LI Congresso Brasileiro de Engenharia Agrícola - CONBEA 2022
Pelotas Parque Tecnológico - Pelotas - RS
27 a 29 de outubro de 2022

DIFERENCIAÇÃO DE MUDAS DE CLONES DE EUCALIPTOS POR MEIO DA ESTATÍSTICA CLÁSSICA E DE MINERAÇÃO DE DADOS

ISABELA C. TOLEDO[1], CINTHYA BERTOLDO[2], RAFAEL M. LORENSANI[3], JOÃO PAULO GIANASI[4],

[1] Estudante, Faculdade de Engenharia Agrícola (FEAGRI/UNICAMP), i218077@dac.unicamp.br

[2]Professora doutora, Faculdade de Engenharia Agrícola (FEAGRI/UNICAMP), cinthyab@unicamp.br

[3]Pesquisador doutor, Faculdade de Engenharia Agrícola (FEAGRI/UNICAMP), rafaelmansini@hotmail.com

[4] Estudante, Faculdade de Engenharia Agrícola (FEAGRI/UNICAMP), j237556@dac.unicamp.br.

Apresentado no
LI Congresso Brasileiro de Engenharia Agrícola - CONBEA 2022
27 a 29 de outubro de 2022 - Pelotas - RS, Brasil

RESUMO: A estatística clássica é a base de grande parte de pesquisas relacionadas à qualidade da madeira e a diferenciação entre clones. Como alternativa à estatística clássica, técnicas de mineração de dados, como o *Machine Learning*, permitem prever e compreender aspectos dos dados observados de uma forma diferente, viabilizando a análise de uma nova perspectiva. O objetivo dessa pesquisa científica foi comparar modelos de diferenciação de mudas de clones de eucalipto gerados através de métodos estatísticos e métodos na ciência de dados. Para a pesquisa foram utilizados mudas de 28 diferentes clones de eucalipto que foram ensaiados por equipamento de ultrassom e transdutores exponenciais de 45 kHz de frequência, também foram mensurados dados de altura e diâmetro dos indivíduos. Os dados dos ensaios foram então avaliados tanto pela estatística clássica como por técnicas de mineração de dados. Os resultados mostraram que o modelo de mineração de dados (Árvore de Decisão), apresentou melhores métricas de classificação quando comparado com análise realizada utilizando a estatística clássica, demonstrando uma oportunidade dessa técnica de avaliação de dados para a indústria de transformação da madeira.

PALAVRAS-CHAVE: plantas jovens; árvore de decisão; ultrassom.

DIFFERENTIATION OF EUCALYPTUS CLONE SEEDLINGS THROUGH CLASSICAL AND DATA MINING STATISTICS

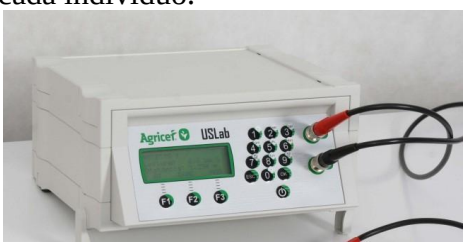
ABSTRACT: Classical statistics are the basis of much of the research related to wood quality and the differentiation between clones. As an alternative to classical statistics, data mining techniques, such as Machine Learning, allow you to predict and understand aspects of the observed data in a different way, enabling analysis from a new perspective. The objective of this scientific research was to compare models of differentiation of seedlings of eucalyptus

clones generated through statistical methods and methods in data science. For the research, seedlings of 28 different eucalyptus clones were used, which were tested by ultrasound equipment and exponential transducers of 45 kHz frequency, height and diameter data of the individuals were also measured. The trial data were then evaluated by both classical statistics and data mining techniques. The results showed that the data mining model (Decision Tree) presented better classification metrics when compared to the analysis performed using classical statistics, demonstrating an opportunity of this data evaluation technique for the wood processing industry.

KEYWORDS: young plants; data mining; ultrasound.

