ALEXANDRE GOMES FERRAZ

EFICIÊNCIA PREDITIVA DE CARACTERÍSTICAS DE QUALIDADE DA MADEIRA DE EUCALYPTUS COM ABORDAGENS DE *MACHINE LEARNING* APLICADAS A DADOS NIR

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento para o título de *Doctor Scientiae.*

Orientador: Cosme Damião Cruz

VIÇOSA - MINAS GERAIS
2022

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
F381e
2022

Ferraz, Alexandre Gomes, 1993-
Eficiência preditiva de características de qualidade da madeira de *Eucalyptus* com abordagens de *Machine Learning* a dados NIR / Alexandre Gomes Ferraz. – Viçosa, MG, 2022.
1 tese eletrônica (79 f.): il. (algumas color.).

Orientador: Cosme Damião Cruz.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Biologia Geral, 2022.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2022.544
Modo de acesso: World Wide Web.

1. *Eucalyptus. 2.* Melhoramento genético. 3. Tecnologia da madeira. 4. Aprendizado do computador. 5. Madeira - Qualidade. I. Cruz, Cosme Damião, 1958-. II. Universidade Federal de Viçosa. Departamento de Biologia Geral. Programa de Pós-Graduação em Genética e Melhoramento. III. Título.

CDD 22. ed. 634.973766

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

ALEXANDRE GOMES FERRAZ

EFICIÊNCIA PREDITIVA DE CARACTERÍSTICAS DE QUALIDADE DA MADEIRA DE EUCALYPTUS COM ABORDAGENS DE *MACHINE LEARNING* APLICADAS A DADOS NIR

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento para o título de *Doctor Scientiae.*

APROVADA: 04 de julho de 2022.

Assentimento:

Alexandre Gomes Ferraz
Autor

Cosme Damião Cruz
Orientador

A Deus,

Luz em meu caminho.

OFEREÇO

Aos meus avós, Faustino de Souza Ferraz (in memoriam) e Terezinha Vieira de Souza Ferraz (in memoriam), pelo incondicional apoio durante toda a minha vida.

DEDICO

AGRADECIMENTOS

Primeiramente, agradeço a Deus, pois sem ele não teria enfrentado todos os desafios para chegar até aqui.

À minha irmã Mariana, meu padrinho Balú, à minha prima-irmã Diana e meu padrinho Juninho e aos agregados por estarem perto sendo meu porto seguro durante toda minha jornada.

Aos meus pais, Eliana e Ulisses, e a todos familiares que torceram e torcem pelo meu sucesso. Às minhas irmãs Luísa e Laís pelo companheirismo e amizade.

À Universidade Federal de Viçosa pela oportunidade de cursar a graduação, mestrado e doutorado e aos órgãos de fomento CNPq, FAPEMIG E CAPES, pelo apoio financeiro durante a realização destes trabalhos ao longo da minha vida acadêmica.

Ao Professor Cosme Damião Cruz pela orientação, pela amizade e pelo profissional dedicado e humilde.

Ao BIOAGRO e a todos os funcionários que me ajudaram de alguma forma.

À empresa Bracell e todos os colaboradores do time da P&D que vem se desenvolvendo juntos e em um ambiente muito prazeroso. Em especial, ao time de melhoramento, que me apoiam e dão suporte para levar a pesquisa ao patamar mais alto de tecnologia.

Ao Professor Glêison e ao grupo GenMFlor por todo o conhecimento técnico e viagens.

Aos amigos que fiz durante todo este período, em especial aos amigos do Laboratório de Bioinformática e LICAE.

Aos meus grandes amigos Daniel Vasconcelos, Paula Izidoro, Evelyn Chaves, Bárbara Cadim, Igor Lorito, Filipe Manoel, Rodrigo Vieira, Luana Cardoso, Guilherme Mendes, Carla Castro, Michelle Brandão, Caio Varonill, Thales Martins, Samuel Dutra, Jéssica Valente, Vinicius Lima, Romulo Castro, Marco Antônio Nakata e Matheus Perdigão.

Aos amigos do intercâmbio em Toronto, no Canadá. Muito obrigado pelos melhores momentos da graduação.

A todos aqueles que colaboraram de alguma forma com incentivo, compreensão e amizade.

MUITO OBRIGADO!

De nada lhe servirá a força

se não existir a perseverança e o

espírito de superação.

Nunca desista!

RESUMO

FERRAZ, Alexandre Gomes, D.Sc., Universidade Federal de Viçosa, julho de 2022. **Eficiência preditiva de características de qualidade da madeira de Eucalyptus com abordagens de *Machine Learning* aplicadas a dados NIR.** Orientador: Cosme Damião Cruz.

A qualidade da madeira é uma das características decisivas na recomendação de clones nos programas de melhoramento de eucaliptos. Essa informação mensurada de forma acurada e precoce, auxilia nas decisões do melhorista e aumenta as chances de obter clones superiores. A mensuração dessa característica no gênero *Eucalyptus* é laboriosa, requer vários dias para determinação em laboratório, é um processo caro, aplicado em um número restrito de indivíduos e, muitas das vezes, demanda a perda total dos indivíduos amostrados. Para contornar essas dificuldades, a técnica de espectroscopia no infravermelho próximo tem sido uma alternativa que possibilita a predição dessas características da associação de comprimentos de ondas e as características avaliadas em laboratório. O principal método usado para predição é o dos mínimos quadrados parciais ou PLS (*Partial Least Squares*) que, apesar de eficiente para algumas características, ainda se mostra limitante no que se refere à acurácia preditiva, sendo necessário testar novas metodologias de predição. Além disso, os métodos de pré-tratamentos usados para limpeza dos dados espectrais são poucos difundidos, gerando muitas dúvidas de qual é o melhor a se usar. O objetivo desse trabalho foi avaliar a importância do uso de pré-tratamento na predição de características tecnológicas da madeira e testar metodologias de *Machine Learning* na predição dessas características, com base em informações de NIR (Near Infrared Spectroscopy), para fins de seleção indireta de indivíduos de Eucalyptus. O material para as análises foi composto por 75 indivíduos de *E. benthamii* 3 de *E. saligna*, 3 de *E. grandis* e 3 de *E. dunnii,* totalizando 87 indivíduos escolhidos por critérios industriais. Para avaliação dos pré-tratamentos e modelos de predição, foram usadas 11 características obtidas por análises laboratoriais. Para a avaliação da importância dos pré-tratamento, usou-se 15 métodos de pré-tratamentos, sendo que alguns apresentam parâmetros variáveis, totalizando 199 possibilidades. Para os modelos de predição, foram testadas 4 metodologias de Machine Learning (Árvores de decisão, Floresta Aleatória, bagging e boosting) e comparadas ao PLS em diferentes cenários (no mesmo background genético, usando diferentes background com os dados pré-

tratados e sem pré-tratamento). A avaliação de pré-tratamento para fins de ajustes de modelos para predição foi indispensável via PLS. Diferentes técnicas de pré-tratamentos se mostram eficientes, considerando informações de diferentes características na população de *E. benthamii*, sendo recomendável estudos prévios para adequação do melhor pré-tratamento. O uso do pré-tratamento envolvendo técnicas de segunda derivada com gap se destacou no conjunto de dados analisados e deve ser enfatizado como uma alternativa vantajosa em estudos de ajuste de modelo. No estudo de predição, conclui-se que diferentes características se identificam com diferentes abordagens e que o procedimento PLS é uma opção de análise a ser considerada, mas seu generalizado não é recomendado, sendo que outras opções podem apresentar resultados comparativamente superiores. O background considerado nos conjuntos de dados de treinamento e validação influenciam os resultados. Validar conjuntos de mesmo background conduz a resultados de eficiência de predição mais elevados.

Palavras-chave: *Eucalyptus*. Aprendizado de Máquina. Qualidade da Madeira. Melhoramento Genético.

ABSTRACT

FERRAZ, Alexandre Gomes, D.Sc., Universidade Federal de Viçosa, July, 2022. **Predictive efficiency of *Eucalyptus* wood quality characteristics with *Machine Leaning* approaches applied to NIR data.** Adviser: Cosme Damião Cruz.

The wood quality is one of the decisive traits in a clonal recommendation in *Eucalyptus* breeding programs. This information, measured accurately and early, helps in the breeder's decisions and increases the chances of obtaining superior clones. The measurement of this trait in the *Eucalyptus* genus is laborious, requires several days to be determined in the laboratory, is an expensive process, applied to a limited number of individuals and, often, requires the total loss of the sampled individuals. To overcome these difficulties, the technique of near-infrared spectroscopy has been an alternative that allows the prediction of these traits of the association of wavelengths and the traits evaluated in the laboratory. The main method used for prediction is the partial least squares or PLS which, despite being efficient for some traits, is still limiting in terms of predictive accuracy, making it necessary to test new prediction methodologies. In addition, the pre-treatment methods used to clean spectral data are not widespread, generating many doubts as to which is the best to use. The objective of this work was to evaluate the importance of the use of pre-treatment in the prediction of technological traits of wood and to test *Machine Learning* methodologies in the prediction of these traits, based on information from NIR (Near Infrared Spectroscopy), for the purpose of indirect selection of Eucalyptus individuals. The material for the analysis consisted of 75 individuals of *E. benthamii* 3 of *E. saligna*, 3 of *E. grandis* and 3 of *E. dunnii*, totaling 87 individuals chosen by industrial criteria. For the evaluation of pre-treatments and prediction models, 11 traits obtained by laboratory analysis were used. To assess the importance of pre-treatment, 15 pre-treatment methods were used, some of which have variable parameters, totaling 199 possibilities. For the prediction models, 4 Machine Learning methodologies were tested (Decision Trees, Random Forest, bagging and boosting) and compared to PLS in different scenarios (in the same genetic background, using different background with pre-treated and non-pre-treated data). The pre-treatment evaluation for the purpose of model adjustments for prediction was indispensable via PLS. Different pre-treatment techniques are shown to be efficient, considering information on different traits in the *E. benthamii* population, and prior studies are recommended for the adequacy of the best pre-

treatment. The use of pre-treatment involving gapped second derivative techniques stood out in the analyzed dataset and should be emphasized as an advantageous alternative in model fit studies. In the prediction study, it is concluded that different characteristics are identified with different approaches and that the PLS procedure is an analysis option to be considered, but its generalization is not recommended, and other options may present comparatively superior results. The background considered in the training and validation datasets influences the results. Validating sets from the same background leads to higher prediction efficiency results.

Keywords: *Eucalyptus*. Machine Learning. Wood Technology. Genetic Breeding.

SUMÁRIO

1. INTRODUÇÃO GERAL

O Brasil é o maior exportador de celulose do mundo, sendo o *Eucalyptus* a cultura usada como principal matéria prima com 6.69 milhões de hectares de área plantadas (IBÁ, 2020). Apesar da grande extensão territorial do Brasil e variações climáticas, as espécies desse gênero apresentam grande adaptação às condições ambientais do país. Este fato desafia os melhoristas que precisam selecionar o melhor material genético para diferentes localidades ou regiões.

O melhoramento de *Eucalyptus* tem foco em dois grupos de características. O primeiro, são as características de crescimento da madeira representadas pelo diâmetro na altura do peito, altura total, volume e incremento médio anual. Já o segundo grupo é composto pelas características tecnológicas da madeira que podem ser químicas, físicas, mecânicas e anatômicas. Durante anos, o melhoramento florestal tem dado mais importância às características do primeiro grupo, entretanto, esse cenário vem mudando e as do segundo grupo têm sido o principal alvo de estudos, com foco em uma avaliação precoce. A seleção precoce em características de qualidade da madeira é muito vantajosa para o melhorista florestal, visto que o ciclo de melhoramento para espécies florestais é em torno de 18 anos, se considerarmos idade de corte de 6 anos para as 3 etapas do ciclo de seleção (teste de progênie, teste clonal e teste clonal ampliado) e essas são avaliadas apenas no final de cada ciclo, pois nas etapas iniciais não se consegue avaliar todas as características, uma vez que são indivíduos únicos.

A importância das características tecnológicas da madeira deve-se principalmente ao grande impacto dessas no produto final, principalmente focando a indústria de Celulose. No entanto, alguns desafios podem ser listados quando se refere à obtenção das características tecnológicas da madeira no gênero Eucalyptus, pois é uma avaliação laboriosa e requer vários dias para ser determinada em laboratório. Além disso, é um processo caro e aplicado em um número restrito de indivíduos e, muitas das vezes, demanda a perda total do indivíduo amostrado, o que o caracteriza como método destrutivo (Nunes, 2015). Ademais, em algumas situações, para se ter uma caraterização mais próxima da que se obtém na indústria, se faz necessário o transporte da árvore toda, e, ainda no total, são avaliadas três árvores para cada clone. Deve-se ter em mente que os métodos destrutivos necessitam que os indivíduos estejam em idade de corte (6 anos) para que se

obtenham informações das características de qualidade da madeira, o que impossibilita a seleção precoce para essas características.

Assim, para obtenção dessas características, métodos rápidos e que contornem esses problemas, sem redução da acurácia na seleção, são fundamentais para o sucesso do programa de melhoramento florestal. Neste contexto, técnicas como a espectroscopia de infravermelho próximo (Near Infrared Spectroscopy – NIR's) tem grande potencial de utilização no melhoramento florestal, pois é um método que contorna as barreiras dos métodos em uso, principalmente quando o enfoque é predizer a qualidade da madeira. Outras áreas de estudos já têm aplicado essa técnica, como, por exemplo, em química de alimentos (Núñez-Sánchez et al 2015), alimentação animal (Decruyenaere et al 2015) e solos (Chodak, 2011).

A NIRS é uma técnica não destrutiva, rápida na mensuração das amostras que pode ser aplicada a qualquer tipo de material biológico, demandando pouca ou nenhuma preparação das amostras (Pasquini, 2003). Essa tecnologia é baseada em espectroscopia vibracional que monitora mudanças nas vibrações moleculares intimamente associadas às mudanças na estrutura molecular. O NIRS tem vantagem substancial sobre outros indicadores, pois os espectros contêm informações sobre todos os constituintes químicos do material orgânico, principalmente CH, OH e NH, que representam a espinha dorsal de todos os compostos biológicos (Baillères et al. 2002). Essa tecnologia também está sendo usada de forma crescente na área de florestais e da madeira para previsão rápida do rendimento da polpa e de outras características de polpação (Downes et al 1997; Hoffmeyer et al 1995; Raymond et al 1993; Schimleck et al 1998; Schimleck et al 1997; Wright et al 1990).

Ferramentas de análise multivariada são amplamente utilizadas na análise e modelagem de dados espectroscópicos (Jin and Xu, 2011; Smith-Moritz et al, 2011; Xu et al., 2013; Lupoi et al., 2014), sendo as principais: a Análise Componente Principal (PCA) e Mínimos Quadrados Parciais (PLS). Entretanto, para determinadas características, essas técnicas apresentam baixa acurácia, principalmente por serem utilizados modelos estatísticos lineares. Portanto, há a necessidade de avaliar a aplicação de novas técnicas não lineares, como técnicas de Machine Learning, a fim de testar técnicas para esses tipos de análises.

As técnicas de *Machine Learning*, as Árvores de Decisão e seus refinamentos, não necessitam de pressuposições sobre o modelo (Sousa et al., 2021). Além do mais, tais metodologias apresentam boa performance preditiva (James et al., 2013),

permitindo a não-linearidade dos dados e também são de fácil interpretação (Prasad et al., 2006) por fornecerem as informações sobre quais atributos são mais importantes para previsão ou classificação (Beiki et al., 2012; Ebrahimi et al., 2011; Hosseinzadeh et al., 2012).

Assim, objetivo desse trabalho foi avaliar a importância do uso de pré-tratamento na predição de características tecnológicas da madeira e predição de características de qualidade da madeira, com base em informações de NIR, por técnicas de *Machine Learning* para fins de seleção indireta de indivíduos de Eucalyptus. Além de avaliar a eficiência de técnicas de *Machine Learning* para realizar predição da qualidade da madeira genótipos de Eucalyptus.

2. REVISÃO DE LITERATURA

2.1. Qualidade da madeira para a produção de celulose

O termo qualidade da madeira expressa o conjunto de características físicas, químicas e anatômicas que uma árvore ou parte possui, e essa qualidade é alterada de acordo com o objetivo final do uso da madeira (Mitchell, 1961).

A produção de celulose de forma simplificada consiste na degradação e remoção da lignina da madeira que une as fibras, possibilitando a separação e individualização delas. Segundo Gomide (1997), para que a produção de celulose seja de alta qualidade, é necessário que a madeira tenha características físicas, químicas, anatômicas e mecânicas com especificações de acordo com o produto final. Além da genética, as práticas silviculturais, manejo dos povoamentos florestais e a espécie são fatores determinantes para características da madeira.

Os programas de melhoramento para a indústria de celulose e papel visam principalmente o aumento no incremento volumétrico, genótipos tolerantes a estresses abióticos e bióticos, e características relacionadas com a qualidade da madeira. Sendo que, a determinação desta última, consiste em grande desafio devido ao fato de que para se obter informações desta característica, é necessária a destruição completa do indivíduo, por meio do cozimento. Outro desafio é a necessidade de a planta estar na idade de corte para determinação das propriedades da madeira, o que ocasiona em maior tempo para a seleção de genótipos superiores (Silva Junior et al., 1997).

De acordo com Foelkel (1997), a densidade básica, o teor de lignina total e insolúvel em ácido, o teor de cinzas e o teor de extrativos são alguns parâmetros importantes relacionados à madeira e necessitam ser analisados e incluídos em programas de melhoramento, principalmente quando o foco são características de qualidades da madeira para celulose.

O teor de cinzas é a fração que permanece como resíduo após a combustão do carvão vegetal. Esta, varia de 0,5% a mais de 5%, dependendo da espécie, da quantidade de casca e da presença de terra e areia na madeira. Teores muito elevados exigem limpezas da caldeira mais frequentes e podem provocar corrosão em equipamentos metálicos. O carvão vegetal de boa qualidade deve ter um teor de cinzas inferior a 3% (Klock et al., 2005).

O teor de extrativos é um dos constituintes menores da madeira, ele é encontrado fora da parede celular, dependente de fatores como espécie, idade e local de ocorrência. Esses constituintes são responsáveis por características de cor, cheiro, resistência ao apodrecimento, sabor e propriedades abrasivas da madeira (D'Almeida, 1998). O teor de extrativos é importante para o desempenho da polpação, já que a presença de compostos fenólicos tende a aumentar o consumo de reagentes químicos durante o cozimento e a reduzir o rendimento (Hilli and Brow, 1978). Os tipos de extrativos podem ser: extrativos em água, extrativos em acetona e extrativos totais.

A lignina é uma substância abundante e importante para a madeira, sendo um agente permanente de ligação entre células, formando uma estrutura resistente ao impacto, compressão e dobra (Klock et al., 2005). Ela é a terceira substância macromolecular componente da madeira, sendo suas moléculas completamente constituídas por diferentes polissacarídeos, formadas por um sistema aromático composto de unidades de fenilpropano. Há maior teor de lignina em coníferas do que em folhosas, e existem algumas diferenças estruturais entre a lignina encontrada nas coníferas e nas folhosas. Na utilização da madeira, na, maioria das situações, a lignina é parte do produto, no entanto, para a polpação, esta deve ser parcialmente ou totalmente retirada (Klock et al., 2005). Quando relacionamos teor de lignina e a densidade básica, observamos uma correlação positiva entre estes (Dias & Junior, 1985; Foelkel, 1978; Barrichelo et al., 1984). Os tipos mais importantes para a polpa de celulose são: lignina klason e lignina total.

O número kappa é definido como o número de mililitros de solução de permanganato de potássio 0,1 N consumido por um grama de pasta de celulose

absolutamente seca sob condições específicas, e corrigido para um consumo relativo de 50% de permanganato. O aumento do número kappa de polpas não-branqueadas pode aumentar o rendimento e o consumo específico de reagentes químicos no branqueamento e aumenta à medida que se reduz o número kappa (Leite & Kivialho, 1988).

