ou não até este período.

Bioindicadores	**6 meses**	**18 meses**	**30 meses**
Área foliar específica	X	X	
Área foliar individual	X	X	
Comprimento foliar	X	X	
Largura foliar	X	X	
Potencial hídrico foliar	X	X	
Indicador de vida da planta			X

Tabela 1. Periódos de coleta das características [bioidentificadores] das plantas

4. Resultados

Nesta seção serão apresentados não só as configurações dos modelos e dados de entrada utilizados para a execução de ambos os algoritmos, como também os resultados dos mesmos.

4.1. Configuração dos modelos

4.1.1. Escolha das variáveis

Os modelos foram executados utilizando dois conjuntos distintos de variáveis. O primeiro conjunto é representado pelas seguintes variáveis: área foliar específica, área foliar individual, comprimento e largura foliar e potencial hídrico foliar. Em contrapartida, o segundo conjunto foi obtido a partir do coeficiente de correlação de Pearson (CRP), o qual foi utilizado para classificar, em ordem de relevância a partir do indicador de vida da planta, todos os atributos do modelo, como pode ser visto na Tabela 2. A partir desta classificação foram selecionadas, para a execução do modelo, as três varíaveis mais relevantes identificadas: potencial hídrico foliar, área foliar específica e comprimento foliar.

Classificação de relevância dos bioidentificadores	**Atributos**	**Valor do CRP**
1º	Potencial hídrico foliar	0.26
2º	Área foliar específica	-0.14
3º	Comprimento foliar	0.13
4º	Área foliar individual	0.11
5º	Largura foliar	0.04

Tabela 2. Relevância dos bioindicadores

4.1.2. Árvore de decisão

A partir do contexto da árvore de decisão foram utilizados os seguintes critérios apresentados na Tabela 3, que foram obtidos através de uma observação empírica.

Variáveis	**Valores**
Porcentagem de dados utilizados para o **conjunto de treinamento**	75%
Porcentagem de dados utilizados para o **conjunto de teste**	25%
Critério de classificação	Entropia
Estado randômico	123
Profundidade da árvore	16

Tabela 3. Variáveis utilizadas na árvore de decisão

4.1.3. Floresta aleatória

Dentro do contexto da floresta aleatória foi utilizado os seguintes critérios apresentados na Tabela 4, que foram obtidos através de uma observação empírica.

Variáveis	Valores
Porcentagem de dados utilizados para o **conjunto de treinamento**	75%
Porcentagem de dados utilizados para o **conjunto de teste**	25%
Critério de classificação	Entropia
Estado randômico	123
Profundidade da árvore	16
Quantidade de árvores	500

Tabela 4. Variáveis utilizadas na floresta aleatória

4.2. Execução dos algoritmos

4.2.1. Árvore de decisão

A Figura 6 apresenta uma parte do modelo final que foi gerado a partir dos bioidentificadores propostos inicialmente pelo trabalho.

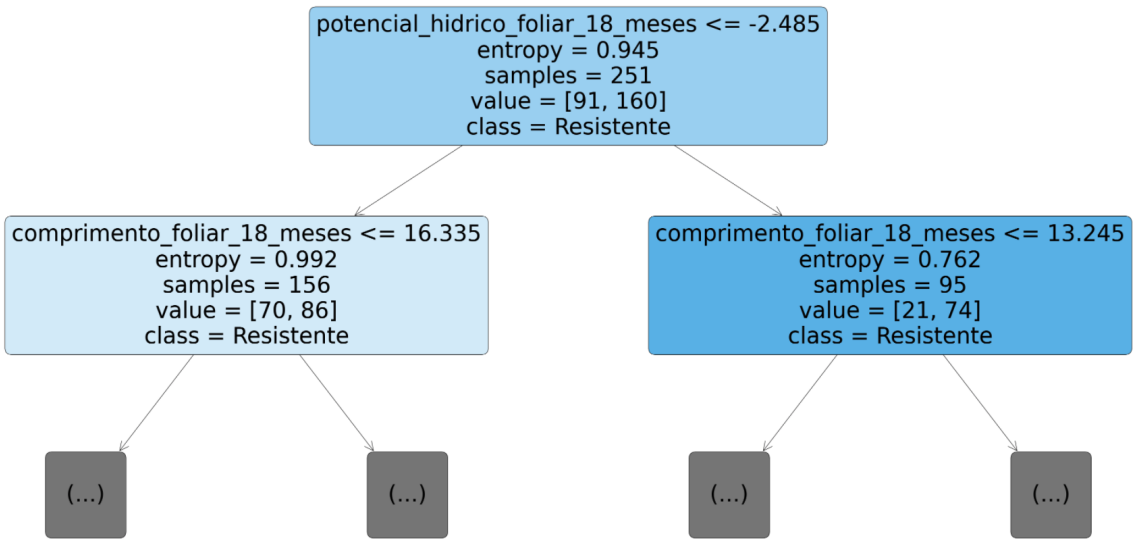

Figura 6. Esquema da árvore de decisão após o treinamento

O resultado da acurácia pode ser visto na Tabela 5, sendo que para o modelo executado com todas as variáveis foi de 61.9% e para o modelo executado com as variáveis mais relevantes foi de 70.2%, no conjunto de teste. Para encontrar a acurácia do modelo, foi utilizada a biblioteca *sklearn.metrics.accuracy_score*. Vale ressaltar que, como visto na Figura 6, o bioidentificador "potencial hídrico foliar" está no papel de raiz da árvore de decisão, sendo que este dado é relativo ao período de amostragem de 18 meses. Logo, pode-se afirmar que, para esse modelo, esta é a variável mais importante para a classificação dos clones.