INTRODUÇÃO: A estatística clássica é a base de grande parte de pesquisas relacionadas às propriedades da madeira e sua diferenciação entre clones. Conforme resultados anteriores do Grupo de Pesquisa em Ensaios Não Destrutivos da Faculdade de Engenharia Agrícola (FEAGRI), ensaios de ultrassom em árvores e mudas possibilitaram, em conjunto com análises baseadas na estatística clássica, a separação de clones de eucalipto por rigidez (GONÇALVES, 2013) e a inferência de propriedades da madeira (GONÇALVES, 2018 e 2019). Embora esse método analítico apresenta suas vantagens, também possui limitações quanto à compreensão da evolução e relação mútua dos parâmetros resultantes dos ensaios de ultrassom. Como alternativa, técnicas de mineração de dados, como o *Machine Learning*, permitem prever e compreender aspectos dos dados observados de uma forma diferente, viabilizando uma análise desses parâmetros sob uma nova perspectiva (PLAS, 2016 e GUIDO & MÜLLER, 2016). Avaliar a diferença entre os resultados dos modelos pode contribuir para uma melhor compreensão e uso dos métodos analíticos. O objetivo dessa pesquisa científica foi comparar modelos de diferenciação de mudas de clones de eucalipto gerados através de métodos estatísticos e métodos na ciência de dados.

MATERIAL E MÉTODOS: Foram avaliados 28 clones de eucalipto, com idades variadas e plantados em diferentes localidades do estado de São Paulo. Os ensaios foram realizados pelo grupo de pesquisa, sendo que os dados precisaram ser tabulados e organizados em planilhas em função da análise (estatística clássica ou mineração de dados) a ser realizada. Esses clones foram ensaiados com um equipamento de ultrassom (Figura 1a) e transdutores exponenciais modificados, para perfurar as mudas de tamanho reduzido (Figura 1b). Por meio do ensaio de ultrassom nas mudas dos clones de eucalipto, foi possível obter o tempo de propagação de onda em cada indivíduo.



(a)

(b)

Figura 1. Equipamento de ultrassom USLab (Agricef, Brasil) (a) e transdutores exponenciais com ponteira modificada (b).

Também foram mensurados a altura até a primeira bifurcação, altura total, diâmetros na base (engastamento com o solo), no ponto central do fuste e na primeira bifurcação. Todos esses dados foram avaliados por meio de software estatístico para a proposição de modelos capazes de separar os clones de acordo com os parâmetros medidos, utilizando conceitos de intervalo de confiança, Multiple Range Test e ANOVA, bem como avaliados utilizando técnicas de

mineração de dados, avaliando os dados pelos seguintes métodos: K-Nearest Neighbor (KNN), Decision Tree, Random Forest, Gradient Boosting, Xtreme Gradient Boosting, Superior Vector Machine e Redes Neurais. Para avaliar os métodos de mineração de dados utilizamos a matriz de confusão (Tabela 1) e os conceitos de acurácia, precisão, recall e f1-score.

Tabela 1. Exemplo de uma matriz de confusão.

		PREVISTO	
		SIM	**NÃO**
REAL	**SIM**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	NÃO	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Acurácia: Representa a performance geral do modelo, mostrando dentre todas as avaliações, quais o modelo acertou: $C = \frac{VP+VN}{VP+VN+FP+FN}$

Precisão: Avaliando todas as classificações positivas quantas estão positivas: $PR = \frac{VP}{VP+FP}$

Recall ou Sensibilidade: Dentre todas as avaliações tidas nos valores previstos, quantas estão corretas: $RE = \frac{VP}{VP+FN}$

F1-Score: Médias harmônica entre Precisão e Recall: $F1 = \frac{2*PR*RE}{PR+RE}$

RESULTADOS E DISCUSSÃO: Uma vez que nem todos os clones possuíam a mesma configuração estrutural, algumas variáveis apresentam lacunas, uma vez que a variação de idade acarretava a falta de uma copa mais definida. Sendo assim, para ambos os métodos foram consideradas as seguintes variáveis: velocidade de propagação do pulso ultrassônico (v), diâmetro de engastamento com o solo (db), altura total (ht) e idade. Todos os clones foram analisados quanto a sua distribuição normal, tendo sido avaliado o intervalo em que se encontravam os parâmetros de Simetria e Curtose de cada conjunto de dados. Todos os conjuntos de dados apresentaram simetria e curtose dentro do intervalo de -2 a 2, garantindo assim a normalidade dos mesmos. Avaliando a variável velocidade de propagação do pulso ultrassônico (v) como variável dependente e as variáveis altura total (ht), idade e diâmetro da base (db) como covariáveis, por meio do *Multiple Range Test* foi possível obter a divisão com uma grande sobreposição entre 16 grupos homogêneos formados, evidenciando uma dificuldade do método estatístico convencional em separar os diferentes clones através da velocidade de ultrassom, mesmo quando utilizamos as demais variáveis na tentativa de melhorar essa separação, com alguns grupos apresentando até 7 clones em seu espectro.
Tais agrupamentos propostos pela estatística clássica não possuem as métricas de avaliação dos algoritmos de aprendizado de máquina, entretanto a sua incapacidade de distinguir entre os 28 clones, e propor agrupamentos com intersecções tão extensas, apresenta-se como uma desvantagem.Todos os métodos de mineração de dados foram configurados de acordo com Plas (2016) e Wes (2017) no que se referem à configuração de hiperparâmetros (variáveis e condições estabelecidas antes dos modelos rodarem que influenciam na performance dos modelos, visando que não ocorram problemas de ajuste). Após rodar cada um dos modelos, foram compilados os resultados da acurácia, precisão, recall e f1 score de cada um dos modelos (Tabela 2).