A densidade básica é uma das características mais importante a ser considerada, pois está relacionada a alguns aspectos tecnológicos e econômicos muito importantes. Características da madeira como, densidade e rendimento de celulose, são importantes para ser inseridas em programas de melhoramento com foco na produção de celulose e papel (Silva et al., 2009; Milagres et al., 2013; Hamilton et al.,2017).

O crescimento, a densidade básica e o rendimento de celulose são características importantes para serem inseridas em programas de melhoramento com foco na produção de celulose e papel (Silva et al., 2009; Hamilton et al., 2017). Essas características são importantes, pois as indústrias concentram-se em resultados de toneladas de celulose produzidas por hectare por ano, e, esse valor, é obtido mediante o produto do incremento médio anual de volume da madeira, da densidade básica e do rendimento de celulose (Raymond et al., 2010; Brawner et al., 2012).

Em populações de programas de melhoramento, a avaliação da qualidade da madeira é um desafio, pois nem sempre é possível derrubar e abater uma árvore superior, devido ao seu valor genético no programa de melhoramento. Assim, métodos de avaliação não-destrutivos são necessários para possibilitar a coleta de uma pequena amostra de madeira sem comprometer a sobrevivência do indivíduo. Outro fator importante, é a idade de avaliação da madeira, visto que as propriedades da madeira só podem ser determinadas após a árvore atingir a idade de corte (Pasquini et al., 2007), o que dificulta o tempo de seleção e o avanço da geração. Portanto, a idade mínima ótima de avaliação que permite uma boa correlação com a qualidade da madeira com 7 anos de idade deve ser determinada para permitir um melhoramento precoce (Schimleck et al., 2005). Diante desses desafios, a técnica de espectroscopia no infravermelho próximo tem sido avaliada para com grande potencial para auxiliar os programas de melhoramento na busca de materiais genéticos superiores em qualidade da madeira.

2.2. Near Infrared Reflectance (NIR)

A descoberta da região do infravermelho próximo foi em 1800, pelo cientista inglês Frederick William Herschel. Os espectros dessa região têm um alto potencial quando aplicados em análises químicas e\ou físicas, pois armazenam muitas informações sobre as amostras avaliadas (Pasquini, 2003).

Na década de 30, já haviam relatos da utilização desta técnica, mas foi na década de 60 que o grupo de pesquisadores de Karl Norris iniciaram as pesquisas com a aplicação da espectroscopia no infravermelho próximo (NIRS) em novo método de determinação da umidade em produtos agrícolas (Bokobza, 1998). Nas décadas de 80 e 90, devido aos avanços de algumas ciências importantes para a técnica NIRS, como a Quimiometria e o desenvolvimento dos microcomputadores, foram possibilitadas forte impulsão no uso desta técnica a partir de métodos matemáticos, estatísticos e informáticos de forma a adquirir informações relevantes dos dados químicos medidos para representar e interpretar essas informações dos espectros (Wold & Sjöström, 1998).

A espectroscopia no NIR tem fornecido resultados eficientes para a determinação de moléculas orgânicas e variáveis quantitativas (Muñiz et al 2012). Devido ao nível que o desenvolvimento tecnológico tem atingido nos últimos anos, a espectroscopia NIR está presente em praticamente todas as áreas, como: áreas agrícolas, alimentícia, médica, petroquímica, farmacêutica e florestal. Sendo o setor florestal o foco deste trabalho.

2.3. Near infrared spectroscopy no setor florestal

A espectroscopia NIR é uma potencial tecnologia que pode fornecer um grande conjunto de informações a respeito das propriedades da madeira de forma a auxiliar na compreensão de como a genética e os fatores do ambiente influenciam na madeira (So et al., 2004). Essa tecnologia, nas últimas décadas, tem sido aceita como uma ferramenta na modelagem local e global para a predição de muitas características de madeira maciça e madeira dura (Meder et al., 2011). Pelo fato da possibilidade rápida de estimação por um simples espectro de NIR, os modelos gerados pelo aparelho é um grande atrativo para os programas de genética e melhoramento (Santos et al., 2015). Na área florestal, a técnica é utilizada na previsão e classificação das

propriedades da madeira (Schimleck et al., 2007), na calibração de modelos para estimar o número Kappa de diferentes tipos de polpas de processos Kraft (Birkett & Gambino 1989), na caracterização da lignina em madeira (Kihara et al. 2002), na predição de propriedades mecânicas da madeira (Schimleck et al. (2001a), Thumm & Meder (2001), Meder et al. (2002), Gindl et al. (2002), Hauksson et al. (2001) e Hoffmeyer & Pedersen (1995)), na avaliação das propriedades energéticas de resíduos de madeiras tropicais (Silva et al., 2014), entre outras. Com o auxílio de modelos estocásticos, as informações obtidas por meio da espectroscopia NIR têm sido utilizadas para classificar madeiras e outros materiais biológicos conforme diferentes critérios (Tsuchikawa & Kobori, 2015).

Em resumo, essa técnica se baseia no desenvolvimento de uma calibração que relaciona os espectros gerados pelo NIR com uma grande quantidade de amostras de madeiras com a sua constituição química já conhecida, por exemplo, rendimento de polpa ou teor de celulose (Raymond, 2001). Esta calibração é então realizada para predizer rendimento de polpa ou teor de celulose apenas com o uso das informações obtidas pelo NIR, sem a necessidade de destruição dos indivíduos por completo (Raymond, 2001). Além dessas, as propriedades químicas, físicas, mecânicas também podem ser preditas usando espectroscopia NIR (Cogdill et al., 2004; Kelley et al., 2004).

2.4. Técnica de espectroscopia no infravermelho próximo

O espectro é a informação que é obtida por meio de instrumentos de espectroscopia. Sendo que, para a obtenção do espectro no infravermelho, é necessário que comprimentos de onda na região do infravermelho incidam sobre os átomos de algumas moléculas. Com isto, parte da radiação é absorvida e outra é transmitida. A leitura é feita na radiação transmitida gerando os espectros (Ferreira, 2018). A região de radiação que permite a obtenção dos espectros no NIR varia de 4000 a 10000 cm^{-1} em comprimento de onda (Bokobza, 1998).

Após obtenção dos dados do NIR estes são organizados em uma matriz X em que as linhas se correspondem a cada amostra (espectro) e as colunas aos números de ondas (variáveis), que podem ser visualizados na figura 1. É possível observar que

apenas uma amostra pode gerar muitas variáveis, sendo, em geral, o número de comprimentos de ondas lidos no NIR muito superior ao de amostras coletadas.

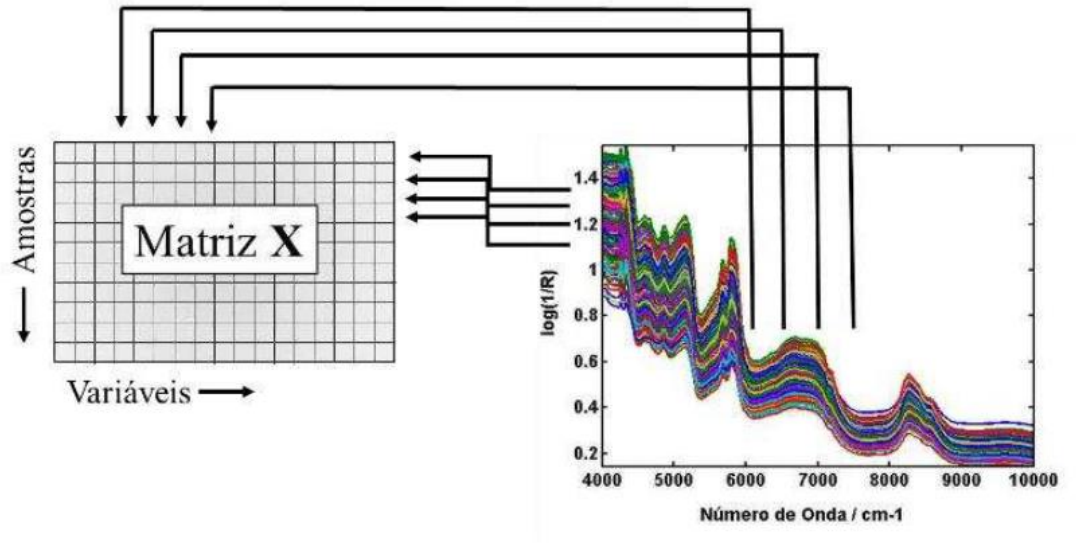

Figura 1- Organização dos dados NIR, FONTE: (Roque, 2015).

Na figura 1 podemos observar a organização dos dados NIR, sendo as amostras (indivíduos avaliados) contidas no eixo **Y**, valores estes obtidos por algum método de referência e as variáveis (comprimento de onda em cada banda espectral) contidas no eixo X. Os dados provenientes do NIR são altamente correlacionados, originando problemas de multicolinearidade (Freund et al., 2006). Assim, métodos como os mínimos quadrados podem não ser os mais indicados nessa situação, sendo necessário empregar métodos específicos de redução de dimensionalidade para o modelo ajustado (Teófilo, 2007).

Para análise e modelagem de dados de espectroscopia NIR, deve-se realizar algumas fases como: pré-tratamentos e a calibração multivariada. O pré-tratamento é realizado na matriz de dados antes da obtenção do modelo estatístico de forma a minimizar os erros sistemáticos presentes nos mesmos (Xu et al., 2008, Souza & Poppi, 2012). Pode-se aplicar aos espectros diferentes pré-tratamentos: centragem na média, normalização, alisamento, derivação e correção multiplicativa de sinal (MSC) (Souza & Poppi, 2012; Ferreira, 2015). Aqueles que apresentarem o melhor ajuste são aplicados aos dados, sendo que pode ser aplicado apenas 1 ou a combinação de mais de um pré-processamento. Logo após, os dados estarão prontos

para análise. Mais detalhes e esclarecimentos sobre pré-tratamentos podem ser obtidos em Ferreira (2015).

As técnicas multivariadas para modelagem e calibração dos dados NIR mais usadas são as de componentes principais (PC) e regressão dos mínimos quadrados parciais (PLS) (Næs et al., 2002).

2.5. Métodos estatísticos utilizados em predições

Para contornar os desafios apresentados em relação à predição, vários métodos têm sido propostos, os quais se diferem pelo tipo de suposição sobre o modelo genético associado ao caráter quantitativo (Sant'anna 2018). Entre eles, os métodos estatísticos como RR-BLUP, Bayes A e Bayes B (Meuwissen et al. 2001), LASSO Bayesiano (Legarra et al. 2008, De Los Campos et al. 2009), ou baseadas em inteligência computacional como as redes neurais artificiais (RNAs), Função de Base Radial (RBF) e Perceptron Multicamadas (MPL) (Glória et al. 2016, Li et al. 2018, Sant'anna 2018, Silva 2018, Azodi et al. 2019) e as Árvores de decisão e seus refinamentos (*bagging*, *random forest* e *boosting*) (Ghafouri-Kesbi et al. 2017, Alves et al. 2020, Souza et al. 2020). As metodologias que irão ser aplicadas no presente projeto são descritas a seguir.

A depender do método estatístico usado para a predição de características de qualidade da madeira por dados de espectroscopia no infravermelho próximo, a acurácia pode aumentar ou reduzir. A escolha deste modelo está ligada às pressuposições acerca do mesmo, tais como: dimensionalidade das matrizes envolvidas, multicolinearidade entre os comprimentos de onda e a complexidade dos caracteres quantitativos em estudo (Sant'anna 2018).

2.6. Árvore de Decisão e seus refinamentos

A Árvore de Decisão (AD) é uma metodologia que particiona o espaço preditor em sub-regiões através de alguns critérios. Para cada sub-região formada é atribuído um valor que será utilizado como valor predito para os novos indivíduos que serão alocados a essas sub-regiões (James et al., 2013). O nó é dito interno quando os dados contidos neste nó são divididos de acordo com um critério de divisão, formando,

assim, dois novos grupos de dados, sendo esses novos grupos ligados ao grupo antigo pelos ramos. Já o nó é dito externo (folha) quando não ocorrem mais divisões dos indivíduos pertencentes a este nó. O aprendizado indutivo de árvores de decisão é geralmente dividido em aprendizado supervisionado e não-supervisionado, embora o aprendizado semi-supervisionado também tenha sido considerado ao longo dos últimos anos (Chapelle et al. 2006).

Árvores de Decisão e seus possíveis refinamentos (*Boosting, Bagging, Random Forest*) demandam menos recurso computacional e apresentam a importância dos marcadores (entrada) de maneira fácil e direta se comparadas às técnicas de inteligência Artificial (James et al. 2013, Souza et al. 2020). Outra diferença, é que não necessitam de pressuposições sobre o modelo (Souza et al. 2020). Além do mais, tais metodologias apresentam boa performance preditiva (James et al. 2013), permitindo a não-linearidade dos dados e são de fácil interpretação (Prasad et al. 2006), por fornecerem as informações sobre quais atributos são mais importantes para previsão ou classificação (Ebrahimi et al. 2011, Beiki et al. 2012, Hosseinzadeh et al. 2012).

2.7. Árvore de Decisão

De acordo com Tan et al. (2006), a Árvore de Decisão pode ser utilizada para os seguintes propósitos: modelagem descritiva (classificação) e modelagem preditiva (regressão). Árvore de Decisão pode ser classificada como árvore de regressão quando a variável resposta é do tipo quantitativa (contínua) ou árvore de classificação quando a variável dependente assume valores qualitativos (categóricos). A árvore de decisão tem como objetivo subdividir diversas vezes o conjunto de observações de tal forma que os subgrupos formados subsequentes sejam cada vez mais homogêneos (Sousa et al., 2020). Entretanto, existem diferentes tipos de critérios de seleção, sendo esta uma das variações entre os diversos algoritmos de indução de árvores de decisão. Esses critérios são definidos em termos da distribuição de classe dos exemplos antes e depois da divisão (Tan et al. 2006).

O valor obtido pelos nós externos nos dados de treinamento é o valor médio das suas observações. Assim, a uma nova observação de dados, atribui-se o valor médio correspondente. Para cada região formada na árvore é atribuído um valor, que será utilizado para predizer o valor da variável resposta de um novo indivíduo, sendo este valor a média de todos os indivíduos pertencentes à região utilizada na construção da respectiva árvore. A estrutura da árvore de decisão é feita pela busca da árvore que

leva a partição dos dados até a formação de grupos homogêneos. Para isso, avalia-se o quão razoável é uma dada árvore T através do erro quadrático médio, como na equação abaixo:

$$P(T) = \sum_{m=1}^{M} \sum_{k \in Rm} (y_k - \hat{y}_{R_m})^2$$

em que: $\hat{y}_{R_m}$ é a média da variável resposta das observações de treinamento pertencente a m-ésima região-- e y_k é o valor verdadeiro da característica de cada indivíduo dentro do grupo k.

Porém, o custo computacional é muito alto, sendo inviável considerar cada partição possível do espaço em M regiões para obter o menor erro quadrático médio. Para contornar o custo computacional, James et al. (2013) recomendam um procedimento baseado em divisões binárias recursivas, no qual o objetivo é obter a variável Xp e o ponto s, que divide o espaço em duas regiões, como:

$$R_1(p, s) = \{X_p \leq s\} \; e \; R_2(p, s) = \{X_p > s\}$$

em que o ponto s divide a p-ésima variável em duas regiões que obtenha o menor erro quadrático médio. Por fim, utilizamos a variável que obteve o menor erro quadrático médio para a primeira divisão, em seguida, repetimos o processo para cada região gerada.

Quando árvores de decisão são construídas, muitas das arestas ou sub-árvores podem refletir ruídos ou erros. Enquanto uma árvore muito grande pode sofrer *overfitting* (super-ajuste) dos dados, uma árvore pequena pode não capturar uma boa estrutura. Para detectar e excluir essas arestas e sub-árvores, são utilizados métodos de poda (*pruning*) da árvore, cujo objetivo é melhorar a taxa de acerto do modelo para novos exemplos, os quais não foram utilizados no conjunto de treinamento (Han 2001). Uma abordagem para a escolha do tamanho da árvore, seria construir uma árvore até que nenhuma região obtenha mais que 5 indivíduos e, em seguida, podá-la usando o custo de complexidade da poda (Hastie et al. 2009). Assim, em uma segunda etapa, é realizada a poda com o objetivo de tornar a árvore de regressão menor e menos complexa, de modo a diminuir a variância deste estimador. Nessa etapa do processo, cada nó é retirado, um por vez, observando-se como o erro de predição varia no conjunto de validação e, posteriormente, baseando-se nas observações, é decidido quais nós permanecerão na árvore (Hastie et al. 2009).

Geralmente, uma única árvore não possui boa precisão preditiva quando comparada com outras abordagens (Souza et al. 2020). Alguns refinamentos com o intuito de melhorar a performance do modelo de árvore de decisão são apresentados na literatura. O pior desempenho da Árvore de Decisão quando comparado com seus refinamentos pode ser explicado porque essa metodologia sofre alta variação em termos de previsão (James et al. 2013). Hastie et al. (2009) enfatizaram que a baixa precisão preditiva da Árvore de Decisão pode ser melhorada pelo uso de métodos de ensemble, como *Bootstrap Aggregation (bagging), random forest e boosting* (Breiman 2001). Essas estratégias combinam múltiplas Árvores de Decisão para reduzir a variabilidade (Souza et al. 2020).

2.7.1. *Bagging*

Uma das variações das árvores de decisão é o *Bagging*. Um dos problemas apresentados pela árvore de decisão é a grande variabilidade entre os resultados obtidos, ou seja, se utilizarmos uma parte de um banco de dados para construirmos uma árvore e em seguida utilizarmos a outra parte do mesmo banco de dados para construir uma segunda árvore, iremos obter duas árvores com estruturas diferentes. Para contornar esse problema, o ideal seria obter várias amostras de uma mesma população, construir várias árvores e em seguida obter a média/moda dos valores preditos.

Como não é uma tarefa fácil obter vários conjuntos de treinamento de uma população, o *Bootstrap Aggregation* (*bagging*) (Breiman 2001) é um método que aplica a técnica de *bootstrap*. O *bootstrap* consiste em obter B amostras com reposição da amostragem disponível, obtendo assim B modelos $\widehat{f^1}(x), \widehat{f^2}(x), \ldots, \widehat{f^B}(x)$ (Efron 1992). A amostragem é feita com a substituição dos dados originais e a formação de novos conjuntos de dados. Os novos conjuntos de dados podem ter uma fração das colunas e das linhas, que geralmente são hiperparâmetros em um modelo. Por fim, utilizam-se os modelos gerados para obter uma média, e, assim, diminuir a variabilidade obtida nas árvores de decisão (Breiman 2001). Essa média desses modelos irá ser o modelo final, e é dada por:

$$\widehat{f_{médio}}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f^B}(x)$$

Dessa forma, o *bagging* é uma técnica usada para reduzir a variância das previsões, que combina o resultado de vários classificadores, modelados em diferentes sub-amostras do mesmo conjunto de dados (Breiman 2001). Tomando frações de linha e coluna menores que 1 ajuda na montagem de modelos robustos, menos propensos a *overfitting*. A quantidade de árvores utilizadas no *bagging* não é um parâmetro que irá resultar num super-ajustamento do modelo. Na prática, é utilizado uma quantidade em que o erro tenha estabilizado (James et al. 2013).

2.7.2. Random Forest

Devido ao fato de sempre utilizarmos todas as variáveis em cada partição no *bagging*, as predições obtidas nas Árvores de Decisão estarão altamente correlacionadas, uma vez que as árvores criadas terão estruturas semelhantes. Além disso, está sujeito a quase sempre uma mesma variável esteja no topo da árvore (Hastie et al. 2009, James et al. 2013). A média de valores altamente correlacionados, não resulta numa grande redução da variância, como ocorre quando é feita com valores não correlacionados (James et al., 2013). Para melhorar a acurácia na classificação dos indivíduos, Ho (1995) propôs o *random forest* (RF). O *random forest* é um método de aprendizagem de máquina versátil capaz de executar tarefas de regressão e de classificação. Essa metodologia também aplica métodos de redução dimensional, trata valores faltantes, valores anómalos ('*outliers*') e outras etapas essenciais da exploração de dados. É um tipo de método de aprendizado no qual um grupo de modelos fracos é combinado para formar um modelo mais forte (Ho 1995).