Acurácia do modelo	Todas as variáveis	Variáveis mais relevantes
Conjunto de treinamento	98%	99.2%
Conjunto de teste	61.9%	70.2%

Tabela 5. Acurácia do modelo Árvore de Decisão construído

4.2.2. Floresta Aleatória

A Tabela 6 contém o resultado da média das acurácias entre todas as 500 árvores de decisão geradas pelo modelo de Floresta Aleatória. Para encontrar a acurácia do modelo, foi utilizada a biblioteca *sklearn.metrics.accuracy_score*. Vale acentuar que, para o modelo executado com todas as variáveis a acurácia foi de 60.7% e para o modelo executado com as variáveis mais relevantes foi de 63.1%, no conjunto de teste.

Acurácia do modelo	Todas as variáveis	Variáveis mais relevantes
Conjunto de treinamento	99.0%	99.2%
Conjunto de teste	60.7%	63.1%

Tabela 6. Acurácia do modelo Floresta Aleatória construído

5. Conclusões e trabalhos futuros

Os algoritmos apresentados neste trabalho demonstraram-se eficientes em determinar, para um conjunto de teste, quais indivíduos são resistentes à seca. Além disso, foi possível identificar com clareza todo o fluxo utilizado para a classificação, sendo essa uma das grandes vantagens destacadas na aplicação destes métodos. Ademais, as acurácias, com maior valor, respectivas de cada modelo para o conjunto de teste foi de 70.2% para árvore de decisão e 63.1% para floresta aleatória. Como pode ser visto o resultado da árvore de decisão foi superior ao resultado da floresta aleatória, isso ocorreu porque a floresta aleatória é recomendada para conjuntos de dados grandes e que possuem inúmeras variáveis de entrada [Ali et al. 2012], o que se difere do trabalho em questão, que possui um conjunto pequeno de dados e poucas variáveis de entrada. Por conseguinte, foi possível determinar, dentre o conjunto de dados de entrada, qual foi o nível de relevância de cada elemento através do coeficiente de correlação de pearson, cabendo ressaltar os três melhores em ordem de importância: potencial hídrico foliar, área foliar específica e comprimento foliar. Por fim, para trabalhos futuros recomenda-se utilizar estes atributos como entrada de novos modelos utilizando outros algoritmos de aprendizagem de máquina.

Referências

Ali, J., Khan, R., Ahmad, N., and Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5):272.

Aluízio Borém, Glauco V. Miranda, R. F.-N. (2021). *Melhoramento de plantas*. Oficina de Textos, 8th edition.

Cerri, R. ; de Carvalho, A. C. P. L. F. (2017). Aprendizado de máquina: breve introdução e aplicações.

Comércio, D. D. (2019). Conservação para um futuro de incertezas. https://diariodocomercio.com.br/economia/conservacao-para-um-futuro-de-incertezas/. Acesso em 13 fev. de 2022.

CropLife (2020). Melhoramento genético de plantas: trabalhando para produzir mais, melhor e de forma sustentável. https://bit.ly/34V9gAZ. Acesso em 17 marc. de 2022.

e Silva F.; Neves I.; Paiva V.; Santiago A.; Ribeiro D., M. A. A. C. A. J. C. (2005). Melhoramento genético do eucalipto: que impacto na realidade?

IBGE (2019). Produção da extração vegetal e da silvicultura. https://biblioteca.ibge.gov.br/visualizacao/periodicos/74/pevs_2019_v34_informativo.pdf. Acesso em 30 jan. de 2022.

Lorena, A., Faceli, K., Almeida, T., de Carvalho, A., and Gama, J. (2021). *Inteligência Artificial: uma abordagem de Aprendizado de Máquina*. LTC Gen, 2th edition.

OLIVEIRA, F. S. (2021). Aspectos morfoanatômicos e metabólicos envolvidos na tolerância à seca em eucalipto.

Opiniões, R. (2020). Adaptabilidade de eucalyptus à seca. https://florestal.revistaopinioes.com.br/revista/detalhes/11-adaptabilidade-de-eucalyptus-seca/#:~:text=Recentemente%2C%20grandes%20varia%C3%A7%C3%B5es%20clim%C3%A1ticas%20t%C3%AAm,%C3%A1reas%20de%20cultivo%20de%20Eucalyptus. Acesso em 03 marc. de 2022.

Ponte, C., Caminha, C., and Furtado, V. (2020). Otimização de florestas aleatórias através de ponderação de folhas em árvore de regressão. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 698–708, Porto Alegre, RS, Brasil. SBC.