Tabela 2. Métricas de avaliação obtidas para cada modelo.

Modelo	Acurácia	Precisão	Recall	F1-Score
K-Nearest Neighbor	0,5623	0,5763	0,4997	0,5062
Decision Tree	0,8283	0,8166	0,8017	0,7979
Random Forest	0,7872	0,7833	0,7687	0,7642
Gradient Boosting	0,8176	0,8002	0,7910	0,7896
Xtreme Gradient Boosting	0,8119	0,7908	0,7919	0,7837
Superior Vector Machine	0,3526	0,5678	0,2738	0,3161
Redes Neurais	0,7356	0,7522	0,7357	0,7302

Podemos notar que alguns dos modelos não performaram muito bem (Tabela 2), apresentando acurácias baixas, como o caso do K-Nearest Neighbor e Superior Vector Machine e isso pode ser devido às limitações inerentes do modelo ou hiperparâmetros mal configurados. Entretanto, os outros modelos apresentaram todas as métricas acima dos 73% (Tabela 2). O modelo com melhor performance foi o Decision Tree (Árvore de Decisão), com métricas acima de 79%, indicando uma alta taxa de acerto na classificação dos clones. Podemos ainda citar que como o recall do modelo é alto, não incorremos em erros do Tipo I (Falso Negativo). Nesse caso o erro do Tipo I é mais prejudicial, do ponto de vista financeiro e logístico, pois se equivocadamente classificarmos um clone como sendo de uma classe que não a dele, podemos incentivar seu cultivo em melhores áreas, esperando melhores resultados, ocupando hortos e consumindo recursos com um material genético que não apresentará os melhores resultados.

CONCLUSÕES: O modelo Decision Tree (Árvore de Decisão) apresentou as melhores métricas de classificação, se comparado aos outros algoritmos de aprendizado, e apresenta uma vantagem sobre a estatística clássica, uma vez que consegue diferenciar todos os clones e apresentar probabilidades de classificação correta individualmente. Modelos matemáticos que se baseiam em conceitos de mineração de dados e *machine learning* não são amplamente utilizados na área de inspeção e classificação de produtos florestais e essa pesquisa mostrou a oportunidade que essa linha de trabalho pode representar para indústrias de transformação da madeira.

REFERÊNCIAS:

GONÇALVES, R.; LORENSANI, R. G. M.; RUY, M.; VEIGA, N. S.; MÜLLER, G.; ALVES, C. S.; and MARTINS, G. A. M. Evolution of Acoustical, Geometrical, Physical, and Mechanical Parameters from Seedling to Cutting Age in Eucalyptus Clones Used in the Pulp and Paper Industries in Brazil. **Forest Products Journal:** 2019, Vol. 69, No. 1, pp. 5-16. DOI: 10.13073/FPJ-D-17-00013.

GONÇALVES, R.; LORENSANI, R. G. M.; MERLO, E.; SANTACLARA, O.; TOUZA, M.; GUAITA, M.; LARIO, F. Modeling of wood properties from parameters obtained in nursery seedlings. **Canadian Journal of Forest Research**. 2018. DOI: 10.1139/cjfr-2017-0393.

GONÇALVES, R.; BATISTA, F. A. F.; LORENSANI, R. G. M. Selecting eucalyptus clones using ultrasound test on standing trees. **Forest Products Journal.** 2013. 63(3/4): 112-118. DOI: 10.13073/FPJ-D-12-00114.

GUIDO, S. & MÜLLER, A. Introduction to Machine Learning with Python. 2016. Disponível em <https://learning.oreilly.com/library/view/introduction-to-machine/9781449369880/ >.

PLAS, J. V. Python Data Science Handbook. 2016. Disponível em: < https://jakevdp.github.io/PythonDataScienceHandbook/>.

WES, M. Python for Data Analysis. 2nd ed, 2017, Oreilly. 613p.