O *random forest* segue a mesma ideia do *bagging*, no entanto, além do conjunto de observações, altera também o número de variáveis preditoras (m<p) utilizadas em cada partição. Dessa forma, no *random forest* obtêm-se os valores preditos mais independentes, o que gera redução da variabilidade encontrada nas árvores de decisão. Hastie et al. (2009) sugerem que o número de variáveis preditoras utilizadas em cada partição seja $m = \sqrt{p}$ para árvore de classificação e $m = \frac{p}{3}$ para árvores de regressão. Assim, as predições das árvores se tornam menos correlacionadas e, ainda, corrige o fato de que apenas uma variável esteja sempre no topo da árvore.

2.7.3. Boosting

O termo *boosting* refere-se a uma família de algoritmos que converte uma aprendizagem fraca (também conhecida como base de aprendizagem) em uma aprendizagem forte. Para converter a aprendizagem fraca em aprendizagem forte, a previsão de cada aprendizagem fraca é combinada por métodos que utilizam a média e média ponderada e/ou que consideram a previsão que apresentar mais "votos". Para encontrar uma regra fraca, aplicam-se algoritmos de base de aprendizagem com uma distribuição diferente. Cada vez que o algoritmo de base de aprendizado é aplicado, ele gera uma nova regra de previsão fraca. Este é um processo iterativo. Após muitas iterações, o algoritmo de *boosting* combina essas regras fracas em uma única regra de predição forte (Martins et al. 2009).

Ao contrário do *bagging* que cria múltiplas árvores independentes, o *boosting* cria árvores sequencialmente utilizando-se de informação prévia da árvore anterior. Ao invés de ajustar um modelo para a variável resposta Y, o *boosting* ajusta um grande número de árvores de decisão, $\widehat{f^1}(x), \widehat{f^2}(x), \ldots, \widehat{f^B}(x)$, para o resíduo atual (Freund and Schapire 1999). Nessa metodologia, a aprendizagem é lenta, necessitando, assim, que o número de modelos (B) seja grande. Entretanto, é necessário ter cuidado para criar um *overfitting* do modelo. Assim, no *boosting* é utilizada a validação cruzada para se escolher o número de árvores que será construída, isso diminuiu a possibilidade de *overfitting*, uma vez que todos os indivíduos participarão do conjunto de validação (Bengio; Grandvalet 2004).

3. REFERÊNCIAS BIBLIOGRÁFICAS

Acquaah G (2012) Priciples of Plant Genetics Breeding. **Oxford**, Blackwell, 740p.

Alkimim ER, Caixeta ET, Sousa TV, Resende MDV, da Silva FL, Sakiyama NS and Zambolim L (2020) Selective efficiency of genome-wide selection in Coffea canephora breeding. **Tree Genetics and Genomes** 16.

Alves AAC, Costa RM, Bresolin T, Fernandes Júnior GA, Espigolan R, Ribeiro AMF, Carvalheiro R and Albuquerque LG (2020) Genome-wide prediction for complex traits under the presence of dominance effects in simulated populations using GBLUP and machine learning methods. **American Society of Animal Science** 1–34.

Azodi CB, Bolger E, McCarren A, Roantree M, de los Campos G and Shiu SH (2019) Benchmarking parametric and machine learning models for genomic prediction of complex traits. **G3: Genes, Genomes, Genetics** 9: 3691–3702.

Batchelor, W.D., X.B. Yang, and A.T. Tschanz. 1997. development of a neural network for soybean rust epidemics. **Trans. ASAE** 40(1): 247.Available:http://elibrary.asabe.org/abstract.asp??JID=3&AID=21237&CID=t1997&v=40&i=1&T=1.

Beiki AH, Saboor S and Ebrahimi M (2012) A New Avenue for Classification and Prediction of Olive Cultivars Using Supervised and Unsupervised Algorithms. **PLOS ONE** 7: 1–9.

Beiki, A. H., Saboor, S., & Ebrahimi, M. (2012). A new avenue for classification and prediction of olive cultivars using supervised and unsupervised algorithms. **PLOS ONE**, [S. l.], v. 7, n. 9, p. 1–9, 2012. DOI: 10.1371/journal.pone.0044164

Bengio Y and Grandvalet Y (2004) No Unbiased Estimator of the Variance of K-Fold Cross-Validation. **Journal of Machine Learning Research** 5: 1089–1105.

Bered F, Barbosa Neto JF and Carvalho FIF (1997) Marcadores moleculares e sua aplicação no melhoramento genético de plantas. **Ciência Rural** 27: 513–520.

Bernardo R (2010) Breeding for quantitative traits in plants, 2nd ed. Stemma Press, **Woodbury**, MN., 260p.

Bernardo R (2020) Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. **Hereditary** 11.

Berry EM, Dernini S, Burlingame B, Meybeck A and Conforti P (2015) Food security and sustainability: Can one exist without the other? **Public Health Nutrition** 18: 2293–2302.

Birkett, J. A.; Gambino, M. J. T. Estimation of Pulp kappa number with near-infrared spectroscopy. Tappi Journal, v. 72, n. 9, p. 193-197, 1989. THUMM, A.; MEDER, R. Stiffness prediction of radiata pine clearwood test pieces using near infrared spectroscopy. **Journal of Near Infrared Spectroscopy**, v. 9, n. 2, p. 117-122, 2001.

Bittencourt G (2006) **Inteligência artificial: ferramentas e teorias**, 3. ed. Editora UFSC, Florianópolis, 372p.

Bokobza, L. Near Infrared Spectroscopy. **Journal of Near Infrared Spectroscopy**, v. 6, n. 1, p.3-17, 1998.

Borsellino V, Schimmenti E and El Bilali H (2020) Agri-food markets towards sustainable

patterns. **Sustainability (Switzerland)** 12.

Braga AP, Carvalho ACPL. and Ludemir TB (2007) **Redes neurais artificiais: teoria e aplicações**, 2.ed. LTC, Rio de Janeiro, 248p.

Bramardi, S.J.; Bernet, G. P.; Asíns, M. J.; Carbonell, E. A. Simultaneous agronomic and molecular characterization of genotypes via the generalized procrustes analysis: an application to cucumber. **Crop Science**, v. 45, p. 1603-1609, 2005.

Brawner, J. et al. Selection of Corymbia citriodora for pulp productivity. **Southern Forests: a Journal of Forest Science**, v. 74, n. 2, p. 121–131, 2012.

Breiman L (2001) random forest. Kluwer Academic Publishers 45: 5–32.

Bueno LCS, Mendes ANG and Carvalho SP (2006) **Melhoramento Genético de Plantas: Princípios e Procedimentos**, 2. ed. UFLA, Lavras:, 319p.

Caixeta ET, Oliveira ACB, Brito GG and Sakiyama NS (2016) Tipos de Marcadores Moleculares. Editora UFV, Viçosa, MG, p. 385. **In Borém A and Caixeta ET (eds) Marcadores Moleculares**.

Capone R, El Bilali H, Debs P, Cardone G and Driouech N (2014) Food System Sustainability and Food Security: Connecting the Dots. **Journal of Food Security** 2: 13–22.

Chan LW and Fallside F (1987) An adaptive training algorithm for back propagation networks. **Computer Speech and Language** 2: 205–218.

Chapelle O, Schölkopf B and Zien A (2006) Semi-Supervised Learning. **Massachusetts Institute of Technology Press.,** London, England, 524p.

Cook NR, Zee RYL and Ridker PM (2004) Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. **Statistics in Medicine** 23: 1439–1453.

Coppin B (2010) Inteligência Artificial. **LTC**, Rio de Janeiro, RJ, 664p.

Craven P and Wahba G (1976) Smoothing noisy data with spline function: estimating the correct degree of smoothing by the method of Generalized Cross-Validaton. **Numerische Mathematik** 31: 317–403.

Cruz CD (2012) **Princípios de genética quantitativa**, 2. ed. Editora UFV, Viçosa, MG, 394p.

Cruz CD (2016) Genes software – extended and integrated with the R, Matlab and Selegen. **Acta Scientiarum - Agronomy** 38: 547–552.

Cruz CD and Nascimento M (2018) **Inteligência Computacional aplicada ao melhoramento genético**. Editora UFV, Viçosa, MG, 414p.

Cruz CD, Carneiro PCS and Regazzi AJ (2014) **Modelos Biométricos Aplicados ao Melhoramento Genético**, 3. ed. v2. Editora UFV, Viçosa, MG, 668p.

Cruz CD, Regazzi AJ and Carneiro PCS (2012) **Modelos biométricos aplicados ao melhoramento genético**., 4. ed. v1. Editora UFV, Viçosa, 514p.

Cruz CD, Salgado CC and Bhering LL (2013) Genômica Aplicada Cruz CD, Salgado CC, and Bhering LL (eds). **Suprema**, Visconde do Rio Branco, MG, 424p.

De Los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K and Cotes JM (2009) Predicting quantitative traits with regression models for dense molecular

markers and pedigree. **Genetics** 182: 375–385.

Desta ZA and Ortiz R (2014) Genomic selection: Genome-wide prediction in plant improvement. **Trends in Plant Science** 19: 592–601.

Dias, V. R . L.; Claúdio-da-silva, JR. E. A influência da densidade básica da madeira de híbridos de Eucalyptus grandis em suas características químicas, e propriedades de polpação e do papel. In: CONGRESSO ANUAL DA ABTCP, 18, 1985, São Paulo. **Anais... São Paulo: ABTCP**, 1985. p. 31-56.

Douglas, T. S. Image processing for craniofacial landmark identification and measurement: a review of photogrammetry and cephalometry. **Computerized Medical ImagingandGraphics**, v. 28, n. 7, p. 401-409, 2004.

Downes, G.M.; Hudson, I.L.; Raymond, C.A.; Dean, G.H.; Mitchell, A.J.; Schimleck, L.R.; Evans, R.; Muneri, A. 1997: "Sampling Plantation Eucalypts for Wood and Fibre Properties". **CSIRO Publishing, Melbourne**. 132 p.

Dudley JW and Moll RH (1969) Interpretation and Use of Estimates of Heritability and Genetic Variances in Plant Breeding 1 . **Crop Science** 9: 257–262.

Ebrahimi Mansour, Lakizadeh A, Agha-Golzadeh P, Ebrahimie E and Ebrahimi Mahdi (2011) Prediction of thermostability from amino acid attributes by combination of clustering with attribute weighting: A new vista in engineering enzymes. **PLoS ONE** 6.

Ebrahimi, M., Lakizadeh, A., Agha-Golzadeh, P., Ebrahimie, E., & Ebrahimi, M. (2011). Prediction of thermostability from amino acid attributes by combination of clustering with attribute weighting: a new vista in engineering enzymes. **PloS one**, 6(8), e23146.

Efron B (1992) Bootstrap Methods: Another Look at the Jackknife. **Springer Series in Statistics (Perspectives in Statistics)**, New York, NY, p. 569–595. In Kotz S and Johnson NL (eds) Breakthroughs in Statistics Volume II Methodology and Distribution.

Entringer GC, Cesar J, Vettorazzi F and Pereira MG (2014) Correlação e análise de trilha para componentes de produção de milho superdoce. **Revista Ceres** 61: 356–361.

Falconer SD and Mackay TFC (1996) Introduction to quantitative genetics. **Edinburgh: Addison Wesley Longman**, 464p.

Fernandes AM da R (2003) Inteligência Artificial - Noções Gerais. **Visual Books**.

Ferreira RADC, Silva GN, Glória LS, Sant'anna IC, Rodrigues HS, Silva FF and Cruz CD (2018) RNA - Aplicação em Estudos de Seleção Genômica Ampla. Editora UFV, Viçosa, MG, p. 414. **In Cruz CD and Nascimento M (eds) Inteligência Computacional Aplicado ao Melhoramento Genético.**

Ferreira, M. M. C. **Quimiometria – Conceitos, Métodos e Aplicações.** Campinas, SP: Editora Unicamp, 2015. 493 f.

Ferreira, Roberta de Amorim. **Comparação de métodos de seleção de variáveis em regressão aplicados a dados genômicos e de espectroscopia NIR**. 2018.

Fisher RA (1941) Average excess and average effect of a gene substitution. **Ann Eugen 11**: 53–63.

Fonseca, S.M.; Resende, M.D.V.; Alfenas, A.C.; Guimarães, L.M.S.; Assis, T.F.; Grattapaglia, D. **Manual prática de melhoramento genético do eucalipto.** Viçosa, MG: UFV, 2010.

Freund Y and Schapire RE (1999) A brief introduction to *boosting*. **International Joint Conference on Artificial Intelligence** 2: 1401–1406.

Freund, R. J.; WILSON, W. J.; SA, P. Regression analysis – Statistical Modeling of a response variable. **Elsevier**, Inc., San Diego, 459p, 2006.

Friedman JH (1991) Multivariate Adaptative regression Splines. **The Annals of Statistics** 19: 1–141.

Fritsche-Neto R (2011) **Seleção genômica ampla e novos métodos de melhoramento do milho.** Universidade Federal de Viçosa, Viçosa, 39p.

Galvão, C. O., Valença, M. J. S., Vieira, V. P. P. B., Diniz, L. S., Lacerda, E. G. M., Carvalho, A. C. P. L. F., & Ludermir, T. B. (1999). **Sistemas inteligentes: aplicações a recursos hídricos e ciências ambientais.** Porto Alegre: Editora Universidade.

Garcia, Silvana Lages Ribeiro. Importância de características de crescimento, de qualidade da madeira e da polpa na diversidade genética de clones de eucalipto. 1998. **Tese de Doutorado.** Universidade Federal de Viçosa.

García-peña, M.; Dias, C. T. S. Análise dos modelos aditivos com interação multiplicativa (AMMI) bivariados. **Revista Brasileira de Biometria**, São Paulo, v. 27, n. 4, p. 586-602, 2009.

Ghafouri-Kesbi F, Rahimi-Mianji G, Honarvar M and Nejati-Javaremi A (2017) Predictive ability of *Random Forest*, *boosting*, Support Vector Machines and Genomic Best Linear Unbiased Prediction in different scenarios of genomic evaluation. **Animal Production Science** 57: 229–236.

Glória LS, Cruz CD, Vieira RAM, de Resende MDV, Lopes PS, de Siqueira OHGBD and Fonseca e Silva F (2016) Accessing marker effects and heritability estimates from genome prediction by Bayesian regularized neural networks. **Livestock Science 191**: 91–96.

Goddard ME and Hayes BJ (2007) Genomic selection. **J. Anim. Breed. Genet**. 124: 323–330.

Gomide, J. L. Utilização da madeira de eucalipto para produção de celulose e papel. **Informe Agropecuário**, Belo Horizonte, v. 18, n. 186, 1997.

González-Camacho JM, de los Campos G, Pérez P, Gianola D, Cairns JE, Mahuku G, Babu R and Crossa J (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. **Theoretical and Applied Genetics** 125: 759–771.

Guedes, T.A.; Ivanqui, I. L. Análise procrustes aplicada à seleção de variáveis. **Acta Scientiarum.** Technology, v. 20, p. 505-509, 1998.

Habier D, Fernando RL, Kizilkaya K and Garrick DJ (2011) Extension of the bayesian alphabet for genomic selection. **BMC Bioinformatics 12**.

Hallauer AR, Carena MJ and Miranda Filho JB (2010) Means and variances. **Springer-Verlag**, New York, p. 33–67. In Hallauer AR and Miranda JB (eds) Quantitative genetics in maize breeding.

Hamilton MG, et al. Independent lines of evidence of a genetic relationship between acoustic wave velocity and kraft pulp yield in *Eucalyptus globulus*. Annals of Forest Science, v. 74, n. 1, p. 17, 2017.

Han J (2001) Data Mining: Concepts and Techniques. **Morgan Kaufmann Publishers** Inc.,

San Francisco, CA, USA, 744p.

Hastie T, Tibshirani R and Friedman J (2009) The elements of statistical learning: Data mining, inference, and prediction, **2. ed. Springer**, New York, NY, USA, 764p.

Hayes BJ, Bowman PJ, Chamberlain AJ and Goddard ME (2009) Invited review: Genomic selection in dairy cattle: Progress and challenges. **Journal of Dairy Science** 92: 433–443.

Haykin S (2008) Neural Networks and Learnig Machines, **3. ed. Prentice Hall**, New York, 936p.

Haykin SS (2001) Redes Neurais: Princípios e Práticas, **2 ed. Bookman**, Porto Alegre, 900p.

Haykin, S. (2008) Neural Networks and Learning Machines (3rd ed.). **Hamilton: Pearson – Prentice Hall.**

Higa, A.R.; Garcia, C.H.; Santos, E.T. Geadas, prejuízos à atividade florestal. **Silvicultura**, v.15, n.58, p.40-43, 1994.

Ho TK (1995) Random Decision Forests. **Proceedings of 3rd International Conference on Document Analysis and Recognition** 278–282.

Holland JB (2004) Implementation of molecular markers for quantitative traits in breeding programs - challenges and opportunities. Brisbane, p. 1–13. In Fischer T (ed) New directions for a diverse planet: Proceedin GWS for the **4th International Crop Science Congress.**

Hospital F, Moreau L, Lacoudre F, Charcosset A and Gallais A (1997) More on the efficiency of marker-assisted selection. **Theoretical and Applied Genetics** 95: 1181–1189.

Hosseinzadeh F, Ebrahimi M, Goliaei B and Shamabadi N (2012) Classification of lung cancer tumors based on structural and physicochemical properties of proteins by bioinformatics models. **PLoS ONE** 7.

Huang X and Han B (2014) Natural Variations and Genome-Wide Association Studies in Crop Plants. **Annual Review of Plant Biology** 65: 531–551.

Iglesias, A., B. Arcay, and J.M. Cotos. 2006. Connectionist Systems for Fishing Prediction. p. 265–296. **In Artificial Neural Networks in Real-Life Applications.**

James G, Witten D, Hastie T and Tibshirani R (2013) An Introduction to Statistical Learning with Applications in R, 1. ed. **Springer**, New York, NY, USA, 426p.

Jha S, Tripathi SK, Singh R, Dikshit A and Pandey A (2020) Global Scenario of Natural Products for Sustainable Agriculture. **Springer**, Singapore, p. 1–14. In Natural Bioactive Products in Sustainable Agriculture.

Kavzoglu, T.; P. Mather. 2003. The use of backpropagation artificial neural networks in land cover classification. **Int. J. Remote Sens**. 24(23): 4907–4938.

Kearsey MJ and Farquhar AGL (1998) QTL analysis in plants; where are we now? **Heredity** 80: 137–142.

Kitetu GM and Ko J-H (2020) Climate Change on Agriculture in 2050 : A CGE Approach. Global Trade Analysis Project (GTAP), Purdue University, West Lafayette, p. 1–23. **In 23rd Annual Conference on Global Economic Analysis (Virtual Conference).**

Klingenberg, C. P. Quantitative genetics of geometric shape: heritability and the pitfalls of the

univariate approach. **Evolution**, v. 57, n. 1, p. 191-195, 2003.

Klock, U., Muñiz, G. I. B., Anzaldo, J. H., & Andrade, A. (2005). Química da Madeira. Manual Didático**. Departamento de Engenharia e Tecnlogia Florestal - Fupef Do** Paraná, 86.

Kobayashi, M. L.; Benassi, M. T. Caracterização sensorial de cafés solúveis comerciais por Perfil Flash. **Semina: Ciências Agrárias**, v. 33, n. 6Supl2, p. 3081-3092, 2013.

Kovács, Z. L. (2006) **Redes Neurais Artificiais: fundamentos e aplicações**. 4ª ed. São Paulo: Editora Livraria da Física.

Krzanowski, W. J. Selection of variables to preserve multivariate data structure, using principal components. **Applied Statistics**, p. 22-33, 1987.

Kuhn M and Johnson K (2013) Applied predictive modeling. **Springer Science+Business Media LLC**, New York, NY, USA, 1–600p.

Leathwick JR, Elith J and Hastie T (2006) Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. **Ecological Modelling** 199: 188–196.

Lee TS and Chen IF (2005) A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. **Expert Systems with Applications** 28: 743–752.

Legarra A, Robert-Granié C, Manfredi E and Elsen JM (2008) Performance of genomic selection in mice. **Genetics** 180: 611–618.

Li B, Zhang N, Wang YG, George AW, Reverter A and Li Y (2018) Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. **Frontiers in Genetics** 9: 1–20.

Lin HY, Wang W, Liu YH, Soong SJ, York TP, Myers L and Hu JJ (2008) Comparison of multivariate adaptive regression splines and logistic regression in detecting SNP-SNP interactions and their application in prostate cancer. **Journal of Human Genetics** 53: 802–811.

Martins R, Pina P, Marques JS and Silveira M (2009) Crater detection by a *boosting* approach. **IEEE Geoscience and Remote Sensing Letters** 6: 127–131.

Matlab (2010) Matlab. **The Math Works Inc.**, Natick, Massachusett.

Mcculloch WS and Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics** 5: 115–133.

Meuwissen THE, Hayes BJ and Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. **Genetics** 157: 1819–1829.

Mitchell, H.L. Development of an adequate concept of wood quality for the guidance of geneticists and forest managers. In: **WORLD FORESTRY CONGRESS**, 5, Seattle, 1960. Proceedings. Washington: University of Washington, 1960. v.3, p.1341-1348.

Motsinger AA, Ritchie MD and Reif DM (2007) Novel methods for detecting epistasis in pharmacogenomics studies. **Pharmacogenomics** 8: 1229–1241.

Moura MM, Carneiro PCS, Carneiro JE de S and Cruz CD (2013) Potencial de caracteres na avaliação da arquitetura de plantas de feijão. **Pesquisa Agropecuaria Brasileira** 48: 417–

425.

Norvig, P. and Russell, S. (2013) **Inteligência Artificial. 3rd ed. CAMPOS**, Rio de Janeiro.

Park J and Sandberg IW (1991) Universal approximation using radial basis function networks, 3. ed. **Neural Comput**., 246–259.p.

Paterson AH, Tanksley SD and Sorrells ME (1991) DNA Markers in Plant Improvement. **Advances in Agronomy** 46: 39–90.

Peixoto LA, Laviola BG, Alves AA, Rosado TB and Bhering LL (2017) Breeding Jatropha curcas by genomic selection: A pilot assessment of the accuracy of predictive models. **PLoS ONE** 12: 1–16.

Pessoni, L. A. Estratégias de análise da diversidade em germoplasma de cajueiro (Anacardium spp. L.). 2007. 159 f. Tese (Doutorado em Genética e Melhoramento) – Universidade Federal de Viçosa, Viçosa, MG, 2007.

Prasad AM, Iverson LR and Liaw A (2006) Newer classification and regression tree techniques: *bagging* and *random forest* for ecological prediction. **Ecosystems** 9: 181–199.

R Core Team (2020) R: A Language and Environment for Statistical Computing. **R Foundation for Statistical Computing**, Vienna, Austria.

Raymond CA; Thomas DS; Henson M. Predicting pulp yield and pulp productivity of Eucalyptus dunnii using acoustic techniques. **Australian Forestry**, v. 73, n. 2, p. 91–97, 2010.

Resende Jr. MFR, Alvez AA, Sánches CFB, Resende MDV and Cruz CD (2013) Seleção Genômica Ampla. **Suprema**, Visconde do Rio Branco, MG, p. 424. In Cruz CD, Salgado CC, and Bhering LL (eds) Genômica Aplicada.

Resende Jr. MFR, Muñoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, Resende MDV and Kirst M (2012) Accelerating the domestication of trees using genomic selection: Accuracy of prediction models across ages and environments. **New Phytologist** 193: 617–624.

Resende MDV (2008) Genômica quantitativa e seleção no melhoramento de plantas perenes e animais. **Colombo: Embrapa Florestas**, 330p.

Resende MDV, Resende MFR, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, Abad JM, Takahashi EK, Rosado AM, Faria DA, Pappas GJ, Kilian A and Grattapaglia D (2012) Genomic selection for growth and wood quality in Eucalyptus: Capturing the missing heritability and accelerating breeding for complex traits in forest trees. **New Phytologist** 194: 116–128.

Roque, J. V. Desenvolvimento de modelos de regressão multivariada para determinação de ésteres de forbol em sementes de jatropha curcas L. usando espectroscopia e quimiometria. 2015. 84 f. **Dissertação (Mestrado em Agroquímica)** - Universidade Federal de Viçosa, Viçosa, 2015.

Sant'anna IC (2018) **Redes neurais artificiais para predição genômica na presença de interações epistáticas.** Viçosa: Universidade Federal de Viçosa, 93p.

Sax K (1923) The Association of Size Differences with Seed-Coat Pattern and Pigmentation in Phaseolus Vulgaris. **Genetics** 8: 552–560.

Schimleck, L. R. et al. Microfibril angle prediction of Pinus taeda wood samples based on tangencial face NIR spectra. **IAWA Journal**, v. 28, n. 1, p. 1–12, 2007.

Schimleck, L. R.; Evans, R.; Ilic, J. Estimation of Eucalyptus delegatensis wood properties by near infrared spectroscopy. Canadian Journal Forestry Resource, v. 31, n. 10, p. 1671-1675, 2001a. GINDL, W. et al. The relationship between Near Infrared Spectra of radial wood surfaces and wood mechanical properties. **Journal Near Infrared Spectroscopy**, v. 9, n. 4, p. 255, 2001.

Searchinger T, Waite R, Hanson C and Ranganathan J (2019) Creating a Sustainable Food Future. **World Resources Report** 1: 558.

Silva GN (2018) **Predição de valos genéticos por abordagens de seleção genômica ampla e de inteligência computacional.** Viçosa: Universidade Federal de Viçosa, 108p.

Silva GN, Tomaz RS, Sant'Anna I de C, Nascimento M, Bhering LL and Cruz CD (2014) Neural networks for predicting breeding values and genetic gains. **Scientia Agricola** 71: 494–498.

Silva JC et al. Genetic parameters for growth, wood density and pulp yield in Eucalyptus globulus. **Tree Genetics & Genomes**, v. 5, n. 2, p. 291–305, 2009.

Silva IN, Spatti HD and Flauzino RA (2010) Redes Neurais Artificiais: para engenharia e ciências aplicadas. **Artliber**, São Paulo, SP, 399p.

Singh BD and Singh AK (2015) **Marker-assisted plant breeding: Principles and practices.** 1–514p.

Souza IC, Nascimento M, Silva GN, Nascimento ACC, Cruz CD, Fonseca F, Almeida DP, Pestana KN, Azevedo CF, Zambolim L and Caixeta ET (2020) Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. **Scientia Agricola** 78: 1–8.

Sudheer KP, Gosain AK and Ramasastri KS (2003) Estimating actual evapotranspiration from limited climatic data using neural computing technique. **Journal of Irrigation and Drainage Engineering** 129: 214–218.

Tan P-N, Steinbach M and Kumar V (2006) Introduction to Data Mining. **Addison-Wesley Longman Publishing Co**., Inc., Boston, MA, USA, 169p.

Taylan P and Weber GW (2019) CG-Lasso Estimator for Multivariate Adaptive Regression Spline. Springer International Publishing AG, p. 121–136. In Tas K, Baleanu D, and Machado JAT (eds) **Mathematical Methods in Engineering: Aplications in Dynamics of Complex Systems**.

Tomaz RS, Alvez DP, Nascimento M and Cruz CD (2018) Inteligência Computacional. Editora UFV, Viçosa, MG, p. 414. In Cruz CD and Nascimento M (eds) **Inteligência Computacional Aplicado ao Melhoramento Genético**.

Trugilho, P. F.; Regazzi, A. J.; Vital, B. R. & Gomide, J. L. Aplicação de algumas técnicas multivariadas na avaliação da qualidade da madeira de eucalyptus e seleção de genótipos superiores para produção de carvão vegetal. **Revista Árvore** , Viçosa-MG, v. 21, N° 1, 113-130 p., 1997.

Van Eenennaam AL, Weigel KA, Young AE, Cleveland MA and Dekkers JCM (2014) Applied Animal Genomics: Results from the Field. **Annual Review of Animal Biosciences** 2:

105–139.

Wade MJ (2000) **Epistasis and evolutionary process**. Oxford University Press, New York.

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. **PLoS computational biology**, 13(6), e1005510.

Wold, S.; Sjöström, M., Chemometrics, present and future success, **Chemometrics and Intelligent Laboratory Systems**, v. 44, p. 3-14, 1998.

Wong CK and Bernardo R (2008) Genomewide selection in oil palm: Increasing selection gain per unit time and cost with small populations. **Theoretical and Applied Genetics** 116: 815–824.

Wright, J. A.; Birkett, M. D.; Gambino, M. J. T. Prediction of pulp yield and cellulose content from wood samples using near infrared spectroscopy. **Tappi Journal**. v. 73, n. 8, p. 164-166, 1990.

Xavier, A.; Borges, R. C. G.; Cruz, C. D. & Cecon, P. R. Aplicação da análise da divergência genética no melhoramento de Eucalyptus spp. **Revista Árvore**, Viçosa-MG, v. 20, N° 4, 495 – 505 p., 1996.

York TP and Eaves LJ (2001) Common Disease Analysis Using Multivariate Adaptive Regression Splines (MARS): Genetic Analysis Workshop 12 Simulated Sequence Data. **Genetic Epidemiology** 21: S649–S654.

York TP, Eaves LJ and van den Oord EJCG (2006) Multivariate adaptive regression splines: A powerful method for detecting disease-risk relationship differences among subgroups. **Statistics in Medicine** 25: 1355–1367.

Zeng, X.; Intelligent sensory evaluation: Concepts, implementations and applications. **MathematicsandComputers in Simulation**, v. 77, n. 5, p. 443-452, 2008.

Zhao Y, Mette MF and Reif JC (2015) Genomic selection in hybrid breeding. **Plant Breeding** 134: 1–10.

Zheng G, Yang P, Zhou H, Zeng C, Yang X, He X and Yu X (2019) Evaluation of the earthquake induced uplift displacement of tunnels using multivariate adaptive regression splines. **Computers and Geotechnics** 113: 103099.

4. ARTIGO 1

Abstract

Importância do pré tratamento espectral no uso da espectroscopia no infravermelho próximo na predição de características tecnológicas da madeira em eucalyptus.

Alexandre Gomes Ferraz[1]; Cosme Damião Cruz[1]; Gleison Augusto dos Santos[2]; Talita Baldin[3],
Osmarino Pires dos Santos[4]; Brígida Maria Teixeira dos Reis Valente[5]

[1]Federal University of Viçosa, Department of General Biology, Viçosa, MG, Brazil;

[2] Federal University of Viçosa, Department of Forest Engineering, Viçosa, MG, Brazil;

[3]Institute of Agricultural Sciences, Federal University of Minas Gerais, Montes Claros, Minas Gerais, Brazil;

[4]Empresa CMPC Celulose Riograndense, Guaíba, RS, Brasil;

[5]Empresa Eldorado Brasil, Três Lagoas, MS, Brasil;

Corresponding author: Alexandre Gomes Ferraz

E-mail: alexandre.g.ferraz@gmail.com, Phone: +55(31) 99578-7194

Declarations

The authors are grateful for the financial support of the Conselho Nacional de Desenvolvimento Científico e Tecnológico- CNPq, the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES, Federal University of Viçosa and the company CMPC – Celulose Riograndense.

Resumo

Este trabalho visa compreender a importância e definir os melhores pré-tratamentos de espectros que levariam a maior eficiência de predição de características tecnológicas da madeira, utilizando a metodologia de mínimos quadrados parciais (PLS), com base em informações do NIR, para seleção indireta de indivíduos de *Eucalyptus benthamii*. O material para as análises foi composto por 75 indivíduos de *E. benthamii* 3 de *E. saligna*, 3 de *E. grandis* e 3 de *E. dunnii* totalizando 87 indivíduos, escolhidos por critérios industriais. Para avaliação dos pré–tratamentos e modelos de predição foram usadas 11 características obtidas por análises laboratoriais. Para a avaliação da importância dos pré-tratamentos, usou-se 15 métodos de pré-tratamentos, sendo que alguns apresentam parâmetros variáveis, totalizando 199 possibilidades. Conclui-se que o pré-tratamento aumenta as acurácias do modelo expressas pelos valores de R^2 (coeficiente de determinação) e REQMv (Raiz do Erro Quadrado Médio de Validação). Diferentes técnicas de pré-processamentos são eficientes considerando informações de diferentes características na população de *E. benthamii*. O uso de pré-tratamento envolvendo técnicas de segunda derivada gap se destacou no conjunto de dados analisado e deve ser enfatizado como uma alternativa vantajosa em estudos de ajuste de modelos.

Palavras-chave: NIRS; Tecnologia da Madeira; Reprodução; Modelos; PLS

Introdução

O NIR (*Near Infrared Spectroscopy*) é um tipo de espectroscopia vibracional que utiliza comprimentos de ondas na faixa do infravermelho próximo (750 a 2500 nm) para determinação de moléculas orgânicas e variáveis qualitativas, sendo utilizado em áreas como agricultura, alimentação, farmacêutica e florestal (Muñiz et al., 2012). Essa tecnologia tem como vantagem ser um método rápido, não invasivo e com aplicação quase universal (Pasquini, 2003; Cen & HE, 2007; Pasquini 2018;).

Estudos utilizando este tipo de espectroscopia têm tido sucesso no melhoramento genético de plantas. Estopa et al. (2017) enfatizaram o potencial do NIR no melhoramento do eucalipto que permite uma avaliação precoce e não destrutiva, reduzindo custos e melhorando o desempenho da matéria-prima na indústria. Li e Altaner (2018) verificaram a confiabilidade da espectroscopia NIR no estudo do conteúdo de extrativos (CE) de *Eucalyptus bosistoana*, sendo uma ferramenta para o melhoramento genético para esta característica. Ferreira et al. (2018) constataram que a espectroscopia NIR, aliada à quimiometria, foi suficiente para prever a composição química dos resíduos da madeira de eucalipto pós-colheita.

As informações presentes nos espectros NIR são complexas e quase nunca são prontamente analíticas, os algoritmos quimiométricos/biométricos visam resolver a falta de seletividade para fins quantitativos (Pasquini, 2018). Além disso, os dados obtidos pela espectroscopia NIR contém ruídos, oriundos das vibrações do aparelho e efeitos externos próximos ao aparelho, (Manley & Baeten, 2018) que devem ser eliminados/ajustados através de pré-tratamento dos dados (Cen & He, 2007). Essa etapa é de grande importância para qualquer análise e, geralmente, vários métodos são testados, para que o melhor pré-tratamento seja utilizado e se obtenha um modelo de predição de qualidade (Ferreira, 2015). De acordo com Ferreira, (2015) e Melo, (2017) o objetivo do pré-tratamento é reduzir/ajustar, uma vez que não podem ser eliminadas, os ruídos/variáveis indesejáveis que podem influenciar negativamente na confecção do modelo. Entretanto, a escolha do pré-tratamento é uma etapa desafiadora e requer conhecimento muito específico de transformação de dados para se obter um modelo de confiança e bem acurado.

Ferramentas de análise multivariada são amplamente utilizadas na análise de dados espectroscópicos (Jin and Xu, 2011; Smith-Moritz et al., 2011; Xu et al., 2013; Lupoi et al., 2014), sendo as principais: a Análise Componente Principal (PCA) e

Mínimos Quadrados Parciais (PLS). Normalmente, um modelo adequado precisa apresentar altos valores do coeficiente de determinação, mas, principalmente, baixos valores do erro de predição REQMv (Zhou et al., 2016; Pasquini, 2018). Entretanto, para determinadas características essas técnicas, PCA e PLS apresentam baixa acurácia, principalmente por serem utilizados modelos estatísticos lineares.

Pelo exposto, o objetivo do trabalho foi otimizar e definir os melhores pré-tratamentos que conduziriam a maior eficiência de predição usando metodologia de mínimos quadrados parciais (PLS) considerando características tecnológicas da madeira, com base em informações de NIR e de química da madeira, para fins de seleção indireta de indivíduos de *E. benthamii*.

Material e Métodos

Material Genético e localização do experimento

O material genético utilizado é proveniente de um teste de progênies de *Eucalyptus benthamii*, de propriedade da empresa CMPC Celulose Riograndense, localizado no município de Encruzilhada do Sul, no estado do Rio Grande do Sul. Foram avaliadas 87 amostras, sendo 75 de *E. benthamii*, e 12 indivíduos de outras três espécies (quatro de *E. saligna*, quatro de *E. grandis* e quatro de *E. dunnii*).

Preparação das amostras

Amostragem

Para a seleção das amostras para avaliação dos dados, inicialmente retirou-se uma alíquota de serragem, pelo método não-destrutivo, de 87 árvores no povoamento (número de árvores vivas e em condições de amostragem) definidas anteriormente. Com o auxílio de uma furadeira e uma broca, a serragem foi coletada em quatro pontos distintos na altura do DAP (diâmetro a altura do peito, sem casca) e acondicionada em sacos de papel.

No laboratório de Espectroscopia de Infravermelho foi medido o espectro das 87 amostras utilizando espectrômetro de infravermelho próximo BRUKER, modelo MPA. Para tal, a madeira, em forma de serragem, foi compactada manualmente em

recipientes adaptados para leitura na esfera de integração. Os recipientes eram providos de fundo de quartzo transparente.

Anteriormente à primeira leitura, o aparelho foi calibrado com os padrões de referência (Background). O processamento e a análise dos dados foram realizados em software OPUS Quant 6.5 (BRUKER).

Determinação das propriedades tecnológicas da madeira

As propriedades químicas da madeira foram coletadas de acordo com as normas apresentadas na Tabela 1.

Tabela 1 - Normativas empregadas para análises químicas de madeira moída. Todas as análises foram feitas com base na madeira seca.

Análise química	Norma utilizada
Teor de cinzas	TAPPI 21 1 OM-02
Extrativos (Totais, em acetona e em água)	TAPPI 280 PM-99
Densidade básica da madeira	SCAN – CM 43:95
Análises Químicas	TAPPI T 257
Lignina Klason e Total	TAPII 222 OM-2 e TAPPI UM-250
Teor de Pentosanas	TAPPI T223 CM-01

Fonte: Centro I+D da CMPC Celulose S.A.

Assim foi possível obter informações de 11 variáveis: Teores de cinzas, Teores de extrativos (totais, acetona e em água), Densidade Básica da Madeira, Lignina Klason, Lignina Total, Teores de Pentosanas, Celulose, Holocelulose, número kappa e rendimento de celulose. Esse conjunto de variáveis pode ser definido como o representativo das características tecnológicas da madeira.

Pré-tratamentos

Foram testados 15 pré–tratamentos básicos, usados em estudos de quimiometria aplicados às técnicas de espectroscopia no infravermelho próximo (Ferreira, 2015). Sendo que, em alguns, foram incorporadas variações em alguns parâmetros. Dessa forma, no estudo testou-se 199 possibilidades de pré-tratamento, como mostrado na Tabela 2. Assim, um algoritmo desenvolvido no software R implementado no Portal Genes, permitiu uma varredura no conjunto de dados sob diferentes métodos, e apresentou todas as possibilidades para encontrar o melhor pré-tratamento.

Na Tabela 2 é descrito todos os métodos que foram aplicados para a análise dos pré-tratamentos. Após a aplicação dos diferentes métodos, o conjunto de dados foi utilizado para fins de predição de características da qualidade da madeira utilizando o procedimento estatístico PLS (Mínimos Quadrados Parciais). Considerou-se que os maiores R^2 e os menores REQMv da predição evidenciavam que o pré-tratamento era o mais apropriado.

Tabela 2. Métodos de pré-tratamento para dados espectrais.

	Pré-tratamento	Variação	Parâmetro*
1	Sem nenhum ajuste	1	
2	Média	8	Tamanho da janela
3	Média Móvel	40	Valor m da janela 2m+1
4	Savitzky-Golay	87	Parâmetros m e p
5	Primeira derivada sem gap	1	
6	Segunda derivada sem gap	1	
7	Primeira derivada com gap	11	Parâmetro g
8	Segunda derivada com gap	11	Parâmetro g
9	Normalização	1	
10	Detrend signal	1	
11	Centragem e escalamento - blockScale type soft	1	
12	Centragem e escalamento - blockScale type hard	1	
13	Centragem e escalamento - blockNorm	1	
14	Transformação de Fourier – reflectância	1	
15	Transformação de Fourier – absorbância	1	

(*) Descrição do método de pré-processamento e dos parâmetros encontram-se descrita em SOUSA (2008) FERREIRA 2015

As avaliações foram feitas para cada característica em análise. Para escolha do melhor pré-tratamento, foi realizada a análise preliminar de componentes principais para identificar a quantidade de variáveis latentes necessárias para explicar pelo menos 80% da variação total de cada característica. Por fim, foram testadas todas as possibilidades de pré-tratamentos e suas variações.

Desempenho em relação às características fenotípicas

A avaliação do desempenho das características foi realizada através da análise descritiva do conjunto de dados em avaliação. Os parâmetros obtidos foram a média, desvio padrão, mínimo, máximo e coeficiente de variação. Essa análise permite ter uma visão geral da variação dos dados e avaliar a representatividade da amostra.

Importância do uso de pré-tratamento

Uma das premissas para realizar a calibração via PLS é realizar o pré-tratamento. Há diversos pré-tratamentos e é necessário um conhecimento aprofundado sobre técnicas quimiomêtricas. Assim, essa avaliação irá permitir se é necessário o uso de pré-tratamento, quais características precisam ou não de pré-tratamento tanto levando em consideração o Erro (REQMv) e o coeficiente de determinação (R^2).

Escolha de pré-tratamentos ótimos para cada característica.

Para essa escolha, foram avaliadas as características de forma individual em 5 diferentes cenários. Em cada cenário foram tomados, de forma aleatória, 80% dos dados para treinamento e 20% para validação. Estatísticas de qualidade de ajuste foram obtidas para cada subconjunto e, de acordo com REQMv e o R^2 do conjunto de dados de validação, foram reconhecidos os melhores pré-tratamentos para cada cenário e com a avaliação conjunta dos 5 cenários foram avaliadas e obtidas medidas de eficiência comparativa destes pré-tratamentos.

Desenvolvimento dos modelos de calibração

Os espectros de refletância das amostras, coletados nos diversos comprimentos de onda pelo espectrômetro NIR, foram importados para o computador (software OPUS Quant 6.2), submetidos a pré-tratamento e associados com os valores determinados para as madeiras de *Eucalyptus* em laboratório.

A calibração do modelo foi estabelecida por análise de Partial Least Squares. Os modelos foram ajustados utilizando um número variáveis latentes (VLs) necessárias para fornecer um adequado ajuste do modelo, sem perdas consideráveis da variância dos dados. O número de variáveis adotadas para cada modelo foi definido em função da Raiz do erro Quadrado Médio da validação (REQMv) e do aumento do coeficiente de determinação (R^2) aferido no conjunto de dados destinados à análise de validação.

Foram excluídas as faixas espectrais acima de 10.000 cm^{-1}, posto que, nessa região, o espectro apresenta repetições de ruídos que não apresentam informação relevante sobre a propriedade de interesse. As amostras anômalas, visivelmente diferentes do restante, foram detectadas não representativas (outliers) e excluídas do modelo.

Todas as análises foram realizadas utilizando o Portal Genes (Cruz, 2016) em integração com o software R.

Resultado e Discussão

Desempenho em relação às características fenotípicas

Foram determinadas as propriedades tecnológicas da madeira avaliadas com base na madeira seca conforme apresentado na Tabela 3. Observou-se grandes variações nas características referentes às propriedades em estudo, principalmente para o teor de cinzas, extrativos em acetona, extrativos em água, extrativos totais e teor de pentosanas, com alto coeficiente de variação para o grupo de genótipos avaliados (Tabela 3). Os valores máximos e mínimos observados foram semelhantes aos observados por outros autores (Abjaud et al., 2017;), indicando a possibilidade de uma adequada calibração do modelo.

Tabela 3 – Descritores estatísticos das propriedades químicas da madeira moída analisadas com base na madeira seca.

Propriedades	Média	DP	Mínimo	Máximo	CV (%)
Rendimento de celulose	49,42	2,24	42,60	54,20	4,53
Teor de Cinzas	0,54	0,26	0,29	2,51	48,14
Extrativos em acetona	0,95	0,35	0,29	1,97	36,84
Extrativos em água	2,59	0,71	0,02	6,72	27,41
Extrativos Totais	3,55	0,85	0,67	7,55	23,94
Teor de Pentosanas	16,83	2,59	12,66	22,01	15,38
Lignina Klason	27,54	1,71	22,87	34,30	6,20
Lignina Total	31,13	1.58	27,35	36,77	5,07
Holocelulose	65,13	2,12	56,20	69,80	3,25
Densidade básica da madeira	414,4	31,72	349,0	502,0	7,65
Número Kappa	18,78	1,75	15,70	23,60	9,31

*DP – Desvio Padrão; CV – Coeficiente de variação.

É importante destacar a amplitude da variação dos dados em estudos que tratam de calibrações espectroscópicas (Hein et al., 2010). Para a calibração de um modelo com base no NIR, as amostras do conjunto de calibração e treinamento precisam ser representativas de forma a garantir as extrapolações dos modelos estabelecidos para outros conjuntos de dados (Costa et al., 2018; Trugilho et al., 2015;

Gomide et al. 2010). Ainda neste sentido, os limites superiores e inferiores de cada propriedade se fazem importante, visto que valores menores ou maiores que os limites da faixa de variação contabilizada no modelo não poderão ser satisfatoriamente preditos em outros conjuntos de dados experimentais.

Importância do uso de pré-tratamento

O pré-tratamento é uma das etapas no processo de calibração em estudos com NIR e existem algumas alternativas de análises que são apresentadas para este fim. A justificativa para essa etapa é relacionada aos ruídos gerados no aparelho de leitura das amostras que são devidos à vibração do aparelho e a efeitos externos nos momentos da obtenção dos dados. Na Tabela 4 estão apresentados os valores de acurácia de predição das 11 características em estudo (R^2 validação) dos dados sem pré–tratamento e quando se aplica o melhor pré–tratamento. Em abordagens biométricas é fundamental que a técnica de pré-tratamento consiga alta correlação entre o valor predito com o observado para que o julgamento da superioridade de um genótipo em relação a outro, sob seleção, não seja comprometido.

Tabela 4. Média e desvio padrão das medidas de acurácia de predição de características R^2v (coeficiente de determinação na validação) relacionadas à qualidade da madeira, considerando conjunto de dados sem e com pré-processamento.

Características	R^2v (sem pré-processamento)		R^2v (com pré-processamento)	
	Média	Desvio Padrão	Média	Desvio Padrão
Rendimento de Celulose	0.41	0.25	0.60	0.16
Teores de Cinzas	0.04	0.04	0.20	0.12
Extrativos em acetato	0.57	0.15	0.66	0.17
Extrativos em água	0.06	0.07	0.29	0.11
Extrativos Totais	0.06	0.10	0.23	0.11
Holocelulose	0.47	0.18	0.63	0.16
Número kappa	0.02	0.02	0.24	0.14
Lignina Total	0.48	0.16	0.68	0.17
Lignina Klason	0.53	0.20	0.73	0.09
Teor de pentosanas	0.17	0.35	0.18	0.08
Densidade Básica	0.49	0.11	0.57	0.09

Nas Tabelas 4 e 5 podemos ver a melhoria da predição com a adoção de procedimentos de pré-tratamento, evidenciando que ele é fundamental para o ajuste de modelos. Entretanto, trabalho como realizado por Milagres (2009) encontrou os melhores resultados de R^2 em características de qualidade da madeira não usando nenhum tipo de pré-tratamento. O mesmo autor, ao avaliar o menor REQMv, verificou que a primeira derivada foi a que apresentou melhores resultados para os parâmetros avaliados. Assim, podemos definir que o uso de pré-tratamento deve ser considerado, mas em algumas situações os dados não tratados podem ser a melhor opção. Além disso, os tipos de pré-tratamento usados para R^2 e REQMv podem ser diferentes, ou seja, nem sempre o método que forneceu o melhor R^2 é aquele que fornecerá o REQMv.

Tabela 5. Média e desvio padrão das medidas de acurácia de predição de características REQMv (raiz do erro quadrático médio) relacionadas à qualidade da madeira, considerando conjunto de dados sem e com pré-processamento

Características	REQMv(sem pré-processamento)		REQMv(com pré-processamento)	
	Média	Desvio Padrão	Média	Desvio Padrão
Rendimento de Celulose	39.45	9.29	14.12	12.68
Teores de Cinzas	2.60	1.82	0.64	0.13
Extrativos em acetato	0.98	0.67	0.21	0.04
Extrativos em água	13.91	6.45	2.45	0.68
Extrativos Totais	20.31	3.40	3.87	0.32
Holocelulose	43.25	7.12	15.66	9.12
Número kappa	13.68	6.97	2.13	0.39
Lignina Total	30.64	6.55	17.35	10.17
Lignina Klason	41.25	4.67	11.66	6.03
Teor de pentosanas	44.96	10.10	16.78	7.58
Densidade Básica	675.97	142.88	102.06	59.74

O erro REQMv é condicionado à dimensão de cada variável e para o ajuste dos modelos é importante que seja o menor possível, pois significa que os valores preditos estão mais próximos do observado permitindo fazer uso de tais valores para melhor tomada de decisão de recursos a serem utilizados e para fazer projeções de desempenho comparativos. Ao comparar os erros, sem aplicação e com pré-tratamentos, pode-se observar que houve redução no valor do erro e que o uso do pré-tratamento é a uma etapa importante e deve ser considerada nas análises.

Escolha de pré-tratamentos ótimos para cada característica

Os resultados anteriores evidenciaram a importância do pré-tratamento, entretanto, um fator complicador é a existência de várias possíveis técnicas de pré-tratamentos que, em alguns casos, envolvem a necessidade de estabelecimento de parâmetros iniciais para ajuste. A aplicação do pré-tratamento é algo primordial para obter melhores resultados usando a técnica de PLS. Dessa forma, é preciso avaliar os melhores de acordo com cada característica. Assim, resumiremos, a seguir, algumas opções de pré–tratamentos para cada característica analisada, considerando o bom desempenho quantificado pelos valores de R^2 e REQMv obtidos em ajustes de modelo em conjuntos de dados de validação.

A seguir, na tabela 6, foram listados os melhores pré-tratamentos para cada característica avaliada referente aos cincos conjuntos de treinamento e validação. Para a característica de rendimento de celulose, o pré-tratamento *segunda derivada com gap* demonstrou ser o mais eficiente em apenas uma das 5 análises, proporcionando valor de R^2 igual a 0,84 bem superior ao obtido sem o uso de pré-tratamento que foi de 0.41. O pré-tratamento *detrend signal* se destacou por ser o mais eficiente em 2 das 5 análises, com valores de R^2 superiores ao 0.41 (obtido sem pré-tratamento), mas inferior ao obtido pelo uso da segunda derivada.

Tabela 6 – Melhores pré-tratamentos em cada um dos 5 cenários com os respectivos R^2 e REQMv considerando as 11 características.

Características	Pré-tratamento	Parâmetro	R^2	REQMv
Rendimento de Celulose	Segunda derivada com gap	g=120	0.84	44.36
	Savitzky-Golay	m=30,p=2	0.41	34.41
	Média Móvel	m=40	0.48	44.27
	Detrend signal		0.49	2.21
	Detrend signal		0.55	7.55
Teor de Cinza	Segunda derivada sem gap		0.12	0.92
	Média	w=16	0.12	0.89
	Média	w=21	0.11	5.05
	Segunda derivada com gap	g=120	0.23	3.33
	Média	w=26	0.30	0.72
Extrativos em Acetato	Primeira derivada com gap	g=110	0.72	0.19
	Primeira derivada com gap	g=105	0.63	0.21
	Primeira derivada sem gap		0.42	0.22
	Segunda derivada com gap	g=105	0.78	0.19
	Segunda derivada com gap	g=90	0.44	0.34
Extrativos em Água	Segunda derivada sem gap		0.20	2.57
	Segunda derivada com gap	g=75	0.40	11.89
	Segunda derivada sem gap		0.28	2.88
	Segunda derivada sem gap		0.11	3.25
	Média Móvel	m = 27	0.32	14.57
Extrativos Totais	First Derivative Without Gap		0.08	6.78
	Segunda derivada com gap	g = 110	0.21	14.46
	Segunda derivada com gap	g = 75	0.20	15.48
	Segunda derivada com gap	g = 100	0.34	15.47
	Média	J = 40	0.33	13.59
Holocelulose	Segunda derivada com gap	g = 110	0.55	55.95
	Detrend signal		0.40	9.14
	Média Móvel	m = 40	0.38	25.09
	Detrend signal		0.72	11.88
	Média Móvel	m = 40	0.75	25.19
Número Kappa	Segunda derivada com gap	g = 90	0.10	3.48
	Segunda derivada com gap	g = 105	0.17	3.35
	Segunda derivada com gap	g = 95	0.21	6.13
	Segunda derivada com gap	g = 105	0.12	7.52
	centering and scaling - blockScale type soft		0.31	39.47
Lignina Klason	Segunda derivada com gap	g = 75	0.71	25.90
	Média Móvel	m = 40	0.74	40.58
	Segunda derivada com gap	g = 100	0.58	25.60
	Segunda derivada com gap	g = 100	0.68	16.04
	Segunda derivada com gap	g = 85	0.82	21.49
Lignina Total	Segunda derivada com gap	g = 75	0.65	27.47
	Detrend signal		0.72	12.74

	Detrend signal		0.64	1.53
	Segunda derivada com gap	g = 70	0.41	25.51
	Segunda derivada com gap	g = 70	0.83	25.93
Pentosanas	Primeira derivada sem gap		0.22	47.13
	Média	w = 40	0.20	50.97
	Segunda derivada com gap	g = 105	0.23	45.58
	Segunda derivada com gap	g = 110	0.20	44.02
	Primeira derivada com gap	g = 90	0.04	62.67
Densidade Básica da Madeira	Primeira derivada sem gap		0.56	360.76
	Segunda derivada com gap	g = 75	0.56	57.11
	Primeira derivada sem gap		0.52	254.89
	Segunda derivada com gap	g = 70	0.45	268.13
	Segunda derivada com gap	g = 75	0.55	62.96

p = ordem do polinômio; w =janela (must be odd); m = m-th derivative; g = segment size;

Para o teor de cinzas (tabela 6), o pré-tratamento *média,* usando janela de 16, 21 e 36, foi o que mais se destacou, ou seja, demonstrou eficiente em 60% das avaliações realizadas. Os valores de R^2 foram relativamente baixos, mas superior ao obtido nas análises realizadas sem o uso de pré-tratamento que foi de 0.04. Com estes resultados, já se pode observar que características diferentes podem se adequar a pré-tratamentos diferentes, cabendo ao pesquisador a busca desta informação para que o modelo ajustado seja o mais eficiente na predição dos indivíduos de interesse. Também nesta análise, observa-se que o pré-tratamento *média* se destacou na maioria das análises realizadas para esta característica, porém o parâmetro demandado pela abordagem, que no caso é o comprimento da janela, deve também ser adequadamente ajustado para se ter o ajuste, no modelo de predição, de maior acurácia.

Para a característica de teor de extrativos em acetato, três pré-tratamentos se mostraram eficientes nas realizadas (tabela 6). A *primeira derivada com gap*, valores de g = 110 e g = 105, apresentou uma eficiência de 40%, assim como o pré-tratamento *segunda derivada com gap* (g = 105 e g = 90). Além dessas, a *primeira derivada sem g*ap teve eficiência de 10%. Quando analisamos o R^2 é possível observar que se teve valores altos, ao se comparar às demais características, chegando até a 0.78 e o menor valor 0.42. Assim, nessa situação, a *primeira derivada com gap* e a *segunda derivada com gap* se mostraram mais eficientes, sendo a *primeira derivada* que apresentou maiores valores de R^2 nas duas análises que se mostraram mais eficientes.

Nas análises mostradas na tabela 6 da característica teor de extrativos em água, o pré-tratamento *segunda derivada sem gap* teve eficiência de 60%. A *segunda derivada com gap*, usando g = 75, teve eficiência de 20%, assim como a *média móvel*, com m = 27. Os valores de R^2 variaram entre 0.11 e 0.40 e os de erro médio de 2.57 a 14.57. Podemos observar que quando o melhor pré-tratamento é a *segunda derivada sem gap,* os valores de erro médio foram os mais baixos em relação aos outros dois métodos. Do exposto, para a característica teor de extrativos em água, a *segunda derivada sem gap* é o pré-tratamento mais adequado.

Extrativos totais é uma característica obtida através da soma do extrativos totais em acetato com os extrativos em água. A correlação fenotípica entre extrativos totais e em água foi significativa e de magnitude 0.90 (Ferraz et al. 2020). Na relação apresentada (tabela 6), pode-se observar que três pré-tratamentos se destacaram, sendo a *primeira derivada sem gap* com 20% de eficiência, a *segunda derivada com gap*, considerando valores de g iguais a 110, 75 e 100, com 60% de eficiência e, por fim, a *média*, com janela igual a 40. Os valores de R^2 e erro médio variaram de 0.083 a 0.34 e 6.78 a 15.48, respectivamente. Podemos definir que a *segunda derivada com gap* é o pré-tratamento mais eficiente para a característica Teor de extrativos totais.

Nas análises do melhor pré-tratamento para a característica holocelulose (tabela 6), dois pré-tratamentos apresentaram a mesma eficiência de 40%, sendo estes o *detrend signal* e a *média móvel*, com janela de amplitude 40. Além desses, a *segunda derivada com gap* apresentou uma eficiência de 20%. Os valores de R^2 apresentaram valores entre 0.38 e 0.75, melhores quando se comparado aos dados sem pré-tratamento (Tabelas 4 e 5). Baldin et al. (2020) encontraram valores de R^2 e erro médio de validação cruzada (RMSEv) de 0.82 e 0.76%, respectivamente. Os valores de erro variaram de 9.14 a 55.95, ou seja, melhores do que quando foram comparados com os dados não processados. Além disso, os menores valores de erro médio foram os referentes ao do pré-tratamento *detrend signal*, mostrando que esse pré-tratamento é o mais eficiente para a característica holocelulose.

De todas as análises com a característica número kappa, o pré-tratamento *segunda derivada com gap*, valores de g iguais a 90, 95 e 105, apresentou uma eficiência de 80% com destaque predominante (tabela 6). Analisando apenas esse pré-tratamento, os dados de R^2 apresentaram valores entre 0.10 e 0.21 e erros médios entre 3.35 e 7.52, valores melhores do que quando os dados não foram pré-tratados. Além desse pré-tratamento, a *centragem e escalamento* apresentou eficiência de 20%

com R^2 de 0.31, mais alto das análises feitas para esta características. No entanto, o valor de erro médio foi alto em comparação com os apresentados pela *segunda derivada com gap*, sendo o valor igual a 39.47, como mostrado na tabela abaixo. Do exposto, a característica *segunda derivada com gap* é o pré-tratamento mais adequado para a característica número kappa.

A característica lignina klason teve resultado muito semelhante ao do número kappa com a *segunda derivada com gap*, com valores de g iguais a 75, 85 e 100, apresentando 80% de eficiência com R^2, variando de 0.58 a 0.82 e erro médio de 16.04 a 25.90 (tabela 6). Além dessa, com 20% de eficiência, a *média móvel*, com janela de amplitude de 40, foi o outro pré-tratamento para a característica lignina klason com R^2 de 0.74 e erro médio 40.58. Analisando os dois pré-tratamentos, os valores de R^2 foram melhores e relativamente altos em todas as análises. Quanto ao erro médio, os valores do pré-tratamento segunda derivada com gap apresentou valores menores. Assim, para a característica lignina klason tem como pré-tratamento mais adequado a *segunda derivada com gap.*

É possível observar na tabela 6 que, para a característica lignina total, o pré-tratamentos *segunda derivada com gap* e *detrend signal* são os dois métodos que se destacaram. O primeiro método apresentou eficiência de 60% com R^2 variando de 0.41 a 0.83 e erro médio entre 25.51 a 27.47. Já para o segundo método, *Detrend signal* teve 40% de eficiência e R^2 com valores de 0.64 e 0.72 e erro médio 1.53 e 12.74. Valores promissores que já eram esperados, pois, segundo Sousa (2008), os valores de lignina total são os que melhores se adaptam a modelos de predição. Andrade et al. (2015) verificaram o método de pré-tratamento da *primeira derivada* como superior a outros métodos para estas duas características, em seu estudo com híbridos naturais de *Eucalyptus urophylla*. Entretanto, para a lignina total no cenário apresentado a *segunda derivada com gap* foi mais eficiente.

Também na tabela 6, para a características de teor de Pentosanas, quatro diferentes pré-tratamentos se destacaram nas análises. A *primeira derivada sem gap* apresentou média com janela igual a 40 e a *primeira derivada com gap* com g igual a 90 e apresentaram, cada um, 20% de eficiência. O pré-tratamento *segunda derivada com g*ap, com valores de g igual 105 e 110, teve eficiência de 40%. Quanto ao R^2 esse variou de 0.035 a 0.23 e o erro médio entre 44.02 a 62.67, apesar de baixos são relativamente bons quando comparados aos dados não tratados. Assim, para a características de teor de Pentosanas, o pré-tratamento *segunda derivada com gap*

foi o mais eficiente. Terdwongworakul (2005), avaliando propriedades químicas da madeira, também obteve suficientes valores com o método da segunda derivada para a análise do conteúdo de pentosanas, sendo o coeficiente do valor de correlação entre valores medidos e preditos, e erro padrão de predição, 0.94 e 0.91, respectivamente.

Por fim, para a característica densidade básica da madeira dois pré-tratamentos se destacaram: o método *segunda derivada com gap*, com valores de g iguais a 70 e 75, e o método *primeira derivada sem gap* (tabela). O primeiro apresentou eficiência de 60% e o segundo 40% de eficiência, respectivamente. No geral os valores de R^2 foram bons, no entanto, apenas no método *segunda derivada com gap* apresentou valores de erro médio mais baixos. Para essa característica, a *segunda derivada com g*ap se mostrou mais eficiente.

Visão geral dos pré-tratamentos mais eficientes por características

Após a análise individual de cada característica, compilou-se na Tabela 7 os pré-tratamentos mais eficientes no conjunto de características em estudos analisando a questão da possibilidade de priorizar um pré-tratamento ou se há um padrão entre elas no que diz respeito ao pré-tratamento que deva ser priorizado. Na Tabela 7 estão relacionadas todas as características analisadas com seus pré-tratamentos mais eficientes. De acordo com o exposto, é possível formar grupos de características de acordo com os pré-tratamentos.

Tabela 7. Relação de pré-tratamento com destaque de maior eficiência para características relacionadas à qualidade da madeira em eucalipto

Característica	**Pré – tratamentos mais eficientes**
Rendimento de Celulose	Detrend signal
Teor de cinzas	Média usando janela de 16, 21 e 36
Teor de extrativos em acetato	Primeira derivada com gap (g = 110 e g = 105) e o Segunda derivada com gap (g = 105 e g = 90)
Teor de extrativos em água	Segunda derivada sem gap
Teor de extrativos totais	Segunda derivada com gap (g = 75 e g = 110)
Holocelulose	Detrend signal
Número Kappa	Segunda derivada com gap (g = 90, g = 95 e g = 105)
Lignina Klason	Segunda derivada com gap (75, 85 e 100)
Lignina Total	segunda derivada com gap (g = 70 e g = 75)
Pentosanas	Segunda derivada com gap (g = 105 e g = 110)
Densidade Básica	Segunda derivada com gap (g = 70 e g = 75)

O pré-tratamento *segunda derivada com gap* foi eficiente para 7 características (Teor de extrativos em acetato, teor de extrativos totais, Número Kappa, Lignina Klason, Lignina Total, Pentosanas e Densidade Básica), ou seja, 64%. De acordo com Milagres (2009), a primeira e a segunda derivada são as que melhor definem os picos presentes em uma amostra que se encontram sobreposto em uma determinada região. Andrade et al., (2015), Li e Altaner (2019), para dados de teor de cinza, extrativos totais e teor de pentosanas, relatam que o método da *segunda derivada* superior a demais métodos de pré-processamentos estudados.

Um segundo grupo pode ser formado com as características de holocelulose e rendimento de celulose que apresentaram, em comum, o método *detrend signal* como pré-tratamento mais eficiente.

As outras duas características Teor de cinzas e Teor de extrativos em água apresentaram os métodos *média* e *segunda derivada sem gap*, respectivamente.

CONCLUSÃO

A avaliação de pré-tratamento para fins de ajustes de modelos de predição considerando características tecnológicas da madeira, com base em informações de NIR e de química da madeira, para fins de seleção indireta de indivíduos de *E. benthamii* permitiu concluir que:

O uso de pré-tratamento é indispensável, pois seu uso proporciona acréscimo nas acurácias preditivas do modelo expressas pelos valores de R^2 e erro quadrático médio, respectivamente.

Diferentes técnicas de pré-tratamento se mostram eficientes considerando informações de diferentes características na população de *E. benthamii.* Estudos prévios para adequação do melhor pré-tratamento é recomendável.

O uso do pré-tratamento envolvendo técnicas de *segunda derivada com gap* se destacou no conjunto de dados analisados e deve ser enfatizado como uma alternativa vantajosa em estudos de ajuste de modelo.

5. ARTIGO 2

Eficiência preditiva de características de qualidade da madeira de Eucalyptus com abordagens de inteligência artificiais aplicadas a dados NIRs

Resumo

As técnicas de como a espectroscopia de infravermelho próximo (*Near Infrared Spectroscopy* – NIR's) têm grande potencial de utilização no melhoramento florestal, pois é um método que contorna as barreiras dos métodos utilizados atualmente, principalmente quando o enfoque é predizer a qualidade da madeira. Os métodos estocásticos são os principais usados nas mais diversas áreas de estudos para predição de características em dados NIR. No entanto, nos últimos anos, os métodos de aprendizado de máquina (*Machine Learning* – ML) vêm sendo utilizados como alternativa para melhorar os valores de acurácia dos modelos, pois captam relações lineares e não lineares entre as variáveis independentes e a variável resposta, além de não necessitarem de pressuposições, como os métodos estocásticos apresentam. Desse modo, o presente trabalho tem por objetivo aplicar métodos de aprendizado de máquinas (*Machine Learning* - ML) na predição de caraterísticas tecnológicas da madeira para avaliar seu potencial em termos de desempenho de predição em comparação com modelos convencionais usados na técnica NIR envolvendo características de qualidade da madeira. Para diferentes características, identifica-se diferentes abordagens de desempenho superior para fins de predição. O procedimento PLS é uma opção de análise a ser considerada, mas seu generalizado não é recomendado. Outras opções podem apresentar resultados comparativamente superiores. O background considerado nos conjuntos de dados de treinamento e validação influenciam resultados de R^2 nos modelos. Validar conjuntos de mesmo background conduz a resultados de eficiência de predição mais elevados. Por fim, analisar dados submetidos a pré-tratamentos proporciona resultados de eficiência de predição mais elevados.

Palavras chave: *Machine Learning*, melhoramento florestal, Predição.

Introdução

Conhecer a base de características de qualidade da madeira é muito vantajosa para o melhorista florestal, visto que o ciclo de melhoramento para espécies florestais é em torno de 15 anos e essas são avaliadas apenas no final de cada ciclo. Neste contexto, técnicas como a espectroscopia de infravermelho próximo (*Near Infrared Spectroscopy* – NIR's) tem grande potencial de utilização no melhoramento florestal, pois é um método que contorna as barreiras dos métodos utilizados atualmente, principalmente, quando o enfoque é predizer a qualidade da madeira. Diversas áreas de estudos já têm aplicado essa técnica, como, por exemplo, em química de alimentos (Núñez-Sánchez et al 2015), alimentação animal (Decruyenaere et al 2015) e solos (Chodak, 2011).

Ferramentas de análise multivariada são amplamente utilizadas na análise de dados espectroscópicos (JinandXu, 2011; Smith-Moritzetal., 2011; Xuetal., 2013; Lupoietal., 2014). Sendo as de maior destaque, a Análise Componente Principal (PCA) e de Mínimos Quadrados Parciais (PLS). Entretanto, para determinadas características, não há trabalhos que mostrem resultados, principalmente por serem utilizados modelos estatísticos que captam relações lineares entre as variáveis independentes e a variável resposta. Portanto, há a necessidade de avaliar a aplicação de novas técnicas como técnicas de inteligência Artificial ou aprendizado de máquina, a fim de buscar melhores soluções para esses tipos de análises e aumentar o entendimento desses métodos nessas análises.

Nos últimos anos, os métodos de aprendizado de máquina (*Machine Learning* – ML) têm sido considerados em diversos contextos. Os métodos ML são modelos não paramétricos que fornecem flexibilidade para se adaptar a associações não lineares e com a capacidade de se adaptar a padrões muito complexos. Métodos *Machine Learning* fazem uso de buscas exaustivas ou heurísticas fundamentadas em amostragens para busca de soluções. O uso de modelos baseados em aprendizado de máquina vem sendo aplicado com sucesso na solução de diversos problemas ligados à genética (Barbosa et al., 2021; Sousa et al., 2021; Tomaz et al., 2018). Estas metodologias se diferem das modelagens estocásticas, por não possuírem pressuposições quanto ao modelo, uma vez que seus resultados dependem do aprendizado e não da distribuição das variáveis em si (Ferreira et al., 2018). As árvores de regressão e seus refinamentos não requerem suposições sobre o modelo.

Além disso, as árvores de regressão permitem captar a não linearidade dos dados e proporcionam interpretações fáceis de resultados, além de fornecer informações sobre quais atributos são mais importantes para previsão.

Desse modo, o presente trabalho tem por objetivo aplicar métodos de aprendizado de máquinas (ML) na predição de caraterísticas tecnológicas da madeira para avaliar seu potencial em termos de desempenho de predição em comparação ao PLS, modelo convencional para modelagem em técnicas NIR.

Materiais e métodos

Material genético e localização do experimento

Os dados utilizados foram provenientes de um teste de progênies de *Eucalyptus benthamii*, da empresa CMPC Celulose Riograndense, localizado no município de Encruzilhada do Sul, no estado do Rio Grande do Sul. Foram avaliadas 87 amostras, sendo 75 de *E. benthamii*, e 12 indivíduos de outras três espécies (quatro de *E. saligna*, quatro de *E. grandis* e quatro de *E. dunnii*).

Preparação das amostras para construção dos modelos de calibração

Amostragem inicial

Para a seleção das amostras que constituíram o modelo de calibração, inicialmente, retirou-se uma alíquota de serragem, pelo método não-destrutivo, de 87 árvores no povoamento (número de árvores vivas e em condições de amostragem) definidas anteriormente. Com o auxílio de uma furadeira e uma broca, a serragem foi coletada em quatro pontos distintos na altura do DAP (diâmetro a altura do peito, sem casca) e acondicionada em sacos de papel.

No laboratório de Espectroscopia de Infravermelho foi medido o espectro das 87 amostras, utilizando espectrômetro de infravermelho próximo BRUKER, modelo MPA. Para tal, a madeira, em forma de serragem, foi compactada manualmente em recipientes adaptados para leitura na esfera de integração. Os recipientes eram providos de fundo de quartzo transparente.

Anteriormente à primeira leitura, o aparelho foi calibrado com os padrões de referência (Background). O processamento e a análise dos dados foram realizados em software OPUS Quant 6.5 (BRUKER).

Determinação das propriedades tecnológicas da madeira

As propriedades químicas da madeira foram coletadas de acordo com as normas apresentadas na Tabela 1.

Tabela 1 - Normativas empregadas para análises químicas de madeira moída. Todas as análises foram feitas com base na madeira seca.

Análise química	Norma utilizada
Teor de cinzas	TAPPI 21 1 OM-02
Extrativos (Totais, Em água e acetato)	TAPPI 280 PM-99
Densidade básica da madeira	SCAN – CM 43:95
Análises Químicas	TAPPI T 257
Lignina Klason e Total	TAPII 222 OM-2 e TAPPI UM-250
Teor de Pentosanas	TAPPI T223 CM-01

Fonte: Centro I+D da CMPC Celulose S.A.

Assim, foi possível obter informações de 11 variáveis: Teores de cinzas, Teores de extrativos (totais, acetona e em água), Densidade Básica da Madeira, Lignina Klason, Lignina Total, Teores de Pentosanas, Celulose, Holocelulose, número kappa e rendimento de celulose. Esse conjunto de variáveis pode ser definido como o representativo das características tecnológicas da madeira.

Pré-tratamentos

Foram testados 15 pré-tratamentos básicos, sendo que, em alguns, foram incorporadas variações em alguns parâmetros, de forma que no estudo testou-se 199 possibilidades de pré-tratamentos, como mostrado na Tabela 2. Assim, um algoritmo computacional foi implementado permitindo fazer uma varredura no conjunto de dados, sob diferentes métodos, e apresentar todas as possibilidades para encontrar o melhor pré-tratamento.

Na Tabela 2 são descritos todos os métodos que foram aplicados aos dados. Após a aplicação dos diferentes métodos o conjunto de dados, foi utilizado para fins de predição de características da qualidade da madeira utilizando o procedimento estatístico PLS (Partial Least Squares). Considerou-se que os maiores R^2 e os menores REQMv da predição evidenciava que o pré-tratamento é o mais apropriado.
Tabela 2. Métodos testados de pré-tratamentos para dos dados espectrais.

	Pré-tratamento Testados	Variação	Parâmetro*
1	Sem nenhum ajuste	1	
2	Média	8	Tamanho da janela
3	Média Móvel	40	Valor m da janela 2m+1
4	Savitzky-Golay	87	Parâmetros m e p
5	Primeira derivada sem gap	1	
6	Segunda derivada sem gap	1	
7	Primeira derivada com gap	11	Parâmetro g
8	Segunda derivada com gap	11	Parâmetro g
9	Normalização	1	
10	Detrend signal	1	
11	Centragem e escalamento - blockScale type soft	1	
12	Centragem e escalamento - blockScale type hard	1	
13	Centragem e escalamento - blockNorm	1	
14	Transformação de Fourier – reflectância	1	
15	Transformação de Fourier – absorbância	1	

(*) Descrição do método de pré-processamento e dos parâmetros encontram-se em SOUSA (2008) e FERREIRA (2015)

As avaliações foram feitas para cada característica em análise e definido o melhor pré-tratamento. No quadro abaixo foram definidos os melhores pré-tratamentos para cada característica.

Característica	**Pré – tratamentos mais eficientes**
Rendimento de Celulose	Detrend signal
Teor de cinzas	Média usando janela de 36
Teor de extrativos em acetato	Primeira derivada com gap (g = 110)
Teor de extrativos em água	Segunda derivada sem gap
Teor de extrativos totais	Segunda derivada com gap (g = 100)
Holocelulose	Detrend signal
Número Kappa	Segunda derivada com gap (g = 105)
Lignina Klason	Segunda derivada com gap (g = 85)
Lignina Total	segunda derivada com gap (g = 70)
Pentosanas	Segunda derivada com gap (g = 105)
Densidade Básica	Segunda derivada com gap (g = 75)

Fonte: Autor;

Desenvolvimento dos modelos de calibração

Os espectros de refletância das amostras, coletados nos diversos comprimentos de onda pelo espectrômetro NIR, foram importados para o computador (software OPUS Quant 6.2).

Foram excluídas as faixas espectrais acima de 10.000 cm^{-1}, posto que, nessa região, o espectro apresenta repetições de ruídos que não apresentam informação

relevante sobre a propriedade de interesse. As amostras anômalas, visivelmente diferentes do restante, foram detectadas como outliers e excluídas do modelo. Em seguida, foram feitas as modelagens usando PLS e as metodologias de ML.

PLS – Partial Least Squat

Os modelos foram ajustados com o número de variáveis latentes (VLs) necessárias para fornecer o melhor ajuste, sem perder a variância dos dados. O número de variáveis adotadas para cada modelo era considerado em função da diminuição do erro padrão médio da validação (REQMv) e do aumento do coeficiente de determinação da validação (R^2).

Árvore de Regressão e seus refinamentos (*Boosting, Bagging e Random Forest*)

A árvore de regressão tem como objetivo subdividir diversas vezes o conjunto de observações de forma que os subgrupos formados subsequentes sejam cada vez mais homogêneos (Breiman et al., 1984). A estrutura da árvore de regressão foi feita pela busca da árvore que levasse a partição dos dados até a formação de grupos homogêneos. Para isso, avaliou-se o quão razoável foi uma dada árvore T através de seu erro quadrático médio, como na equação abaixo:

$$P(T) = \sum_{R} \sum_{k \in R} (y_k - \hat{y}_R)^2$$

em que: $\hat{y}_R$ é o valor predito para a resposta fenotípica da característica e y_k é o valor verdadeiro da característica de cada indivíduo dentro do grupo.

Geralmente, uma única árvore não possui boa precisão preditiva quando comparada com outras abordagens (Rogan et al., 2008). Alguns refinamentos com o intuito de melhorar a performance do modelo de árvore de regressão são apresentados na literatura e apresentam desempenhos superiores (Balta e Topal, 2020; Sousa et al., 2021). Dessa forma, também foi testado a performance preditiva dos modelos *bagging*, *random forest* (RF) e *boosting*.

Um dos problemas apresentados pela árvore de regressão é a grande variabilidade entre os resultados obtidos. Para contornar esse problema o *bagging* é um método que aplica a técnica de bootstrap. Assim, obtém-se B amostras do conjunto de observações, com reposição, adquirindo assim um número de B modelos $\hat{f}^1(x), \hat{f}^2(x), \ldots \hat{f}^B(x)$ (Breiman, 1996). A média aritmética desses modelos irá ser o valor predito no modelo final.

O RF segue a mesma ideia do *bagging*, no entanto, além do conjunto de observações, altera-se também o número de variáveis preditoras ($m = \sqrt{p}$) utilizadas em cada partição. O modelo RF funciona das seguintes maneiras: (1) ele produz subfases dos dados anteriores usando a ferramenta de reamostragem de *bootstrap*, assim diferentes variáveis podem ocupar o topo da árvore; (2) ele gera árvores de decisão aplicando as subfases; e (3) em última análise, ele produz a saída ao fundir os resultados da previsão de todas as árvores de decisão (Chen et al., 2019).

Já o *boosting* cria árvores sequencialmente utilizando informações das árvores anteriores, ao contrário do *bagging* que cria múltiplas árvores independentes, e é realizado com um processamento sequencial. O *boosting* é empregado combinando modelos preditores fracos para produzir melhor precisão preditiva. Os dados incorretos da previsão anterior são classificados como dados "difíceis" e serão usados para o próximo processo de previsão para que o valor de precisão alcance um ponto máximo. Depois que todo o processo de previsão é realizado, todos os modelos são mesclados. O impulso transforma um modelo de preditor fraco em um preditor complexo confiável. As etapas deste processo de aprendizagem são a previsão para regressão, cálculo de erros do resíduo e processo de aprendizagem para processar o resíduo (Syahrani, 2019).

Treinamento e validação

Para a predição das características da madeira, foram gerados 5 cenários. Cada cenário consiste na partição da população com 75 indivíduos sendo 80% para treinamento e 20% para validação. Para cada cenário, foram combinações diferentes. E, para cada cenário, foram aplicadas as 5 metodologias anteriormente descritas. As técnicas foram comparadas com base na média aritmética e no erro padrão médio das estimativas de desempenho dos conjuntos de validação. Os grupos de treinamento e validação foram os mesmos para todos os métodos avaliados, a fim de evitar a influência da variação de grupos aleatórios nos resultados entre um método e outro.

Foram considerados 3 situações, sendo a primeira a predição de indivíduos dentro da mesma espécie, a segunda, a predição de características em espécies diferentes e, a terceira, a predição de indivíduos de espécies diferentes sem o pré-tratamento.

Comparação da eficiência das metodologias

Para avaliar a eficiência das metodologias, foram utilizados os parâmetros de raiz do erro quadrático médio de validação ($REQM$) e o Coeficiente de determinação (R^2).

A raiz do erro quadrático médio é adotada para expressar a acurácia preditiva dos modelos, pois apresenta a vantagem de apresentar os valores do erro na mesma escala da variável de interesse, e é descrita conforme a seguir:

$$REQM = \sqrt{\frac{\sum(\hat{y}-y)^2}{n}}$$

A acurácia seletiva é medida pelo quadrado da correlação entre os valores estimados ($\hat{y}$) e os valores verdadeiros (y), ou seja, mede o quanto a estimativa obtida é relacionada ao valor real do parâmetro, que, em genética quantitativa, expressa a herdabilidade da característica (RESENDE et al., 2012). A acurácia foi dada pela seguinte equação:

$$R^2 = \left(cor(\hat{y}, y)\right)^2$$

A comparação dos resultados para cada cenário foi realizada a partir da média de todos os resultados que estavam contidos na avaliação de cada nível deste efeito, desconsiderando qualquer efeito de interação dos resultados de R^2 e REQMv entre os diferentes cenários.

Aspectos computacionais

As metodologias de PLS, árvore de regressão, *boosting* (iteração 1 e iteração 2), *bagging* e *random forest* foram realizadas com auxílio do software Genes em integração com o software R (R CORE TEAM, 2019; Cruz, 2016).

Resultado e Discussão

No presente trabalho avaliamos a eficiência comparativa da predição de valores de características da madeira, em eucalipto, considerando várias situações e as informações de espectros NIR em um conjunto de genótipos representativos de algumas espécies. A seguir, detalhamos os resultados obtidos em diferentes cenários, considerando valores de eficiência de predição em conjuntos de dados de validação.

Predição dentro da mesma espécie - *Eucalyptus benthamii*

Para este cenário, tanto as informações utilizadas para treinamento ou calibração do modelo quanto para validação, eram provenientes de genótipos representativos da espécie *Eucalyptus benthamii*. Todos apresentados nos gráficos da tabela 3.

A espectroscopia no infravermelho próximo (NIR) é uma técnica que tem sido amplamente aplicada à madeira (Tsuchikawa e Kobori 2015; Schimleck e Tsuchikawa 2020), não destrutiva e pode fornecer uma estimativa do rendimento de celulose. O rendimento de celulose é crítico para a economia da indústria de celulose e papel (Greaves e Borralho 1996) e é muito caro para medir (Meder et al. 2011). Com isso, é importante buscar técnicas de predição cada vez mais precisas para uma maior assertividade. Assim, na Tabela 3 (pag. 67), verifica-se que, para a característica de Rendimento de Celulose, os valores de R^2 variaram de 0.42 (Árvore de Decisão) a 0.22 (*boosting* com uma iteração). Valores relativamente altos em comparação com as demais características. Outros dois métodos que se destacaram, foram os métodos *bagging* com 0.36 de R^2 e o Floresta Aleatória com 0.39, considerados os mais eficientes. A metodologia considerada como tradicional (PLS) teve eficiência de 0.32, similar ao método *boosting* com 2 iterações.

.

Tabela 3. Eficiência de predição, expressas R^2, considerando diferentes características da madeira, mensuradas em *Eucalyptus benthamii*, e obtidas pelas abordagens PLS e de árvores de decisão e seus refinamentos.

2

3

Na literatura, os esforços para melhorar o desempenho da calibração através da utilização de técnicas avançadas ainda é escasso. Mora e Schimleck (2008) utilizaram três técnicas diferentes de seleção de amostras (algoritmos CADEX, DUPLEX e SELECT) para identificar amostras mais representativas de seu conjunto de dados para o desenvolvimento de calibrações de rendimento de polpa. Eles mostraram que o desempenho da calibração foi melhorado utilizando apenas amostras selecionadas e recomendaram que esses métodos fossem empregados para identificar amostras únicas antes de fazer qualquer determinação de propriedades da madeira utilizando modelos baseados em espectros NIR. Mais recentemente, Li et al. (2018) utilizaram uma abordagem de otimização de máquina de vetor de suporte (SVM) e observaram uma previsão de densidade aprimorada para quatro espécies chinesas comercialmente importantes.

O teor de cinzas é um material prejudicial ao processo de polpação da madeira, causando corrosão e incrustações no equipamento industriais, reduzindo o poder calorífico e diminuindo a produtividade industrial (Jardim et al.2017). Para essa característica, os valores de R^2 foram baixos variando entre 0.11 e 0.04, sendo o método de Árvore de Decisão o que proporcionou maior valor e o PLS com menor valor. Como será destacado a seguir, a abordagem fundamentada em árvore de decisão se destacou em quatro (rendimento, teor de cinzas, kappa e pentosanas), das 13 características mensuradas neste trabalho, indicando ser um procedimento útil e que deve ser considerado em trabalhos de melhoramento. Porém, se destacou como pior desempenho (holocelulose, lignina total, lignina Klason e densidade básica) em outras 4 características mensuradas. Este resultado já indica que não se pode apontar um determinado procedimento biométrico como aquele de uso mais apropriado, pois cada situação tem a sua particularidade que deve ser considerada para aplicação de determinada técnica.

O conteúdo dos extratos totais pode ser subdividido em extrativos em acetona e extrativos em água. Em geral, o problema desta variável está relacionado com a sua viscosidade e aderência ao equipamento, exigindo paradas de todo o processo para limpeza (Jardim et al. 2017). Extrativos totais é uma característica com alta correlação entre extrativos em acetato e extrativos em água (Ferraz et al. 2020). Entretanto, tiveram divergências quanto ao método mais eficiente para o cenário em estudo. Para extrativos totais, o PLS com R^2 igual a 0.28 e a metodologia menos eficiente, foi a *bagging* com duas iterações com R^2 igual a 0.04. Para a característica Teor de

Extrativos em Acetato, foi a Floresta Aleatória com 0.35 de R^2 e o método menos eficiente foi o *boosting* com 0.12. A metodologia *bagging* teve um desempenho bom com 0.34 de R^2. Os métodos PLS e Árvore Aleatória tiveram a mesma eficiência quando se refere a R^2 com 0.29. Para extrativos em água, a árvore de decisão foi o melhor método com R^2 igual a 0.22 e o *boosting* com 1 iteração foi o menos eficiente (R^2 = 0.06).

Para a produção de celulose, quanto maior o teor de holocelulose, maior o rendimento e qualidade da celulose, enquanto para a produção de carvão vegetal ocorre o contrário (Protásio et al. 2012). Os resultados de R^2 para a característica variaram entre 0.14 e 0.31, sendo o método PLS, mais tradicional para predição de características por NIR e o método mais eficiente foi o de Árvore de Decisão. Os métodos *bagging*, Floresta Aleatória e *boosting* com iteração igual a 1 tiveram valores de R^2 muito similares.

O número kappa é uma característica relevante em cozimento a lenha, sendo definido como o número de mililitros de solução de permanganato de potássio 0,1 N consumido por grama de polpa de celulose absolutamente seca, sob condições específicas, e corrigido para um consumo 50% de permanganato (D'Almeida 1988). O número kappa apresentou valores de R^2 baixo, de forma geral, apenas a metodologia Árvore de Decisão com 0.15. Os demais métodos apresentaram R^2 abaixo de 0.02 da metodologia Floresta Aleatória.

O teor de lignina é importante para o desempenho da polpação, pois a presença de compostos fenólicos tende a aumentar o consumo de reagentes químicos durante o processo de cozimento e para reduzir o rendimento (Sousa et al. 2019). Para lignina total, as metodologias *bagging* e Floresta aleatória apresentaram a mesma eficiência com R^2 igual a 0.32. Já a metodologia menos eficiente, com R^2 igual a 0.18, foi a de Árvore de Decisão. Os métodos Floresta Aleatória e PLS apresentaram, ambos, 0.24 de R^2. Já a lignina Klason apresentou resultados de R^2 entre 0.38 (PLS) e 0.16 e Árvore de Decisão. A metodologia PLS, que é a tradicional nesse tipo de análise, foi a mais eficiente.

A importância do conteúdo de pentosanas da madeira no processo de polpação está ligado à relevância do teor de hemicelulose (Garcia 1998). Hemiceluloses contribuem para o rendimento e têm benefícios interfibras reações e resistência à polpa celulósica (Souza 2016). Segundo Gomes e cols. (2008), a maioria processos de obtenção de polpa de celulose buscam remover o mínimo possível deste material

devido aos seus benefícios, mas, em casos de produção de celulose solúvel, é importante uma menor quantidade nesse processo. Assim, materiais com alto teor de pentosanas são desejáveis no processo de polpação quando se trata de celulose para papel. Já quando o produto final é voltado para a indústria farmacêutica e alimentícia, essa característica não é desejável, caso esteja em baixa porcentagem. A Árvore de Decisão foi a metodologia mais eficiente para a característica pentosanas com r^2 igual a 0.20. Os demais métodos foram abaixo de 0.02, sendo o mais baixo, o método de Floresta Aleatória com R^2 igual a 0.02.

A densidade básica da madeira foi a característica que apresentou boa eficiência para a maioria dos métodos, variando de 0.44(PLS) a 0.24(Árvore de Decisão). Os valores de R^2, com exceção da Árvore de Decisão, foram muito similares próximos a 0.40. Esse resultado é interessante pela importância dessa característica para a qualidade da madeira já que é considerada uma característica chave na determinação da qualidade da madeira inteira porque apresenta uma forte correlação com outras propriedades da madeira (Panshin; De Zeeuw, 1980), devido a sua importância. A densidade (densidade básica) é normalmente a primeira propriedade da madeira a ser avaliada em um programa de melhoramento de árvores (Lima et al., 2000). De forma geral, a densidade é definida como a massa de madeira seca em estufa por unidade de volume de madeira em condição verde, é um traço crítico de qualidade da madeira para a produção de celulose, papel e serrado madeira (ZOBEL; VAN BUIJTENEN, 1989). No entanto, de acordo com Hein et al. (2010), a densidade não explica por si só o comportamento mecânico da madeira.

De maneira geral, os resultados encontrados revelam que para cada característica analisada tem-se uma técnica que pode ser mais adequada que outra e, portanto, deve-se fazer o processamento de todas elas e, somente diante de seus resultados, proceder a escolha da melhor opção parece ser o mais apropriado. Atualmente, existem aplicativos computacionais que permitem rapidamente processar conjuntos de dados sob diferentes modelos de forma que o pesquisador possa escolher aquele que melhor lhe atende.

A abordagem tradicionalmente utilizada (PLS) se destacou como de eficiência superior em quatro das características analisadas (Teor de extrativos totais, holocelulose, lignina Klason e densidade básica). Porém, teve pior resultado comparativo em relação ao teor de cinzas e Kappa. Deve-se ter em mente que técnicas como PLS evita problemas estatísticos de análises, tais como:

multicolinearidade e dimensionalidade, pois elas tendem a captar apenas relações lineares entre a variável resposta e as variáveis preditoras. Por outro lado, as abordagens fundamentadas em aprendizado de máquina, representadas pela árvore de decisão e seus refinamentos, seriam mais sensíveis a possíveis efeitos não lineares e de interação entre os preditores conduzindo a melhores resultados nesta situação.

Por fim, merece destaque a abordagem de floresta aleatória que, apesar de não ter sido a melhor em muitas situações, esteve como a melhor ou a segunda melhor opção em 7 das 13 situações consideradas.

Estudos sobre correlações genéticas são importantes para a seleção genética, pois verificam se os genes são pleiotrópicos e quais características são correlacionados positiva ou negativamente uma com as outras (Gion et al., 2011; Hung et al., 2015). Ferraz et al. 2020, através do estudo de correlação, conseguiu identificar grupos de características de alta similaridade. Extrativos Totais, Extrativos em água e extrativos em acetato são altamente correlacionados formando um grupo (grupo 1), um outro grupo (grupo 2), formado por lignina total e lignina klason, pois também são altamente correlacionados e um terceiro grupo (grupo 3) pode ser formado pelas características do rendimento de celulose e holocelulose. Já as demais características apresentaram moderada a baixa correlação (Ferraz et al. 2020).

Definido esses grupos de similaridades, é interessante avaliar se os métodos entre as características agrupadas são semelhantes e assim fazer uma predição indireta de outras características através de uma mais fácil ou barata de se obter. O grupo 1, dos teores de extrativos, não apresentou nenhum método de predição em comum quanto a melhor eficiência. Para o grupo 2, com características ligadas ao teor de lignina, também foi similar ao que ocorreu no grupo 1. No grupo 3, constituído por características de produtividade de celulose, Rendimento de celulose e Holocelulose, a diferença dentre essas duas é apenas o conteúdo de pentosanas que está presente na segunda. No entanto, mesmo com essa similaridade os métodos de predição foram diferentes.

Assim, nesse cenário podemos considerar que cada o método e predição é variável de acordo com cada conjunto de dados e precisa ser avaliado individualmente e em cada novo cenário. As correlações nos permitiram observar que não há similaridade entre métodos em relação características altamente correlacionadas.

Uma análise alternativa é saber se o background genético influência na predição e nos métodos.

Predição com informações de diferentes espécies - *Eucalyptus benthamii* (treinamento) e *E. saligna*, *E. grandis* e *E. dunnii* (validação)

Para este cenário, as informações utilizadas para treinamento foram as mensuradas em indivíduos da espécie *Eucalyptus benthamii* mas, para fins de validação, foram utilizadas informações de indivíduos de outras espécies. Deve ser feita a ressalva de que se trata de uma geração de informações preliminares tendo em vista o tamanho amostral para representar as espécies de validação representadas por 12 indivíduos, sendo quatro de *E. saligna*, quatro de *E. grandis* e quatro de *E. dunnii*. No entanto, como é uma ferramenta ainda no processo de introdução nos programas de melhoramento, essa é uma realidade que acorre geralmente nas empresas.

O método mais eficiente para o Rendimento de Celulose foi a *boosting* com iteração igual a 1 com R^2 igual a 0.26. O método PLS, a tradicional, foi a metodologia menos eficiente com R^2 menor que 0.01. Este resultado já demonstra algumas particularidades importantes que devem ser consideradas pelo pesquisador em sua análise de dados. O primeiro fato é que o R^2 da melhor técnica foi inferior ao obtido para a mesma característica, quando o background genético utilizado para fins de validação era o mesmo do utilizado para fins de treinamento, ou seja, ambos eram representados por genótipos de *Eucalyptus benthamii.* O segundo fato é a inconsistência da superioridade de uma técnica biométrica. Na Tabela 3, verificamos que a melhor técnica era Árvore de Decisão, mas na Tabela 4, para a mesma característica, destaca é o *boosting* com 1 iteração.

A tabela 4 mostra os valores de R^2 para todas as características em análise, mostrando os mais eficientes e o menos eficiente para cada variável. Para o Teor de cinzas, se destacou o método PLS com R^2 de 0.27. A situação foi similar para a característica Extrativos em Acetato, sendo PLS com R^2 igual a 0.15. Já o mais eficiente método para a característica extrativos em água foi a *bagging* com R^2 igual a 0.09. O método de Floresta Aleatória foi o mais eficiente para Extrativos Totais com R^2 de 0.31 a fim de predizer características de outras espécies. Na caraterística Holocelulose, o método PLS se destacou na eficiência com R^2 igual 0.68. A característica número kappa apresentou R^2 com valor de 0.21 no método PLS. Para

Lignina Total, o método mais eficiente com R^2 igual a 0.21 foi o método *boosting* com uma iteração. Para lignina klason, com 0.09 de R^2 o Árvore de Decisão foi o método mais eficiente. Para Teor de Pentosanas, o valor de R^2 foi de 0.26 usando se o método PLS e, para as demais características, os valores foram abaixo de 0.10. Por fim, para a Densidade Básica o método mais eficiente foi o PLS com R^2 igual a 0.14, relativamente alto em relação aos demais métodos.

Tabela 4. Eficiência de predição, expressas pelo coeficiente de determinação R^2, considerando 11 características da madeira, de Eucalyptus benthamii (dados de treinamento) e em E. saligna, E. grandis e E. dunnii (dados de validação), por abordagens PLS e de Árvores de Decisão e seus refinamentos.

A tabela 4 mostra os valores de eficiência de predição obtidos em ajuste de modelo usando mesmo background genético ou diferentes para treinamento e validação. Em questão de eficiência dos valores pelo método de R^2, o cenário em que as validações foram realizadas no mesmo background apresentaram melhores valores de R^2 com cerca de 55% das características sendo Rendimento de Celulose, Teor de Extrativos em Acetato e Teor de extrativos totais, lignina total, lignina klason e densidade básica da madeira. Ao se avaliar predição de diferentes backgrounds, as caraterísticas Teor de Cinzas, Teor de Extrativos Totais, holocelulose, número kappa e pentosanas foram as características que apresentaram melhores R^2.

Tabela 5. Valores de eficiência de predição obtidos em ajuste de modelo usando mesmo background genético ou diferentes para treinamento e validação

	Mesmo background/ou dentro da mesma espécie		Diferentes background/usando espécies diferentes	
Característica	R2	Técnica	R2	Técnica
Rendimento de Celulose	0.42	Árvore de Decisão	0.26	boosting_1
Teor de cinzas	0.11	Árvore de Decisão	0.27	PLS
Teor de extrativos em acetato	0.35	Floresta Aleatória	0.15	PLS
Teor de extrativos em água	0.22	Árvore de Decisão	0.09	Bagging
Teor de extrativos totais	0.28	PLS	0.31	Floresta Aleatória
Holocelulose	0.31	PLS	0.68	PLS
Número Kappa	0.15	Árvore de Decisão	0.21	PLS
Lignina Total	0.32	Floresta Aleatória	0.21	boosting_1
Lignina Klason	0.38	PLS	0.09	Árvore de Decisão
Pentosanas	0.20	Árvore de Decisão	0.26	PLS
Densidade Básica	0.44	PLS	0.14	PLS

Quanto aos métodos de predição, apenas as características de Holocelulose e Densidade Básica da madeira apresentaram o mesmo método de predição que foi o PLS. Já as demais características não apresentaram essa mesma consistência.

Afim de entender e relacionar os métodos testados com as características, será feito uma comparação destes com as herdabilidade de cada uma. A herdabilidade é um parâmetro genético imperativo, uma vez que quantifica a fração hereditária da variação fenotípica, que pode ser explorada na seleção (Resende, 2004). As magnitudes de herdabilidade individual são classificadas como baixas ($0,01 \leq h2a \leq 0,15$), moderado ($0,15<h2a<0,50$) e alto ($h2a \geq 0,50$) (Resende, 2007).

Há vários trabalhos que mostram as herdabilidades de características da madeira (Gallo et al., 2018; Hein et al., 2010; Stackpole et al., 2011; Varghese et al., 2017). Algumas características têm herdabilidade bem definidas, como Rendimento de Celulose, que possui herdabilidade moderada, teor de lignina que possui alta herdabilidade e Teor de extrativos totais com baixa herdabilidade. Assim, podemos relacionar os métodos de predição com herdabilidade de cada características.

Primeiramente para cada tipo de herdabilidade houve perda no R^2 quando a validação foi em um background de base genética diferente como observado na tabela 5. Quando se avalia o R^2 no mesmo background para características de alta herdabilidade, se destacou o método Floresta aleatória, para a de moderada herdabilidade o método de Árvore de decisão e a baixa herdabilidade a metodologia de PLS. Ao se analisar o R^2 para diferentes backgrounds, observamos a redução nos valores e que para baixas herdabilidades o método Floresta Aleatória foi mais eficiente. Já para moderada herdabilidade, foi o *boosting* com iteração igual a 1 e com altas herdabilidade que também apresentou este resultado.

Os efeitos genéticos e ambientais sobre as técnicas utilizadas podem ser apreciados considerando os valores de acurácia nos cenários de diferentes herdabilidade. O impacto da herdabilidade sobre os resultados de R^2 foram destacados em outros trabalhos (Alves et al., 2020; Barbosa et al., 2021). De acordo com Barbosa et al., (2021) o *boosting* não é uma boa estratégia, principalmente para dados com baixa herdabilidade e com maior influência da variância residual, uma vez que é treinada repetidamente na mesma amostra para que a cada iteração, uma medida de erro de previsão seja calculada para cada indivíduo, e, na próxima iteração, indivíduos com maiores erros recebam maior peso no treinamento do modelo.

Predição em outras espécies sem uso de pré-tratamento

Nesse tópico foram realizadas análises semelhantes ao apresentado anteriormente envolvendo a predição em outras espécies que constituíam o conjunto de dados de validação de forma a entender se o background influencia na predição entre espécies diferentes, além de comprovar a superioridade dos pré-tratamentos. Os resultados das análises anteriores foram obtidos a partir de informações submetidas a pré-tratamento, ou seja, nesse caso, foram utilizados os dados brutos. Foram avaliados os valores de R^2 (Tabela 6) para cada característica, avaliando os

parâmetros mais eficientes para cada um. Por fim, definiu de acordo com R^2 qual o método de predição mais eficiente para cada variável.

Para Rendimento de Celulose, o melhor valor de R^2 foi de 0.19 no método de Árvore de Decisão e o menor valor obtido foi pelo método de *boosting,* com iteração igual a 1 com valor de R^2 abaixo de 0.01. O PLS, método mais tradicional, obteve valor de 0.08, ficando em quarto lugar ao ranquearmos os do método mais eficiente para o menos. A partir desta informação básica, verifica-se que o pré-tratamento impacta tanto a eficiência da predição quanto influencia o desempenho das abordagens biométricas.

Na característica Teor de Cinzas, os métodos PLS e *boosting* com iteração igual a 1 foram os mais eficientes com 0.32 de R^2. O destaque na característica teor de extrativos em Acetato vai para os métodos de *bagging* e *boosting* com iteração igual a 1. O método Árvore de Decisão foi o método de predição mais eficiente para a característica Teor de extrativos em Água com 0.07 de R^2. As características Teor de extrativos Totais e Holocelulose apresentaram PLS como método de melhor desempenho com R^2 de 0.21 e 0.13, respectivamente. Para as duas características, o restante dos métodos apresentou valores de R^2 muito baixos. A Árvore de Decisão com R^2 igual a 0.21 foi o método mais eficiente para a características número kappa. Para os valores de R^2, a características lignina total variaram de 0.39, *boosting* com iteração igual a 2, à 0.19, Árvore de Decisão e PLS, sendo respectivamente método mais eficiente e método menos eficiente. Com R^2 igual 0.44, Floresta Aleatória foi o método mais eficiente para lignina klason, sendo que o método *bagging* teve desempenho bem próximo com R^2 igual a 0.43. *Boosting* com iteração igual a 2 teve R^2 de 0.15 apresentando maior eficiência dentre os métodos para a característica pentosanas. Por fim, a densidade básica teve o PLS como método mais eficiente com R^2 igual a 0.57 e o método menos eficiente foi o *boosting* com iteração igual a 1 com R^2 igual a 0.16.

Tabela 6. Eficiência de predição, expressas pelo coeficiente de determinação R2, considerando diferentes características da madeira, mensuradas em *Eucalyptus benthamii* (dados de treinamento) e em *E. saligna, E. grandis e E. dunnii* (dados de validação), e obtidas pelas abordagens PLS e de árvores de decisão e seus refinamentos.

Na tabela 7 podemos observar a superioridade das análises com pré-tratamentos. É valido lembrar que as análises foram orientadas de acordo com essa metodologia. Quando não se fez o uso de pré–tratamento, as metodologias de boosting e Árvore de Decisão se destacaram e em alguns casos até melhoraram o R^2 das características. Assim, em casos do não conhecimento das técnicas de pré-tratamento dos dados podem se optar por esses modelos.

Tabela 7. Valores de eficiência de predição obtidos em ajuste de modelo usando conjunto de dados sem e com pré-tratamento.

	Sem pré-tratamento		Com pré-tratamento	
Característica	R2	Técnica	R2	Técnica
Rendimento de Celulose	0.19	Árvore de Decisão	0.26	boosting_1
Teor de cinzas	0.32	boosting_1	0.27	PLS
Teor de extrativos em acetato	0.12	boosting_1	0.15	PLS
Teor de extrativos em água	0.07	Árvore de Decisão	0.09	Bagging
Teor de extrativos totais	0.21	PLS	0.31	Floresta Aleatória
Holocelulose	0.13	PLS	0.68	PLS
Número Kappa	0.11	Árvore de Decisão	0.21	PLS
Lignina Total	0.39	boosting_2	0.21	boosting_1
Lignina Klason	0.44	Árvore de Decisão	0.09	Árvore de Decisão
Pentosanas	0.15	boosting_2	0.26	PLS
Densidade Básica	0.57	PLS	0.14	PLS

Conclusão

Para diferentes características, identifica-se diferentes abordagens de desempenho superior para fins de predição. O procedimento PLS é uma opção de análise a ser considerada, mas seu generalizado não é recomendado. Outras opções podem apresentar resultados comparativamente superiores. O background considerado nos conjuntos de dados de treinamento e validação influenciam nos resultados. Validar conjuntos de mesmo background conduz a resultados de eficiência de predição mais elevados, necessitando aumentar o tamanho amostral para melhor essa eficiência. O pré-tratamento de dados influencia os resultados. Analisar dados submetidos a pré-tratamento proporciona resultados de eficiência de predição mais elevados.

6. CONCLUSÃO GERAL

O uso de pré-tratamento é indispensável, pois proporciona acréscimo nas acurácias seletivas e preditivas do modelo expressas pelos valores de R^2 e erro quadrático médio, respectivamente. Diferentes técnicas de pré-tratamento se mostram eficientes considerando informações de diferentes características na população de *E. benthamii*. Estudos prévios para adequação do melhor pré-tratamento é recomendável.

Para diferentes características, identifica-se diferentes abordagens de desempenho superior para fins de predição. O procedimento PLS é uma opção de análise a ser considerada, mas seu generalizado não é recomendado. Outras opções podem apresentar resultados comparativamente superiores. O background considerado nos conjuntos de dados de treinamento e validação influenciam nos resultados. Validar conjuntos de mesmo background conduz a resultados de eficiência de predição mais elevados.

7. CONSIDERAÇÕES GERAIS

A principal contribuição científica deste trabalho refere-se à ampliação da base de conhecimento relacionada ao comportamento e potencialidade das metodologias baseadas em aprendizado de máquinas comparadas a metodologias tradicionalmente aplicadas. Esses conhecimentos podem fornecer subsídios para a escolha de metodologias mais apropriadas para a predição de características tecnológicas da madeira, tornando mais eficientes o uso destas dentro de um programa de melhoramento genético.

A execução deste trabalho permitiu adquirir conhecimentos para um melhor entendimento da influência dos diferentes cenários e suas consequências para os resultados de R^2 e REQMv. Além disso, permitiu avaliar diferentes métodos de aprendizado de máquinas para 11 diferentes características tecnológicas da madeira, nunca realizado em um trabalho científico desta área. Espera-se ainda que os resultados possam ser motivadores, no sentido de outros pesquisadores optarem pelo uso dos métodos, de forma que possam agregar conhecimento e aumentar a eficiência e ampliação do uso dessas técnicas nos programas de melhoramento.

Os conhecimentos gerados por este trabalho poderão ser utilizados como literatura para auxiliar no desenvolvimento de cientistas em diversas áreas de estudo, como na genética, estatística, bioinformática, melhoramento de plantas, entre outros. Esse trabalho fornece, ainda, oportunidades para novas perguntas e questionamentos que devem ser resolvidos em pesquisas futuras, a fim de obter um ganho contínuo no conhecimento científico.

8. REFERÊNCIAS BIBLIOGRÁFICAS

Alves AAC, Costa RM, Bresolin T, Fernandes Júnior GA, Espigolan R, Ribeiro AMF, Carvalheiro R and Albuquerque LG (2020) Genome-wide prediction for complex traits under the presence of dominance effects in simulated populations using GBLUP and machine learning methods. **American Society of Animal Science** 1–34.

Balta, B., & Topal, M. (2020). describing factors affecting birth weight and growth traits in hemsin lambs using decision tree methods. japs, journal of Animal and Plant Sciences, 30(3), 560-567.

Barbosa, I. D. P., da Silva, M. J., da Costa, W. G., de Castro Sant'Anna, I., Nascimento, M., & Cruz, C. D. (2021). Genome-enabled prediction through machine learning methods considering different levels of trait complexity. **Crop Science**, 61(3), 1890-1902.

Breiman, L. (1996). *Bagging* predictors. **Machine learning**, 24(2), 123-140.

Breiman, L., & Ihaka, R. (1984). **Nonlinear discriminant analysis via scaling and ACE**. Davis One Shields Avenue Davis, CA, USA: Department of Statistics, University of California.

Chen, H., Qiao, H., Xu, L., Feng, Q., & Cai, K. (2019). A fuzzy optimization strategy for the implementation of RBF LSSVR model in vis–NIR analysis of pomelo maturity. **IEEE Transactions on Industrial Informatics**, 15(11), 5971-5979.

Chodak, M. (2011). Near-infrared spectroscopy for rapid estimation of microbial properties in reclaimed mine soils. **Journal of Plant Nutrition and Soil Science**, 174(5), 702-709.

Cruz CD (2012) **Princípios de genética quantitativa**, 2. ed. Editora UFV, Viçosa, MG, 394p.

Cruz CD (2016) Genes software – extended and integrated with the R, Matlab and Selegen. **Acta Scientiarum - Agronomy** 38: 547–552.

D'Almeida MLO (1988) Tecnologia de fabricacão de pasta celulósica. Instituto de Pesquisa Tecnológicas do Estado de São Paulo 1:45–106

Decruyenaere, A., Decruyenaere, P., Peeters, P., Vermassen, F., Dhaene, T., & Couckuyt, I. (2015). Prediction of delayed graft function after kidney transplantation: comparison between logistic regression and machine learning methods. **BMC medical informatics and decision making**, 15(1), 1-10.

Ferraz, A. G., Cruz, C. D., dos Santos, G. A., Nascimento, M., Baldin, T., dos Santos, O. P., ... & dos Santos, C. E. M. (2020). Potential of a population of Eucalyptus benthamii based

on growth and technological characteristics of wood. Euphytica, 216(6), 1-15.

Ferreira RADC, Silva GN, Glória LS, Sant'anna IC, Rodrigues HS, Silva FF and Cruz CD (2018) RNA - Aplicação em Estudos de Seleção Genômica Ampla. Editora UFV, Viçosa, MG, p. 414. **In Cruz CD and Nascimento M (eds) Inteligência Computacional Aplicado ao Melhoramento Genético.**

Ferreira, M. M. C. Quimiometria: Conceitos, Métodos e Aplicações. 1 ed. Campinas: Unicamp, 2015. 496p.

Ferreira, Roberta de Amorim. **Comparação de métodos de seleção de variáveis em regressão aplicados a dados genômicos e de espectroscopia NIR**. 2018.

Gallo, R., Pantuza, I. B., dos Santos, G. A., de Resende, M. D. V., Xavier, A., Simiqueli, G. F., ... & Valente, B. M. D. R. T. (2018). Growth and wood quality traits in the genetic selection of potential *Eucalyptus dunnii* Maiden clones for pulp production. **Industrial Crops and Products**, 123, 434-441.

Garcia, Silvana Lages Ribeiro. Importância de características de crescimento, de qualidade da madeira e da polpa na diversidade genética de clones de eucalipto. 1998. **Tese de Doutorado.** Universidade Federal de Viçosa.

Gierlinger, N., Schwanninger, M., Hinterstoisser, B., & Wimmer, R. (2002). Rapid determination of heartwood extractives in Larix sp. by means of Fourier transform near infrared spectroscopy. **Journal of Near infrared spectroscopy**, 10(3), 203-214.

Gion, J. M., Carouché, A., Deweer, S., Bedon, F., Pichavant, F., Charpentier, J. P., ... & Plomion, C. (2011). Comprehensive genetic dissection of wood properties in a widely-grown tropical tree: Eucalyptus. **Bmc Genomics**, 12(1), 1-19.

Gomes FJB, Gouveˆa AFG, Colodette JL, Gomide JL, Carvalho AMML, Trugilho PF, Rosado AM (2008) Influence of content and S/G relation of the wood lignin on kraft pulping performance. **O Papel** 12:95–105

Greaves, B. L., & Borralho, N. M. G. (1996). The influence of basic density and pulp-yield on the cost of eucalyptus kraft pulping: a theoretical model for tree breeding. In **50th Appita Annual General Conference & 1996 Pan Pacific Conference** (Vol. 50, pp. 859-864).

HEIN PRG, BRANCHERIAU L, TRUGILHO PF, LIMA JT, CHAIX G (2010) Ressonância e espectroscopia de infravermelho próximo para avaliar as propriedades dinâmicas da madeira Journal Near Infrared Spectroscopy, 18, pp. 443 – 454.

Hung, T., Pratt, G. A., Sundararaman, B., Townsend, M. J., Chaivorapol, C., Bhangale, T., ... & Behrens, T. W. (2015). The Ro60 autoantigen binds endogenous retroelements and

regulates inflammatory gene expression. Science, 350(6259), 455-459.

Jardim JM, Gomes FJB, Colodette JL, Brahim BP (2017) Avaliação da qualidade e desempenho de clones de eucalipto na producão de celulose. O Papel 78(11):122–129

Li B, Zhang N, Wang YG, George AW, Reverter A and Li Y (2018) Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. **Frontiers in Genetics** 9: 1–20.

Lima, J. T., Breese, M. C., & Cahalan, C. M. (2000). Genotype-environment interaction in wood basic density of Eucalyptus clones. **Wood Science and Technology**, 34(3), 197-206.

Lupoi, J. S., Singh, S., Simmons, B. A., & Henry, R. J. (2014). Assessment of lignocellulosic biomass using analytical spectroscopy: an evolution to high-throughput techniques. **BioEnergy Research**, 7(1), 1-23.

Ma, T., Schimleck, L., Inagaki, T., & Tsuchikawa, S. (2021). Rapid and nondestructive evaluation of hygroscopic behavior changes of thermally modified softwood and hardwood samples using near-infrared hyperspectral imaging (NIR-HSI). **Holzforschung**, 75(4), 345-357.

Mora, C. R., Schimleck, L. R., & Isik, F. (2008). Near infrared calibration models for the estimation of wood density in Pinus taeda using repeated sample measurements. **Journal of Near Infrared Spectroscopy**, 16(6), 517-528.

Muñiz GIBD, Magalhães WLE, Carneiro ME and Viana LC (2012) Fundamentos e estado da arte da espectroscopia no infravermelho próximo no setor de base florestal. **Ciência Florestal**, 22, 865-875.

Panshin, A. J., & De Zeeuw, C. (1980). Textbook of wood technology: structure, identification, properties, and uses of the commercial woods of the United States and Canada.

Protásio TDP, Tonoli GHD, Guimarães Júnior M, Bufalino L,Couto AM, Trugilho PF (2012) Correlações canônicas entre as características químicas e energéticas de resíduos lignocelulósicos. **Cerne** 18(3):433–439

R Core Team (2020) R: A Language and Environment for Statistical Computing. **R Foundation for Statistical Computing**, Vienna, Austria.

RESENDE, M. D. V. (2004). Métodos estatísticos ótimos na análise de experimentos de campo. **Embrapa Florestas**-Documentos (INFOTECA-E).

Resende, M. D. V., & Duarte, J. B. (2007). Precisão e controle de qualidade em experimentos

de avaliação de cultivares. **Embrapa Florestas**-Artigo em periódico indexado.

Resende, M. D., Resende Jr, M. F., Sansaloni, C. P., Petroli, C. D., Missiaggia, A. A., Aguiar, A. M., ... & Grattapaglia, D. (2012). Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. **New Phytologist**, 194(1), 116-128.

Resende, R. T., Resende, M. D. V., Silva, F. F., Azevedo, C. F., Takahashi, E. K., Silva-Junior, O. B., & Grattapaglia, D. (2017). Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in Eucalyptus. **New Phytologist**, 213(3), 1287-1300.

Rogan, J., Franklin, J., Stow, D., Miller, J., Woodcock, C., & Roberts, D. (2008). Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. **Remote Sensing of Environment**, 112(5), 2272-2283.

Schimleck, L. R., & Tsuchikawa, S. (2021). Application of NIR Spectroscopy to Wood and Wood-Derived Products. In **Handbook of Near-Infrared Analysis** (pp. 759-780). CRC Press.

Smith-Moritz, A. M., Chern, M., Lao, J., Sze-To, W. H., Heazlewood, J. L., Ronald, P. C., & Vega-Sánchez, M. E. (2011). Combining multivariate analysis and monosaccharide composition modeling to identify plant cell wall variations by Fourier transform near infrared spectroscopy. **Plant Methods**, 7(1), 1-13.

Sousa, I.C.; Nascimento, M.; Silva, G.N.; Nascimento, A.C.C.; Cruz, C.D.; Silva, F.F.E.; DE Almeida, D.P.; Pestana, K.N.; Azevedo, C.F.; Zambolim, L.; Caixeta, E.T. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. **Scientia Agricola**, 78(4), 2021.

Sousa, R. R. D., Gouveia, J. R., Nacas, A. M., Tavares, L. B., Ito, N. M., Moura, E. N. D., ... & Santos, D. J. D. (2019). Improvement of polypropylene adhesion by kraft lignin incorporation. **Materials Research**, 22.

Souza, G. S. L. B. D. (2016). Efeito da impregnação prolongada dos cavacos no rendimento e branqueabilidade da polpa kraft de eucalipto.

Stackpole, D. J., Vaillancourt, R. E., Alves, A., Rodrigues, J., & Potts, B. M. (2011). Genetic variation in the chemical components of Eucalyptus globulus wood. G3: Genes| **Genomes| Genetics**, 1(2), 151-159.

Syahrani, I. M. Comparation Analysis of Ensemble Technique With boosting(Xgboost) and Bagging (Randomforest) For Classify Splice Junction DNA Sequence Category. J. Penel.

Pos dan Inform. 9, 27–36, 2019.

TAPPI T-21 OM-2 (2002) Effect of cooking temperature on kraft pulping of hardwood. TAPPI Press, Atlanta, GA

TAPPI T-222 OM-2 (2002) Acid-insoluble lignin in wood and pulp. TAPPI Press, Atlanta, GA

TAPPI T-223 CM-01 (1999) Pentosans in wood and pulp. TAPPI Press, Atlanta, GA

TAPPI T-280 PM-99 (1999) Acetone extractives of wood and pulp. TAPPI Press, Atlanta, GA

TAPPI UM250 (1991) ‘‘Acid-soluble lignin in wood and pulp’’,TAPPI Useful Method 250. TAPPI Press, Atlanta, GA

Tomaz RS, Alvez DP, Nascimento M and Cruz CD (2018) Inteligência Computacional. Editora UFV, Viçosa, MG, p. 414. In Cruz CD and Nascimento M (eds) Inteligência Computacional Aplicado ao Melhoramento Genético.

Tsuchikawa, S., & Kobori, H. (2015). A review of recent application of near infrared spectroscopy to wood science and technology. Journal of Wood Science, 61(3), 213-220.

Tumari, M. M., Saat, S., Kasno, A., Syahrani, M., Bahari, M. F., & Ahmad, M. A. (2019). Single input fuzzy logic controller for liquid slosh suppression. Int. J. Electr. Eng. Appl. Sci., 2(1), 45-52.

Varghese, M., Harwood, C. E., Bush, D. J., Baltunis, B., Kamalakannan, R., Suraj, P. G., ... & Meder, R. (2017). Growth and wood properties of natural provenances, local seed sources and clones of Eucalyptus camaldulensis in southern India: Implications for breeding and deployment. **New Forests**, 48(1), 67-82.

Xu, F., Yu, J., Tesso, T., Dowell, F., & Wang, D. (2013). Qualitative and quantitative analysis of lignocellulosic biomass using infrared techniques: a mini-review. Applied energy, 104, 801-809.

Zobel, B. J., & Buijtenen, J. P. V. (1989). Wood variation and wood properties. In Wood variation (pp. 1-32). **Springer**, Berlin, Heidelberg.